Universidad de Oviedo

Programa de Doctorado en Ingeniería de Producción,
Minero-Ambiental y de Proyectos

TESIS DOCTORAL

Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning

Manuel J. García Rodríguez

Oviedo, mayo de 2022

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning	Inglés: Public procurement announcements: data analysis and forecasting systems using machine learning methods

2.- Autor	
Nombre: Manuel José García Rodríguez	DNI/Pasaporte/NIE:
Programa de Doctorado: Programa de Doctorado en Ingeniería de Producción, Minero-Ambiental y de Proyectos	
Órgano responsable: CIP	

RESUMEN (en español)

La contratación pública es el gasto público para proveer de bienes, servicios o trabajos a una entidad pública. Es un campo de gran importancia por representar un porcentaje significativo del gasto sobre el PIB de los Estados, un 16% según algunas estimaciones oficiales. Sin embargo, es un campo muy poco estudiado por los investigadores del ámbito científico-tecnológico porque hasta hace pocos años no se disponía de datos de contratos públicos (licitaciones) con la información estructurada y accesible para su descarga masiva por cualquier ciudadano.

Esta Tesis aplica la ciencia de datos a la contratación pública. La ciencia de datos es la unión de 3 disciplinas: conocimiento matemático-estadístico, habilidades de programación y el conocimiento del objeto de estudio (la contratación). En particular, se analizan los datos de las licitaciones, tanto de España como del extranjero, y se abordan algunos de los problemas que hay en este campo aplicando los algoritmos de aprendizaje automático, Machine Learning (ML), que se encuadran dentro de la Inteligencia Artificial (IA). Es una investigación innovadora en el campo académico, en la Administración y en el sector privado.

Se introduce la contratación, la ciencia de datos y sus evoluciones históricas. Se describen los retos actuales de la contratación, qué organismos públicos de España están ligados a ella, la transparencia y datos en abierto de España y las iniciativas de la Comisión Europea (CE) asociadas a los datos de contratación. Los datos en abierto son fundamentales para poder desarrollar este tipo de investigación, realizada por una persona externa a la Administración y que sólo dispone de información pública. Se van a mencionar las causas de por qué la contratación es ahora un área digital y cuantitativa, su legislación, tecnologías aplicables a la contratación y algunos casos de uso. Se describe en detalle la Plataforma de Contratación del Sector Público (PLACSP) desde la perspectiva de los datos en abierto. Además, se ha recopilado literatura académica asociada a la contratación: datos en abierto y calidad del dato, innovación y gestión en la contratación, predicción en la contratación y colusión y corrupción.

Se explican los 4 artículos que forman la Tesis: “Licitaciones en España: regulación, análisis de datos y estimador del importe de adjudicación usando ML”, “Estimador del importe de adjudicación de licitaciones usando ML: caso de estudio con licitaciones de España”, “Recomendador de licitadores usando ML: análisis de datos, algoritmo y caso de estudio con licitaciones de España” y “Detección de colusión en licitaciones aplicando algoritmos de ML”. Los artículos presentados son ejemplos útiles y prácticos de cómo la ciencia de datos y el ML pueden ayudar a los diferentes actores de la contratación (Administración, operadores económicos, investigadores, legisladores, etc.) y, en última instancia, a la sociedad.

Finalmente, la Tesis trata de ser un puente de unión entre dos ámbitos bastante distantes entre

sí hasta la fecha actual: la contratación pública y la ciencia de datos. El primero son expertos de las ciencias sociales y el segundo son expertos del área científico-tecnológica. Ambos colectivos se verán beneficiados si trabajan de manera conjunta, colaborando para conseguir una contratación más eficiente, segura y digitalizada.

RESUMEN (en Inglés)

Public procurement is public spending to provide goods, services or jobs to a contracting authority. It is a field of great importance as it represents a significant percentage of spending on the States, 16% of GDP according to some official estimates. However, public procurement has been little studied by researchers in the scientific-technological field because public procurement data were not available with structured information and accessible for mass download by any citizen until a few years ago.

This PhD applies data science to public procurement. Data science is the union of 3 disciplines: mathematical-statistical knowledge, programming skills and knowledge of the object of study (public procurement). In particular, data from tenders will be analysed, both from Spain and abroad, and some of the problems in this field will be addressed by applying Machine Learning algorithms (Artificial Intelligence). It is an innovative research in the academic field, in the Public Administration and in the private sector.

It will be introduced the public procurement, data science and their historical evolutions. The current challenges of contracting will be described, which public agencies in Spain are linked to it, transparency and open data in Spain and the initiatives of the European Commission associated with public procurement data. Open data is essential to be able to carry out this research by a person external to the Administration and who only has public information. It will mention the reasons why public contracting is now a digital and quantitative area, its legislation, technologies applicable to contracting and some use cases. The Public Contracting Platform from Spain will be described in detail from the perspective of open data. In addition, academic literature associated with public procurement has been compiled: open data and data quality, innovation and management, forecasting and collusion and corruption.

The 4 research papers of the PhD are: “Public procurement in Spain: regulations, data analysis, and award price estimator using ML”, “Award price estimator for public procurement using ML algorithms: case study with tenders from Spain”, “Bidders recommender for public procurement using ML: data analysis, algorithm, and case study with tenders from Spain”, and “Collusion detection in public procurement with ML algorithms”. These papers are useful and practical examples of how data science and ML can help the stakeholders in public procurement (Administration, economic operators, researchers, policy makers, etc.) and, ultimately, the society.

Finally, the PhD could be an excellent bridge between two areas quite distant: public procurement and data science. The first are experts from the social sciences and the second are experts from the scientific-technological field. Both groups will benefit if they work together, collaborating to achieve more efficient, secure and digitised public procurement.

SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN INGENIERÍA DE PRODUCCIÓN, MINERO-AMBIENTAL Y DE PROYECTOS

TESIS DOCTORAL

Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning

Autor: Manuel J. García Rodríguez
Director: Vicente Rodríguez Montequín
Codirector: Ramiro Concepción Suárez

Resumen

La contratación pública es el gasto público para proveer de bienes, servicios o trabajos a una entidad pública. Es un campo de gran importancia por representar un porcentaje significativo del gasto sobre el PIB de los Estados, un 16 % según algunas estimaciones oficiales. Sin embargo, es un campo muy poco estudiado por los investigadores del ámbito científico-tecnológico porque hasta hace pocos años no se disponían de datos de contratos públicos (licitaciones) con la información estructurada y accesible para su descarga masiva por cualquier ciudadano.

Esta Tesis aplica la ciencia de datos a la contratación pública. La ciencia de datos es la unión de 3 disciplinas: conocimiento matemático-estadístico, habilidades de programación y el conocimiento del objeto de estudio (la contratación). En particular, se analizarán los datos de las licitaciones, tanto de España como del extranjero, y se abordarán algunos de los problemas que hay en este campo aplicando los algoritmos de aprendizaje automático, Machine Learning (ML), que se encuadran dentro de la Inteligencia Artificial (IA). Es una investigación innovadora en el campo académico, en la Administración y en el sector privado.

Se introducirá la contratación, la ciencia de datos y sus evoluciones históricas. Se describirán los retos actuales de la contratación, qué organismos públicos de España están ligados a ella, la transparencia y datos en abierto de España y las iniciativas de la Comisión Europea (CE) asociadas a los datos de contratación. Los datos en abierto son fundamentales para poder desarrollar este tipo de investigación, realizada por una persona externa a la Administración y que sólo dispone de información pública. Se van a mencionar las causas de por qué la contratación es ahora un área digital y cuantitativa, su legislación, tecnologías aplicables a la contratación y algunos casos de uso. Se describirá en detalle la Plataforma de Contratación del Sector Público (PLACSP) desde la perspectiva de los datos en abierto. Además, se ha recopilado literatura académica asociada a la contratación: datos en abierto y calidad del dato, innovación y gestión en la contratación, forecasting en la contratación y colusión y corrupción.

Se explicarán los 4 artículos que forman la Tesis: *"Licitaciones en España: regulación, análisis de datos y estimador del importe de adjudicación usando ML"*, *"Estimador del importe de adjudicación de licitaciones usando ML: caso de estudio con licitaciones de España"*, *"Recomendador de licitadores usando ML: análisis de datos, algoritmo y caso de estudio con licitaciones de España"* y *"Detección de colusión en licitaciones aplicando algoritmos de ML"*. Los artículos presentados son ejemplos útiles y prácticos de cómo la ciencia de datos y el ML pueden ayudar a los diferentes actores de la contratación (Administración, operadores económicos, investigadores, legisladores, etc.) y, en última instancia, a la sociedad.

Finalmente, la Tesis trata de ser un puente de unión entre dos ámbitos bastante distantes entre sí hasta la fecha actual: la contratación pública y la ciencia de datos. El primero son expertos de las ciencias sociales y el segundo son expertos del área científico-tecnológica. Ambos colectivos se verán beneficiados si trabajan de manera conjunta, colaborando para conseguir una contratación más eficiente, segura y digitalizada.

Palabras clave: *análisis de datos*, *inteligencia artificial*, *aprendizaje automático*, *contratación pública*, *licitación*, *derecho administrativo*, *colusión*.

Abstract

Public procurement is public spending to provide goods, services or jobs to a contracting authority. It is a field of great importance as it represents a significant percentage of spending on the States, 16 % of GDP according to some official estimates. However, public procurement has been little studied by researchers in the scientific-technological field because public procurement data were not available with structured information and accessible for mass download by any citizen until a few years ago.

This PhD applies data science to public procurement. Data science is the union of 3 disciplines: mathematical-statistical knowledge, programming skills and knowledge of the object of study (public procurement). In particular, data from tenders will be analysed, both from Spain and abroad, and some of the problems in this field will be addressed by applying Machine Learning algorithms (Artificial Intelligence). It is an innovative research in the academic field, in the Public Administration and in the private sector.

It will be introduced the public procurement, data science and their historical evolutions. The current challenges of contracting will be described, which public agencies in Spain are linked to it, transparency and open data in Spain and the initiatives of the European Commission associated with public procurement data. Open data is essential to be able to carry out this research by a person external to the Administration and who only has public information. It will mention the reasons why public contracting is now a digital and quantitative area, its legislation, technologies applicable to contracting and some use cases. The Public Contracting Platform from Spain will be described in detail from the perspective of open data. In addition, academic literature associated with public procurement has been compiled: open data and data quality, innovation and management, forecasting and collusion and corruption.

The 4 research papers of the PhD are: *"Public procurement in Spain: regulations, data analysis, and award price estimator using ML"*, *"Award price estimator for public procurement using ML algorithms: case study with tenders from Spain"*, *"Bidders recommender for public procurement using ML: data analysis, algorithm, and case study with tenders from Spain"*, and *"Collusion detection in public procurement with ML algorithms"*. These papers are useful and practical examples of how data science and ML can help the stakeholders in public procurement (Administration, economic operators, researchers, policy makers, etc.) and, ultimately, the society.

Finally, the PhD could be an excellent bridge between two areas quite distant: public procurement and data science. The first are experts from the social sciences and the second are experts from the scientific-technological field. Both groups will benefit if they work together, collaborating

to achieve more efficient, secure and digitised public procurement.

Keywords: *data analysis*, *artificial intelligence*, *machine learning*, *public procurement*, *auction*, *administrative law*, *collusion*.

Agradecimientos

En todo camino, y más en los intelectuales, se tienen compañeros y guías que te arropan, iluminan y enseñan. Quiero dedicar estas líneas de reconocimiento y gratitud a esas personas que me han acompañado en este viaje lleno de curiosidad científica y descubrimiento personal.

Este viaje comenzó hace 5 años gracias al impulso de los profesores Vicente Rodríguez Montequín y Fran Ortega. Vicente se convirtió en mi Director de Tesis, siendo un guía excepcional que con sus conocimientos, consejos y ayuda me ha permitido afrontar y acabar con éxito el camino emprendido. Además, he tenido el privilegio de conocer y aprender de bastantes investigadores y profesionales durante estos años. Algunos de ellos son coautores de los artículos que forman esta Tesis. Les estaré siempre agradecido por compartir su tiempo conmigo y dedicar sus esfuerzos a llevar un paso más allá esta investigación. Especialmente al profesor Pablo Ballesteros Pérez, por su compromiso con la investigación y la excelencia, un referente para mí.

Este viaje investigador he tenido que recorrerlo en paralelo a mis obligaciones laborales en la empresa privada. El trabajo es más fructífero y gratificante cuando estás rodeado de buenos profesionales y compañeros y yo he tenido esa suerte. Solamente este último año he podido alinear mi labor profesional con la investigación en contratación pública, reforzándose ambas con tal vigor que me ha permitido alcanzar caminos y foros que creía imposibles.

Ni mucho menos el haber llegado hasta aquí es mérito mío, es un largo trayecto vital de varias generaciones que con su humildad, entrega y ejemplo han forjado en mí la persona que soy. Por eso el mayor agradecimiento es a mi familia asturiana y gallega (abuelos, padres, hermana, tíos, etc.), por todo el cariño, paciencia y educación que me han regalado. Y a mis amigos, fieles acompañantes en el enrevesado camino de la vida.

Espero que esta Tesis arroje luz a los lectores e investigadores interesados en atravesar estos caminos y que consigamos vivir en una sociedad más transparente, justa y libre.

Índice

Índice de figuras

Índice de tablas

Glosario

AA.PP. Administraciones Públicas

AGE Administración General del Estado

AIReF Autoridad Independiente de Responsabilidad Fiscal

CC.AA. Comunidades Autónomas

CE Comisión Europea

CNMC Comisión Nacional de los Mercados y la Competencia

CODICE Componentes y Documentos Interoperables para la Contratación Electrónica

CPV Common Procurement Vocabulary

DGRCC Dirección General de Racionalización y Centralización de la Contratación

DOUE Diario Oficial de la Unión Europea

IA Inteligencia Artificial

IGAE Intervención General de la Administración del Estado

LCSP Ley 9/2017 de Contratos del Sector Público

MAPE Mean Absolute Percentage Error

MdAPE Median Absolute Percentage Error

ML Machine Learning

OIReScon Oficina Independiente de Regulación y Supervisión de la Contratación

PLACSP Plataforma de Contratación del Sector Público

TED Tenders Electronic Daily

UE Unión Europea

Capítulo Primero

Introducción

1.1. Motivación

La contratación pública es la denominación común del gasto público para proveer de bienes, servicios o trabajos a una entidad pública [1]. La contratación pública es un intensivo y complejo proceso que consume cuantiosos recursos. Por ejemplo, la Unión Europea (UE) gasta alrededor del 16 % de su PIB en contratos públicos [2]. Por tanto, los mayores adjudicadores de contratos de un país, tanto por número como por importe, son los organismos públicos y cualquier pequeña eficiencia que se logre en sus procesos supondrá un ahorro significativo.

Esta Tesis aplica la ciencia de datos (data science) a la contratación pública, combinación innovadora tanto en el campo académico como en el sector privado o en las Administraciones Públicas (AA.PP.) ¿Qué es la ciencia de datos? Porque toda ciencia involucra datos. No es concepto fácil de definir y más si cabe en los últimos años donde se ha vuelto omnipresente. En la actualidad, es un campo multidisciplinar que utiliza métodos científicos (procesos, algoritmos, etc.) para extraer conocimientos de los datos mediante la informática. La ciencia de datos se puede definir como la unión de 3 campos: conocimiento matemático y estadístico, conocimientos y habilidades de programación y el conocimiento experto del objeto de estudio (en el caso que nos ocupa, la contratación) [3]. En particular, en esta Tesis se estudiarán los datos de los contratos públicos (licitaciones) y se abordarán algunos de los problemas que hay en este campo aplicando la ciencia de datos y, en particular, los algoritmos de aprendizaje automático, Machine Learning (ML) (se explicará en el capítulo 3), que se encuadran dentro de la IA.

Siempre que nos refiramos a la contratación, se sobrentenderá que es contratación pública. Esta Tesis no tiene por objeto estudiar las particularidades de la contratación privada, esto es, la realizada por empresas dentro de sus actividades económicas. Sin embargo, ambos tipos de contratación comparten muchas características [4] porque solamente se diferencian en la naturaleza, pública o privada, de la entidad contratante. La gran mayoría de problemas enunciados, metodologías propuestas, aplicaciones desarrolladas y conclusiones plasmadas en esta Tesis se pueden adaptar y aplicar a la contratación privada.

La contratación pública no ha sido ajena a la ciencia de datos o a la IA. Los estudios y opiniones sobre su aplicabilidad a la contratación son cada vez más comunes. Sin embargo, las aplicaciones de este tipo de tecnologías en la práctica administrativa de la compra pública son, por el momento, escasas, dispersas y veladas [5]. Es decir, hay una gran carencia de herramientas innovadoras que utilicen la IA y esta Tesis trata de paliarlo usando uno de sus subcampos más relevantes: el ML. En otros ámbitos de las AA.PP., no limitándose a la contratación, cada vez van teniendo más ejemplos exitosos que explotan los datos que manejan gracias a la IA [6].

Esta Tesis trata de ser un puente entre dos ámbitos bastante distantes entre sí hasta la fecha: la contratación pública y la ciencia de datos. El primero está liderado por expertos del ámbito de las ciencias sociales (juristas, economistas, politólogos, etc.) y el segundo por expertos del ámbito científico-tecnológico (ingenieros, matemáticos, informáticos, etc.). Ambos colectivos se verán beneficiados y recompensados si trabajan de manera conjunta, colaborando para conseguir una contratación más eficiente, segura y digitalizada. Por tanto, se puede denominar a esta Tesis de frontera, donde el autor hará especial esfuerzo en que ambos grupos de personas entiendan el campo de conocimiento del otro. Es decir, que la Tesis sea comprensible para todos y sirva de ejemplo para futuros proyectos transversales y multidisciplinares.

En los siguientes apartados se resumirá los antecedentes de la ciencia de datos y la contratación, la estructura de la Tesis (6 capítulos) y las publicaciones que se han llevado a cabo.

1.2. Antecedentes

En este apartado se va a hacer un breve resumen de la **evolución histórica de la ciencia de datos y la contratación pública** para tener una visión amplia de ambos campos. Hay dos diferencias sustanciales entre ambos, una de carácter geográfico y otra temporal. La ciencia de datos se ha desarrollado a nivel mundial, cooperando personas de distintos países, a partir del último tercio del siglo XX y de manera muy intensa en los últimos años hasta la actualidad. Sin embargo, la contratación pública tiene una evolución propia y diferencia para cada país, por formar parte de su legislación (derecho administrativo), que en el caso español nace ya en el siglo XIX.

El nacimiento de cualquier campo del saber es muy difícil de ubicar y fechar y el caso de la ciencia de datos no es diferente. Un hecho primigenio y relevante tuvo lugar en 1962 cuando el profesor John W. Tukey[1] (Univ. de Princeton) escribió *"The Future of Data Analysis"* [7] donde esboza una nueva ciencia sobre el aprendizaje de los datos, instando a los académicos a reducir su enfoque en la teoría estadística y a participar en todo el proceso de análisis de datos:

> *"Durante mucho tiempo pensé que era un estadístico interesado en inferencias de lo particular a lo general. Pero a medida que observé la evolución de las estadísticas matemáticas, tuve motivos para preguntarme y dudar [...] Llegué a sentir que mi interés central está en el análisis de datos. El análisis de datos, y las partes de las estadísticas que se adhieren a él, deben [...] asumir las características de la ciencia en lugar de las matemáticas [...] el análisis de datos es intrínsecamente una ciencia empírica."*

Ahí se menciona la evolución de la estadística matemática como ciencia de datos. Sin embargo, no sería hasta 1974 cuando Peter Naur [8], científico danés pionero en la informática, acuñara el término que actualmente conocemos. El desarrollo de las ciencias de la computación (software), la microelectrónica (hardware) y las redes de telecomunicaciones a finales del siglo XX y comienzos del XXI, gracias a la inversión económica realizada por el sector privado (grandes multinacionales) y público (Universidades y centros tecnológicos de investigación), han conseguido un altísimo nivel de informatización, de digitalización de la sociedad, que ha supuesto una revolución social y económica. Esto ha sentado las bases para que en los últimos 10 años haya habido una eclosión de proyectos basados en ciencia de datos en muchos sectores de la economía (industria, banca, telecomunicaciones, energía, transporte, sanidad, etc.) y, algo más tarde, también en las AA.PP.

La evolución histórica de la contratación pública en España se remonta al Real Decreto de 27 febrero de 1852, publicado por Bravo Murillo y que decía *"proyecto de ley de contratos sobre*

[1] Estadístico experto en topología y conocido, entre otras aportaciones, por el cálculo de la transformada rápida de Fourier (FFT) y el diagrama de cajas (boxplot).

servicios públicos, con el fin de establecer ciertas trabas saludables, evitando los abusos fáciles de cometer en una materia de peligrosos estímulos, y de garantizar la Administración contra los tiros de la maledicencia...” [9]. Es el nacimiento de la contratación administrativa en España y ya menciona 3 elementos cruciales que aún hoy siguen estando presentes: limitar los abusos fáciles (de autoridades y funcionarios) contra peligrosos estímulos (cohechos, prebendas, sobornos) y evitar los tiros de la maledicencia (desconfianza en los servidores públicos).

Después, se dictaron otras disposiciones de igual rango, hasta que en 1911 se promulgó la Ley de Contabilidad de la Hacienda Pública, en cuyo capítulo V se trataba la contratación. La única forma de adjudicación de los contratos era mediante la subasta (la proposición más económica, sin importar la calidad, plazos u otras características del suministro). Estas disposiciones estuvieron vigentes hasta la Ley de Bases de Contratación del Estado de 1963 y su texto articulado de 1965, finalizándose la regulación de esta materia con la promulgación del Reglamento de Contratación del Estado en 1967. En 1975 se actualizó mediante el Decreto 3410/1975 y, tras la incorporación de España en la Comunidad Económica Europea en 1986, también hubo una serie de adaptaciones a la legislación comunitaria europea. En 1995 se derogan expresamente las leyes anteriores de contratación y se crea la Ley de contratos de las Administraciones Públicas (Ley 13/1995). En los años posteriores se sucederán varias leyes de contratación que buscan incrementar la concurrencia y aumentar la transparencia y objetividad en los procedimientos de adjudicación, así como simplificar en lo posible los procedimientos de contratación, adaptarlos a la digitalización de las AA.PP. La actual ley en vigor es la Ley 9/2017 de Contratos del Sector Público (LCSP) y abre las puertas a poder utilizar los datos de las licitaciones de manera masiva, aplicando la ciencia de datos y las técnicas de ML.

1.3. Estructura de la Tesis

Se ha hecho un esfuerzo en tratar el objeto de estudio, la contratación pública, no sólo como un mero conjunto de datos en donde aplicar la ciencia de datos, esto es, la experimentación con algoritmos de ML y técnicas cuantitativas, sino en comprender su naturaleza desde distintas perspectivas (legal, de gestión pública, económica, etc.). Con esta visión más holística se puede investigar con conocimiento de causa, aportando soluciones realistas e innovadoras mediante el ML. La investigación se ha estructurado en 6 capítulos, resumidos a continuación.

Capítulo 1: Introducción. Se describe la motivación para haber realizado la investigación, así como los antecendentes y relevancia que tiene la contratación pública.

Capítulo 2: Retos y objetivos. Se explican los retos actuales de la contratación y los organismos públicos con competencias en esta materia. Además, se resumen los conceptos de la transparencia en la contratación, los datos en abierto (open data) y las iniciativas de la CE ligadas a los datos de contratación a nivel europeo, por ser elementos necesarios para desarrollar esta Tesis. Es decir, la investigación se ha basado fundamentalmente en datos públicos, accesibles a cualquier ciudadano. Finalmente, se enumeran los objetivos y metodología que se han marcado en la Tesis.

Capítulo 3: Fundamentos de la investigación. Se describe la contratación desde el punto de vista cuantitativo, legislativo, de tecnologías software y sus posibles aplicaciones (casos de uso). Por ser la principal fuente de datos empleada, se detalla la Plataforma de Contratación del Sector Público (PLACSP): los órganos que publican y adjudican las licitaciones y el origen de los datos así como su formato y calidad. Se hace una recopilación de la literatura utilizada

en la contratación para que se tengan referencias del estado del arte actual. Además, se formulan varias métricas de error típicas para evaluar los algoritmos de ML, en particular para los problemas de regresión y clasificación (los tratados en la Tesis).

Capítulo 4: Artículos de la Tesis. La Tesis Doctoral se ha desarrollado en la modalidad de compendio de publicaciones. Se describe un programa informático, desarrollado por el autor, que detecta licitaciones irregulares y que ha sido probado con éxito en España. Por otro lado, se resumen los 4 artículos que forman el compendio de publicaciones de la Tesis y son el grueso de la investigación y desarrollo informático efectuado:

- "Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning."
- "Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain."
- "Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain."
- "Collusion detection in public procurement auctions with machine learning algorithms."

Capítulo 5: Discusión de los resultados. Análisis de los resultados más relevantes de los anteriores artículos, con una visión de conjunto y transversal. Se analiza primero desde una perspectiva técnica, algorítmica, para evaluar la aplicación de los algoritmos de ML en el campo de la contratación. Después, se analiza desde una perspectiva cuantitativa, económica, para abrir el campo a futuros estudios económicos. Así, esta Tesis se puede tomar como un ejemplo pionero en la utilización de la ciencia de datos para hacer estudios económicos en la contratación.

Capítulo 6: Conclusiones. Conclusión de la investigación y futuras líneas que abre la Tesis.

Finalmente, se anexan los 4 artículos (resumidos en el capítulo 4) que forman el compendio de publicaciones de la Tesis y se detalla la bibliografía referenciada a lo largo del trabajo. Esta memoria se ha elaborado de acuerdo a lo establecido en el artículo 28 del Reglamento de los Estudios de Doctorado de U. de Oviedo (BOPA núm. 185 del 9 de Agosto de 2018). Para mantener la coherencia con las tablas y figuras de los artículos de investigación publicados en inglés, cuando se haga uso en la memoria de dichos elementos se mantendrán en su formato original en inglés.

1.4. Publicaciones realizadas en la Tesis

A continuación, se citan las publicaciones realizadas dentro del marco de la investigación de la Tesis llevada a cabo durante los últimos 5 años. En la Tabla 1.1 se citan los artículos que forman parte del compendio de publicaciones de la Tesis y en qué revistas se publicaron (todas pertenecen al Science Citation Index Expanded, SCIE).

- [10] M. J. García Rodríguez y col., "Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning," *Complexity*, vol. 2019, n.º v, 2019, issn: 10990526. doi: 10.1155/2019/2360610
- [11] M. J. García Rodríguez y col., "Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain," *Studies in Informatics and Control*, vol. 30, n.º 4, págs. 67-76, dic. de 2021, issn: 12201766. doi: 10.24846/v30i4y202106

Revista	Editor	Artículo	Factor de impacto	Categorías y ranking por factor de impacto
Automation in construction	Elsevier	[13]	7.700	Civil engineering (2020) = 2/137 (Q1); Construction & Building technology (2020) = 3/67 (Q1)
Complexity	Wiley-Hindawi	[10]	2.462	Mathematics, interdisciplinary applications (2019) = 28/106 (Q2); Multidisclipinary sciences (2019) = 31/71 (Q2)
Complexity	Wiley-Hindawi	[12]	2.833	Mathematics, interdisciplinary applications (2020) = 31/108 (Q2); Multidisclipinary sciences (2020) = 30/72 (Q2)
Studies in informatics and control	National Institute for R&D in Informatics, ICI Bucharest	[11]	1.649	Operations research & management science (2020) = 61/84 (Q3); Automation & control systems (2020) = 43/63 (Q3)

Tabla 1.1: Revistas de los artículos publicados para la Tesis.

- [12] M. J. García Rodríguez y col., "Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain," *Complexity*, vol. 2020, T. C. Silva, ed., págs. 1-20, nov. de 2020, issn: 1099-0526. doi: 10.1155/2020/8858258
- [13] M. J. García Rodríguez y col., "Collusion detection in public procurement auctions with machine learning algorithms," *Automation in Construction*, vol. 133, pág. 104 047, ene. de 2022, issn: 09265805. doi: 10.1016/j.autcon.2021.104047
- [14] M. J. García Rodríguez y col., "Spanish Public Procurement: Legislation, open data source and extracting valuable information of procurement announcements," *Procedia Computer Science*, vol. 164, págs. 441-448, 2019, issn: 18770509. doi: 10.1016/j.procs.2019.12.204
- [15] D. Arosa Otero y col., "La contratación pública en España: fuentes de datos, normativa y aplicaciones tecnológicas," *Revista de la Escuela Jacobea de Posgrado*, vol. 21, págs. 87-112, 2021. dirección: https://www.jacobea.edu.mx/revista/numero21.php
- [16] M. J. García Rodríguez, "Tecnologías digitales para el control de la contratación pública," en *Auditoria pública*, ASOCEX, ed., vol. 79, Sevilla, 2022, isbn: 1136-517 X. dirección: http://auditoriapublica.com

Además, hasta la fecha actual se ha participado en los siguientes congresos para compartir y debatir las investigaciones llevadas a cabo:

- Comunicación titulada *"La Contratación Pública en España: fuentes de datos, normativa y aplicaciones tecnológicas"* [15] y conferencia impartida en el *Congreso Internacional online: temas clave de la contratación pública*. Organizado por la Univ. de Vigo el día 5/10/2021.
- Comunicación titulada *"Mejorar la gestión y supervisión de la contratación: los algoritmos de IA y los datos en abierto de PLACSP"* y conferencia impartida en el *XI Congreso Internacional sobre contratación pública*. Organizado por la Univ. de Castilla-La Mancha los días

25-26/1/2022 en Cuenca. La comunicación ha sido seleccionada para publicarse en el libro *Observatorio de los contratos públicos 2022*, editorial Aranzadi (pendiente de publicación).

- Conferencia online titulada *"La implantación de la IA en la contratación pública: un diálogo entre tecnología y derecho"* impartida en el *II Seminario de expertos sobre derecho administrativo e innovación tecnológica*. Organizado por la Univ. de Castilla-La Mancha el 15/3/2022.
- Conferencia online titulada *"Mejorar la supervisión de la contratación pública: los algoritmos de IA aplicados a los datos en abierto de PLACSP"* impartida en la *"II Jornada del Observatorio Sector Público e IA de la Facultad de Derecho de la Univ. de Cádiz"* el 20/4/2022.
- Conferencia titulada *"Tecnologías digitales para velar por la contratación: el Big Data y la IA"* impartida en el *"XIII Seminario de Contratación Pública"*. Organizado por la Univ. de Zaragoza los días 22-23-24/6/2022 en Panticosa (Huesca).

Capítulo Segundo

Retos y objetivos

2.1. Retos actuales en la contratación pública

La contratación, como elemento principal del sector público para la prestación de servicios, no está exenta de los procesos de transformación y retos de la administración. Las AA.PP. han asimilado los procesos de transformación digital hasta el momento sin mayor inconveniente, aprovechando los instrumentos digitales para transformar la provisión de servicios y atención a la ciudadanía [17]. Sin embargo, todavía existen retos y áreas por cubrir, especialmente en el ámbito de la contratación.

Este estudio no trata de hacer un análisis exhaustivo de los retos que debe abordar la contratación pública, sino que identifica alguna de las problemáticas que se pueden abordar mediante la introducción de tecnologías. Es decir, utilizando las nuevas tecnologías como un medio o palanca de transformación para superar determinadas problemáticas, no como un fin en sí mismo. Podemos agrupar los retos en dos grandes bloques:

1. El aprovechamiento de los datos en la contratación para la mejora de las funciones de supervisión, monitorización, transparencia, planificación estratégica o la elaboración de políticas públicas.
2. La necesidad de realizar el proceso de compra pública de una forma más eficiente mediante la mejora operativa de la contratación.

En torno al primer bloque de retos, uno de los principales retos a los que se enfrenta la contratación es la calidad y la accesibilidad de los datos [18, 19]. En este sentido, la CE [20] destaca la falta de armonización de datos a nivel europeo como una de las grandes problemáticas para entender cómo están comprando y licitando las AA.PP. europeas. Esta falta de armonización se refleja en la imposibilidad de conocer, con precisión, cuánto están gastando las AA.PP. en la contratación de servicios, obras y suministros, o en la concesión de servicios públicos. Además, en España este problema se aprecia en la existencia de diversas fuentes de datos de contratación pública, con diferentes estándares de datos y de calidad y disponibilidad de la información. Esto se ha tratado de abordar en los últimos años mediante la obligatoriedad del uso de Plataforma de Contratación del Sector Público (PLACSP), pero todavía existe margen de mejora para garantizar el uso adecuado de la misma. En esta línea, el informe anual de la OIReScon sobre la contratación pública destaca el reto de la falta de información que permita la monitorización de la ejecución de contratos públicos [21], asignatura pendiente de las AA.PP. no sólo en España, sino en la mayoría de los estados europeos [22].

La supervisión y el control de la contratación es un pilar fundamental para garantizar el proceso de contratación, la buena ejecución de las compras públicas, el control de gasto público de manera

eficaz y eficiente, elaboración de políticas públicas, así como prevenir y abordar problemas de corrupción y prácticas fraudulentas. En el sector privado se entiende que la contratación es una herramienta fundamental para conseguir los objetivos estratégicos de la organización y, sin embargo, esta idea apenas ha sido abordada en el sector público [4]. Para abordar este reto, diversas iniciativas se han implementado a nivel europeo, utilizando sistemas tecnológicos o innovadores [23]. En este sentido, la mayoría de los sistemas desarrollados hasta el momento han abarcado la detección de riesgos, asociado a la problemática de la corrupción. Se citan algunos ejemplos:

- El proyecto de Open Tender permite la búsqueda, análisis y supervisión de contratos públicos de los países Europeos. Ver un ejemplo para la supervisión del caso español en la Figura 2.1.
- El proyecto de Red Flags, liderado por Transparencia Internacional, permite la supervisión de contratos públicos en Hungría y la detección automática de alertas de riesgo.
- DoZorro, liderado por varios actores internacionales, para la mejora del sistema de supervisión de la contratación pública de Ucrania.
- Iniciativas llevadas a cabo en Chile [24]. A través de su observatorio de compras públicas ha implementado diversas herramientas de medición de riesgo en las compras públicas de cara a realizar auditorías de contrato.

Otro reto es la transparencia en la gestión de recursos públicos ligada a la contratación. La transparencia es uno de los mecanismos fundamentales para garantizar las normas de funcionamiento de la UE con respecto al mercado único [25]. La introducción de reformas para conseguir transparencia en el gasto en contratación es crucial, facilitando el acceso a la información de precios, competencia o mecanismos de adjudicación [26]. Abordar este reto desde la adopción de reformas en el ámbito tecnológico es muy importante. Los beneficios que aportará la transparencia al proceso de contratación no están únicamente relacionados con la reducción de corrupción o prácticas ilícitas sino también mejorar la buena gobernanza de las administraciones públicas [27].

En cuanto al segundo bloque de retos, el más relevante es la mejora operativa de los procesos de compra, teniendo en cuenta que el principal objetivo de la compra pública es la propia adquisición de bienes y servicios. Es un reto para las AA.PP. el tener información y un conocimiento extenso de temas tan relevantes para la consecución de una buena compra como son el diseño de la estrategia de contratación, la definición de los requisitos del servicio a prestar o el suministro a proveer, el conocimiento del propio mercado y los proveedores o la información sobre los precios de mercado [24]. La propia eficiencia operativa del proceso de contratación es algo perseguido por las AA.PP., de cara a ser más eficientes en la gestión de sus recursos. En este sentido, la automatización de procesos repetitivos o sin valor añadido o la reducción de la carga administrativa son materias pendientes de la contratación.

2.2. Organismos públicos españoles relacionados con contratación

La innovación en las herramientas analíticas, basadas en datos como se proponen en esta Tesis, ayudarán a los diferentes organismos públicos relacionados con la gestión, fiscalización y supervisión de la contratación e, incluso, a la investigación policial. La ciudadanía cada vez es más consciente de la importancia de los contratos públicos, exigiendo mayores niveles en la gestión, transparencia y rendición de cuentas, sobre todo a los políticos que son la cabeza visible de la Administración. A continuación se enumeran organismos públicos españoles que gestionan o utilizan información de contratos públicos:

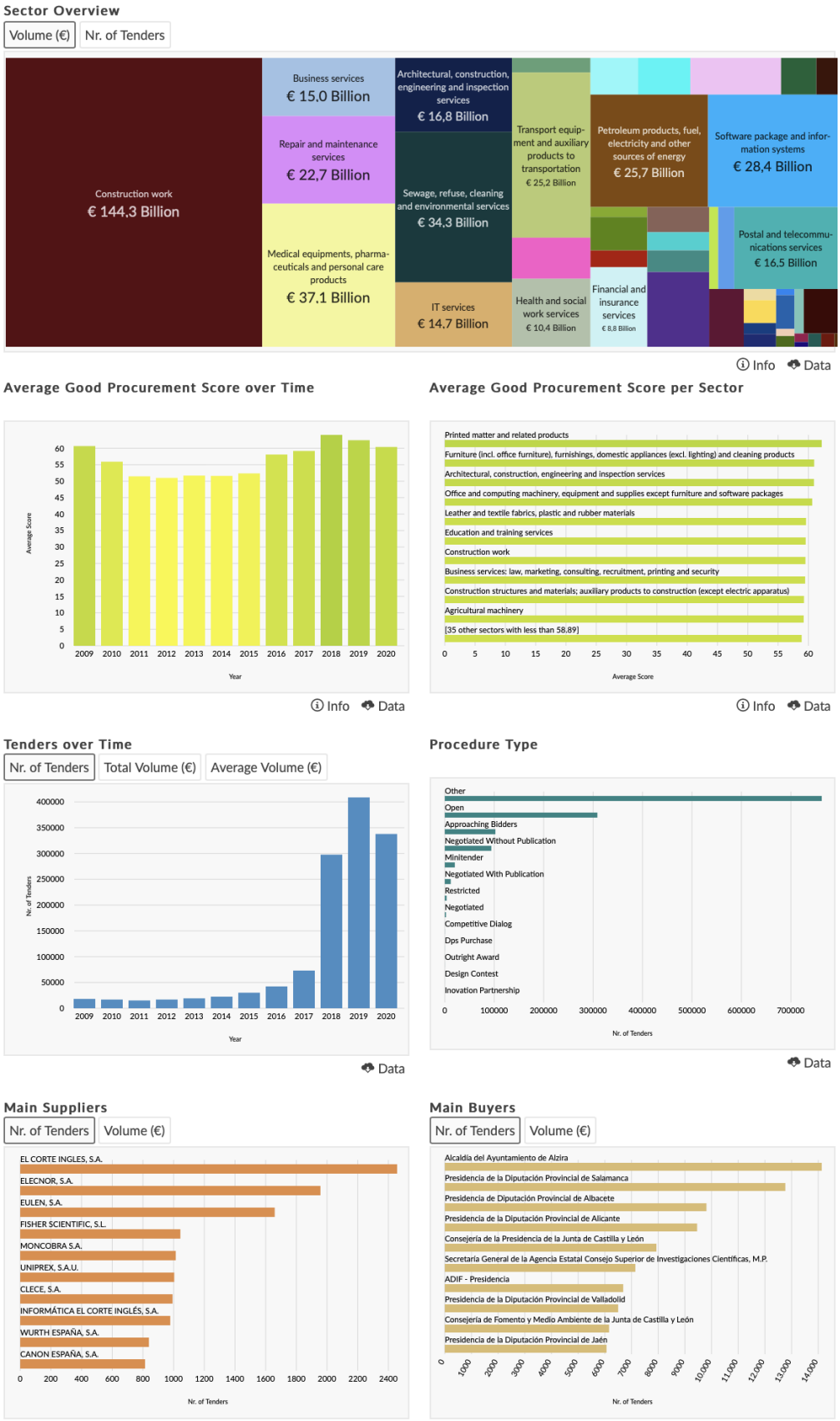


Figura 2.1: Ejemplo de supervisión de la contratación de España. Proyecto Open Tender.

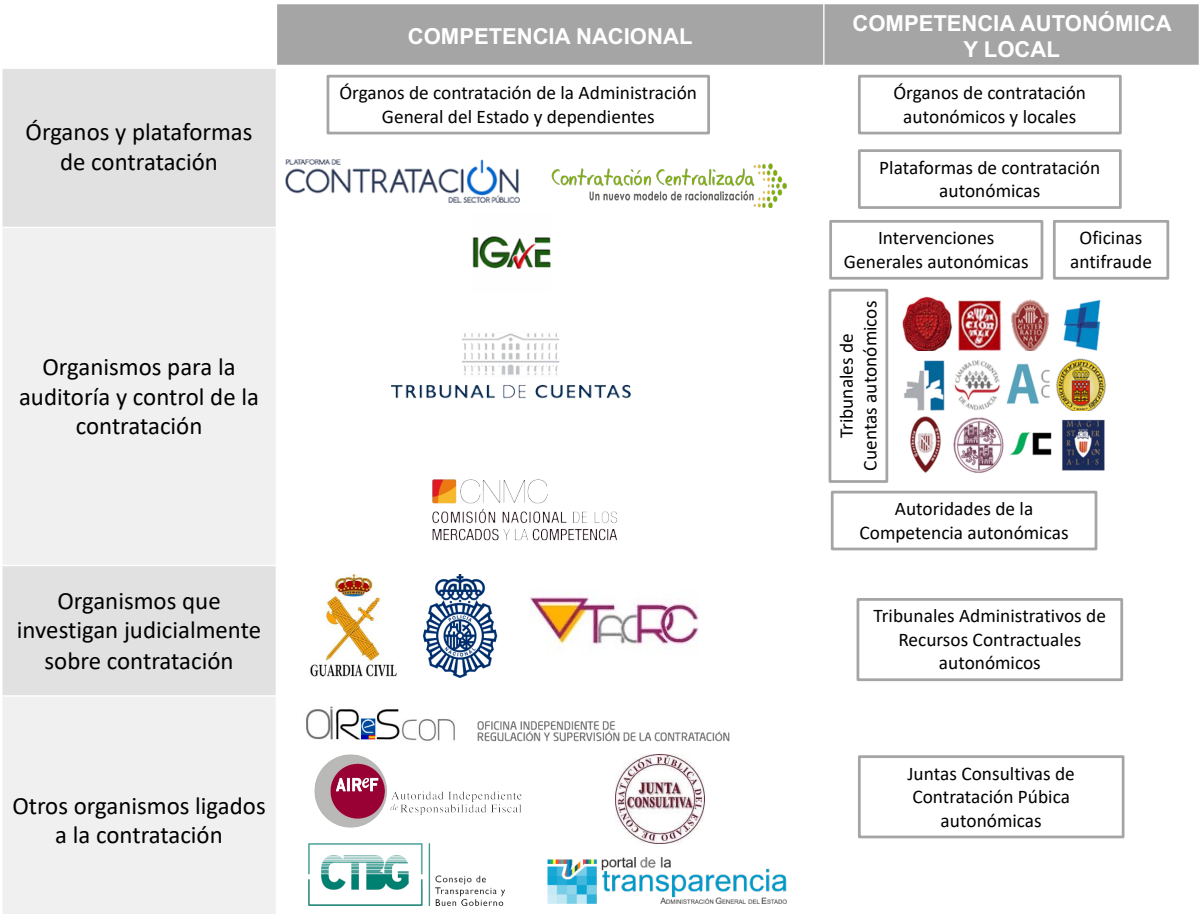


Figura 2.2: Mapa de los organismos públicos españoles ligados a la contratación.

- **Órganos y plataformas de contratación**:
  - Órganos de contratación en todos los niveles de la administración (local, autonómica y estatal) y las entidades dependientes del sector público. La contratación española está muy fragmentada, basta con dar el siguiente dato ilustrativo: en 2021 estaban registrados en Plataforma de Contratación del Sector Público (PLACSP) 20.500 órganos con competencias para realizar contratos.
  - PLACSP, plataformas de contratación autonómicas y la Dirección General de Racionalización y Centralización de la Contratación (DGRCC).
- **Organismos para la auditoría y control de la contratación**:
  - Intervención General de la Administración del Estado (IGAE) y las Intervenciones Generales autonómicas.
  - Oficinas Antifraude autonómicas.
  - Órgano de Control Externo Nacional (Tribunal de Cuentas) y autonómicos (Sindicaturas, Cámaras y Consejos de Cuentas).
  - Comisión Nacional de los Mercados y la Competencia (CNMC), así como Autoridades de la Competencia autonómicas.
- **Organismos que investigan judicialmente sobre contratación**:
  - Fuerzas y cuerpos de seguridad del Estado (Guardia Civil, Policía Nacional, etc.) que investigan a instancia de órganos judiciales.
  - El Tribunal Administrativo Central de Recursos Contractuales y los autonómicos.
- **Otros organismos con intereses en la contratación** y su transparencia son: OIReScon, Junta Consultiva de Contratación del Estado y las juntas autonómicas, Autoridad Independiente de Responsabilidad Fiscal (AIReF), Consejo de Transparencia y Buen Gobierno (CTBG), Portal de la Transparencia de la AGE, etc.

Por tanto, hay varias decenas de organismos (sin contar los miles de órganos de contratación) que se verían beneficiados si se dotasen de buenas herramientas analíticas, basadas en datos. Como en el mercado casi no existen programas informáticos para estos fines, las distintas AA.PP. tiene que diseñar y comprar aplicaciones informáticas ad hoc. Si se comprasen de manera conjunta y coordinada, se crearían muchas sinergias y se reducirían los costes de desarrollo, mantenimiento y evolución. Cuesta mucho esfuerzo diseñar y desarrollar aplicaciones informáticas pero luego son fácilmente adaptable a las distintos organismos y necesidades. Por ejemplo, las Comunidades Autónomas (CC.AA.) deberían de comprar conjuntamente aplicaciones informáticas específicas para sus órganos de contratación, Intervenciones Generales autonómicas, Órganos de Control Externo autonómicos, Comisiones de la Competencia autonómicas, etc. Como estos organismos tienen unas competencias autonómicas equivalentes, tendrán necesidades muy parecidas o iguales.

2.3. Transparencia y datos en abierto de contratación en España

La transparencia en la contratación ha sido objeto de estudios jurídicos [28, 29, 30, 31, 26] y tiene tres aspectos fundamentales [32]. Primero, dar a conocer información relativa a los contratos públicos para rendir cuentas ante la sociedad. Segundo, contribuir a crear un mercado europeo de compras públicas. Y tercero, se obtendría información valiosa para incrementar la calidad de

la gestión y supervisión, gracias a las técnicas de análisis masivo de datos. Además, una ventaja derivada es que tanto los órganos de contratación como los licitadores se sienten más vigilados, por lo que tienen más incentivos a hacer un trabajo más honesto y ajustado a derecho. En resumen, la transparencia actúa como principal *"escudo"* contra la corrupción [33]. Hay otros actores como los interventores, supervisores de la contratación, auditores de cuentas, autoridades de la competencia, etc. que también se benefician indirectamente de una mayor transparencia por convertirse en un ámbito de estudio mayor.

Para que haya transparencia se necesitan datos en abierto, llamados open data en su denominación inglesa. Son datos que pueden ser utilizados, compartidos y procesados libremente por cualquier persona, en cualquier lugar y para cualquier propósito (definición de la Open Knowledge Foundation). Los datos abiertos se basan en 8 principios: garantizar que están completos, primarios, oportunos, accesibles, procesables por una máquina, licencia libre, el acceso debe ser no discriminatorio y los formatos no deben ser propietarios. El estudio de temáticas asociadas a open data ha crecido muy fuertemente en el ámbito académico [34], sobre todo en la última década [35]. Esto es debido a la utilidad (directa o indirecta) de dichos análisis para obtener nueva información valiosa [36, 37] para los distintos involucrados. En el ámbito de la contratación pública, es cada vez mayor la cantidad y calidad de los repositorios de datos abiertos disponibles, tanto a nivel local, autonómico y estatal como europeo.

El open data asociado a la contratación pública está también evolucionando intensamente debido, principalmente, a factores tecnológicos (desarrollo de modelos y software para el e-Procurement [38]), burocráticos (estandarización del lenguaje de la contratación y digitalización de las AA.PP.), políticos (mayor transparencia en la toma de decisiones políticas), económicos (globalización, empresas compitiendo cada vez en mercados más lejanos a su origen) y sociales (menor tolerancia a la gestión pública y política, ineficaz y el interés por tener más información sobre las AA.PP. y sus contrataciones con empresas privadas). Para hacerse una idea de la importancia del open data en Europa, la Comisión Europea (CE) ha estimado [39] que en 2019 hubo un millón de trabajadores que trabajaron directamente con datos en abierto, el tamaño del mercado es de unos 180 mil millones de € y las previsiones para los próximos años son de crecimientos significativos.

Un ejemplo ilustrativo del beneficio que aporta el open data es la licitación[1] que publicó la OIReScon en abril de 2021 cuyo objeto es el *"Desarrollo de herramientas ETL (Extract, Transform and Load) y modelo de datos de la información de múltiples plataformas de contratación"*. Para realizar sus estudios y trabajos de supervisión, la OIReScon contrata a una empresa para desarrollar un sistema informático que extraiga periódicamente los datos de las plataformas de contratación de España utilizando los datos en abierto. Dichos datos se deben validar, transformar y almacenar para que puedan ser explotados fácilmente por el personal de la OIReScon. Esta licitación indica la ausencia de mecanismos que hay para el intercambio de información entre las propias AA.PP., teniendo la OIReScon que recurrir a los datos en abierto. Y como la propia OIReScon indica en su último informe anual de 2021 [21]: *'se considera que debiera avanzarse a la generación de datos abiertos mucho más completos, en formatos homogéneos que faciliten su carga y tratamiento conjunto y respaldados por documentación acreditativa y técnica clara."*

[1] El detalle de la licitación se puede consultar en: https://contrataciondelestado.es/wps/poc?uri=deeplink:detalle_licitacion&idEvl=Uwt%2FOsOe%2B5ymq21uxhbaVQ%3D%3D.

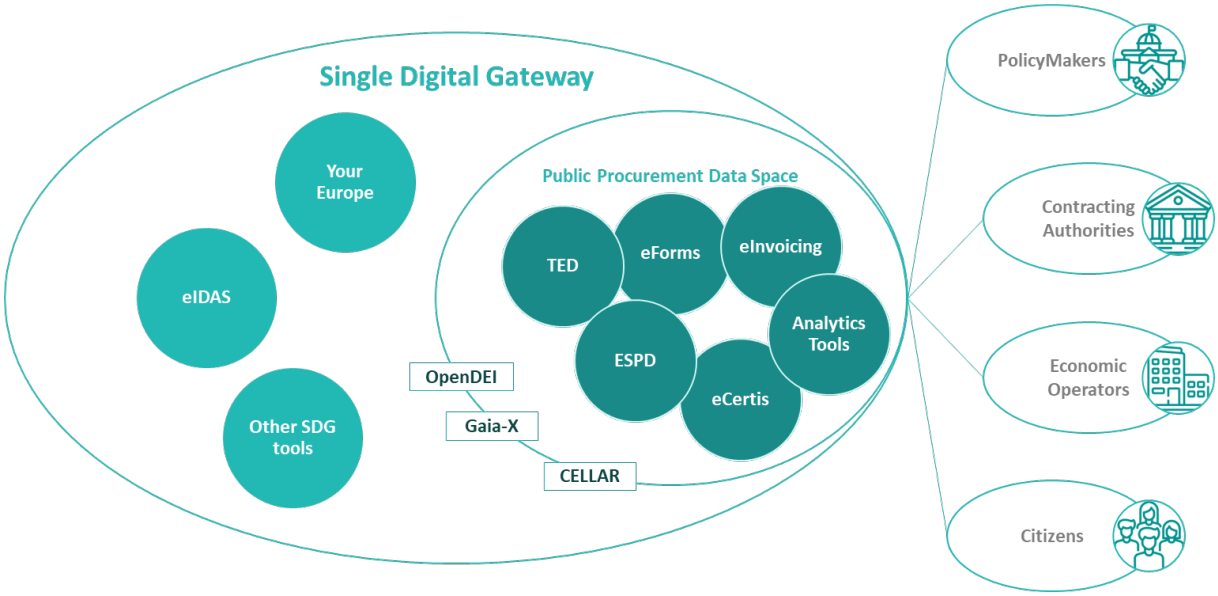


Figura 2.3: Elementos para tener un espacio de datos de contratación pública europeo unificado.

2.4. Iniciativas de la Comisión Europea asociadas a datos de contratación

En el primer apartado se mencionaba el reto de aprovechar los datos de contratación para mejorar las funciones de supervisión, monitorización, transparencia, planificación estratégica o la elaboración de políticas públicas. En la actualidad, la Comisión Europea (CE) tiene su portal de publicaciones de contratación, llamado Tenders Electronic Daily (TED), pero tienen carencias y limitaciones en relación a la publicación de licitaciones porque no contienen la misma información y, por tanto, no se asegura la homogeneidad y calidad de sus datos. Consecuentemente, la CE está trabajando en varias iniciativas para tener una verdadera estructura unificada de datos de contratación de los países miembros de la UE. El ecosistema digital promovido por la CE es un aspecto crucial para avanzar hacia un verdadero marco de contratación pública interoperable que garantice los principios del Mercado Único Digital (Single Digital Market) y, en particular, los del Portal Único Digital (Single Digital Gateway, SGD). Además, la alineación con eIDAS, habilitador clave para las transacciones electrónicas transfronterizas en la UE, debe considerarse como una oportunidad para facilitar la adopción de servicios digitales transfronterizos en el campo de la contratación pública.

Actualmente la CE está definiendo el espacio de datos de contratación pública (Public Procurement Data Space) (ver Figura 2.3) para que todos los países miembros de la UE vuelquen de manera estandarizada sus licitaciones y se pueda hacer un análisis de los datos a nivel europeo. Hasta ahora es muy difícil hacer análisis de datos de contratación a nivel europeo (en general, internacionalmente) porque cada país tiene su propia legislación y plataformas digitales de contratación. Por tanto, la CE es consciente del valor de los datos de contratación (al igual que el autor de esta Tesis) y para ello necesita previamente diseñar y crear un repositorio común de datos, alineando a todos los países miembros. Así, los distintos actores de la contratación podrán explotarlos convenientemente.

A continuación se enumeran los sistemas o iniciativas más relevantes para conseguir el espacio

de datos de contratación público de la UE:

- Tenders Electronic Daily (TED). Es el portal de publicación de licitaciones de la UE.
- eForms. Es una normativa[2] de la CE para crear un estándar, formularios normalizados a rellenar por los órganos de contratación, en la publicación de las licitaciones en TED.
- European Single Procurement Document (ESPD). Documento estandarizado de autodeclaración, según una normativa[3], que es solicitado por los órganos de contratación a los operadores económicos para declarar que cumplen los requisitos legales y administrativos de la licitación en particular (si no los cumplen, serán excluidos del proceso de licitación).
- eCertis. Sistema de información que ayuda a los operadores económicos y órganos de contratación en la identificación y reconocimiento de los documentos y certificados más solicitados en los procedimientos de contratación de los Estados miembros. Persigue el objetivo de proporcionar interoperabilidad normativa y facilitar la contratación transfronteriza (órganos de contratación y operadores económicos de países diferentes).
- eProcurement Ontology (ePO). Crear una ontología de contratación electrónica, comúnmente acordada, que codifique formalmente los procesos administrativos de contratación. Así se conseguirá un formato estructurado de datos de contratación legible por máquinas de manera automatizada. El objetivo es cubrir el proceso de contratación de principio a fin: publicación, proceso de licitación, adjudicación, pedido, facturación y pago. De esta manera, ePO conseguiría unificar las prácticas existentes de contratación electrónica, facilitando el intercambio, acceso y reutilización de datos.
- CELLAR. Es un repositorio semántico construido por la Oficina de Publicaciones de la CE. Es un servicio de infraestructura masivo que incluye grandes cantidades de contenido legal, publicaciones de la UE y conjuntos de datos de referencia. La plataforma se ha diseñado para ser escalable y admitir la interoperabilidad de los servicios, al proporcionar un marco común para el almacenamiento de contenido y metadatos.
- eInvoicing. Normativa europea[4] para la facturación electrónica, debido a los distintos formatos de factura utilizados en los países miembros de la UE. Los órganos de contratación deberán aceptar las facturas que cumplan con la norma europea, pero seguirán siendo válidas las facturas que cumplan las diferentes normativas nacionales.

También hay iniciativas en el ámbito privado como, por ejemplo, Gaia-X que quiere crear una infraestructura de datos abierta, federada e interoperable, constituida sobre los valores de soberanía digital y disponibilidad de los datos. La CE tiene otra iniciativa llamada OpenDEI, para la creación de plataformas de datos comunes basadas en una arquitectura unificada y un estándar establecido. En conclusión, se está en un momento de organización y estandarización de la información a nivel europeo, tanto en materia de contratación como en otros campos.

[2]Reglamento de Ejecución (UE) 2019/1780 de la Comisión, de 23 de septiembre de 2019, por el que se establecen formularios normalizados para la publicación de anuncios en el ámbito de la contratación pública

[3]Reglamento de Ejecución (UE) 2016/7 de la Comisión, de 5 de enero de 2016, por el que se establece el formulario normalizado del documento europeo único de contratación

[4]Directiva 2014/55/UE del Parlamento Europeo y del Consejo, de 16 de abril de 2014, relativa a la facturación electrónica en la contratación pública

2.5. Objetivos y metodología de la investigación

En los siguientes 5 puntos se recogen los **principales objetivos de la Tesis**. Los 3 primeros se marcaron desde el inicio de la Tesis y los 2 últimos han sido consecuencia directa al ir desarrollándose los anteriores.

- **Adquisición y uso de los datos de contratación**. En 2017, cuando se comenzó la Tesis, era un reto en sí mismo el conseguir datos de licitaciones y poder manipular sus datos. En estos últimos 5 años en España se ha mejorado la cantidad, calidad y facilidad de acceso a los datos públicos de licitaciones. Por otro lado, esta Tesis hace hincapié en cruzar distintas fuentes de datos para abordar eficazmente los problemas en la contratación.
- **Empleo de innovadoras técnicas de Machine Learning (ML) a la contratación**. Al ser un campo donde se ha experimentado poco con técnicas de IA y ML, hay una incertidumbre al buscar problemas en los que se pueda aplicar con éxito estas técnicas.
- **Crear aplicaciones útiles en la contratación**. Es el objetivo más importante. Hay que desarrollar herramientas software que tengan interés para los agentes públicos y privados, obteniendo beneficios por su empleo, tanto para ellos como indirectamente para la sociedad. Se ha podido llevar a cabo combinando los dos puntos anteriores: utilizar fuentes de datos ligadas a la contratación y aplicar algoritmos de ML.
- **Detección de licitaciones irregulares**. Desarrollar programas informáticos que detecten licitaciones potencialmente fraudulentas para conseguir una contratación más segura y con menos corrupción. Y no sólo abordar el problema desde la visión técnica, sino también sacándolo a la luz pública para generar un debate mediático y señalar a los infractores. De esta manera se podrán corregir los problemas de fondo que afectan a la contratación en su conjunto.
- **Difusión de la investigación**. Al ser la contratación un tema de interés general para la sociedad y transversal para investigadores muy diversos, es recomendable hacer una labor de divulgación de la Tesis. Que la ciudadanía, funcionarios, expertos, investigadores en contratación (juristas, economistas, polítogos, etc.) vean el potencial y utilidad del data science (ciencia de datos) y el ML. Para conseguir este fin, el autor ha impartido conferencias, dado entrevistas en medios de comunicación y presentado sus aplicaciones a premios.

Todos los objetivos se han cumplido satisfactoriamente. En la actualidad se sigue con la labor de divulgación en foros especializados, con una cálida acogida por los expertos en contratación pública, principalmente juristas que son los que tradicionalmente han liderado esta disciplina.

A continuación, se resume la **metodología general** llevada a cabo en los artículos de la Tesis. Estas etapas son características del data science. Cada artículo ha requerido un desarrollo software ad hoc, realizado en el lenguaje de programación Python [3, 40, 41]. Para el detalle particular metodológico, léase el capítulo 4 donde se explica cada artículo.

1. **Formulación del problema y enfoque analítico**. Se sientan las bases, hipótesis, del objeto de estudio y se plantea el problema a resolver. En el enfoque analítico se identifica cuál sería el procedimiento que nos puede ayudar para obtener los resultados esperados.
2. **Adquisición de las fuentes de datos** (de contratación pública y otras fuentes asociadas). En general, se han utilizado datos públicos (PLACSP, TED) pero también datos no públicos (datos empresariales del Registro Mercantil y bases de datos de licitaciones colusivas).

3. **Almacenamiento** y estructuración de los datos. Se almacenan y estructuran los datos para poder ser analizados.

4. **Tratamiento de los datos**. Extracción, limpieza, transformación y filtrado de los datos. Enriquecimiento de los datos con fuentes complementarias cuando fuese posible.

5. **Comprensión de los datos**. Es decir, un análisis descriptivo y gráfico de los datos, no sólo para entenderlos sino también para conocer las carencias existentes. Se hace una verificación para asegurar la integridad y calidad de lo datos que serán utilizados en la siguiente etapa. Esta etapa se itera, retroalimenta, a la etapa anterior.

6. **Desarrollo del modelo de predicción** (regresión o clasificación) utilizando algoritmos de ML. Dos de los artículos predicen el importe de adjudicación de la licitación. Otro artículo predice un grupo de empresas que pueden llevar a cabo la licitación. Otro artículo clasifica la licitación en competitiva o colusiva (anticompetitiva, los ofertantes forman un cártel).

7. **Validación del modelo**. Se evalúan los algoritmos de ML mediante unas métricas de error (definidas en el capítulo 3) para seleccionar aquellos que más aciertan en la predicción.

8. **Discusión de los resultados y conclusiones**.

Capítulo Tercero

Fundamentos de la investigación

3.1. La contratación pública

Actualmente la contratación pública en España es un área mucho más digital y cuantitativa que hace años. Es decir, los contratos tienen su información más relevante accesible digitalmente para su descarga masiva, de manera estructurada, pudiendo ser analizados dichos datos con herramientas propias de las disciplinas cuantitativas o científicas. Esta cuantización es debida a varios factores de origen legal:

- El primer paso básico es la digitalización de AA.PP. La Ley 39/2015 del Procedimiento Administrativo Común de las AA.PP. y la Ley 40/2015 de Régimen Jurídico del Sector Público vienen a configurar un escenario en el que la tramitación electrónica debe constituir la actuación habitual de las AA.PP.
- La LCSP unifica la publicidad y transparencia de los contratos públicos. También alude a que la publicación ha de realizarse en formatos abiertos y reutilizables, para que el ingente volumen de información pueda ser manejada por terceros.
- Transparencia y acceso a datos públicos por parte de los ciudadanos. La Ley 19/2013 de Transparencia, Acceso a la Información Pública y Buen Gobierno amplía y refuerza la transparencia de las AA.PP., regula y garantiza el derecho de acceso a la información relativa a sus actividades y establece las obligaciones de buen gobierno.
- La disposición de datos públicos crea un mercado que incentiva a empresas y personas a crear herramientas reutilizando dichos datos, obteniendo un beneficio económico por ello. Esto lo articula la Ley 18/2015 sobre la reutilización de la información del sector público.

Por tanto, estos cuatro factores legales posibilitan unas AA.PP. digitalizadas, publicidad y transparencia en la contratación, libre acceso a los datos de contratación y la posibilidad de reutilización de dichos datos con fines comerciales. La naturaleza descentralizada de la contratación en España, su gran dispersión y todavía el uso del papel en sus procesos no facilita el análisis digital y masivo de los datos. Sin embargo, si se echa la vista atrás en el tiempo el avance ha sido espectacular. Por ejemplo, en el año 2000 la manera de consultar las licitaciones era leyendo los Boletines Oficiales de las CC.AA. y del Estado. Hoy en día se pueden consultar de manera masiva y estructurada en internet a través de las diferentes plataformas de contratación autonómicas y nacional. El procesamiento masivo de datos se puede asociar al concepto tan manido de Big Data [42].

La contratación es ya un área cuantitativa como demuestra PLACSP (se explicará posteriormente en detalle), especialmente a partir de 2018. En 2021 se han publicado aproximadamente

200.000 licitaciones (considerando sólo perfiles del contratante en PLACSP y excluyendo contratos menores) y tiene dados de alta unos 20.500 órganos de contratación. Si se quiere conseguir una contratación moderna, eficiente y que esté bien controlada y supervisada, los organismos públicos que tienen asociadas estas responsabilidades deberían contar con equipos multidisciplinares. Desde juristas y economistas hasta ingenieros e informáticos que crean los sistemas (programas informáticos, aplicaciones web, bases de datos, etc.) que permiten analizar los datos de contratación de manera automática.

Las AA.PP. son cada vez más conscientes de la importancia del uso de la información para aumentar la eficiencia en sus procesos, ahorrar costes, aumentar la calidad de los servicios que prestan y para su toma de decisiones. La contratación genera un gran volumen de información, es decir, datos estructurados pero también documentos con información no estructurada (pliegos técnicos, administrativos, resoluciones de adjudicaciones, etc.). Es muy difícil de manejar ese volumen ingente de información por una persona u organismo que no tenga herramientas software especializadas.

3.2. Legislación en la contratación pública

Tanto en el ámbito europeo como en el español, se han desarrollado leyes relacionadas con la reutilización de la información del sector público y la contratación en el sector público que se resumen en la Tabla 3.1. Un tema particularmente importante en la contratación son las medidas anticorrupción, para así dificultar y frenar actuaciones fraudulentas, cuya normativa europea es objeto de la Tesis Doctoral de Javier Miranzo Díaz [43].

En el ámbito español, según la Ley 20/2013, disposición adicional tercera, y la LCSP, artículo 347, en la web de PLACSP se deben publicar todas las convocatorias de licitaciones y sus resultados por parte de todos los órganos de contratación que pertenecen al Sector Público estatal. Según la LCSP, las CC.AA. podrán optar por publicar sus perfiles del contratante a través de sus propios servicios de información o directamente en PLACSP. Si optaran por publicar su información contractual en medios ajenos a PLACSP, deberán publicar en la misma mediante mecanismos de agregación las publicaciones de las licitaciones y sus resoluciones. Las Administraciones locales, así como sus entidades vinculadas o dependientes, podrán optar por alojar la publicación de sus perfiles de contratante en el servicio de información que a tal efecto estableciera la Comunidad Autónoma de su ámbito territorial, o bien por alojarlos en PLACSP Si optaran por alojarlos en el servicio de la Comunidad Autónoma, esta deberá publicar mediante mecanismos de agregación las convocatorias de licitaciones y sus resultados en PLACSP.

La información mínima que deben contener los anuncios de las licitaciones está definida en la LCSP, Anexo III *"Información que debe figurar en los anuncios"*. PLACSP dispone de un apartado de Open Data para la reutilización de información de las licitaciones publicadas, en cumplimiento de las obligaciones de publicidad establecidas en la LCSP.

Respecto al anuncio oficial de las licitaciones que requieren publicidad fuera del ámbito español, el artículo 135 de la LCSP establece que cuando las licitaciones estén sujetas a regulación armonizada (es decir, aquellas licitaciones con un importe mayor a un determinado umbral o con ciertas características, estipuladas en los artículos del 19 al 23 de la LCSP), la licitación deberá publicarse, además, en el Diario Oficial de la Unión Europea (DOUE). Cuando el órgano de contratación lo estime conveniente, se podrán anunciar en el DOUE las licitaciones no sujetas a regulación armonizada.

La UE tiene un portal oficial de datos abiertos que se creó en 2012, de conformidad con la

Ley	Descripción	Ámbito	Enlace a la legislación
Directiva 2014/24/EU	Contratación pública	Europa	http://data.europa.eu/eli/dir/2014/24/oj
Directiva 2014/23/EU	Adjudicación de contratos de concesión	Europa	http://data.europa.eu/eli/dir/2014/23/oj
Directiva 2014/25/EU	Contratación por entidades que operan en los sectores del agua, la energía, los transportes y los servicios postales	Europa	http://data.europa.eu/eli/dir/2014/25/oj
Ley 9/2017	Transposición al Derecho español de las anteriores Directivas europeas 2014/23/UE y 2014/24/UE	España	https://boe.es/eli/es/l/2017/11/08/9
Directiva 2003/98/EC	Reutilización de la información del sector público	Europa	http://data.europa.eu/eli/dir/2003/98/oj
Directiva 2013/37/EU	Modificación de la anterior Directiva 2003/98/CE	Europa	http://data.europa.eu/eli/dir/2013/37/oj
Directiva 2007/2/EC	Establecimiento de una Infraestructura de Información Espacial en la Comunidad Europea (INSPIRE)	Europa	http://data.europa.eu/eli/dir/2007/2/oj
Ley 37/2007	Transposición al Derecho español de la Directiva Europea 2003/98/CE	España	https://boe.es/eli/es/l/2007/11/16/37
Real Decreto 1495/2011	Desarrollo de la Ley española 37/2007	España	https://boe.es/eli/es/rd/2011/10/24/1495
Decisión de la CE 2011/833/EU	Sobre la reutilización de los documentos de la CE	Europa	http://data.europa.eu/eli/dec/2011/833
Ley 19/2013	Transparencia, acceso a la información del sector público y buen gobierno	España	https://boe.es/eli/es/l/2013/12/09/19
Ley 20/2013	Garantía de la unidad de mercado	España	https://boe.es/eli/es/l/2013/12/09/20
Ley 18/2015	Transposición al Derecho español de la Directiva Europea 2013/37/UE	España	https://boe.es/eli/es/l/2015/07/09/18

Tabla 3.1: Legislación sobre contratación pública y el uso de datos.

Decisión de la CE 2011/833/EU sobre la reutilización de documentos de la CE, e invita a todas las instituciones europeas a que pongan sus datos a disposición del público siempre que sea posible. Además, existe un portal denominado llamado TED dedicado a la contratación pública europea. En este portal, se publican diariamente todos los anuncios de licitación y adjudicación, además de anuncios de información previa, para contratos sujetos a regulación armonizada (S.A.R.A.), es decir, que son susceptibles de ser de interés transfronterizo y para los cuales se debe garantizar la libre concurrencia de operadores de todos los estados miembro de la UE. En cuanto a la reutilización de los datos de TED de cara a producir informes sobre el estado de la contratación pública en Europa, es la Dirección General de Comercio (DG GROW) la que se encarga de la elaboración de un informe de indicadores anuales sobre contratación pública[1] y el mantenimiento de un portal de indicadores comparativos a nivel europeo[2]. Además, se realizan estudios ad-hoc[3] sobre el estado del arte de la contratación que tienen como objetivo influir en la elaboración de políticas públicas, proporcionando un mayor conocimiento sobre el mercado público y su funcionamiento.

Queda fuera del alcance de la Tesis explicar los principios generales del derecho de la contratación pública internacional y la comparación de las legislaciones nacionales (el llamado Derecho comparado). Esta legislación define el sistema de contratación de cada país y, por tanto, debería articular los mecanismos de publicación de los datos de los contratos públicos (transparencia). Para ahondar en esta complejo tema jurídico se puede consultar [44].

Un debate que está en boga en el ámbito jurídico es la ética de los algoritmos de IA en su funcionamiento, produciendo sesgos o búsquedas dirigidas [45]. La realidad es que los algoritmos aplicados a la contratación están en un estado incipiente, poco evolucionado. Por lo tanto, la ética de la IA en este campo está de momento en un plano teórico, filosófico, hasta que dichos algoritmos no se sofistiquen y tengan un comportamiento más complejo, más humano.

3.3. Tecnologías en la contratación pública

A continuación, se enumeran una serie de relevantes tecnologías software de reciente desarrollo o difusión en el ámbito de las AA.PP. Son la base tecnológica para aumentar la eficiencia de la contratación, facilitando nuevas herramientas que permitan dar un salto cuantitativo y cualitativo. Se han agrupado en 3 bloques: data, IA y automatización de procesos:

- **Data:**
  - **Procesamiento masivo de datos (Big Data)**. Se ocupa de las actividades relacionadas que manipulan grandes conjuntos de datos o que requieren gran capacidad de computación: extracción masiva de datos, almacenamiento y búsqueda, software optimizado para trabajar en paralelo con gran cantidad de información, computación en la nube, etc... El límite de procesamiento ha ido creciendo a lo largo de los años, no es un concepto estático. Los ordenadores de consumo actuales son supercomputadores si se comparan con los de hace 25 años. Los sistemas informáticos de contratación no son especialmente intensivos en el almacenamiento o intercambio de datos como pueden ser otros (banca, compañías telefónicas, empresas de contenidos audiovisuales bajo demanda, etc...). Por tanto, las tecnologías software llamadas Big Data no son estrictamente

[1] La última versión de los indicadores aporta información sobre la contratación del año 2017: https://ec.europa.eu/docsroom/documents/38003

[2] Accesible en https://ec.europa.eu/internal_market/scoreboard/performance_per_policy_area/public_procurement/index_en.htm

[3] Accesibles en https://ec.europa.eu/growth/single-market/public-procurement/studies-networks_en

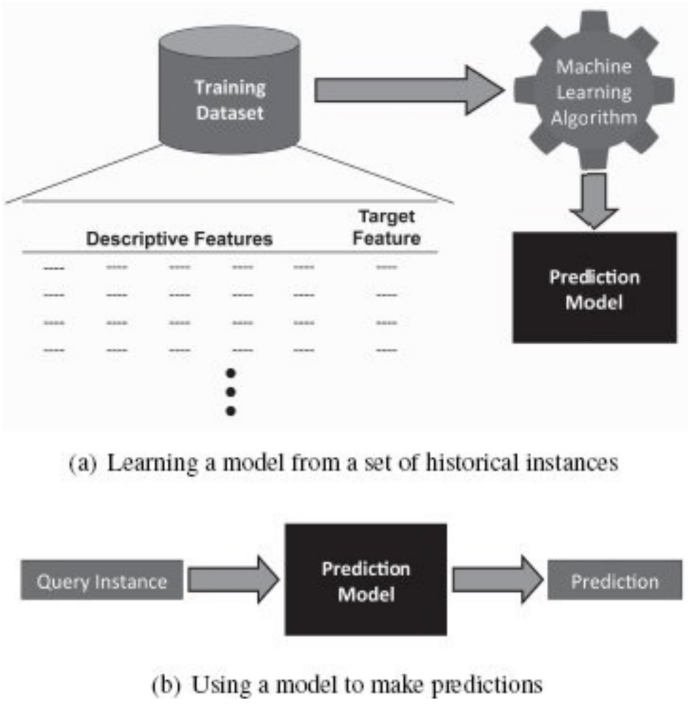


Figura 3.1: Funcionamiento de un algoritmo de ML: entrenamiento (a) y uso (b). Fuente [46].

necesarias en contratación por el volumen de información manejada, de momento, pero sí recomendables para tener un sistema escalable que funcione con una mayor cantidad de datos.

- **Análisis de datos (Data Analytics)**. Es un proceso que consiste en inspeccionar, limpiar y transformar datos con el objetivo de que un usuario pueda visualizar y comprender la información para detectar o resaltar información útil, dando lugar a conclusiones que apoyen la toma de decisiones. Ejemplos comerciales de este tipo de herramientas software son Microsoft Power BI, Tableau o Kibana. En nuestro caso, se pueden utilizar para transformar los datos de contratación y crear cuadros de mando con indicadores y gráficas para ayudar en la operación y supervisión de la contratación.

• **Inteligencia Artificial (IA):**

- **Machine Learning (ML)**. Son algoritmos informáticos para el descubrimiento de nuevo conocimiento a partir de grandes cantidades de datos. Es decir, es un proceso automático para extraer patrones de los datos [46] (ver Figura 3.1). El "aprendizaje" entra en juego cuando damos a estos algoritmos parámetros adaptables a los datos observados (el programa está "aprendiendo" de los datos). Una vez que estos modelos se han entrenado, ajustado, a esos datos, se pueden usar para predecir y comprender aspectos de otros datos nuevos [3]. Se considera una de las innovaciones más disruptivas y un fuerte factor para crear ventajas competitivas. Si bien el ML ha existido durante más de 60 años, recientemente ha mostrado un potencial significativo para cambiar las economías y sociedades [47]. Es la principal tecnología usada en esta Tesis.
- **Procesamiento del Lenguaje Natural (NLP)**. Se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio del lenguaje natural. En la contratación pública una gran can-

tidad de información (con mucho valor) se almacena en documentos no estructurados (pliegos, actas, memorias, etc.), en lenguaje natural. Esta tecnología es especialmente relevante para obtener información estructurada/clasificada a partir de textos, comprensión del lenguaje natural por parte de la máquina (NLU) para poder hacer búsquedas por tópicos y no por palabras exactas, respuestas elaboradas por una máquina a preguntas de una persona (chatbot), traducción automática (textos en idiomas distintos), etc.

- **Automatización de Procesos:**
  - **Minería de Procesos (Process minning)**. Análisis de los procesos de la información utilizando el registro de eventos entre las distintas etapas del flujo de información. Esta tecnología permite analizar la traza de los procesos en estudio, incluyendo información de los actores que lo realizan, los tiempos involucrados, volumen que pasa por cada etapa, cuellos de botella, los caminos típicos que se recorren, etc. Uno de los objetivos es llevar el control de los procesos, pero además tiene como objetivo permitir el descubrimiento y optimización de procesos, controles, información y estructuras organizacionales partiendo de dichos registros de eventos. Los sistemas informáticos de contratación pública son buenos destinatarios de esta tecnología, les ayudaría a mejorar sus procesos de licitación e, incluso, etapas posteriores: seguimiento del contrato, supervisión y facturación.
  - **Automatización Robótica de Procesos (RPA)**. Es la automatización de los procesos/etapas de negocio que realiza una máquina replicando las acciones que hace una persona interactuando con la interfaz de usuario (IU) del sistema informático. Es decir, el robot software opera en la IU de la misma manera que un ser humano. Esto es una diferencia significativa a las formas tradicionales que se basan en Interfaces de Programación de Aplicación (API). El RPA es una tecnología que aplicada convenientemente permitiría aumentar la eficiencia de los órganos de contratación, automatizando validaciones, cálculos o tareas repetitivas en las distintas etapas del proceso de contratación.

3.4. Aplicaciones en la contratación pública

En este apartado se explican una serie de casos de uso relevantes para la contratación utilizando las nuevas tecnologías anteriormente descritas. Se ha dividido en 3 bloques (ver Figura 3.2): fuentes de datos, procesamiento y almacenamiento y posibles aplicaciones.

1. **Fuentes de datos**. El primer paso es disponer de un gran repositorio de información de contratación que permita tener una visión completa. Para ello se identifican diferentes fuentes de datos, tanto públicas como privadas, que pueden contribuir a generar dicho repositorio. Se pueden dividir en tres grupos según su origen:
   - **D1. Datos públicos de contratación**. Principalmente, en España serían las plataformas de contratación de las CC.AA., la nacional PLACSP y la DGRCC (gestiona acuerdos marco, sistemas dinámicos de adquisición y contratos centralizados). A nivel europeo, la plataforma de contratación de los países miembros es TED.
   - **D2. Otros datos públicos relacionados con contratación**. Por ejemplo, resoluciones de la Junta Consultiva de Contratación Pública del Estado, el Tribunal Administrativo Central de Recursos Contractuales, la Central de Información Económico-Financiera de las AA.PP., el Portal de Transparencia de la AGE, el Registro de Contratos o la Base Nacional de Subvenciones para rastrear las ayudas concedidas a personas jurídicas o físicas.

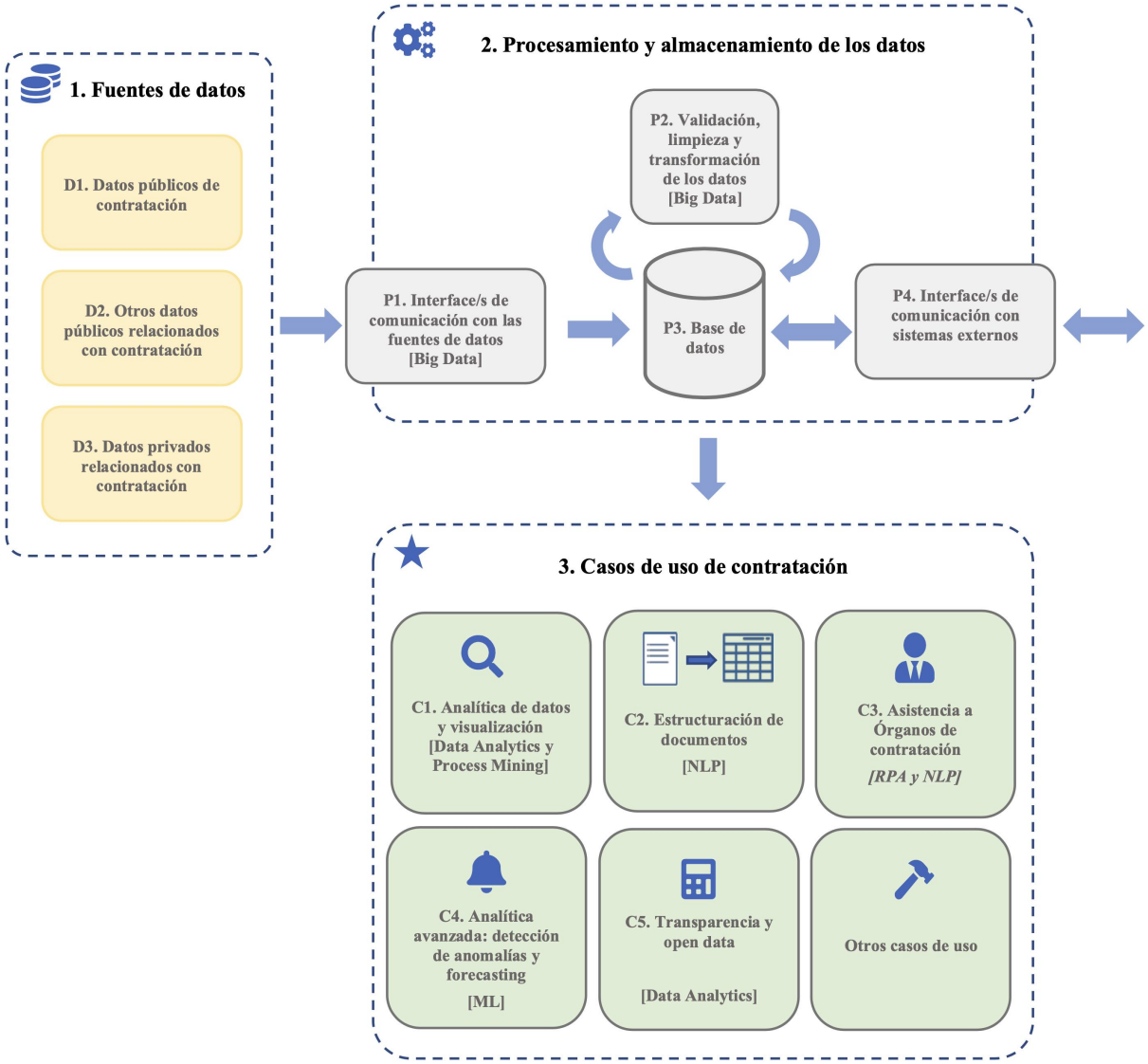


Figura 3.2: Diagrama de posibles aplicaciones para la contratación pública.

- **D3. Datos privados relacionados con contratación**. Por ej., información operacional de acceso privado de las plataformas de contratación antes mencionadas o los datos almacenados en el Registro Mercantil para obtener información empresarial y económica de los adjudicatarios.

2. **Procesamiento de los datos de contratación**. Los componentes fundamentales serían:

   - **P1. Interface de comunicación con las fuentes de datos**. La función principal es la automatización de la ingesta de los datos. Para las fuentes de tipo D1 serían los datos estructurados de la licitación y los documentos de la licitación (pliegos técnicos y administrativos, anexos, etc.).
   - **P2. Validación, limpieza y transformación de los datos**. Hay que validar la coherencia e integridad de los datos, limpiarlos y enriquecerlos gracias a las distintas fuentes. En este proceso se harían todas las tareas necesarias para normalizar y transformar los datos.
   - **P3. Base de datos**. Almacena la información recibida de las distintas fuentes (datos brutos) y, tras procesarlos en el componente P2, almacena también los datos netos (transformados, enriquecidos) que se utilizarán en los diferentes casos de uso. De esta manera, se unifica toda la información en un único repositorio. En cuanto a la tecnología, puede ser desde una base de datos de tipo SQL tradicional a específicas de Big Data si el volumen de datos lo requiere.
   - **P4. Interface de comunicación con sistemas externos**. En el futuro la intercomunicación de los sistemas será mucho mayor que en la actualidad y se crearán mayores sinergias por la interconexión entre los diferentes organismos con competencias en contratación.

3. **Aplicaciones**. Gracias al procesamiento y almacenamiento de los datos realizado en la fase anterior, ya se tiene la infraestructura necesaria para llevar a cabo los casos de uso. A continuación se ponen varios ejemplos aunque el abanico de casos es muy grande (más ejemplos en [48]), tantos como necesidades tengan los involucrados de la contratación.

   - **C1. Analítica de datos y visualización**. Visualización de información analítica de contratación (tablas, gráficos, indicadores, informes): análisis geográfico de los lugares de ejecución de las licitaciones, sectores con mayor contratación, plazos en la resolución de la adjudicación, competencia empresarial (número de empresas que participan en una licitación), características de las empresas que más licitaciones ganan (tamaño, origen, sector principal de negocio, tipo de sociedad mercantil, etc.), precios máximos y mínimos de las ofertas, desviación del importe de licitación respecto al importe de adjudicación (ahorro obtenido), cómo y qué contratan los órganos contratantes, impacto de la división en lotes de los contratos (y la fragmentación ilícita en contratos menores), comparación de parámetros de contratación con años pasados y con otras Administraciones, etc.
   - **C2. Estructuración de documentos**. El proceso de contratación en gran parte se basa en la gestión documental. Esto supone que la información relevante se encuentra en documentos (pliegos, formularios, anexos, etc.) de manera desestructurada. Los documentos deben ser correctamente interpretados, clasificados y estructurados para almacenarse en la base de datos y ser explotados posteriormente. De esta manera se tiene una visión completa y se podrán crear aplicaciones que ayuden en el análisis y toma de decisión.

Figura 3.3: Web de la Plataforma de Contratación del Sector Público (PLACSP).

- **C3. Asistencia a órganos de contratación**. Los órganos de contratación suelen trabajar de manera independiente sin mucha interrelación, realizando consultas a otros departamentos de asistencia en materia de contratación. Es decir, en las AA.PP. suele haber una fragmentación de la contratación coexistiendo muchos órganos que se enfrentan licitaciones que nunca antes había redactado ni tienen herramientas comunes para trabajar colaborativamente. Ejemplos de aplicaciones que les ayuden serían la automatización de las etapas en el proceso de contratación (RPA), chatbot para dar respuestas a preguntas recurrentes, herramientas colaborativas para reaprovechar el conocimiento de los órganos, generación automatizada de un repositorio común de proveedores u otra información de interés, etc.
- **C4. Analítica avanzada: detección de anomalías y forecasting**. El control de la contratación se hace generalmente a partir de la revisión de expedientes de contratos individuales, lo que dificulta la detección de eventuales prácticas fraudulentas. No se pueden hacer estimaciones a futuro por no tener herramienta que analicen de manera agregada las licitaciones. Sin embargo, un tratamiento masivo de la información disponible de contratación permite hacer forecasting, identificar patrones y detectar anomalías que ayuden a identificar prácticas irregulares. Los artículos de investigación realizados en esta Tesis son precisamente de este tipo, son buenos ejemplos del uso de algoritmos de IA y ML.
- **C5. Transparencia y open data**. Hay pocos portales de transparencia donde se muestre información detallada de contratación (tanto a nivel agregado como desagregado) a los ciudadanos como ejercicio de transparencia y rendición de cuentas. Para resolver esto, habría que simplificar el primer caso de uso (C1), mostrando datos y visualizaciones generales que se pueden descargar libremente para que ciudadanos o usuarios externos a las AA.PP. puedan hacer sus propios análisis e investigaciones.

3.5. La Plataforma de Contratación del Sector Público (PLACSP)

El matemático Clive Humby afirmó que *"Los datos son el nuevo petróleo. Los datos son valiosos pero tienen que refinarse como el petróleo, si no carecen de utilidad"*. La contratación pública genera un gran volumen de información, datos en crudo que necesitan refinarse para aportar utilidad. La mayor plataforma de datos de contratación en España es PLACSP (ver Figura 3.3), así como en la UE es el TED. En los siguientes subsecciones se detallará PLACSP desde una perspectiva del origen de los datos, su formato y su uso.

3.5.1. Órganos de contratación que publican en PLACSP

La PLACSP, regulada en el artículo 347 de la ley LCSP, es: *"...una plataforma electrónica que permita la difusión a través de Internet de sus perfiles de contratante, así como prestar otros servicios complementarios asociados al tratamiento informático de estos datos"*. Este mismo artículo define a la PLACSP como un instrumento de publicidad obligatorio para todo el Sector Público Estatal. Define la posibilidad que tienen las CC.AA. y las Ciudades Autónomas para establecer servicios de información similares pero que deberán también publicar las convocatorias de licitaciones y sus resultados en PLACSP. Finalmente, en el último párrafo del punto 3 del artículo 347, se indica dónde deben estar alojados los perfiles de contratante de las administraciones locales y sus entidades vinculadas o dependientes.

En resumen, se establecen tres ámbitos con distintos niveles de obligatoriedad según la legislación vigente de la LCSP:

- **Sector Público Estatal**. Todos los órganos de contratación están obligados a publicar su perfil del contratante en PLACSP, lo que incluye todos los documentos e informaciones referentes a su actividad contractual, y en concreto: anuncios de licitación, pliegos, anuncios de adjudicación, anuncios de formalización, anuncios de renuncia o desistimiento u otros documentos.
- **Comunidades Autónomas (CC.AA.) y Ciudades Autónomas**. Todos los órganos de contratación, así como sus entes, organismos y entidades vinculadas deberán alojar sus perfiles de contratante en el sistema de información que disponga la CC.AA. o Ciudad Autónoma y comunicar a PLACSP mediante el mecanismo de agregación un conjunto reducido de datos sobre la publicación de la convocatoria y su resolución. En caso de no disponer de dicho sistema de información, los perfiles de contratante deberán alojarse en PLACSP.
- **Administraciones locales**. Pueden optar por adherirse a PLACSP o a la que disponga su CC.AA. o Ciudad Autónoma.

Las licitaciones publicadas en PLACSP aumentaron significativamente a partir de 2018 debido a la entrada en vigor de la LCSP en 2017. Progresivamente se han ido incorporando más órganos de contratación a dicha plataforma.

3.5.2. Origen de los datos de PLACSP

La PLACSP tiene varios conjuntos de datos abiertos referentes a las licitaciones publicadas en el portal de transparencia del Ministerio de Hacienda. Todos estos conjuntos de datos abiertos son accesibles[4] y diariamente se publican las actualizaciones del día anterior. Concretamente, se publican 3 conjuntos de datos abiertos (ficheros XML con extensión .atom[5]):

1. **Sector Público**. Expedientes de contratación publicados en los perfiles del contratante ubicados en PLACSP, excluyendo los contratos menores.
2. **Agregadas**. Licitaciones publicadas en PLACSP mediante mecanismos de agregación, excluyendo los contratos menores.

[4] Accesible en https://www.hacienda.gob.es/es-ES/GobiernoAbierto/Datos%20Abiertos/Paginas/licitaciones_plataforma_contratacion.aspx

[5] La estructura del fichero XML se conforma siguiendo las especificaciones descritas para un "Atom Feed Document" en la RFC 4287 de la IETF. Accesible en https://www.ietf.org/rfc/rfc4287.html

3. **Menores del Sector Público**. Contratos menores publicados en los perfiles del contratante ubicados en PLACSP.

Las administraciones autonómicas que opten por mantener los perfiles del contratante en su propia plataforma de contratación deberán publicar las convocatorias de licitación y sus resultados en PLACSP mediante mecanismos de agregación. En la actualidad, las plataformas de contratación que van al conjunto de datos 2 (agregadas) son las 7 siguientes CC.AA.: Cataluña, País Vasco, La Rioja, Comunidad de Madrid, Galicia, Andalucía y Navarra. El resto de CC.AA., incluyendo a Ceuta y Melilla, están alojadas en PLACSP. Es decir, van al conjunto de datos 1 (Sector Público) o 3 (contratos menores del Sector Público). Hay más organismos públicos (Administración General del Estado, Mutuas, Entidades Locales, Universidades, etc.) que también están integrados en sendos conjuntos de datos. El artículo 63 de la LCSP en su apartado cuarto establece que la información del perfil de contratante se publicará en formatos abiertos y reutilizables, permaneciendo accesible al público durante un tiempo no inferior a 5 años.

3.5.3. Formato y calidad de los datos en abierto de PLACSP

Todo el detalle referente al formato de los ficheros de datos abiertos y a los campos disponibles en las licitaciones se puede consultar en el *"Formato de sindicación y reutilización de datos sobre licitaciones publicadas en la Plataforma de Contratación del Sector Público"*, siendo su última versión[6] de fecha 19/10/2021. Dicho documento lo elabora la Subdirección General de Coordinación de la Contratación Electrónica (D. G. del Patrimonio del Estado). Para más detalle, consultar el documento *"Resumen de contenido en conjuntos de datos abiertos"* donde especifica más de 100 campos disponibles[7] que se agrupan dentro de las siguientes áreas: datos generales del expediente, lugar de ejecución, lotes, procesos de licitación, entidad adjudicadora, plazo de obtención de pliegos, extensión del contrato, condiciones de licitación, garantías requeridas, requisitos de participación, criterios de evaluación técnica y económica-financiera, subcontratación permitida, criterios de adjudicación, limitación del número de licitadores a invitar, información sobre el contrato, adjudicatario, importe de adjudicación, condiciones de subcontratación, justificación del proceso, modificaciones del contrato, publicaciones oficiales y otros documentos publicados.

La información sobre los expedientes de licitación se van agregando de forma incremental utilizando la arquitectura Componentes y Documentos Interoperables para la Contratación Electrónica (CODICE)[8]. Ésta proporciona una biblioteca de componentes estándar, reutilizables, y extensibles o adaptables a diversos contextos o necesidades de contratos públicos específicos, para satisfacer las necesidades de información de los documentos y mensajes intercambiados a lo largo del ciclo completo de los procedimientos electrónicos de contratación. Las especificaciones CODICE están incluidas en el catálogo de estándares del Esquema Nacional de Interoperabilidad.

Tras manejar los datos de PLACSP durante los últimos 5 años para poder elaborar esta Tesis, se ha llegado a la conclusión que PLACSP no tiene mecanismos para validar el formato y coherencia de los datos que se introducen en cada licitación. La plataforma es un contenedor de información, siendo los órganos de contratación los únicos responsables legales en la veracidad y completitud de la información de las licitaciones que suben a la plataforma. Este extremo se menciona por la OIReScon [21] y ha sido confirmado por los funcionarios responsables de PLACSP[9]. No se ha

[6] Accesible en https://contrataciondelsectorpublico.gob.es/datosabiertos/especificacion-sindicacion.pdf

[7] Accesible en https://contrataciondelsectorpublico.gob.es/datosabiertos/DGPE_PLACSP_ResumenDatosAbiertos.pdf

[8] Más información de CODICE en https://contrataciondelestado.es/wps/portal/codice

[9] María Lafuente Fernández, Subdirectora General Adjunta de Coordinación de la Contratación Electrónica y

encontrado legislación que castigue o penalice a los órganos de contratación que cometan errores en los datos publicados de sus licitaciones. Este hecho será muy dañino para los investigadores que traten estos datos. Para más detalle sobre la transparencia y los datos en abierto de la PLACSP y las plataformas autonómicas de contratación, consultar [21].

Como se acaba de decir, que un campo exista (mencionados anteriormente) no significa que se haya rellenado, que tenga el formato correcto o que el valor se haya introducido correctamente. Todos estos casos se han encontrado en PLACSP, especialmente campos no rellenados, y consecuentemente ha dificultado la realización de los artículos de esta Tesis. Ciertos campos deben tener una estructura fija (NIF, código postal, fecha, etc,) pero algunos valores no siguen dicha estructura. Además, se han detectado valores anómalos. Por ejemplo, que el número de ofertantes tenga un valor altísimo (más de 1.000 ofertantes, suceso extremadamente improbable) o que el importe de adjudicación sea muy superior al de licitación (50 veces superior al de licitación). Consecuentemente, la calidad de los datos en PLACSP es baja, se deben de "limpiar" (data cleaning explicado en [10]) para conseguir estudios rigurosos, sin información errónea que desvirtúe el análisis de datos agregados.

Hay que dotar a PLACSP de mecanismos de validación automático cuando el funcionario del órgano de contratación introduzca los datos de la licitación. Es comprensible los errores humanos y muchos de ellos se podrían corregir de manera sencilla. Por otro lado, se deberían de incluir en las leyes de contratación (LCSP y afines) una regulación sobre la calidad del dato y las responsabilidades legales que se deriven si los órganos de contratación las incumplen.

La falta de calidad del dato, interoperabilidad, poca estandarización, etc. no es exclusivo de PLACSP o del ámbito de la contratación, son carencias generalizadas en las AA.PP. españolas. Para paliar esto se crea la Oficina del Dato[10] en 2020, dependiente de la Secretaría de Estado de Digitalización e IA, siendo sus principales líneas estratégicas *"el diseño, coordinación y seguimiento del modelo de referencia arquitectónico para fomentar la recolecta, gestión e intercambio de datos públicos"*. España necesita progresar mucho en esta materia y esta oficina jugará un papel capital si se le dota del personal y recursos adecuados. Hay que tomar referentes como, por ejemplo, el Reino Unido que publicó una guía gubernamental de calidad del dato[11] en 2020.

3.5.4. Programa OpenPLACSP para analizar los datos en abierto

La Subdirección General de Coordinación de la Contratación Electrónica ha puesto a disposición del público en el verano de 2021 el programa OpenPLACSP[12] (ver captura de pantalla en la Figura 3.4) para facilitar la transformación de los ficheros de datos abiertos en una tabla (hoja de cálculo). Con esta herramienta se pretende que cualquier interesado pueda trabajar de una forma rápida y sencilla con las licitaciones publicadas. Esto no era así hasta la aparición de este programa. Se debían de tener conocimientos de programación para convertir los ficheros XML (.atom) a una base de datos u hoja de cálculo y poder explotar los datos. Como los artículos de esta Tesis que utilizan datos de PLACSP son de fecha anterior al verano de 2021, se tuvo que desarrollar un script ad-hoc para convertir dicheros ficheros .atom.

encuadrada en la D.G. de Patrimonio del Estado (Ministerio de Hacienda y Función Pública), como responsable de PLACSP mencionó esta cuestión en el encuentro *"Contratación Electrónica Inteligente: situación y retos"* en el Instituto Nacional de Administración pública celebrado el día 20/1/2022.

[10]Orden ETD/803/2020 publicada en el BOE: https://www.boe.es/diario_boe/txt.php?id=BOE-A-2020-10008

[11]"The Government Data Quality Framework: principles, guidance and case studies". Accesible en https://www.gov.uk/government/publications/the-government-data-quality-framework

[12]Accesible para su descarga en la web de PLACSP en el menú "Datos abiertos".

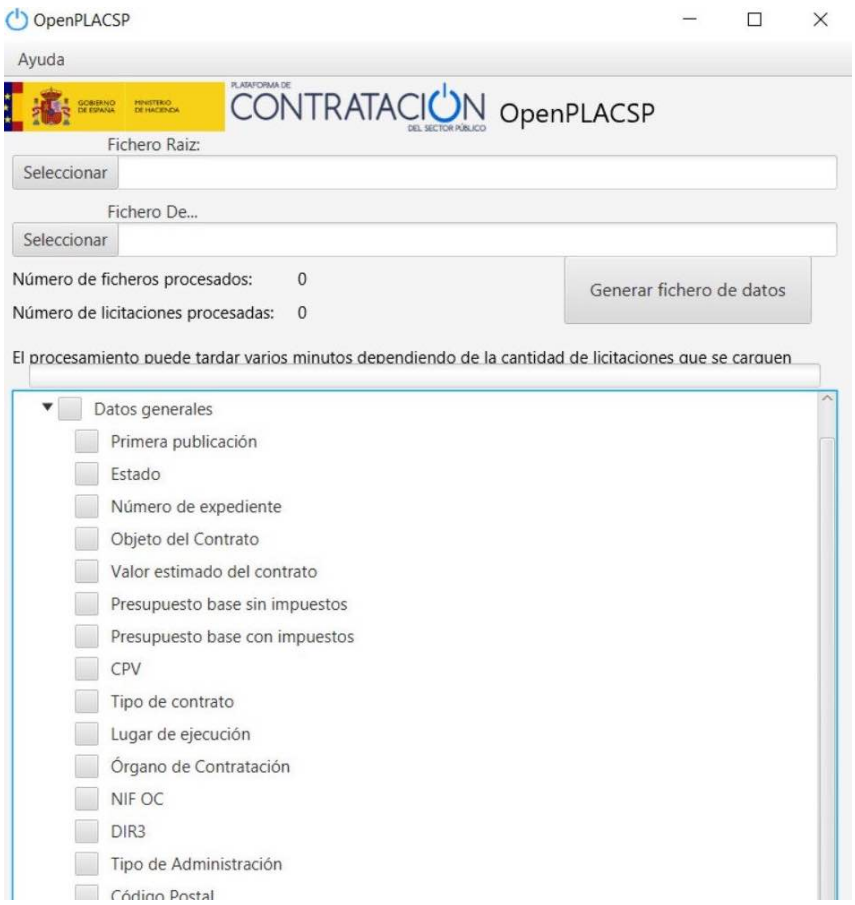


Figura 3.4: Programa OpenPLACSP para extraer licitaciones de Plataforma de Contratación del Sector Público (PLACSP).

Este programa es un gran avance para organizaciones, investigadores o ciudadanos que quieran realizar análisis y estudios de contratación sin tener conocimientos de programación. Es un elemento necesario para que cada vez haya más estudios jurídico-económico de tipo cuantitativo. ¿Qué tipo de procedimiento produce mayores bajas económicas? ¿Y cuántas ofertas se tienen de media? ¿Favorece la competencia o no? Se podrían hacer un sinfín de preguntas.

3.6. Literatura académica sobre la contratación pública

Las tecnologías como el ML y el big data se pueden (y deben) aplicar al campo económico por cosechar éxitos notables [49, 50, 51]. Esta Tesis particulariza su aplicación al ámbito de la contratación, campo muy relevante por su peso en el PIB y por ser un gasto público, que concierne a toda la sociedad. Hay abundante bibliografía para conocer en detalle los algoritmos de ML y la matemática que hay detrás de ellos [52, 53]. Sin embargo, no hay tantas referencias sobre su aplicación en la contratación por ser un campo todavía poco explorado e investigado desde el ámbito científico-tecnológico.

En las Tablas 3.2 y 3.3 se recopila una parte de los artículos consultados para llevar a cabo la investigación, sin pretender hacer una revisión exhaustiva de la literatura. Esto se debe a que cada artículo publicado en esta Tesis tiene su correspondiente revisión de la literatura, con más referencias bibliográficas. La mayoría de los artículos consultados se han publicado en los últimos 6 años. No solo en el campo académico, también organismos como la CNMC empiezan a utilizar algoritmos de IA para detectar prácticas anticompetitivas mediante el procesamiento masivo de datos PLACSP[13].

Las Tablas se han dividido en 4 áreas temáticas:

- Datos en abierto y calidad del dato. Los datos en abierto gubernamentales y, especialmente, los ligados a la contratación y su calidad, son el pilar para la investigación de esta Tesis.
- Innovación y gestión en la contratación. Artículos que desarrollen estas áreas y que engarzan con la finalidad de esta Tesis.
- Forecasting en la contratación. Algoritmos de ML aplicados a la predicción (estimación) para resolver problemas en el ámbito de los contratos públicos: estimar el número de ofertantes, el importe de adjudicación, etc.
- Colusión y corrupción en la contratación. La colusión (en inglés, collusion o bid-rigging) es un pacto ilegal (coordinado) entre los ofertantes para no competir en la licitación e incrementar su margen de beneficios [87]. La corrupción es el abuso del poder público para obtener beneficios privados [79]. Se recopilan artículos donde la IA y, particularmente, el ML sea el mecanismo para detectar la colusión o corrupción.

3.7. Métricas de evaluación para los algoritmos de ML

Las métricas de evaluación, también llamadas métricas de error, tienen como finalidad evaluar el performance del algoritmo de predicción. Es decir, analizar cómo se comporta la predicción del

[13]Para más información léase el capítulo *"La utilización del big data y la IA para la detección de ilícitos de competencia"* [6] y véase el vídeo en el que la Unidad de Inteligencia Económica de la CNMC habla de los algoritmos y herramientas que utilizan: https://www.youtube.com/watch?v=KQgDnvrRAGA&t=8159s

Área	Ref.	Título del artículo o libro
Datos en abierto y calidad del dato	[54]	"Data Quality Barriers for Transparency in Public Procurement"
	[39]	*The Economic Impact of Open Data Opportunities for value creation in Europe*
	[55]	"Modeling of Open Government Data for Public Sector Organizations Using the Potential Theories and Determinants—A Systematic Review"
	[56]	"Open government data portals in the European Union: Considerations, development, and expectations"
	[57]	*Open Data Exposed*
	[22]	"DIGIWHIST Recommendations for the Implementation of Open Public Procurement Data An Implementer's Guide"
	[58]	"Open data: Quality over quantity"
	[59]	*Datos Abiertos: Guía estratégica para su puesta en marcha Conjuntos de datos mínimos a publicar*
	[60]	"Exploring the economic value of open government data"
	[61]	"A systematic review of open government data initiatives"
	[62]	"The exploitation of Business Register data from a public sector information and data protection perspective: A case study"
Innovación y gestión en la contratación	[63]	"AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings"
	[64]	"Innovation and public procurement: Terminology, concepts, and applications"
	[65]	"Project procurement management: A structured literature review"
	[66]	*Global Public Procurement Theories and Practices*
	[67]	"Research perspectives on public procurement: Content analysis of 14 years of publications in the journal of public procurement"

Tabla 3.2: Recopilación de la literatura relacionada con los datos en abierto y calidad del dato, la innovación y gestión en la contratación.

Área	Ref.	Título del artículo o libro
Forecasting en la contratación	[68]	"Predicting The Number of Bidders in Public Procurement"
	[69]	"Big Data with deep learning for benchmarking profitability performance in project tendering"
	[70]	"Predicting bid prices by using machine learning methods"
	[71]	"Predicting distresses using deep learning of text segments in annual reports"
	[72]	"Optimized artificial intelligence models for predicting project award price"
Colusión y corrupción	[73]	"Deep learning for detecting bid rigging: Flagging cartel participants based on convolutional neural networks"
	[74]	"Anomaly Detection in Public Procurements using the Open Contracting Data Standard"
	[75]	"Artificial Intelligence Techniques to Detect and Prevent Corruption in Procurement: A Systematic Literature Review"
	[76]	"Transnational machine learning with screens for flagging bid-rigging cartels"
	[77]	"A Machine Learning Approach for Flagging Incomplete Bid-rigging Cartels"
	[78]	"Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach"
	[79]	"Corruption Red Flags in Public Procurement: New Evidence from Italian Calls for Tenders"
	[80]	"Preventing rather than punishing: An early warning model of malfeasance in public procurement"
	[81]	"Machine learning with screens for detecting bid-rigging cartels"
	[82]	"Prediction of Public Procurement Corruption Indices using Machine Learning Methods"
	[83]	"Detecting Fake Suppliers using Deep Image Features"
	[84]	"Predicting Public Procurement Irregularity: An Application of Neural Networks"
	[85]	"Detecting the collusive bidding behavior in below average bid auction"
	[86]	"Detecting fraud, corruption, and collusion in international development contracts: The design of a proof-of-concept automated system"

Tabla 3.3: Recopilación de la literatura relacionada con el forecasting en la contratación y la colusión y corrupción.

algoritmo (salida) para el conjunto de datos de entrada. Los problemas que se analizan en esta Tesis son del tipo aprendizaje supervisado: un modelo aprende a relacionar unas variables de entrada y predice, estima, una variable de salida. En particular, hay dos tipos de predicciones o problemas:

- **Problemas de regresión**: se predice un valor numérico. Los artículos de la Tesis que tratan este tipo de problema son:
  - *Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning* [10]. Se va a estimar el importe de adjudicación de una licitación gracias al algoritmo de ML llamado Random Forest.
  - *Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain* [11]. Se va a estimar el importe de adjudicación de una licitación, mejorando el artículo anterior por usarse varios algoritmos, no sólo Random Forest, y obtener el algoritmo que mejor predice (menos error tiene).
- **Problemas de clasificación**: se predice una clase, una categoría, de entre varias previamente definidas. Los artículos de la Tesis que tratan este tipo de problema son:
  - *Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain* [12]. Se va a clasificar el perfil de empresa ganadora de una licitación y, a partir de ella, recomendar un grupo de empresas similares a la estimada.
  - *Collusion detection in public procurement auctions with machine learning algorithms* [13]. La colusión se puede simplificar a una clasificación binaria: clase 0 (no hay colusión; es una licitación honesta, competitiva) y clase 1 (sí hay colusión; es una licitación deshonesta, pactaron precios los ofertantes).

En varios de los artículos de la Tesis se va a utilizar el diagrama de caja (denominado boxplot en inglés). Es un método estandarizado para representar gráficamente, de manera compacta, una serie de datos numéricos a través de sus cuartiles, media, mediana y valores atípicos. Ver la Figura 3.5 que identifica sus elementos característicos.

Cada tipo de problema tiene asociado unas métricas de error, las cuales se explican en los dos siguiente subapartados.

3.7.1. Métricas de evaluación para problemas de regresión

Para conocer la precisión de los algoritmos de ML que hacen prediciones de variables continuas, se necesitan formular una serie de métricas de error. De esta manera se podrá saber qué algoritmos estiman mejor, es decir, resuelven mejor el problema de predicción. Sea r_i las observaciones actuales (valores reales), $\overline{r}$ es la media de las observaciones actuales, p_i las observaciones estimadas y N el número de observaciones.

En las siguientes 10 ecuaciones se define el error absoluto (absolute error, AE), error porcentual absoluto (absolute percentage error, APE), error absoluto medio (mean absolute error, MAE), error absoluto relativo (relative absolute error, RAE), error porcentual absoluto medio (mean absolute percentage error, MAPE), error absoluto mediano (median absolute error, MdAE), error porcentual absoluto mediano (median absolute percentage error, MdAPE), error cuadrático medio (root mean

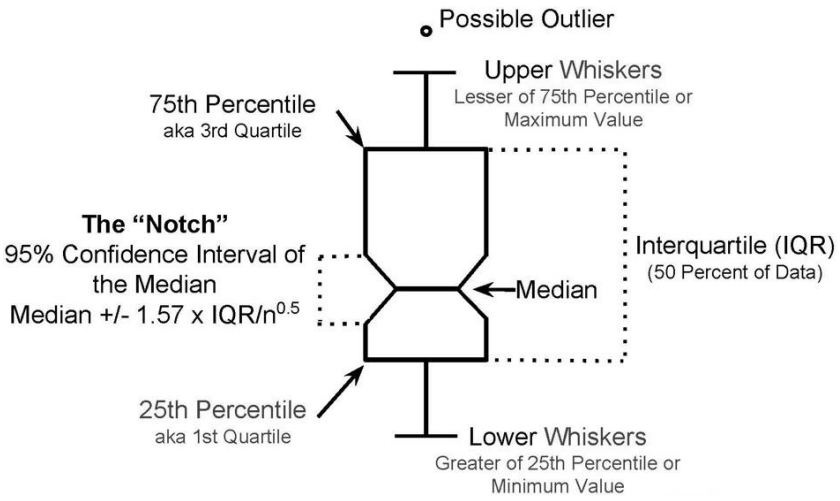


Figura 3.5: Elementos que componen un diagrama de cajas (boxplot).

squared error, RMSE), el error cuadrático medio normalizado (normalized root mean squared error, NRMSE) y el coeficiente de determinación[14] (coefficient of determination, R^2).

$$AE_i = |r_i - p_i| \tag{3.1}$$

$$APE_i = 100 \left| \frac{AE_i}{r_i} \right| = 100 \left| \frac{r_i - p_i}{r_i} \right| \tag{3.2}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} AE_i = \frac{1}{N} \sum_{i=1}^{N} |r_i - p_i| \tag{3.3}$$

$$RAE = \frac{\sum_{i=1}^{N} |r_i - p_i|}{\sum_{i=1}^{N} |r_i - \overline{r}|} \tag{3.4}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} APE_i = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{r_i - p_i}{r_i} \right| \tag{3.5}$$

$$MdAE = \frac{1}{N} median(|r_1 - p_1|, |r_2 - p_2|, \ldots, |r_N - p_N|) \tag{3.6}$$

$$MdAPE = \frac{100}{N} median \left(\left| \frac{r_1 - p_1}{r_1} \right|, \left| \frac{r_2 - p_2}{r_2} \right|, \ldots, \left| \frac{r_N - p_N}{r_N} \right| \right) \tag{3.7}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (r_i - p_i)^2} \tag{3.8}$$

$$NRMSE = \frac{RMSE}{max(r_i) - min(r_i)} \tag{3.9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} |r_i - p_i|^2}{\sum_{i=1}^{N} |r_i - \overline{r}|^2} \tag{3.10}$$

[14] Es la proporción de la variación de la variable dependiente que es predicha por las variables independientes.

3.7.2. Métricas de evaluación para problemas de clasificación

En los problemas de clasificación hay 5 errores importantes para analizar el buen desempeño del algoritmo clasificador: accuracy, precision, recall, balanced accuracy y F1 score. Sea r_i las observaciones actuales (valores reales), p_i las observaciones estimadas, N el número de observaciones y L el número de clases a clasificar ($1 \leq l \leq L$). Los problemas tratados en esta Tesis serán clasificaciones binarias, luego $L = 2$.

El *accuracy* es definido como la proporción de predicciones correctas respecto al total de predicciones hechas:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} 1(p_i = r_i) \tag{3.11}$$

donde $1(p_i)$ es la función indicador: devuelve 1 si la clase coincide entre p_i y r_i y 0 en otro caso.

El *precision*, también llamado observaciones predictivas positivas, es la habilidad del clasificador para no clasificar una clase como positiva cuando realmente es negativa. El *recall*, también llamada sensibilidad, es la habilidad del clasificador para descubrir todos los valores positivos. Sea r_l el subconjunto de valores positivos de la clase l y p_l el subconjunto de predicciones positivas en la misma clase l:

$$Precision_l = \frac{|r_l \cap p_l|}{|r_l|} \tag{3.12}$$

$$Recall_l = \frac{|r_l \cap p_l|}{|p_l|} \tag{3.13}$$

El *balanced accuracy* se utiliza para evitar desviaciones del *accuracy* en aquellos datos con clases muy desbalanceadas. Es decir, que una de las clases aparece mucho más que otras clases. Su definición es:

$$Balanced\ accuracy = \frac{1}{L}\sum_{l=1}^{L} recall_l = \frac{1}{L}\sum_{l=1}^{L} \frac{|r_l \cap p_l|}{|p_l|} \tag{3.14}$$

El $F1\ score$ es interpretado como la media ponderada del $precision$ y $recall$, siendo 1 su mejor valor y 0 el peor:

$$F1\ score = 2\frac{precision \cdot recall}{precision + recall} \tag{3.15}$$

En particular para nuestra investigación, 0 es la clase de licitaciones competitivas (honestas, donde no hubo alteraciones del importe por los ofertantes) y la clase 1 son licitaciones colusivas (deshonestas, donde los ofertantes formaron un cártel para alterar el importe). Es decir, es una clasificación es binaria: clase 1 o 0. Se llama True Positive (TP) a las observaciones de clase 1 correctamente clasificadas como 1 y True Negative (TN) a las de clase 0 también correctamente clasificadas. El False Positive (FP) son las observaciones de clase 0 que se clasifican incorrectamente, asignándoles la clase 1. Análogamente, el False Negative (FN) son las observaciones de clase 1 que se clasifican incorrectamente, asignándoles la clase 0. Lógicamente, el número total de observaciones debe ser la suma de TP + TN + FP + FN. Estos valores se pueden representar en lo que se llama matriz de confusión (ver Figura 3.6).

Las fórmulas anteriores se simplifican para la clasificación binaria:

$$Accuracy = \frac{TP + TN}{N} \tag{3.16}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.17}$$

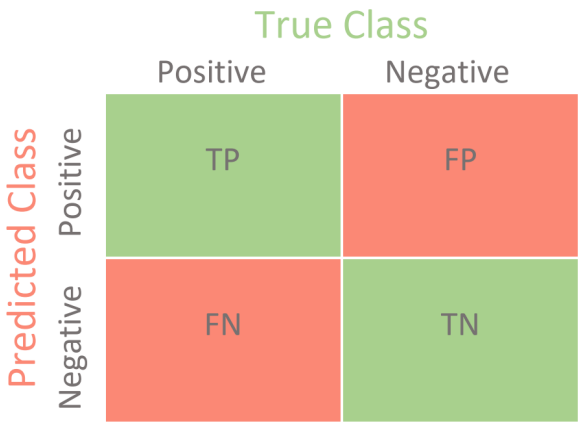


Figura 3.6: Matriz de confusión para la clasificación binaria.

$$Recall = \frac{TP}{TP + FN} \tag{3.18}$$

$$Balanced\ accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{3.19}$$

Capítulo Cuarto

Artículos de la investigación

4.1. Introducción

Hay pocos investigadores y artículos académicos que se dediquen a aplicar la ciencia de datos, el análisis masivo de datos y la IA a la contratación pública, ni en España ni en otros países. Esto se debe fundamentalmente a que se necesitan grandes volúmenes de datos, no disponiéndose de ellos hasta hace pocos años, gracias a la publicación masiva de contratos en las plataformas de contratación. Contratos públicos de cualquier sector, importe y ámbito geográfico. Tradicionalmente, los estudios similares a los de esta Tesis se centraban solamente en el sector de la construcción pública e ingeniería civil, debido a que son los de mayor importancia en la contratación por su alto coste económico y gran impacto para la sociedad.

Este capítulo está organizado en los siguientes apartados, cada uno correspondiéndose con una investigación en particular. Se va a explicar de la manera más sencilla posible los artículos del compendio de publicaciones de la Tesis. Es decir, que no sea demasiado complejo para que un lector no técnico, no experto en IA, pueda entenderlo. Así se consigue abrir la Tesis a investigadores de otros ámbitos, como el jurídico o económico. El lector técnico, especializado en IA y programación, debería leer los artículos (anexos al final de la Tesis) para profundizar en cómo se resuelve el problema, mediante qué técnicas cuantitativas y de ML.

- **Licitaciones en España: regulación, análisis de datos y estimador del importe de adjudicación usando ML** [10]. Primer artículo del compendio de la Tesis cuyo objetivo es introducir la analítica de datos a la contratación y crear un estimador del precio de adjudicación con un algoritmo de ML.

- **Estimador del importe de adjudicación de licitaciones usando ML: caso de estudio con licitaciones de España** [11]. Segundo artículo y extensión del anterior, consiguiendo un mejor estimador del importe de adjudicación. Se testearon varios algoritmos de IA, batiendo alguno de ellos al algoritmo anterior.

- **Recomendador de licitadores usando ML: análisis de datos, algoritmo y caso de estudio con licitaciones de España** [12]. El tercer artículo de la Tesis crea un buscador de empresas que puedan llevar a cabo una licitación dada. Gracias al histórico de licitaciones y a la fuente de datos empresariales, se ha entrenado un algoritmo de ML que permite a un usuario introducir una licitación y obtener qué empresas son las más capacitadas, a priori, para acometer el trabajo.

- **Detección de colusión en licitaciones aplicando algoritmos de ML** [13]. Se presenta

el grave problema de la colusión en la contratación pública, es decir, aquellas empresas que forman un cártel para pactar ofertas económicas en las licitaciones. Se comparan 11 algoritmos de ML para detectar si una licitación es competitiva (los ofertantes no pactaron precios) o colusiva.

Además, se explicará una **aplicación informática para detectar licitaciones irregulares**. Esta investigación no ha dado lugar, de momento, a un artículo. Por tanto, no forma parte del compendio de la Tesis. Tiene como objetivo detectar licitaciones donde ha podido existir actuaciones ilegales, fraudulentas. Se detectaron varias licitaciones de este tipo en España.

Las *fuentes de datos* usadas en los artículos son de dos tipos:

- **Datos de contratación**. Se ha utilizado PLACSP en todos los artículos salvo en el último artículo (detección de colusión). Además, en uno de los artículos se han utilizado licitaciones europeas obtenidas de TED. Estas dos fuentes de información son datos en abierto (públicas y gratuitas). Por otro lado, en el último artículo sobre colusión se han utilizado 6 datasets de 5 países distintos, de acceso restringido en su mayoría (no son datos en abierto), solicitándolos directamente a organizaciones públicas (Autoridades de la Competencia y Polícia) o a investigadores.

- **Datos empresariales**. Se ha conseguido la información más relevante de las cuentas anuales de 1,3 millones de empresas españolas, principalmente entre los años de 2014 a 2019, provenientes del Registro Mercantil. Estos datos son de acceso público pero de pago.

Los artículos se han programado en el lenguaje de programación Python [3, 40, 41] con la librería de ML llamada *Scikit-learn* [88, 89]. Para el segundo artículo, que emplea redes neuronales, se ha utilizado la plataforma de ML llamada WEKA [90, 91] y Tensorflow [92].

Todos los artículos son de publicación open access. Así se facilita su lectura y difusión entre los distintos involucrados (directa o indirectamente) en la contratación, como son las AA.PP., organizaciones públicas con competencias en contratación, operadores económicos, investigadores, ciudadanos, etc. Muchos de estos involucrados son ajenos a la Academia y necesitan un acceso libre y gratuito. Es más, gran parte de los investigadores interesados en la contratación (juristas, economistas, politólogos, etc) no tienen suscripciones a las revistas del ámbito científico-tecnológico donde se han publicado estos artículos. La investigación que se va a exponer son ejemplos prácticos, reales, que aportan valor a los organismos públicos que gestionan, controlan y velan por la contratación. Además, los operadores económicos que son contratados por las AA.PP. también obtendrían ventajas y beneficios al usar estas herramientas innovadoras.

4.2. Licitaciones en España: regulación, análisis de datos y estimador del importe de adjudicación usando ML

Este artículo tiene por título original en inglés "Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning" [10]. Es el primer artículo del compendio de publicaciones de la Tesis y tiene como objetivo principal hacer un estimador (predictor) del importe de adjudicación de una licitación. Es decir, predecir cuál es el precio de mercado, la oferta económica de la empresa ganadora del contrato. Esto tiene gran importancia porque las AA.PP., al publicar una licitación, estiman el importe de licitación a veces sin mucha información de mercado, errando en su estimación inicial. Por eso hay diferencias (más o menos

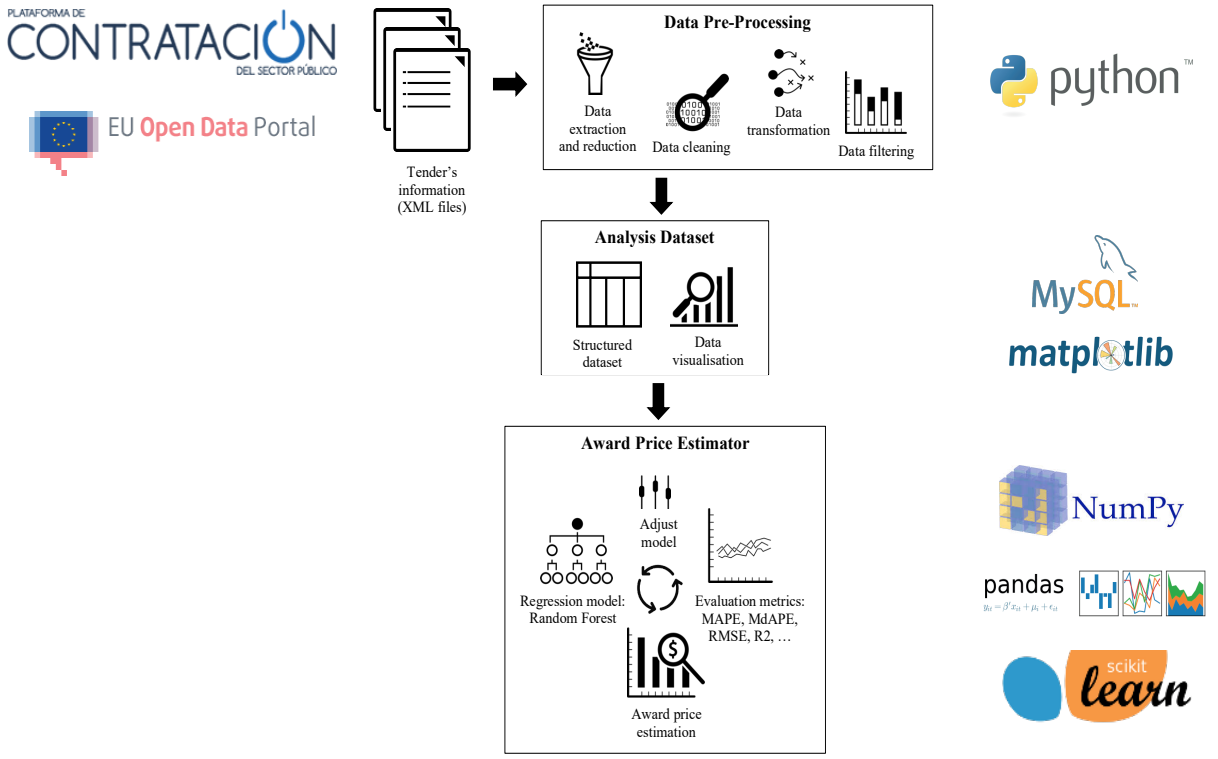


Figura 4.1: Flujograma del análisis de datos de contratación y el estimador del precio de adjudicación. A la derecha, las librerías software utilizadas.

significativas) entre el importe de licitación y el de adjudicación. A esa diferencia se le suele llamar comúnmente *"baja"* en el argot de la contratación.

Como es un artículo de una temática novel (ciencia de datos aplicado a las licitaciones), se hace una presentación de la legislación de la contratación pública (más detallado en esta Tesis en el apartado correspondiente del capítulo 3), se describe PLACSP, los campos de las licitaciones que se han utilizado y cómo se han procesado los datos.

La metodología del artículo se muestra en la Figura 4.1, así como el software utilizado (Python, sus librerías típicas de ciencia de datos y MySQL). Como fuente de datos se utiliza la española PLACSP y la europea TED. Ambas fuentes se utilizan de manera independiente, por separado, para dotar al artículo de mayor generalidad, verificándose que se puede crear un estimador del importe de adjudicación tanto a nivel español como europeo. Después se describe cómo se han procesado el conjunto de datos (extracción, reducción, limpieza, transformación y filtrado) y el análisis de datos realizado (análisis estadístico, matriz de correlación y dispersión, histogramas, etc.). Finalmente, se crea el estimador del importe de adjudicación mediante un algoritmo de ML, llamado Random Forest, que se evalúa y ajusta gracias a las métricas de evaluación del error (MAPE, MdAPE, RMSE, etc.). En este caso, el error es la diferencia entre la estimación (valor predicho) del importe de adjudicación y el valor real adjudicado.

Se puede crear el estimador del importe de adjudicación gracias a que el algoritmo de ML se entrena con un gran volumen de licitaciones, incluyendo sus importes de licitación y adjudicación, y es capaz de estimar el precio de mercado para una licitación dada. Para el caso de España, se entrenó con casi 60.000 licitaciones multisectoriales publicadas entre 2012 y 2018. Para el caso de Europa, se entrenó con aproximadamente 41.500 licitaciones multisectoriales publicadas en 2017

	España	Europa
Nº licitaciones	58.337	41.556
Ubicación	Toda España	Francia, Alemania, Italia, Hungría, Letonia, Croacia, Eslovenia y Bulgaria
Fechas licitaciones	2012-2018	2017
Nº órganos de contratación	3.544	6.163
Nº empresas adjudicatarias	17.305	19.100
Algoritmo de predicción	Random Forest	Random Forest
Variables de entrada	Importe licitación, Fecha, Duración, Organismo licitador, CPV, CPV Agregado, Código procedimiento, Código tipo, Código subtipo, Código urgencia, Código Postal, CC.AA., Provincia y Municipio	Importe licitación, Fecha, Organismo licitador, Actividad principal, CPV, CPV Agregado, Código procedimiento, Código tipo, Código Postal y País
Variable salida (predicción)	Importe adjudicación	Importe adjudicación
MdAPE entre importe de adjudicación y licitación (azul) y MdAPE entre adjudicación y predicción (gris)	11,84 % 9,26 % (-2,58 % menos de error)	4,17 % 6,48 % (+2,31 % más de error)
MAPE entre importe de adjudicación y licitación (azul) y MAPE entre adjudicación y predicción (gris)	39,79 % 28,60 % (-11,19 % menos de error)	27,49 % 23,57 % (-3,92 % menos de error)

Tabla 4.1: Datasets utilizados y métricas de error del estimador del precio de adjudicación.

de 8 países europeos.

En la Tabla 4.1 se resumen las principales magnitudes de los datasets de España y Europa, las variables de entrada para el algoritmo Random Forest y los resultados que obtiene el estimador del importe de adjudicación. Para el caso de España, el MdAPE (error porcentual absoluto mediano) entre el importe adjudicación y de licitación es del 11,84 % y, sin embargo, el MdAPE entre el importe de adjudicación y la predicción es del 9,26 %. Es decir, el estimador es más fiable por tener un -2,58 % menos de error. Si lo analizamos desde la perspectiva del MAPE (error porcentual absoluto medio), el estimador sigue siendo más fiable por tener un -11,19 % menos de error. Para el caso de Europa, el estimador obtiene mejor resultado en media (el MAPE es menor) y en mediana es un poco peor (el MdAPE es ligeramente superior). La causa de que el estimador obtenga peores resultados con licitaciones europeas es porque la *"baja"* es muy pequeña. Es decir, en Europa no hay grandes diferencias entre el importe de licitación y adjudicación, hecho que sí sucede en España. Esto se observa en el MdAPE, el de Europa es casi 3 veces superior al de España. Este suceso es sorprenderte y debería de analizarse con más licitaciones, de otros años y más países.

Para comprender de manera gráfica cómo predice el algoritmo, observar la Figura 4.2 (datos de España). En la gráfica superior, se muestra el importe de licitación (eje x) y el importe de adjudicación (eje y). Si las AA.PP. españolas acertasen totalmente en los importes de adjudicación que dan cuando se publica la licitación, esta gráfica tendría que ser una línea recta a 45 grados del eje x. Es decir, las AA.PP. acertarían siempre y el importe de licitación coincidiría con el precio de mercado (el importe finalmente adjudicado). Sin embargo, se aprecia como hay una nube de puntos, habiendo significativas diferencias entre ambos importes. En la gráfica inferior, se muestra la estimación del importe de adjudicación (eje x) y el importe real de adjudicación (eje y). Se

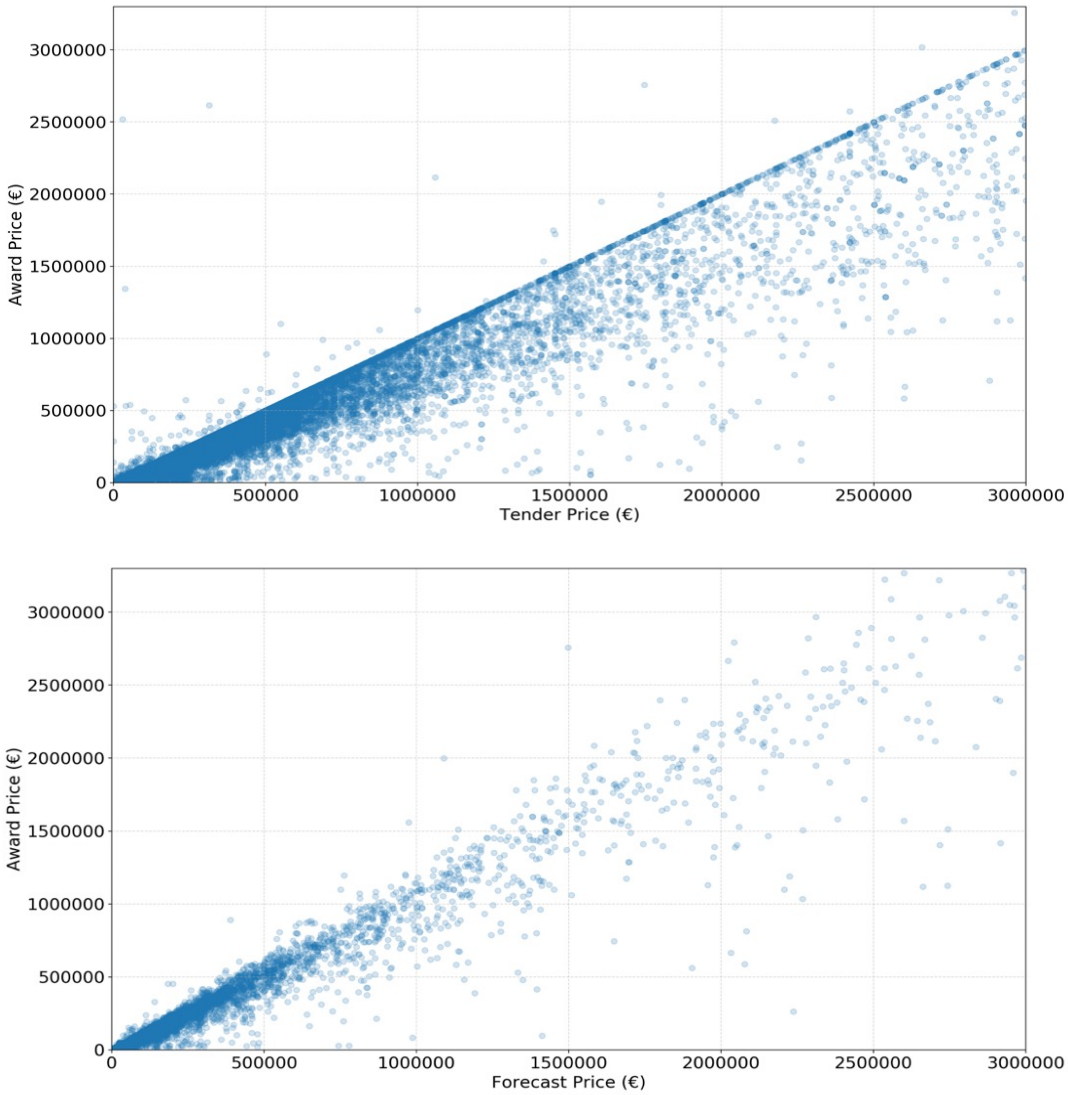


Figura 4.2: En la gráfica superior se compara el importe de licitación (eje x) frente al importe de adjudicación (eje y). En la inferior, la predicción del importe de adjudicación (eje x) frente al importe de adjudicación (eje y). Datos de España.

observa cómo sigue existiendo una nube de puntos porque, como ya vimos, el estimador no es perfecto, tiene error. Sin embargo, los puntos están más concentrados respecto a la línea recta descrita anteriormente que sería la ideal, donde no habría error.

El algoritmo Random Forest se usará también en el resto de artículos de la Tesis. Así que, por su importancia, se hace aquí una breve resumen (en los artículos está su pseudocódigo y más información). Random forest, introducido por Breiman [93] en 2001, es un método de aprendizaje de conjuntos para regresión o clasificación que opera mediante la construcción de una multitud de árboles de decisión en el entrenamiento y la salida es el moda de las clases (si es un problema de clasificación) o la predicción media (si es un problema de regresión) de los árboles individuales. Los dos primeros artículos de la Tesis son problemas de regresión (estimar un valor numérico) y los dos últimos artículos son problemas de clasificación. Es un popular algoritmo de ML que ofrece excelente performance [94, 95], no overfitting [96], una versatilidad para aplicarse a problemas reales con gran volumen de datos y con diferentes tipos de datos [94, 97]. En este artículo (y en los demás de la Tesis) se coge un 80 % de los datos como dataset de entrenamiento y un 20 % para la validación (test) del algoritmo.

4.3. Estimador del importe de adjudicación de licitaciones usando ML: caso de estudio con licitaciones de España

Este artículo tiene por título original en inglés "Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain" [11]. Es una extensión del anterior artículo, utilizando el mismo conjunto de datos de España (el de Europa no se utiliza), cuyo objetivo es utilizar más algoritmos de ML para evaluar cuál resuelve mejor el problema de predicción del importe de adjudicación. Por tanto, la metodología es análoga al artículo anterior, no teniendo que hacer las primeras fases (procesado el conjunto de datos y análisis) por ser el mismo dataset de España.

Para la experimentación computacional de los algoritmos se ha utilizado WEKA [90, 91], la plataforma de ML desarrollada por U. de Waikako que da soporte a un gran número de algoritmos de ML. Se han utilizado 4 tipos de algoritmos de ML:

- Random Forest. Se toma como modelo base para establecer la comparación por ser el utilizado en el primer artículo.
- Regresión lineal. Relaciona de manera lineal (una recta) las variables de entrada y la variable de salida (predicción)[1]. Es un modelo clásico, siendo la primera referencia documentada por el matemático Legendre en 1805.
- Regresión isotónica (o monotónica). Hace una aproximación de una función no decreciente (monótona) minimizando el error cuadrático medio (MSE) en los datos de entrenamiento[2].
- Redes neuronales artificiales. En inglés se denominan Artificial Neuronal Networks (ANNs) y forma parte del subcampo del ML llamado deep learning (aprendizaje profundo). Una ANN es un modelo computacional inspirado en las redes neuronales biológicas. Consiste en una colección nodos (neuronas artificiales) organizados en capas interconectadas (capa de entrada, capas ocultas y capa de salida). Los parámetros del modelo son los pesos y sesgos asociados a las conexiones. La información se procesa desde la capa de entrada a la capa de salida. El

[1] Consultar el artículo para su definición matemática.
[2] Consultar el artículo para su definición matemática.

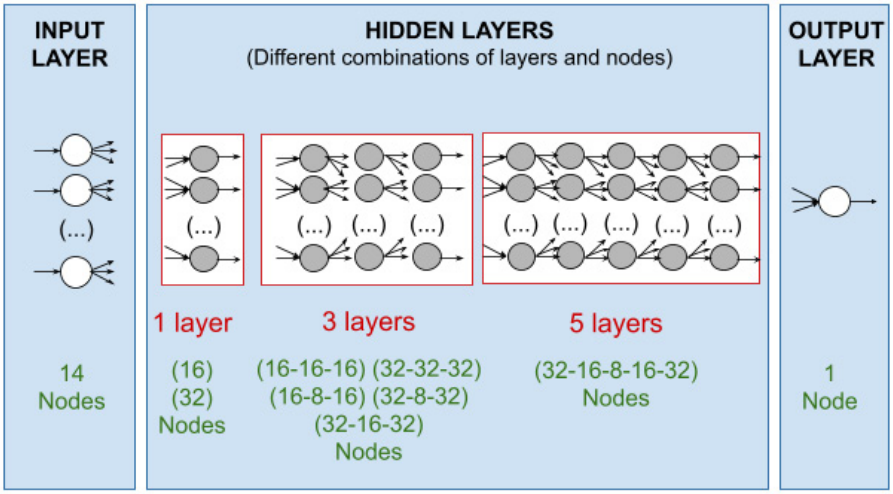


Figura 4.3: Arquitectura de las capas y nodos de las redes neuronales artificiales (ANN) utilizadas.

proceso de aprendizaje se basa en minimizar una función de coste (conocida como función de pérdida) que evalúa el error de la ANN para resolver el problema.

Se han probado muchas configuraciones distintas de ANNs. Por un lado se ha utilizado el MLP (Multi-layer perceptron) [52]. Por otro lado, se han definido 8 arquitecturas distintas de capas y nodos (ver Figura 4.3). Cada configuración viene definida por: número de capas y nodos, función de activación (rectified linear unit (ReLU) o scaled exponential linear unit (SeLU)), función de pérdida (MAE o MSE) y optimizador (Adam, Adamax o Adagrad). La combinación de las 8 arquitecturas diferentes, 2 funciones de activación, 2 funciones de pérdida y 3 optimizadores ha dado lugar a 96 configuraciones distintas de ANN. Las arquitecturas de ANNs se han optimizado usando Tensorflow [92]. Tras la validación de las ANNs, se han seleccionado 4 configuraciones diferentes (ANN1, ANN2, ANN3 y ANN4) que tenían un error más pequeño.

Los resultados de la experimentación realizada (descrita en detalle en el artículo) se muestran en la Tabla 4.2, resaltándose en negrita qué algoritmo tiene el mejor error (el más pequeño) para cada uno de las 4 métricas de error evaluadas (MAE, RMSE, RAE y RRSE). En realidad, son dos errores distintos (MAE y RMSE) y sus respectivos errores relativos (RAE y RRSE). Por eso quien obtenga el mejor MAE, obtendrá consecuentemente el mejor RAE. Análogamente ocurre para el RMSE y RRSE. Las redes neuronales (ANNs) son las que obtiene los mejores errores, en particular el MLP (mejor RMSE y RRSE) y el ANN2 (mejor MAE y RAE).

4.4. Recomendador de licitadores usando ML: análisis de datos, algoritmo y caso de estudio con licitaciones de España

Este artículo tiene por título original en inglés "Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain" [12]. Se ha creado un buscador de empresas que puedan llevar a cabo un contrato. Es decir, el buscador (un algoritmo desarrollado ad hoc) recomienda un grupo de empresas en base a la información del contrato: órgano de contratación, importe de licitación, código CPV (Common Public Vocabulary), tipo de procedimiento y tramitación, fecha de publicación, etc.

	MAE	RMSE	RAE	RRSE
Random Forest (baseline model)	179,247.80€	6,621,784.24€	31.11%	74.86%
Linear regression	228,491.36€	15,535,231.61€	39.66%	175.63%
Isotonic regression	136,971.39€	5,648,693.54€	23.76%	63.86%
ANN (MLP)	270,953.50€	**1,974,981.24€**	47.03%	**22.33%**
ANN1	140,763.27€	7,416,004.50€	23.03%	83.84%
ANN2	**123,570.91€**	5,110,687.50€	**20.22%**	57.78%
ANN3	157,181.00€	9,543,883.00€	25.71%	107.90%
ANN4	124,035.82€	3,304,259.20€	20.29%	37.36%

Tabla 4.2: Métricas de error del estimador del precio de adjudicación para los algoritmos de ML.

Se han utilizado dos fuentes de datos: licitaciones y empresas. Por un lado, se utilizaron 102.000 licitaciones (de cualquier naturaleza) de PLACSP publicadas entre 2014 y 2020. Por otro lado, se utilizaron 1,3 millones de empresas (ubicación, facturación anual, EBITDA número de empleados, clasificación nacional de actividades económicas, etc.) obtenidas de las cuentas anuales presentadas en el Registro Mercantil de España. En el artículo se describen ambos datasets en sendas tablas de manera detallada. La metodología para extraer y manejar los datos de PLACSP se ha descrito en el primer artículo. Para los datos empresariales, no se ha tenido que hacer ninguna metodología por haberse obtenido de una base de datos con toda la información ya estructurada y validada.

Los datos de la PLACSP son de acceso libre y gratuito, como ya se ha mencionado. Sin embargo, actualmente los datos empresariales del Registro Mercantil son de acceso libre pero de pago. Esto supone un handicap para poder desarrollar investigaciones o herramientas en este campo. Algunos papers han propuesto herramientas similares a ésta pero sólo analizando las licitaciones [98, 99], no incluyendo también las características de los proveedores.

Además, se ha desarrollado una aplicación web (ver Figura 4.4) para demostrar de manera práctica, real, el algoritmo de búsqueda planteado en este artículo. A la izquierda de la Figura se muestra el formulario con los campos de la licitación a rellenar y en la parte derecha se arrojan las empresas recomendadas por el buscador. Esta aplicación web fue la ganadora en 2020 del premio Open Data del Gobierno Vasco[3] (dotado con 10.000€). Además, actualmente se encuentra entre los candidatos finalistas al premio EU Datathon 2022, organizado por la Comisión Europea (CE), siendo la temática de esta edición utilizar los datos en abierto de la contratación pública.

El funcionamiento del buscador se describe en la Figura 4.5. El algoritmo que estima la empresa ganadora de la licitación es un Random Forest de tipo clasificación (más detalle en el artículo). La fuente de datos de licitaciones se utiliza solamente para entrenar el modelo de predicción pero la fuente de datos empresarial se utiliza también en fase de ejecución, para buscar las empresas.

En la fase de búsqueda (3), se buscan empresas que sean similares a la empresa predicha (forecast company). Esta similitud se basa en unas serie de reglas (ecuaciones) utilizando la facturación, EBIT, EBITDA, número de empleados, distancia entre la ubicación de la licitación y la sede de la empresa (el factor geográfico es determinante en la contratación) y una similitud los códigos de actividad empresarial (se utilizarán 4 tipos distintos de clasificaciones: NACE2, IAE, SIC y NAICS). Dichas magnitudes tienen que ser mayor o igual a los de forecast company ponderados por una

[3] Concesión en https://euskadi.eus/y22-bopv/es/bopv2/datos/2020/12/2005583a.shtml.

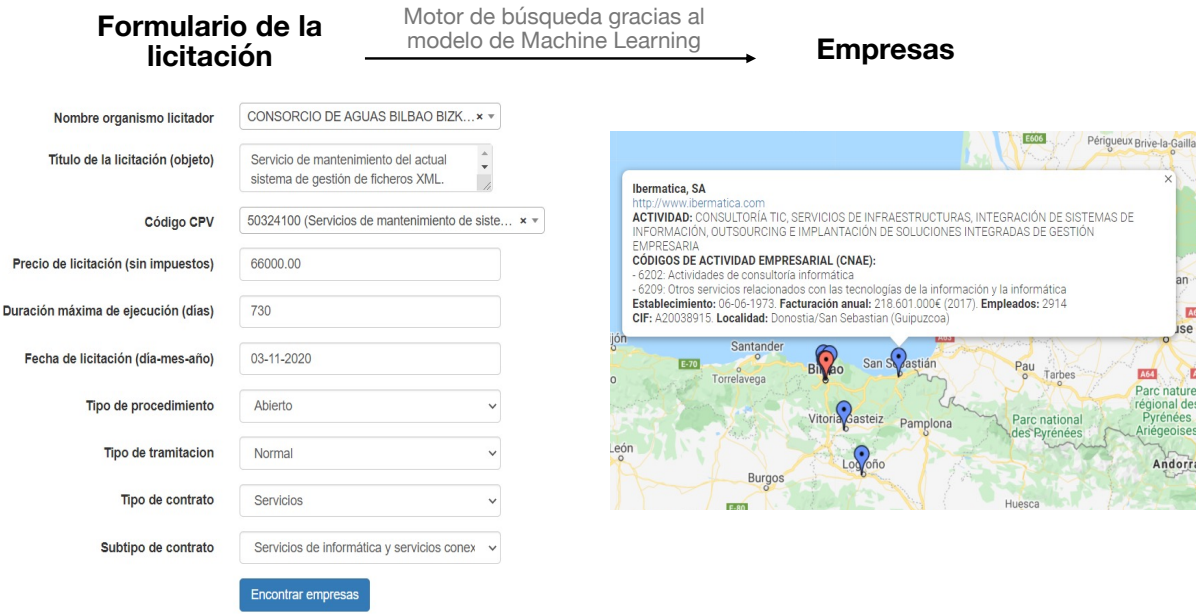


Figura 4.4: Funcionamiento del recomendador de licitadores (buscador de empresas) mediante dos capturas de pantalla de la aplicación web.

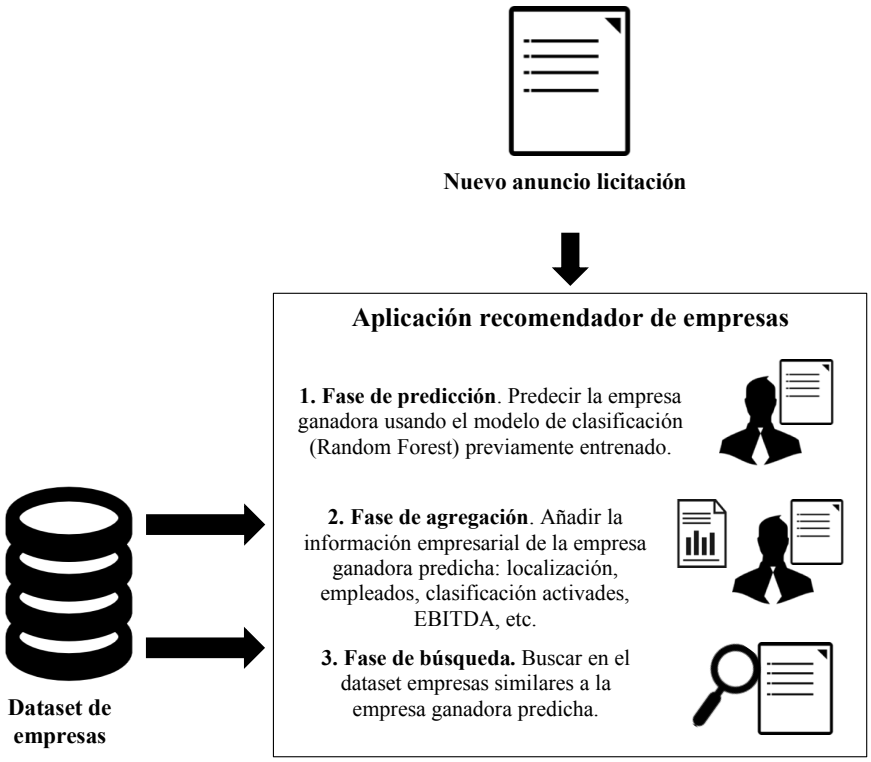


Figura 4.5: Flujograma del funcionamiento del recomendador de licitadores (buscador de empresas) para una nueva licitación de entrada.

constante F_x cuyo valor está comprendido entre 0 y 1.

$$Operating\_income_{company} \geq F_{OI} \cdot Operating\_income_{forecast\ company} \tag{4.1}$$

$$EBIT_{company} \geq F_{EBIT} \cdot EBIT_{forecast\ company} \tag{4.2}$$

$$EBITDA_{company} \geq F_{EBITDA} \cdot EBITDA_{forecast\ company} \tag{4.3}$$

$$Employees_{company} \geq F_E \cdot Employees_{forecast\ company} \tag{4.4}$$

$$Distance_{tender-company} \geq F_D \cdot Distance_{tender-forecast\ company} \tag{4.5}$$

$$\sum_{i=1}^{C} 1[Code_{company} = Code_{forecast\ company}] \geq F_{CEA} \cdot C \tag{4.6}$$

donde $1[Code]$ es la función indicador (devuelve 1 si coinciden los códigos y si no devuelve 0), C es el número total de códigos de la forecast company y $Code$ es el código de clasificación de actividades económicas de la compañía (entre los 4 tipos distintos de los que se dispone).

Es necesario entrenar el modelo para que el buscador esté operativo. Para ello, se muestra en la Figura 4.6 el flujograma para entrenar dicho algoritmo Random Forest de clasificación. Finalmente, se evalúa el performance del algoritmo con un 20 % de las licitaciones.

Los resultados han sido exitosos. La Tabla 4.3 muestra la batería de pruebas realizadas: dos escenarios y 5 configuraciones distintas de las constantes F_x (de menos a más reestrictiva). Por ejemplo, para el escenario 2 y el $Accuracy_{n=5}$, la empresa ganadora está en el grupo de 5 empresas recomendadas en el 31,58 % de las veces. Para el escenario 2 dicho valor cae al 23,12 %, influenciado seguramente por la dificultad en predecir las licitaciones más recientes que las más antiguas. Por tanto, es una aplicación que ejemplifica la utilidad que tienen este tipo de herramientas de IA en la contratación. Especialmente para los órganos de contratación que redactan y gestionan contratos de tipo negociado (con o sin publicidad) y necesitan buscar empresas que puedan acometer licitaciones pero también para las propias empresas licitadoras que pueden sondear qué competencia tienen en el mercado.

4.5. Detección de colusión en licitaciones aplicando algoritmos de ML

Este artículo tiene por título original en inglés "Collusion detection in public procurement auctions with machine learning algorithms" [13]. La colusión es una práctica ilegal mediante la cual algunas empresas competidoras acuerdan en secreto las ofertas económicas que presentarán a una licitación. La colusión es un fenómeno generalizado en la contratación pública de todo el mundo (pudiendo también aparecer en la contratación privada). Estas prácticas colusorias suelen adoptar la forma de aumentos de precios coordinados (no competitivos) establecidos entre las empresas (denominados cárteles) [100]. Por tanto, se socavan los beneficios del mercado competitivo, malgastando el dinero de los contribuyentes. Generalmente, los órganos de contratación no pueden identificar ofertas no competitivas, es decir, adjudicando contratos a precios más altos de los que habrían tenido sin colusión.

Se van a comparar 11 algoritmos de ML para detectar colusión utilizando datasets obtenidos de Brasil, Italia, Japón, Suiza (de dos regiones llamadas Ticino y St Gallen & Graubünden) y Estados Unidos (ver la Tabla resumen 4.4 y para más detalle consultar el propio artículo). Son aproximadamente 10.000 licitaciones de todo el mundo, de distinta naturaleza (obras públicas, infraestructuras

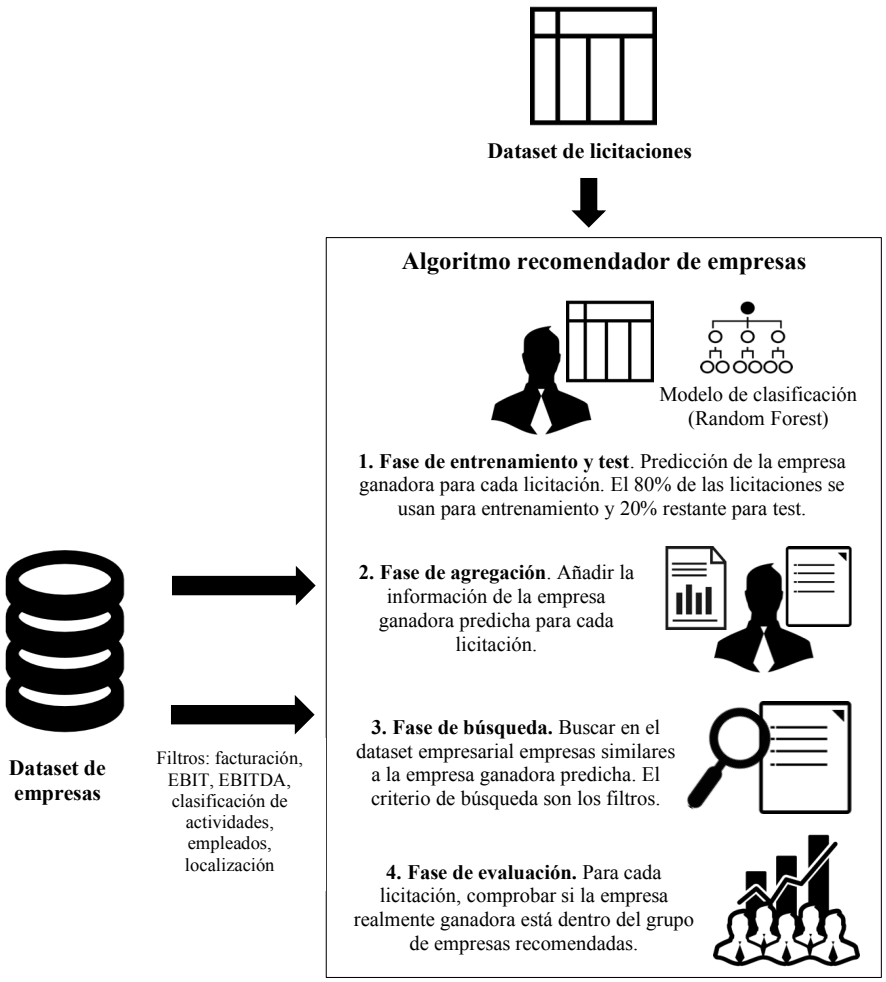


Figura 4.6: Flujograma para crear el algoritmo de ML recomendador de empresas.

Description		Different bidders recommender settings				
		Very low	Low	Medium	High	Very high
Bidders recommender factors for the settings	F_{OI}: operating income factor	0.25	0.5	0.65	0.75	1.0
	F_{EBIT}: EBIT factor	0.25	0.5	0.65	0.75	1.0
	F_{EBITDA}: EBITDA factor	0.25	0.5	0.65	0.75	1.0
	F_E: employees factor	0.15	0.25	0.25	0.35	0.45
	F_{CEA}: classification economic activities factor	0.125	0.15	0.14	0.175	0.2
	F_D: distance tender-company factor	1.6	1.4	1.4	1.2	1
Results of scenario 1: testing subset is the 20% of the dataset randomly chosen	$\mathrm{Accuracy}_{n=1}$: winner company is the forecast company			17.07%		
	$\mathrm{Accuracy}_{n=5}$: winner company is within the top 5 forecast companies			31.58%		
	$\mathrm{Accuracy}_{n=M}$: winner company is within the recommended companies group	38.52%	36.20%	35.92%	34.04%	33.25%
	Mean and median number of the recommended companies of each tender	877.43; 86	469.69; 35	430.48; 31	226.07; 11	145.97; 9
Results of scenario 2: testing subset is the last 20% of the dataset ordered by tender's date	$\mathrm{Accuracy}_{n=1}$: winner company is the forecast company			10.25%		
	$\mathrm{Accuracy}_{n=5}$: winner company is within the top 5 forecast companies			23.12%		
	$\mathrm{Accuracy}_{n=M}$: winner company is within the recommended companies group	30.52%	28.00%	27.73%	25.55%	24.79%
	Mean and median number of the recommended companies of each tender	900.64; 95	470.41; 37	430.33; 33	210.92; 11	132.10; 9

Tabla 4.3: Métricas de error del recomendador de licitadores para distintas configuraciones en dos escenarios diferentes.

Description	Brazil	Italy	Japan	Swiss–Ticino	Swiss–SG&GR	US
Scope	Oil infrastructure projects	Road construction	Building constr. and civil eng.	Road construction	Road construction and civil engineering	School milk market
Time period	2002–2013	2000–2003	2003–2007	1999–2006	14 years (over 2005)	1980–1990
N° auctions	101	278	1080	224	4344	3754
N° bids	683	20,286	13,515	1629	21,231	7004
Awarding criteria	Lowest bid	Average Bid Method	Lowest bid	Lowest bid	Lowest bid	Lowest bid
Avg. n° of bids per auction	6.76	72.97	12.51	7.27	4.89	1.91

Tabla 4.4: Descripción de los datasets (licitaciones competitivas y colusivas) de Brasil, Italia, Japón, Suiza-Ticino, Suiza-SG&GR y USA.

civiles, suministro de leche a escuelas, etc.), de distintos periodos temporales y provenientes de fuentes oficiales por lo que la fiabilidad y calidad del dato es excelente[4]. Por tanto, es un gran conjunto para experimentar en la detección de colusión aplicando los algoritmos de ML, siendo éste el primer artículo que reúne juntos tantos datasets. Nótese que son extremadamente difíciles de conseguir por no ser información accesible al público. Generalmente está restringida a las Autoridades de la Competencia u otras instituciones públicas que investigan prácticas anticompetitivas, demostrando la existencia de los cárteles ante los tribunales de justicia.

La metodología del artículo se muestra en la Figura 4.7. Las dos primeras fases son las tareas típicas de ciencia de datos y la tercera fase (calcular las variables de cribado) es propia para la colusión. Las variables de cribado (screening variables o, simplemente, screens) son variables secundarias o derivadas, calculadas a partir de las ofertas económicas, que ayudan a identificar licitaciones potencialmente colusivas [81]. Estas variables han sido propuestas por diferentes investigadores en el campo de la detección de colusión [87, 76, 77, 101, 102]. Se han utilizado las siguientes screens[5] (denominaciones en inglés): Coefficient of Variation (CV_t), Spread (SPD_t), la diferencia relativa entre las dos licitaciones más bajas en la licitación ($DIFFP_t$), la Relative Distance (RD_t), el Skewness ($SKEW_t$), el Excess Kurtosis ($KURT_t$) y el Kolmogorov-Smirnov test ($KSTEST_t$).

La detección de colusión se mide bajo diferentes escenarios y cantidad de información de entrada. Se supone que cada licitación se puede clasificar como *"colusión"* o *"competitiva"*. Por tanto, el algoritmo de ML tiene que realizar una clasificación binaria para cada licitación. Se van a utilizar y comparar los siguientes 11 algoritmos (denominación en inglés):

- Linear models: SGD (Stochastic Gradient Descent) [103].
- Ensemble methods: Extra Trees (Extremely Randomized Trees) [104], Random Forest [93], Ada Boost [105] and Gradient Boosting [106].
- Support Vector Machines: SVC (C-Support Vector Classification) [107].
- Nearest Neighbors: K Neighbors [108].
- Artificial Neural Network: MLP (Multi-Layer Perceptron) [52].

[4]Agradecer al Dr. David Imhof (investigador y trabajador en la Autoridad Suiza de la Competencia) y al Dr. Regis Signor (investigador e Inspector en la Policía Federal de Brasil) por facilitar varios de los datasets. Además, reconocer al Dr. Imhof por sus artículos pioneros en este campo (detección de colusión mediante ML) y al Dr. Signor por haber colaborado en este artículo.

[5]Consultar el artículo para sus definiciones matemáticas.

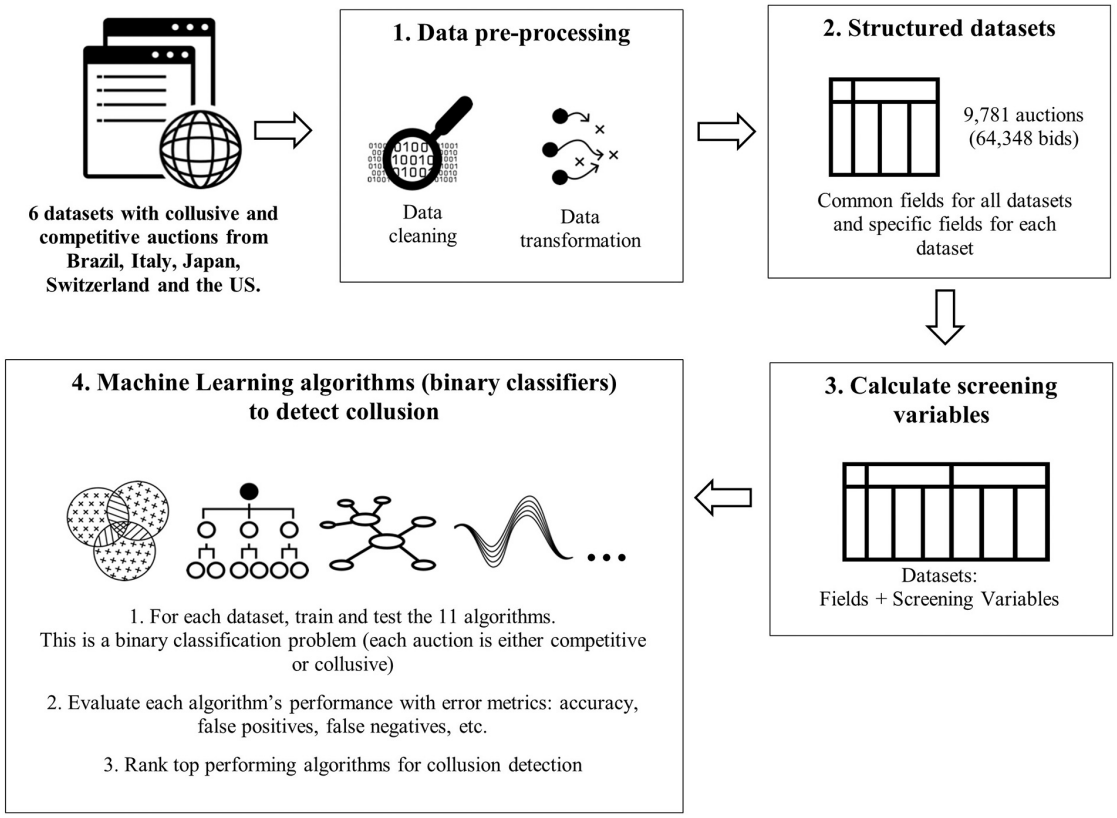


Figura 4.7: Flujograma para la detección de colusión mediante ML.

- Naive Bayes: Bernoulli Naive Bayes and Gaussian Naive Bayes [52].
- Gaussian Process [109].

La detección de colusión se va a testear bajo 4 escenarios distintos de información de entrada (settings). Es decir, el algoritmo de ML dispondrá de 4 configuraciones distintas de variables (campos) de entrada para poder hacer la detección:

- Setting 1. Se utilizan todas las campos disponibles.
- Setting 2. Se utilizan todas las campos disponibles más los screens.
- Setting 3. Se utilizan las campos comunes a todos los datasets.
- Setting 4. Se utilizan las campos comunes a todos los datasets más los screens.

Los resultados para cada dataset se muestra en la Tabla 4.5. La última columna denominada *all datasets*, se refiere a que todos los datasets se han aglutinado en un único conjunto. Así se comprobará si el algoritmo es capaz de entrenarse con licitaciones de todo el mundo y detectar la colusión. Las filas que tienen el topic *Fields* se refieren a los campos que tiene cada dataset (no todos poseen los mismos campos) y que posteriormente se utilizarán como variables de entrada al algoritmo. En la siguiente fila *Results* se muestran el mejor porcentaje de acierto y el algoritmo que lo consiguió para los 4 settings. En las filas *Average accuracy* se muestra el incremento de la precisión media (calculada para los 4 mejores algoritmos) al añadir los screens a las variables de entrada, es decir, el incremento de la precisión del setting 2 respecto al 1 y del setting 4 respecto al 3. Por último, la fila *Detection rates reported in the literature* menciona la precisión y algoritmo obtenidos por otros artículos de la literatura, para que sirva como elemento de comparación del dataset utilizado.

En general viendo la Tabla 4.5, los mayores porcentajes de acierto se dan en el setting 2 y 4 (salvo para Japón y US que todos los settings están muy igualados). Esto significa que las variables de cribado (screens) ayudan a los algoritmos a aumentar la detección de colusión. En general, los algoritmos con mayor porcentaje de acierto son de tipo ensemble (Gradient Boosting, Ada Boost, Extra Trees y Random Forest). Para un mayor análisis de cada algoritmo, dataset, setting y métrica de error (precisión, falsos positivos, falsos negativos y precisión balanceada), consultar las Tablas 2 y 3 del artículo. En dichas tablas se observa que:

- La precisión es generalmente superior al 80 %.
- Los falsos positivos (FP) y falsos negativos (FN) son generalmente inferiores al 10 %.
- La precisión balanceada es generalmente superior al 70 %.

La Figura 4.8 identifica otras tres métricas de error (precision, recall y F1 score) para los settings 3 y 4 del conjunto de datos denominado *all datasets*. Algunos algoritmos no aparecen en la gráfica por estar por debajo del 50 % de precision y recall. Para que los resultados no sean sensibles a una sola iteración (recordemos, 80 % de los datos para entrenamiento y 20 % para test), se han realizado 500 iteraciones con cada algoritmo y, por tanto, hay 500 valores de precision y otros 500 de recall. Dichas métricas de error se indican con una cruz: el punto de corte es la mediana de la precision y recall y los puntos finales de la cruz son los valores mínimo y máximo de la precision y recall. Consecuentemente, los valores de precision, recall y F1 score permanecerán dentro del rectángulo formado por la cruz con un alto grado de confianza. Se observa como 3

		Datasets						
Topic	Description	Brazil	Italy	Japan	Swiss - Ticino	Swiss - SG&GR	US	All datasets
Fields	**Common fields**	Auction code, bid values, winning bid and number of bids per auction						
	All fields in the dataset	Common fields, PTE, difference Bid/PTE, location, Brazilian State and date	Common fields, PTE, difference Bid/PTE, location, type and size of bidding companies	Common fields, PTE, difference Bid/PTE, location and date	Common fields and consortium composition	Common fields, contract type and date	Common fields, bid value with and without inflation and date	Common fields only
	Num. of variables	9	9	8	5	6	7	4
	Screens	Coefficient of variation (CV), spread (SPD), percentage difference between the two lowest bids (DIFFP), relative distance (RD), skewness statistic (SKEW) and Kolmogorov–Smirnov test (KSTEST)						
Results. Best accuracy and top-performing algorithm	**Setting 1** All fields from each dataset	85.2% Gradient Boosting	84.4% Extra Trees	94.7% Extra Trees	81.6% Bernoulli Naive Bayes	84.1% Ada Boost	84.1% Extra Trees	N/A
	Setting 2 All fields from each dataset + screens	92.4% Gradient Boosting	87.4% Extra Trees	94.6% Gaussian Naive Bayes	91.4% Ada Boost	85.3% Extra Trees	84.8% Extra Trees	N/A
	Setting 3 Common fields	87.9% Ada Boost	79.9% Random Forest	94.5% Extra Trees	82.0% Gaussian Naive Bayes	80.2% MLP	83.8% Extra Trees	82.0% Extra Trees
	Setting 4 Common fields + screens	89.6% Extra Trees	86.8% Extra Trees	94.5% Extra Trees	91.4% Ada Boost	81.1% Extra Trees	83.7% Extra Trees	86.3% Extra Trees
Average accuracy increase on including screens (for the four top-performing algorithms)	**Best algorithms**	Ensemble methods: Extra Trees, Random Forest, Ada Boost and Gradient Boosting						
	Setting 2 from 1	+6.0%	+2.3%	−0.5%	+12.1%	+0.1%	−0.1%	N/A
	Setting 4 from 3	+1.6%	+2.5%	−0.9%	+14.1%	+1.9%	−0.7%	+1.1%
Detection rates reported in the literature	**Paper/s**	[30,34]	[36]	[38]	[40]	[40]	[19]	N/A
	Method	Probabilistic methods	Standard hierarchical clustering algorithm	ML methods: Random Forest & Ensemble Method	ML method: Random Forest	ML method: Random Forest	N/A	N/A
	Accuracy	81% - 96%	N/A	88% - 93%	77% - 86%	61% - 84%	N/A	N/A

Tabla 4.5: Resumen de los resultados de detección de colusión para los diferentes datasets (Brasil, Italia, Japón, Suiza-Ticino, Suiza-SG&GR, EE.UU. y en conjunto).

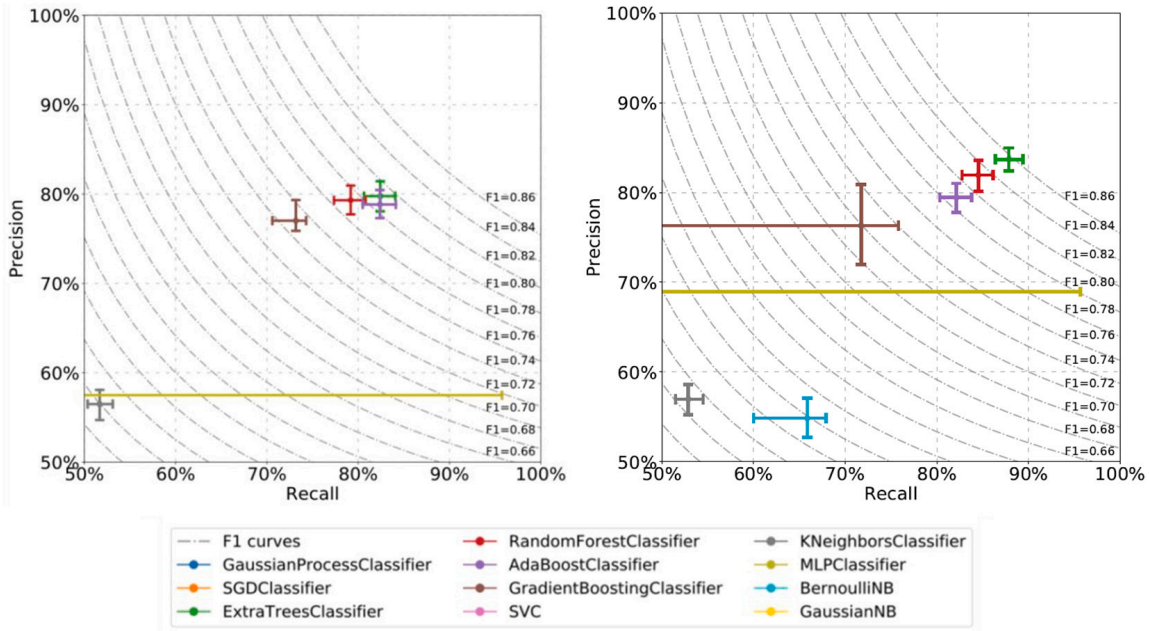


Figura 4.8: Gráficas del error (precision, recall y F1 score) para el dataset conjunto. A la izquierda para el setting 3 (campos comunes) y a la derecha para el setting 4 (campos comunes + screens).

de los ensemble methods (Ada Boost, Extra Trees, Random Forest) son los que mejor porcentaje obtienen (superiores al 80 %) y, algo peor, el otro ensemble method (Gradient Boosting). El resto de algoritmos tienen unos porcentajes de precisión y recall muy inferiores.

En conclusión, los porcentajes de aciertos son notables y confirman que los algoritmos de ML se pueden aplicar de manera exitosa a la detección de colusión. En el material suplementario del artículo se adjunta para cada dataset sus histogramas, gráficas de error como el de la anterior Figura 4.8 y los boxplots de los screens. Reseñar que los datasets y el código programado también se han liberado públicamente como material suplementario, para que cualquier investigador pueda reproducir la experimentación o probar otros nuevos. Se ha tomado esta decisión por el hecho ya mencionado, la extrema dificultad en conseguir datasets que contengan licitaciones colusivas. Con esta decisión se quiere fomentar y facilitar el desarrollo de esta disciplina.

4.6. Aplicación informática para detectar licitaciones irregulares

Finalmente, se comentará una investigación derivada de los conocimientos y experiencia acumulada gracias a la Tesis. Se ha desarrollado una aplicación que detecta pliegos técnicos irregulares y que, por tanto, es de sumo interés para los órganos públicos que velan por la contratación así como para la ciudadanía que quiere una contratación más justa y transparente. Aunque esta herramienta no ha derivado en la redacción de un artículo académico (por falta de tiempo), pone de manifiesto la utilidad de esta Tesis, su aplicación práctica y el gran interés por parte de los medios de comunicación y de la sociedad en general.

Esta herramienta fue desarrollada desde cero y totalmente ad hoc en el verano de 2021 por mí junto a un compañero (José Carlos Montes Luna, científico de datos). Sin entrar en sus particularidades técnicas, se resume a alto nivel los pasos llevados a cabo:

1. La aplicación descargó masivamente documentos asociados a licitaciones publicadas de 2015 hasta 2021, gracias a los datos en abierto de PLACSP. Alrededor de un millón de documentos (pliegos técnicos, pliegos administrativos y otros documentos) fueron descargados.

2. De cada documento se leían sus propiedades (también llamados metadatos) que incluye el autor, la organización (empresa) y fecha de creación del documento. Estos metadatos pasan inadvertidos para personas poco cuidadosas o inexpertas en informática, por no estar visibles en el propio documento digital. Nótese que este estudio tiene una importante limitación debido a que muchos de los documentos de PLACSP están escaneados (no es el documento digital original), perdiéndose los metadatos al escanearse.

3. Se compara la organización que aparece en los metadatos del pliego técnico con el nombre del adjudicatario (razón social) de la licitación[6]. No sólo se compara con el adjudicatario sino también con el grupo empresarial al que pertenece. Dicha relación se establece gracias a la información empresarial obtenidas del Registro Mercantil (la otra fuente de información utilizada). Estas comparaciones se han realizado mediante un algoritmo de IA que es capaz de relacionar dos palabras, aunque no estén escritas exactamente igual.

No tiene sentido que un pliego técnico de una licitación con procedimiento abierto o abierto simplificado contenga el nombre de la empresa ganadora[7]. Esto indicaría que el redactor de dicho pliego podría haber tenido alguna vinculación con el adjudicatario previamente a la publicación de la licitación. Que la relación entre el órgano de contratación y el ganador sea fraudulenta o no es algo que los metadatos no pueden demostrar. Lo que sí se puede afirmar es que el redactor del pliego utilizó como base un documento del adjudicatario. Por tanto, merecería la pena investigar si hubo vinculaciones irregulares entre ambos. Se detectaron varias licitaciones donde el nombre de la empresa ganadora figuraba en los metadatos del pliego y algunas de ellas fueron investigadas por el periódico *El País* que publicó una noticia[8] el 10 de diciembre de 2021 (léase en la Figura 4.9).

Si esta exitosa herramienta de detección ha sido desarrollada por un ciudadano con datos en abierto y escasos recursos, imaginemos los programas informáticos que podrían desarrollar los organismos que velan por la contratación. Organizaciones altamente especializadas que disponen de grandes recursos económicos, técnicos, humanos y que además tienen acceso a bases de datos privadas de las AA.PP.

[6]No se ha utilizado el autor que figura en las propiedades del documento porque hoy en día no se puede relacionar el nombre de una persona con la empresa donde trabaja. Sí se podría automatizar con un programa informático que tenga acceso a las bases de datos de la Agencia Tributaria o la Seguridad Social que sí disponen de la relación persona-empresa.

[7]La LCSP en su artículo 115 (Consultas preliminares del mercado) establece el mecanismo de cómo los órganos de contratación pueden realizar consultas: *"Los órganos de contratación podrán realizar estudios de mercado y dirigir consultas a los operadores económicos que estuvieran activos en el mismo con la finalidad de preparar correctamente la licitación"*. Estas consultas deberán publicarse en el perfil del contratante.

[8]Accesible en https://elpais.com/economia/2021-12-10/los-pliegos-de-tres-concursos-publicos-incluian-al-ganador-en-los-metadatos-antes-de-adjudicarse.html

Los pliegos de tres concursos incluían el ganador en los datos digitales

Una investigación revela anomalías en al menos una convocatoria de RTVE, otra del Ministerio para la Transición Ecológica y una tercera de un consorcio catalán de hospitales

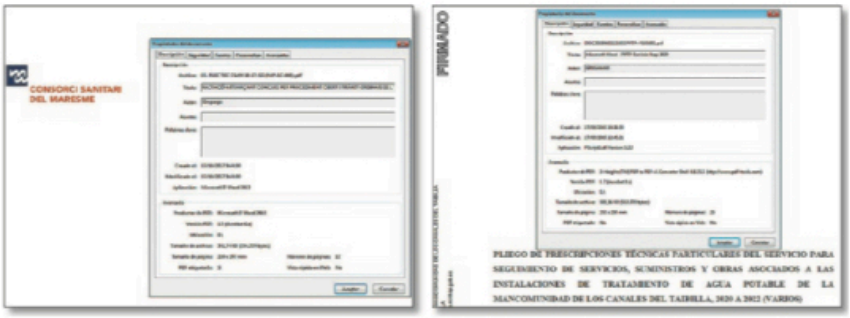
JOAQUÍN GIL, **Madrid**

Los pliegos de tres concursos públicos incluyeron en las propiedades del documento informático el nombre de la empresa ganadora. Así sucedió en Radiotelevisión Española (RTVE), en la Mancomunidad de los Canales del Taibilla, que depende del Ministerio para la Transición Ecológica, y en el Consorci Sanitari del Maresme, una entidad controlada por la Generalitat catalana que gestiona una decena de centros de salud de Barcelona.

Los datos se desprenden del estudio de tres adjudicaciones por valor de 1,5 millones concedidas entre 2018 y 2020 y subidas a la Plataforma de Contratación del Sector Público por los organismos licitantes. Los casos analizados eran procedimientos abiertos, donde puede concurrir cualquier empresa, según establece la Ley de Contratos del Sector Público de 2017.

El complejo Torrespaña de Radiotelevisión Española, en Madrid. / JAIME VILLANUEVA

Una investigación de EL PAÍS a través del sistema de análisis masivo de expedientes desarrollado por el experto en Contratación Pública de la Universidad de Oviedo Manuel García Rodríguez y el científico de datos José Carlos Montes Luna ha hallado las coincidencias. Las propiedades de los archivos de los pliegos técnicos —el documento que recoge los requisitos que debe cumplir el contrato— revelan en unos casos que el nombre de la empresa ganadora aparecía en el campo "título" de las propiedades del PDF. Y en otros, que un equipo informático con el mismo nombre de la firma agraciada creó, alteró o guardó el archivo del pliego técnico meses antes de la presentación oficial de ofertas, la adjudicación o la formalización de los contratos. Estos son los casos analizados:



Capturas de pantalla con las propiedades de los archivos informáticos que incluían los pliegos de dos de los concursos.

Mancomunidad de los Canales del Taibilla. Es un organismo adscrito al Ministerio de Transición Ecológica que abastece de agua potable a 2,4 millones de habitantes en Murcia, Alicante y Albacete. En el mes de junio de 2019 anunció una licitación para contratar a una firma de ingeniería para los años 2020 a 2022. El archivo del pliego técnico, un PDF de 35 páginas, contenía la palabra "Grusamar" en el campo autor. Fue guardado el 17 de mayo de 2019 a las 11.45. Cinco meses después, Grusamar Ingeniería y Consulting SAU, una de las dos firmas candidatas, obtenía esta adjudicación por 1.061.042 euros.

La Mancomunidad esgrimió que el vencedor presentó el proyecto con la mejor relación calidad-precio. Sin embargo, la mención a Grusamar figuraba en los metadatos (huella digital) del expediente tres semanas antes de que la publicación oficial de la licitación se enviara al *Diario Oficial de la UE*, dos meses antes del fin del plazo de recepción de ofertas y tres de que se abriesen los proyectos de los aspirantes.

Un portavoz del Ministerio de Transición Ecológica justificó así la coincidencia: "Es posible que el técnico encargado de la tramitación del contrato utilizara como plantilla un pliego ya existente que podía arrastrar metadatos que le pasaron inadvertidos". Añadió que Grusamar trabaja para la Mancomunidad de los Canales del Taibilla "desde hace muchos años" como "apoyo en la parte administrativa" y precisó que el pliego fue elaborado por un técnico del organismo público. EL PAÍS ha intentado sin éxito recabar la versión de Grusamar.

RTVE. La dirección de compras de la corporación audiovisual

> En dos de los casos, la adjudicataria de la licitación figuraba en el campo autor
>
> Las firmas y los entes afectados niegan irregularidades en los procesos

anunció el 27 de mayo de 2020 la licitación de un contrato abierto "de transporte por satélite de la señal para Norteamérica y Centroamérica de RTVE". Y calculó un valor de 272.683 euros. El organismo público divulgó ese mismo día el pliego técnico, un documento de 34 páginas que recogía los requisitos que debía cumplir el ganador. El documento informático, que fue creado dos días antes del anuncio oficial de la licitación, contenía en sus propiedades los términos "Servicios Audiovisuales Overon, SL". Dos meses después, esta compañía de telecomunicaciones por satélite con 290 empleados ganaba una adjudicación de 246.996 euros para prestar el servicio. Fue la única empresa que acudió al concurso. Servicios Audiovisuales Overon SL es una firma del grupo Imagina Media Audiovisual SA, que está administrada por el empresario y productor de cine Jaume Roures, según el registro mercantil.

Un portavoz de RTVE asegura que la corporación desconoce a qué responde el hecho de que el nombre de la empresa figure en las propiedades del archivo del pliego técnico. Y dice que la convocatoria se comunicó también a las compañías Eurovisión, Eutelsat, Eurona, Intelsat, Telefónica, además de a la ganadora. "Las empresas no conocían el pliego técnico", defiende el portavoz. Preguntada sobre la coincidencia, una representante de Mediapro, el grupo audiovisual de Roures, respondió: "No tenemos ni idea del PDF ni de los metadatos. Nosotros no hemos creado ese documento".

Consorci Sanitari del Maresme. Se trata de un organismo que depende de la Consejería de Sanidad de la Generalitat catalana y que gestiona una decena de centros de salud, entre ellos el Hospital de Mataró. De sus servicios se benefician un total de 713.014 ciudadanos. La entidad sacó a concurso en 2017 un contrato bienal para adjudicar la explotación de máquinas expendedoras de bebidas en las instalaciones sanitarias. Las empresas aspirantes, en este caso, ofertaban cuánto estaban dispuestas a abonar al organismo público a cambio de colocar sus máquinas de venta de café o de agua.

El archivo del pliego técnico de la convocatoria se creó el 17 de octubre de 2017 —dos meses antes de la apertura oficial de plicas— e incluía en el campo autor de las propiedades la palabra "Gruparpa". Se trata —según el registro mercantil— del nombre de la matriz de la firma de *vending* (venta con máquinas expendedoras) Arbitrade SA, que medio año después ganaba una adjudicación de 387.200 euros para prestar este servicio y superaba a otras seis que aspiraban también a hacer negocio con la venta de café y botellas de agua en los hospitales.

El Consorci Sanitari del Maresme niega que la entidad informara al ganador del pliego técnico antes de la convocatoria. "El contenido del concurso de 2018 se elaboró íntegramente desde el Consorcio, sin la contribución de ningún agente externo y siguiendo escrupulosamente lo que marca la ley en este sentido. Por lo tanto, no se informó al adjudicatario de su contenido en ningún momento hasta su publicación", responde un portavoz de este organismo por correo electrónico.

El vencedor del concurso de las máquinas expendedoras atribuye que su nombre aparezca como autor del pliego técnico a un error informático. "Las administraciones y los adjudicatarios intercambian documentos constantemente y aseveramos que la aparición del nombre de nuestra matriz en la autoría del archivo es pura anécdota. Las empresas facilitamos información en archivos informáticos constantemente sobre máquinas, productos, listas, precios, características o cualquier otro detalle que precise la administración", añade un representante de la empresa Gruparpa.

Figura 4.9: Aplicación informática para detectar licitaciones irregulares recogida en la noticia de *El País* publicada el 10/12/2021.

Capítulo Quinto

Discusión de los resultados

Este capítulo pretende hacer una discusión de los artículos en su conjunto, subrayando los elementos comunes y transversales de las técnicas de ML aplicadas al campo de la contratación pública. Para un mayor detalle técnico, léase el apartado de discusión de resultados de cada artículo. Además, se va a realizar ciertos análisis desde la perspectiva económica, estudios que apenas se ha hecho en los artículos y que, por tanto, los complementa. Así, se apreciarán los beneficios de aplicar la ciencia de datos a los contratos públicos. Es decir, los académicos de las ciencias sociales y económicas deben incorporar las herramientas analíticas, cuantitativas, para profundizar en sus investigaciones ligadas a la contratación.

Los artículos de la Tesis ponen de manifiesto la capacidad que tiene el Machine Learning (ML) para resolver problemas característicos de la contratación. En particular, se han tratado tres tipos de cuestiones: estimar el precio de adjudicación, buscar empresas que puedan llevar a cabo una licitación y detectar prácticas colusorias (anticompetitivas) en las licitaciones. Se podrían haber escogido otros problemas porque hay muchas preguntas a resolver en este ámbito. La aplicación de ML a este campo es algo todavía novel, incipiente, que están descubriendo las AA.PP. y el sector privado. Esta Tesis pretende servir de ejemplo y abrir las puertas a futuros investigadores y estudios. Todo lo que rodea a la contratación tiene un gran impacto económico y mucha relevancia dentro de la sociedad.

Se va a tratar dos cuestiones capitales para la contratación y que, consecuentemente, tienen especial interés para las AA.PP. y relevancia para la sociedad:

1. **Mejorar la asignación de los recursos económicos** mediante un cálculo del presupuesto global de contratación más realista. Es decir, que las AA.PP. sean capaces de hacer una mejor planificación económica de sus compras públicas utilizando precios de mercados.

2. **La concurrencia en los contratos y sus efectos económicos**. Es decir, analizar cuantitativamente el impacto que tiene la concurrencia (el número de ofertantes) en los importes de las licitaciones. Por qué el aumento de la competencia tiene un efecto beneficioso.

Comencemos por la cuestión del primer punto. Un presupuesto de contratación global lo más realista posible es que la AA.PP. correspondiente sea capaz de presupuestar todas las licitaciones lo más aproximado al precio final de adjudicación, de mercado. Para ello, se necesitan herramientas como el estimador del importe de adjudicación, objeto de estudio en los artículos [10] y [11] de la Tesis. Por tanto, el elemento fundamental de análisis es la diferencia económica entre el importe de licitación (presupuesto) y el importe de adjudicación. A esto se le llama baja económica.

Se muestra en la Figura 5.1 un diagrama de cajas (boxplot; para más detalle consultar la Figura 3.5) de la baja económica (en azul), es decir, el error porcentual absoluto (APE) entre

el importe de adjudicación y de licitación. Además, en gris se muestra la baja económica entre el importe de adjudicación y la predicción del estimador. Se muestran desagregadas por CPV (Common Procurement Vocabulary), es un código de 8 dígitos que clasifica la naturaleza de las licitaciones (el tipo de trabajo o servicio a realizar). Se han utilizado solamente la agregación de los dos primeros dígitos del CPV para tener un grupo de clasificación manejable y así poder representar gráficamente. Por ejemplo, el CPV 45 es *"trabajos de construcción"* y el 71 es *"Servicios de arquitectura, construcción, ingeniería e inspección"*[1]. Finalmente, se muestra el error porcentual absoluto medio (MAPE) y el error porcentual absoluto mediano (MdAPE). La Figura ha sido extraída del primer artículo [10], por tanto se basa en 58.337 licitaciones de España (para más detalle consultar dicho artículo).

Se observa en la ya mencionada Figura 5.1 que el rango intercuartil (IQR) es generalmente más pequeño si se utiliza el estimador del importe de adjudicación (en gris). Esto se confirma al ver que el MAPE y MdAPE suelen ser también inferiores (cuantificado en la Tabla 4.1). El MdAPE se reduce un -2,58 % [10], esto significa que las AA.PP. españolas podrían haber reducido unos 811 millones de € en sus presupuestos iniciales, pudiendo destinar ese importe a otras partidas. Recordar que el primer artículo [10] utiliza el algoritmo Random Forest, pero éste fue posteriormente superado por la ANN y la regresión isotónica en el segundo artículo [11]. Por tanto, el IQR en gris sería incluso mejor (más pequeño) que el IQR en azul si se empleasen los algoritmos del segundo artículo [11].

Un estimador del importe de adjudicación, además de la utilidad para las AA.PP., es una herramienta útil para los ofertantes (empresas), ayudándolas en su toma de decisiones y tratando de reducir la incertidumbre económica. Sin embargo, es irreal pensar que se pueda diseñar un estimador del importe de adjudicación perfecto, sin error. El mercado es un lugar abierto, libre, competitivo, sin un intercambio de información perfecta, donde los operadores económicos llevan a cabo sus acciones y decisiones. Además, el importe de adjudicación no es el finalmente pagado a los adjudicatarios. Después de la firma del contrato puede haber modificaciones durante la ejecución del contrato, repercutiendo en un cambio del importe (sobrecostes). Actualmente, en España no se publican los datos en abierto de las modificaciones del contrato y, por tanto, no se puede hacer un estudio masivo de licitaciones aplicando la ciencia de datos. Es una pena que las modificaciones no se publiquen de manera masiva y estructurada porque de su estudio y análisis se derivaría un mejor gasto público al caracterizar dichos contratos públicos.

La otra cuestión que se va a estudiar a través de los artículos de la Tesis versa sobre la concurrencia en los contratos y sus efectos económicos en España (extrapolable a otros países). En la Figura 5.2 se representa en un gráfico de barras cuántas empresas han ganado el mismo número de licitaciones [12]. Por ejemplo, el mayor número de empresas (casi 10 mil) han ganado una única licitación. Cerca de 4 mil empresas han ganado 2 licitaciones y así sucesivamente. Esto indica que en España hay una gran fragmentación de proveedores para las AA.PP., es decir, hay una considerable rotación de proveedores. En general, un mismo proveedor no es contratado por las AA.PP., no generándose sinergias por la recurrencia. Esto puede ser a causa de varios factores: los contratos públicos no son atractivos para las empresas, hay una alta competencia en el mercado (posteriormente se demostrará que esta afirmación es falsa), los adjudicatarios trabajan una vez con la AA.PP. y luego no repiten por las razones que sean (mala experiencia en el negocio, pocos beneficios, exceso de regulación, etc.), los adjudicatarios están débiles empresarialmente (no pueden competir en el mercado privado) y recurren como última instancia a llevar a cabo una licitación pero acaban quebrando, etc. Se debería hacer un análisis económico-financiero de los adjudicatarios y una encuesta a las AA.PP. y empresarios para conocer las verdaderas causas.

[1] La descripción completa de códigos CPV está en el Reglamento (CE) núm. 213/2008 de la CE, de 28 de noviembre de 2007, por el que se aprueba el vocabulario común de contratos públicos

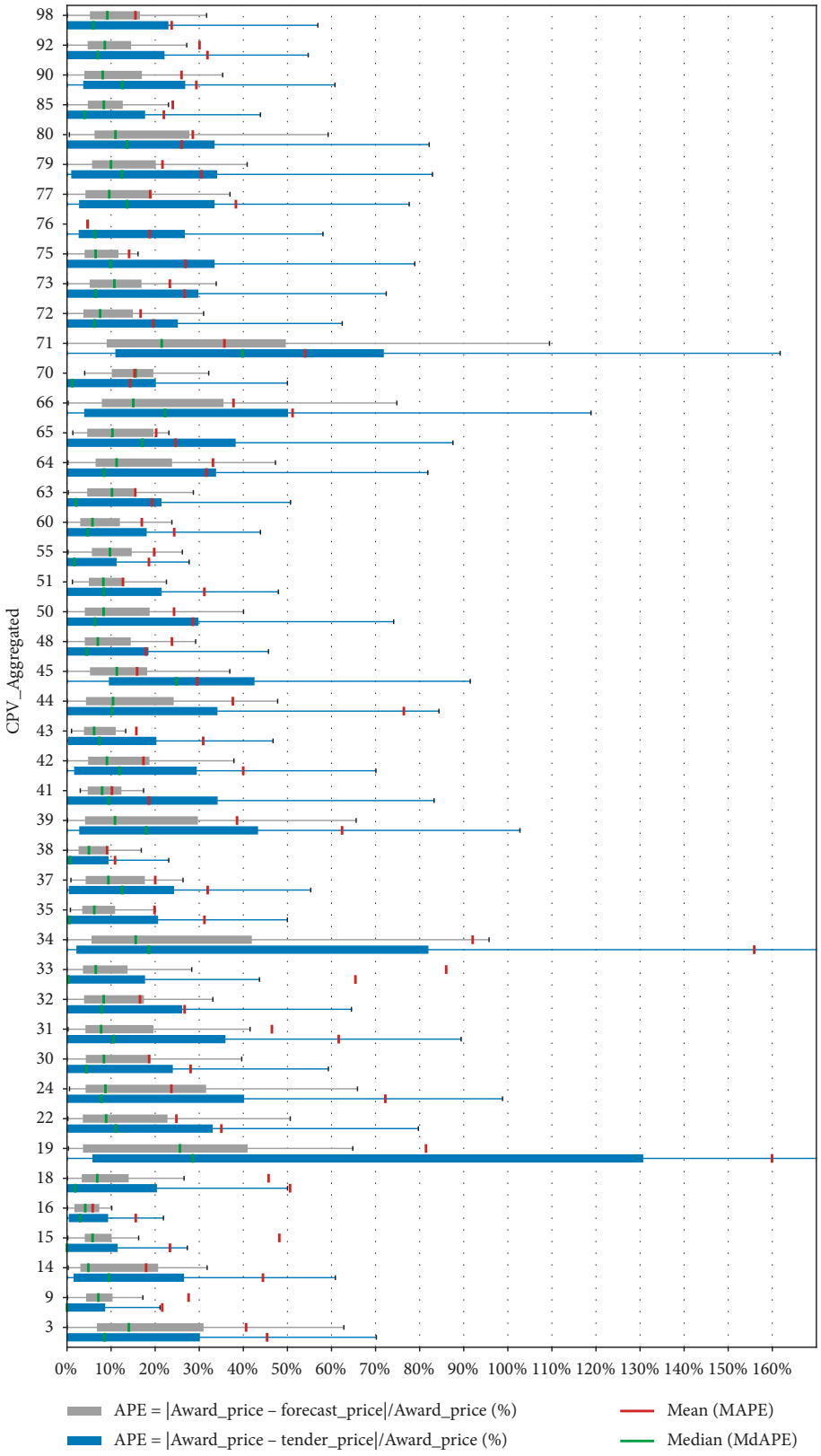


Figura 5.1: Diagrama de cajas (boxplot) del error porcentual absoluto (APE) entre el importe de adjudicación y la predicción (gris) y el APE entre el importe de adjudicación y licitación (azul). Agrupados por CPV para los datos de España. Fuente: elaboración propia en [10].

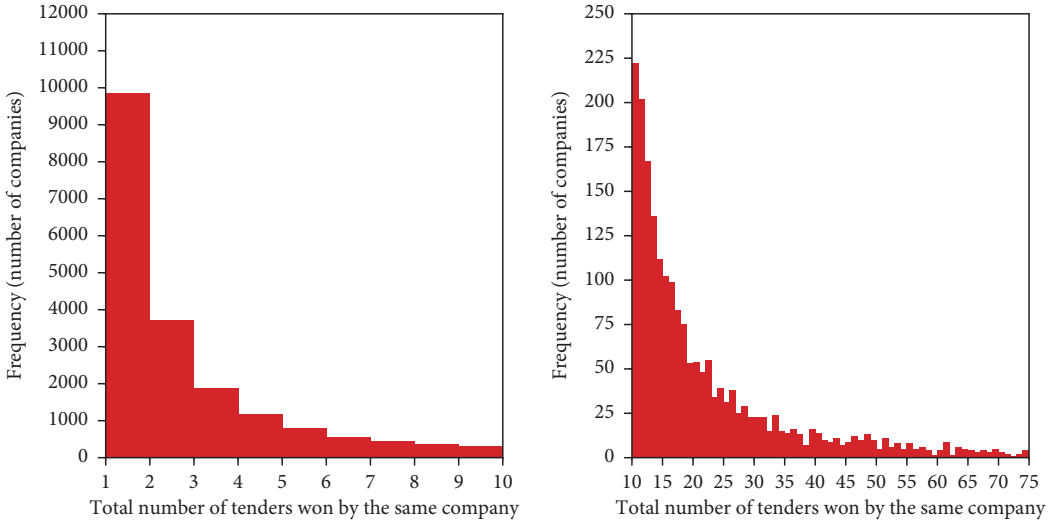


Figura 5.2: Histograma que representa el número de empresas (eje y) que han ganado el mismo número de licitaciones (eje x). Se ha dividido el gráfico en dos para una mejor visualización. Fuente: elaboración propia en [12].

En la Tabla 5.1 se dividen las licitaciones (analizadas en [10]) en 4 grupos:

1. Licitaciones que sólo han recibido una oferta (no hay concurrencia).
2. Licitaciones que han recibido entre 2 y 4 ofertas (baja concurrencia).
3. Licitaciones que han recibido entre 5 y 10 (media concurrencia).
4. Licitaciones que han recibido más de 10 (alta concurrencia).

De esta manera, se puede evaluar fácilmente la competencia que ha habido en la contratación. Se aprecia que las licitaciones tienen un importe mediano (tanto el importe de licitación como el importe de adjudicación) creciente, es decir, cuanto mayor es el grupo de concurrencia, mayor es el importe mediano. Lo mismo sucede con la baja económica media (el MAPE): aumenta cuanto mayor es el grupo de concurrencia. Para el grupo 1 (sin concurrencia) la baja es el 10,02 %, para el grupo 2 es 25,65 %, para el grupo 3 es 54,48 % y para el grupo 4 es 77,98 %. Por tanto, se aprecia claramente el efecto de que a una mayor concurrencia, competitividad, las licitaciones se logran adjudicar con una baja mucho mayor, consiguiendo las AA.PP. un importante ahorro presupuestario.

Continuando con el mismo análisis, en la Figura 5.3 aparecen dos gráficas que representan el número de ofertas (concurrencia) y la baja económica media. En la gráfica superior (elaboración propia en [12]) se observa como la baja crece hasta estabilizarse entre el 30 % y 40 %, aproximadamente. Esta estabilización ocurre a partir de las 20 ofertas por licitación. La gráfica inferior es de la Oficina Independiente de Regulación y Supervisión de la Contratación (OIReScon) publicada en su *"Informe Anual de Supervisión de la contratación pública en España"* de 2022 [21] y, por tanto, está elaborada con una mayor muestra de licitaciones. Ambas gráficas tienen una tendencia similar y corroboran que a una mayor concurrencia, una mayor baja. Dicha baja se estabiliza a partir de un número de ofertas, 20 aproximadamente.

Por tanto, el fenómeno económico de que en un mercado libre con una gran competencia produce unos bienes y servicios de mejor calidad a menores precios, se confirma también en el mercado

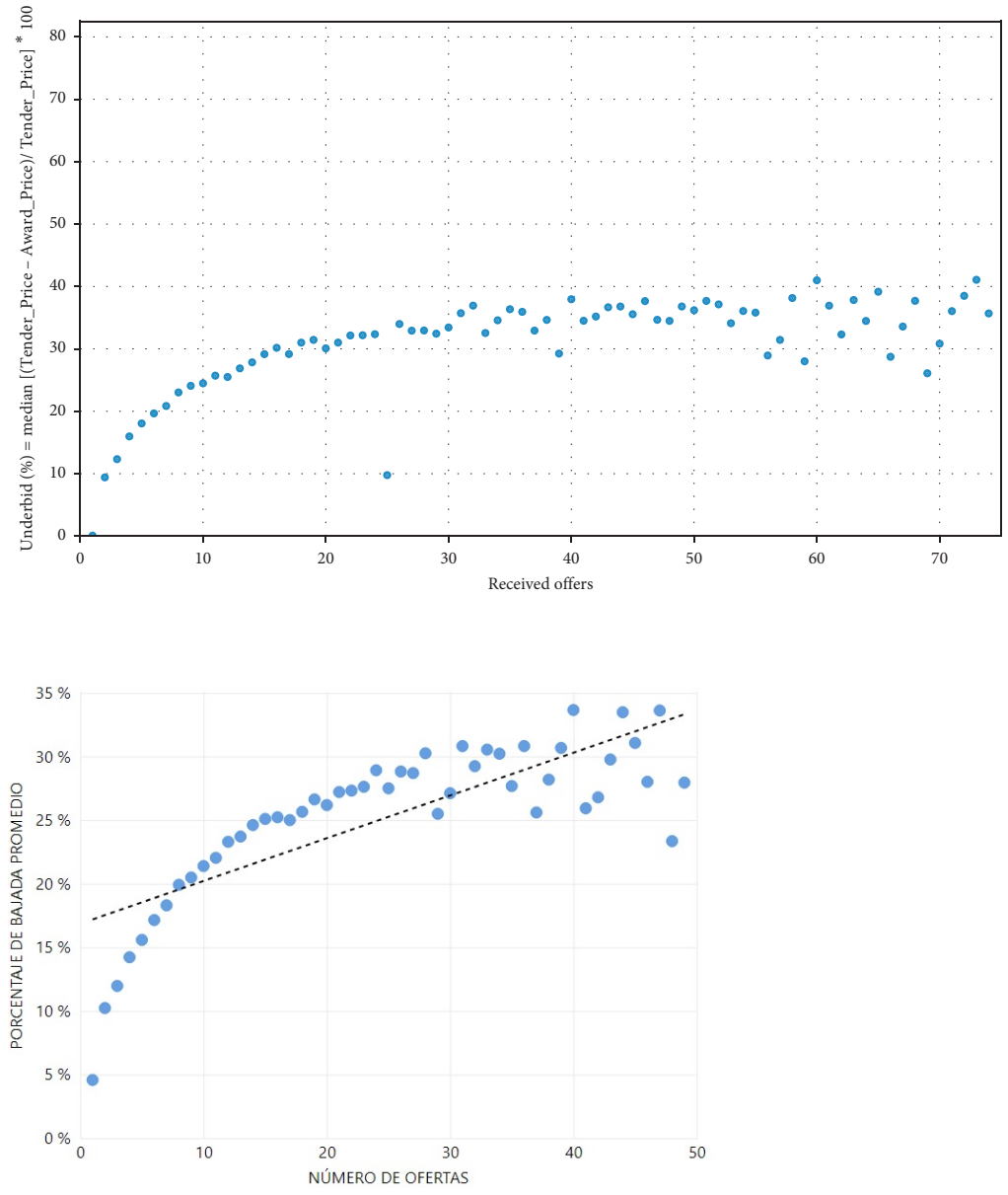


Figura 5.3: Diferencia porcentual entre el importe de licitación y adjudicación (baja económica) (eje y) según el número de ofertas recibidas en la licitación (eje x). La gráfica superior es la baja mediana (elaboración propia en [12]) y la gráfica inferior es la baja media (elaborada por la OIReScon en [21]).

Description	Groups by competitiveness			
	No competitiveness	Low Received offers (2–4)	Medium Received offers (5–10)	High Received offers >10
Total number of tenders in the dataset	18,790	22,714	11,553	5,271
Total number of tendering organisations	1,956	2,553	2,135	1,053
Total number of winning/award companies	7,550	9,555	5,222	2,402
Mean received offers by tender	1.0	2.80	6.73	20.01
Mean duration of tender's works	401.07 days	396.65 days	370.95 days	277.50 days
Mean tender price	€354,882.49	€388,526.27	€785,455.49	€1,301,031.70
Median tender price	€60,500.00	€75,000.00	€121,000.00	€254,376.00
Mean award price	€341,874.79	€323,611.87	€460,548.68	€836,188.79
Median award price	€58,984.50	€64,833.00	€90,689.00	€174,986.00
Median absolute error (MdAE)	€93.50	€7,661.50	€22,854.00	€76,420.00
Median absolute percentage error (MdAPE)	0.12%	13.39%	29.63%	45.94%
Mean absolute error (MAE)	€13,966.65	€68,244.60	€326,698.33	€464,907.75
Mean absolute percentage error (MAPE)	10.02%	25.65%	54.48%	77.98%

Tabla 5.1: Diferencia entre el importe de licitación y adjudicación (baja económica) para diferentes grupos de ofertas recibidas. Fuente: elaboración propia en [10].

de la contratación pública. Esta relación se ha demostrado cuantitativamente en artículos de otros investigadores, por ejemplo, para Italia [110] o República Checa [111, 112]. Y se menciona expresamente *"de mejor calidad"* porque las licitaciones no sólo tienen como criterio de adjudicación el precio, sino también valoran la calidad a través de los criterios de adjudicación técnicos y subjetivos. Sin embargo, aproximadamente el 32 % de licitaciones tienen solamente una única oferta, como se ha observado en la Tabla 5.1 (primera fila, expresado en porcentaje). Este porcentaje también se corrobora en la Figura 5.4, elaborada por la Comisión Europea (CE)[2] con datos oficiales de TED, donde España tiene un 28 % y ha aumentado desde el 2017 al 2020. Se deberían de implementar políticas de contratación que fomenten la participación de empresas porque si no se lastra la contratación. Por esa razón, se diseñó el buscador/recomendador de licitadores [12], herramienta que sirve para fomentar la competitividad. Además, en los escenarios de alta participación es más difícil que se produzca corrupción o colusión por ser un entorno donde, al operar muchas empresas, no se puede manipular la licitación fácilmente.

Se podrían haber estudiado más aspectos económicos analizando los datos de contratación pero hay que poner un límite y, además, queda fuera del alcance de la Tesis. Se han descrito los anteriores ejemplos para resaltar cómo se pueden hacer análisis cuantitativos en la contratación y los beneficios que aportan.

[2]Accesible en la web de la CE Single market scoreboard - public procurement.

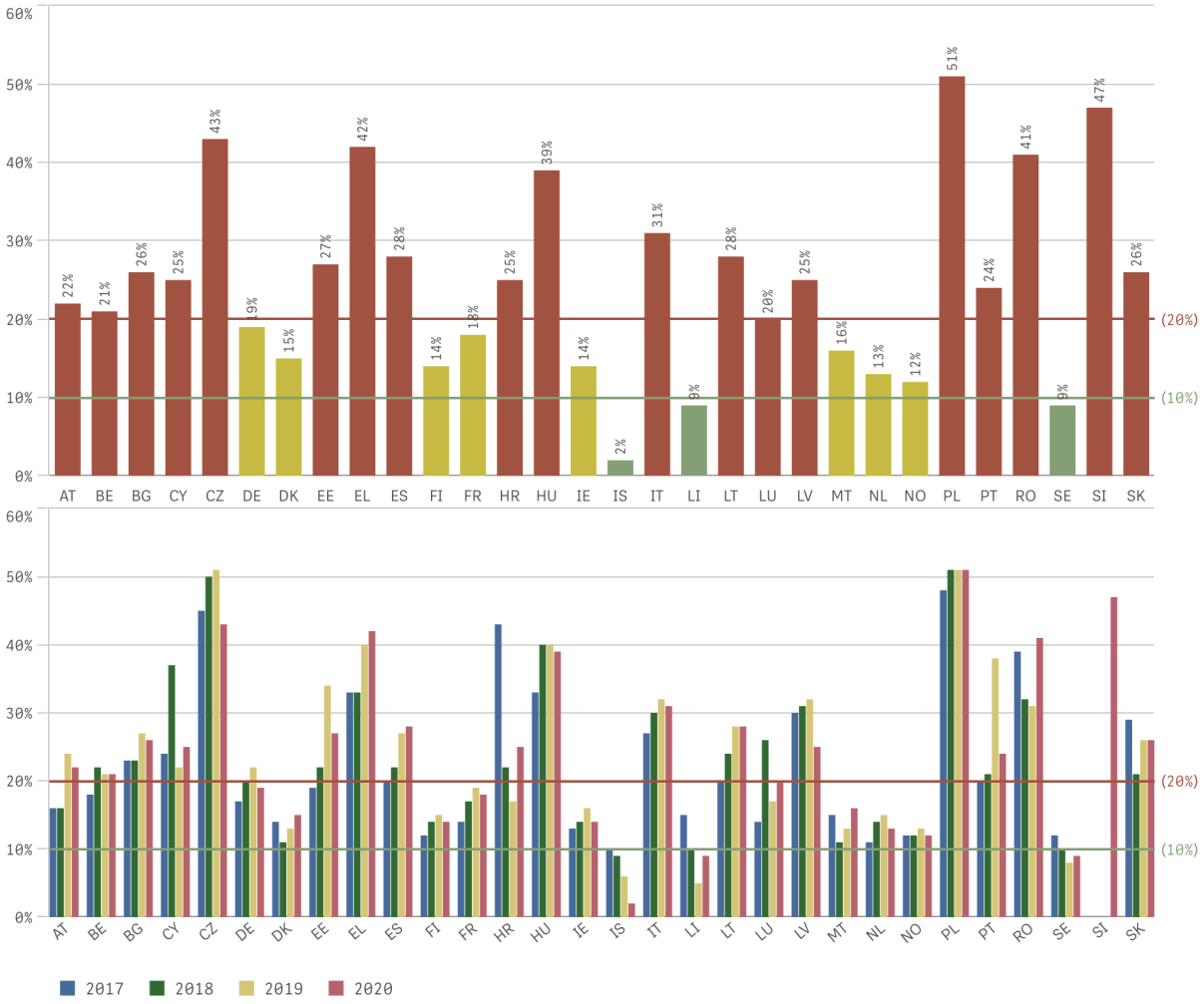


Figura 5.4: Porcentaje de licitaciones que han tenido una única oferta, dividido por países y años. Fuente: CE.

Capítulo Sexto

Conclusiones

6.1. Conclusiones

La contratación pública es un campo de gran importancia por representar un porcentaje significativo del gasto sobre el PIB de los Estados, un 16 % según algunas estimaciones oficiales [2] pero puede ser incluso más. La contratación, vista solamente como las leyes de contratos del Sector Público, ha sido estudiada por juristas, expertos en Derecho Administrativo. Sin embargo, es un campo muy poco estudiado por los investigadores del ámbito científico-tecnológico, presumiblemente porque hasta hace pocos años no se disponían de datos de licitaciones, tanto a nivel español como de la UE u otros países.

Esta Tesis es pionera en la investigación científica-tecnológica sobre la contratación pública, al abordarse desde la ciencia de datos y desarrollar nuevas herramientas gracias a las tecnologías digitales como el procesamiento masivo de datos y el Machine Learning (ML). No se conocen Tesis de similares características en España ni tampoco se han encontrado fuera. Esto queda de manifiesto en la revisión de la literatura académica mencionada en los 4 artículos de la Tesis, que es muy escasa a nivel internacional y prácticamente inexistente a nivel nacional. Tampoco abundan informes técnicos que utilicen dichas tecnologías, redactados por expertos técnicos o por las propias Administraciones Públicas (AA.PP.) No obstante, en los últimos años han surgido iniciativas y proyectos innovadores dentro de las AA.PP., desarrolladas en su mayoría por el sector privado, que aplican las nuevas tecnologías digitales para crear funcionalidades innovadoras que mejoren la gestión y supervisión de la contratación.

A lo largo de los capítulos de la Tesis, se ha introducido la importancia de la contratación, la ciencia de datos y brevemente sus evoluciones históricas. Se han descrito los retos actuales de la contratación, qué organismos públicos de España están ligados a ella, la transparencia y datos en abierto de España y las iniciativas de la Comisión Europea (CE) asociadas a los datos de contratación. Se han enumerado las causas de por qué la contratación es ahora un área digital y cuantitativa, su legislación, tecnologías aplicables a la contratación (Big Data, Data Analytics, IA, NLP, Process minning y RPA) y casos de uso que ponen en práctica dichas tecnologías. Se ha descrito en detalle la Plataforma de Contratación del Sector Público (PLACSP) desde la perspectiva de los datos en abierto: órganos de contratación que publica en PLACSP, origen, formato y calidad de los datos y el programa OpenPLACSP. Se ha recopilado literatura académica asociada a la contratación sobre 4 temáticas: datos en abierto y calidad del dato, innovación y gestión en la contratación, forecasting en la contratación y colusión y corrupción. Además, se formulan varias métricas de error típicas para resolver los problemas tratados en la Tesis (regresión y clasificación) y así poder evaluar y comparar los algoritmos de ML.

Se han descrito los 4 artículos que forman la Tesis de manera asequible a un lector no técnico. De esta manera se abre la Tesis a investigadores de otros ámbitos, como el jurídico o económico. El lector técnico puede profundizar más leyendo en detalle cada artículo, para analizar cómo se resuelve el problema y mediante qué técnicas de ML. Los artículos se titulan en español: *"Licitaciones en España: regulación, análisis de datos y estimador del importe de adjudicación usando ML"*, *"Estimador del importe de adjudicación de licitaciones usando ML: caso de estudio con licitaciones de España"*, *"Recomendador de licitadores usando ML: análisis de datos, algoritmo y caso de estudio con licitaciones de España"* y *"Detección de colusión en licitaciones aplicando algoritmos de ML"*. Además, se ha descrito la investigación titulada *"Aplicación informática para detectar licitaciones irregulares"*, que de momento no ha dado lugar a un artículo académico pero sí ha detectado licitaciones españolas donde ha podido existir actuaciones ilegales, fraudulentas.

Los artículos presentados son ejemplos útiles y prácticos de cómo la ciencia de datos y el ML ayudan a los diferentes actores de la contratación (AA.PP., empresas, investigadores, etc.) y, en última instancia, a la sociedad que sufraga con sus impuestos la compra pública y los padece si son de mala calidad. En los artículos se emplearon dos tipos de fuentes de datos: datos de contratación (PLACSP, TED y datasets de colusión) y las cuentas anuales de empresas (alojadas en el Registro Mercantil de España). Por tanto, los datos en abierto son fundamentales para poder desarrollar este tipo de investigación, realizada por una persona externa a las AA.PP. y que sólo dispone de información pública. La transparencia en la contratación y las leyes que obligan a publicar los datos en formatos abiertos y reutilizables juegan un papel fundamental en este sentido.

Además de la discusión de resultados en los propios artículos, principalmente sobre las técnicas de ML aplicadas y los porcentajes de acierto, en esta memoria de la Tesis se ha incorporado una discusión desde un punto de vista económico. Se han tratado dos cuestiones relevantes. Por un lado, mejorar la asignación de los recursos económicos mediante un cálculo del presupuesto global de contratación más realista. Por otro lado, cómo el aumento de la concurrencia en los contratos tiene un beneficio económico positivo, es decir, cuantificar el ahorro obtenido debido a una mayor competitividad. Sin embargo, también se cuantifica el bajo número de ofertantes en las licitaciones, que impide la competencia efectiva en las licitaciones. Las conclusiones particulares de cada investigación se detallan en dichos artículos (influencias de las variables, comparativa detallada de los modelos, etc.), no se ha hecho una mención explícita y repetitiva en esta memoria.

Finalmente, subrayar que la Tesis pretende servir de puente entre la contratación pública y la ciencia de datos, dos campos bastante distantes. El primero liderado por investigadores de las ciencias sociales (juristas, economistas, politólogos, etc.) y el segundo por investigadores del ámbito científico-tecnológico (ingenieros, matemáticos, informáticos, etc.). Por tanto, a esta Tesis se le denominaría de frontera, porque que ambos grupos pueden comprenderla y asimilarla para sus propias investigaciones. Se espera que sirva de ejemplo para futuros proyectos en colaboración, transversales y multidisciplinares.

6.2. Líneas de investigación futuras

Como se ha mencionado en el apartado anterior, la investigación pretende servir de ejemplo para que otros investigadores del ámbito científico-tecnológico generen nuevos estudios, debido a que hay una gran carencia. Además, por la propia naturaleza de la contratación pública, las investigaciones tendrán un impacto significativo en las AA.PP., el mercado y la sociedad. Por tanto, hay un futuro prometedor y más si se tiene en cuenta que las AA.PP. están en un periodo de profunda digitalización, es decir, están aumentando la cantidad y calidad de los datos asociados a sus procesos administrativos y, particularmente, los de contratación.

Las líneas de investigación futuras que se deberían de llevar a cabo en España y en los países extranjeros pasan por seguir profundizando en el análisis de datos de contratación, tanto en la fase de licitación (pre-award) como en la fase de adjudicación y ejecución del contrato (post-award). Por ejemplo, analizar las desviaciones en plazos, costes y calidad de la ejecución de los contratos con respecto al proyecto inicial (presupuesto económico estimado y planificación de los trabajos). ¿Existe una relación entre proyectos iniciales mal realizados y, como consecuencia, se producen sobrecostes u otras desviaciones en la ejecución del proyecto?

A continuación, se enumeran acciones que se deberían realizar para promover y facilitar investigaciones futuras:

- Mayor transparencia en la contratación, abarcando todo el ciclo de compra. Es decir, que se publiquen en formato de datos en abierto (públicos y reutilizables) las modificaciones de los contratos y las facturas de los trabajos realizados. Con esta información se llevaría a una nueva dimensión la investigación relacionada con la contratación.

- Aumentar los campos disponibles de contratación. Por ejemplo, que cuando se adjudique una licitación se publique, de manera estructurada, la identidad de los ofertantes y sus respectivas ofertas económicas. Esto no se hace en la actualidad y es una información valiosa. Así se podrán elaborar estudios económicos sobre la competencia o desarrollar herramientas de detección de colusión más potentes aplicando ML.

- Mejorar la calidad del dato de la contratación. No sólo sucede en este campo, es un problema generalizado a las AA.PP. españolas porque actualmente no hay una cultura del dato. Faltan profesionales, aplicaciones y normativas que estandaricen cómo estructurar, almacenar y gestionar los datos.

- Las fuentes de datos públicas deben ser también gratuitas. Por ejemplo, para realizar estudios de contratación es fundamental conocer a los adjudicatarios, caracterizar a las empresas. Estos datos provienen del Registro Mercantil y, en la actualidad, son públicos pero de pago. De poco sirve que la ley disponga que los datos sean públicos si tienen un coste.

- Aumentar la interoperabilidad de los sistemas informáticos de las AA.PP. Así se conseguiría realizar estudios más elaborados y complejos, de más valor añadido, al incorporar fuentes de datos de distintas organizaciones o naturaleza.

- Colaboración entre los expertos de las ciencias sociales y del científico-tecnológico para abordar los retos de la contratación de una manera integral. Es decir, investigaciones realizadas por equipos multidisciplinares que trabajen de manera coordinada y conjuntamente.

Publicación A

Licitaciones en España: regulación, análisis de datos y estimador del importe de adjudicación usando ML

Hindawi
Complexity
Volume 2019, Article ID 2360610, 20 pages
https://doi.org/10.1155/2019/2360610

WILEY | Hindawi

Research Article

Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning

Manuel J. García Rodríguez, Vicente Rodríguez Montequín, Francisco Ortega Fernández, and Joaquín M. Villanueva Balsera

Project Engineering Area, University of Oviedo, Oviedo 33004, Spain

Correspondence should be addressed to Vicente Rodríguez Montequín; montequi@uniovi.es

Received 27 June 2019; Revised 13 September 2019; Accepted 27 September 2019; Published 14 November 2019

Guest Editor: Benjamin M. Tabak

The largest project managers and adjudicators of a country, both by number of projects and by cost, are public procurement agencies. Therefore, knowing and characterising public procurement announcements (tenders) is fundamental for managing public resources well. This article presents the case of public procurement in Spain, analysing a dataset from 2012 to 2018: 58,337 tenders with a cost of 31,426 million euros. Many studies of public procurement have been conducted globally or theoretically, but there is a dearth of data analysis, especially regarding Spain. A quantitative, graphical, and statistical description of the dataset is presented. Mainly, the analysis is of the relation between the award price and the bidding price. An award price estimator is proposed that uses the random forest regression method. A good estimator would be very useful and valuable for companies and public procurement agencies. It would be a key tool in their project management decision making. Finally, a similar analysis, employing a dataset from European countries, is presented to compare and generalise the results and conclusions. Hence, this is a novel study which fills a gap in the literature.

1. Introduction

Every year, public authorities in European countries spend around 14% of GDP on public procurement, about 1.9 trillion euros [1], which is the latest estimate (2017) not including spending by utility companies. Spain is also similar, which spends around 10% to 20% of GDP [2]. Public procurement is very important in sectors such as civil construction, energy, transport, defence, IT, or health services. Therefore, it is crucial to analyse the public procurement notices, also called requests for tenders or simply tenders, to understand their behaviour in terms of prices, bidding companies, duration of projects, types of work, etc.

The growing awareness of public procurement as an innovative policy tool has recently sparked the interest of both policy makers and researchers [3]. The open data associated with public procurement and other open government data initiatives [4] are increasing mainly due to the following factors:

(i) Technological factors: software tools to manipulate big data and machine learning algorithms to analyse data (e.g., to make predictions) [5, 6].

(ii) Bureaucratic factors: standardisation of contracting language e-procurement [7, 8] and the benefits of the digitalisation of public procurement agencies [9].

(iii) Political factors: greater transparency in political decision making and design of methods of selecting suppliers for public procurement [10].

(iv) Economic factors: accuracy of the estimation of the cost [11], contract renegotiation [12], risk and uncertainty in the contracts [13], estimation of bidder participation in tenders [14] and its impact on prices [15], and globalisation—companies competing in markets far away from their origin [1].

(v) Social factors: less tolerance for inefficient political management or political irregularities in the procedure [16] and greater transparency and flexibility

in award mechanisms between public procurement agencies and private companies [17].

The layout of this paper is connected with the method employed in the research, as depicted in Figure 1. Section 2 summarises the legislation regarding public procurement notices. A tender is organised in fields, but nevertheless, it is necessary to preprocess the information to produce the dataset. The data fields involved in the process as well as how the data are preprocessed are described. Section 3 analyses the dataset (main characteristic values, correlation, dispersion, etc.), lists the evaluation metrics used (types of errors), and makes a quantitative and graphical analysis of two fundamental fields: the tender price and the award price. The competition in public tenders and its impact on savings have been analysed: how the award price is affected by the competitiveness of the companies. In Section 4, an estimator of the award price is proposed using the machine learning algorithm random forest for regression. Several fields of the tender (the name of the public procurement agency, type of contract, geographical location, type of work or service, duration, date, etc.) have been used to make the prediction. The success of the estimator is analysed based on the evaluation metrics defined previously. Furthermore, a similar analysis employing a dataset from other European countries is presented. Lastly, some concluding remarks and avenues for future research are presented in Section 5.

As far as we know, this article is the first attempt to provide an award price estimator for all types of tenders in a country using machine learning algorithms. Similar articles dealing with this topic [18, 19] have been published recently but only for construction projects and small datasets. It is typical to find literature only applied to construction projects; this is mainly because they are the biggest public procurement projects. On the contrary, the approach of this article is from a multidisciplinary perspective, and it analyses a large volume of data using machine learning techniques.

2. Spanish Public Tenders (2012–2018): Description of the Dataset

In this section, the origin and nature of the Spanish public procurement processes are analysed. Section 2.1 presents a summary of the legislation associated with public procurement and the reuse of public information. Section 2.2 lists the fields of the public procurement notice with information that appears in the announcement. Section 2.3 explains how the original information has been preprocessed to finally obtain a dataset which is valid for statistical and mathematical analysis.

2.1. European and Spanish Legislation on Public Procurement and on the Reuse of Public Information. At the European and Spanish levels, laws have been developed related to the reuse of public sector information and procurement or contracting in the public sector. They are summarised in Table 1. According to *Spanish Law 20/2013*, the website of the Public Sector Contracting Platform (P.S.C.P.) of Spain has to publish the public procurement notices and their resolutions of all contracting agencies belonging to the Spanish Public Sector.

With regard to official announcements of Spanish tenders outside Spain, Article 135 of *Law 9/2017* establishes that when tenders are subject to harmonised regulations (those with an amount greater than a threshold or with certain characteristics, stipulated in Articles 19 to 23), tenders have to also be published in *The Official Journal of the European Union* (OJEU) [20]. When the public contracting authority considers it appropriate, tenders not subject to harmonised regulations can be announced in the OJEU. The Europe Union (EU) has an Open Data Portal [21] which was set up in 2012, following *Commission Decision 2011/833/EU* on the reuse of commission documents. All EU institutions are invited to make their data publicly available whenever possible.

Furthermore, there is a portal called Tenders Electronic Daily (TED) [22] dedicated to European public procurement. It provides free access to business opportunities in the EU, the European Economic Area, and beyond.

2.2. Data Fields of Spanish Public Procurement Notices. The information of public procurement notices is defined in *Spanish Law 9/2017*, Annex III "Information that has to appear in the announcements." P.S.C.P. has an open data section for the reuse of this information (in compliance with the publicity obligations established in *Law 9/2017*) which will be used in this article to generate the dataset. The information is provided by the Ministry of Finance (link in the Data Availability section) and has been published as open data since 2012 and updated monthly in XML format.

The fields of the public procurement notices are numerous, and they can completely define the tender. The most important fields are as follows (more details in Table 2):

(i) Announcement fields: tender status, contract file number, object of the contract, tender price (budget), duration of the contract, CPV classification, contract type, contract subtype, place of execution, lots, type of procedure, contracting system, type of processing, contracting body, place and deadline for submission of tenders, participation requirements, award criteria, subcontracting conditions, contract modifications, etc.

(ii) Award fields: award result, identity of the winning company (CIF and company name), award price, number of received offers, maximum and minimum received bids, etc.

Not all fields have been selected (last column in Table 2) to mathematically analyse the tenders for several reasons:

(1) Some fields are usually empty or have inconsistent data or errors.

(2) Not all fields have the same importance. For example, the tender price is more important than the language of the tender document.

(3) The content of many of these fields is textual, which makes their mathematical modelling very complex.

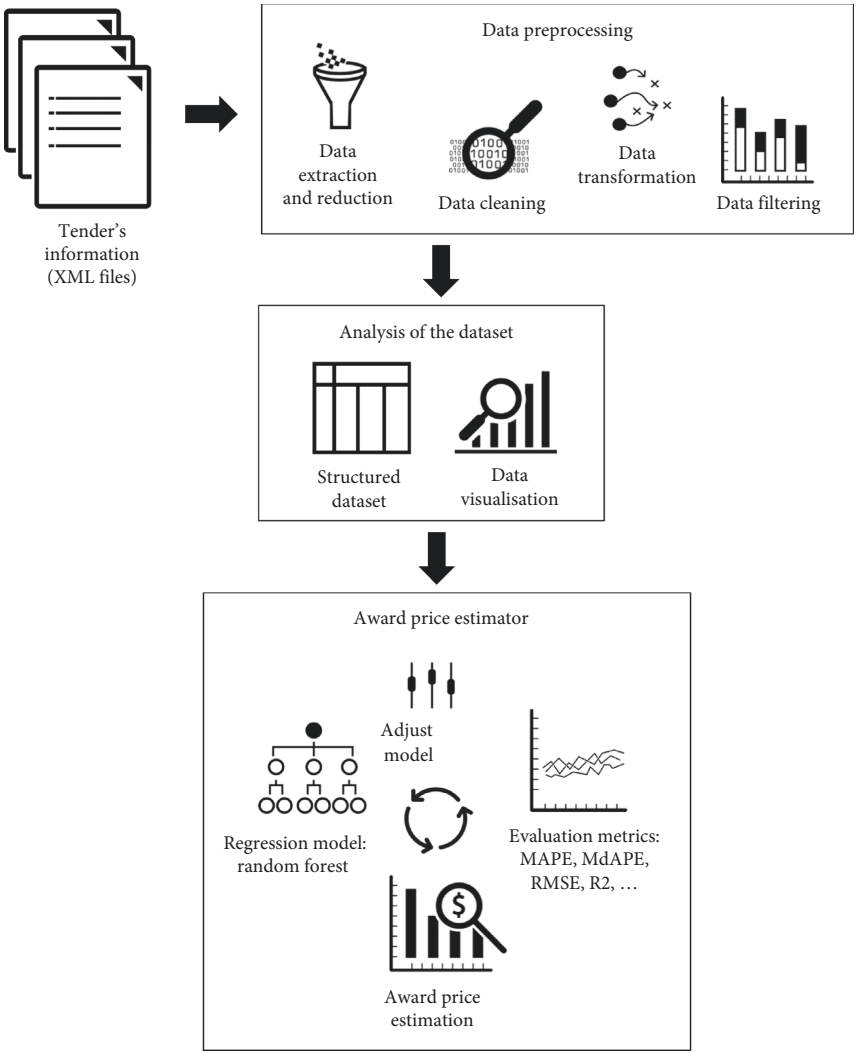


Figure 1: Flowchart of the data analysis and award price estimator.

2.3. Data Preprocessing. It is necessary to carry out several steps to preprocess the data. This is a laborious task because the tender's information has not been verified automatically to correct human errors. The preprocessing can be divided into the following 5 consecutive tasks:

(1) *Data Extraction*. Structured data are stored in text files (XML format). A script has been created to read the fields recursively, saving in the database one tender per row and as many columns as there are fields to be stored.

(2) *Data Reduction*. Around 60 fields are selected; a priori they are interesting for the performance of a statistical and mathematical analysis.

(3) *Data Cleaning*. The data are cleaned. For example, deleting spaces, punctuation marks, and special characters, conversion to capital letters, deleting data with fixed structure (postal code, CPV, CIF, etc.) which do not obey the structure's rules, etc.

(4) *Data Transformation*. Basically, four types of transformations are carried out:

   (a) *Normalisation*. This consists of homogenising the fields. For example, converting dates to time stamps.

   (b) *Aggregation*. This consists of adding new useful fields for the analysis. For example, creating a new field which is the first two numbers of the CPV classification (common procurement vocabulary).

   (c) *Data Enhancement*. It serves to create fields with external information and thus enables checking the consistency of the extracted data. For example, employing the postal code of the tender, it has generated its geographical

Table 2: Continued.

Name	Description	Name column dataset
Contract execution place	Contract's execution has a place through the Nomenclature of Statistical Territorial Units (NUTS), created by Eurostat [23]	Not used (assumed equal to Postalzone)
Type of procedure	Procedure by which the contracts was awarded: open, restricted, negotiated with advertising, negotiated without publicity, competitive dialogue, internal rules, derived from framework agreement, project contest, simplified open, association for innovation, derivative of association for innovation, based on a system dynamic acquisition, bidding with negotiation, or others	**Procedure_code**
Contracting system	The contracting system indicates whether it is a contract itself or a framework agreement or dynamic acquisition system	
Type of processing	Type of processing: ordinary, urgent, or emergency	**Urgency_code**
Award result	Type of results: awarded, formalised, desert, resignation, and withdrawal	**Result_code**
Winner identifier	Identifier of the winning bidder (called CIF in Spain) and its province (region)	**CIF_Winner Winner_Province**
Award price	Amount offered by the winning bidder of the contract (taxes included)	**Award_Price**
Date	Date of agreement in the award of the contract	**Date**
Number of received offers	Number of received offers (bidders participating) in each tender	**Received_Offers**

location (latitude and longitude), the municipality, the province, and the autonomous community.

(d) *Conversion*. This consists of converting fields from one format to another. For example, conversions of text fields (strings) to a unique numeric identifier (integers) because the regression algorithm used only works with numeric variables: string_1 > 1, string_2 > 2, ..., string_N > N.

(5) *Data Filtering*. The data are filtered to discard useless data for our analysis. Basically, this involves the following:

(a) Only formalised or awarded tenders are selected.
(b) A tender is removed when it has one or several empty fields.
(c) A tender is removed when it has an abnormally large positive price (award price or tender price) to remove outliers.
(d) A tender which is formed by several different contracts (called lots) is removed. This is because it does not give the tender price for each contract, and this is a fundamental field for further analysis.

At first, there were 232,175 tenders. After data preprocessing, there were 58,337 tenders.

3. Statistical Analysis of the Dataset

In Section 3.1, a quantitative description of the dataset and a correlation analysis between fields of dataset are presented. In Section 3.2, nine evaluation metrics are defined. In Section 3.3, they are used to calculate the error between two very important fields: tender price versus award price.

3.1. General Description. These data preprocessing operations prepare a structured and organised dataset ready for the data analysis. There are 58,337 tenders from 2012 to 2018 spread across Spain. Table 3 shows the quantitative description of the dataset: total numbers, means, medians, maximum, etc. The dataset has 19 fields or variables: 15 announcement fields and 4 award fields. Special emphasis is placed on Tender_Price and Award_Price. The amount is one of the most important variables in any project. Furthermore, the amount is fundamental in this article because an award price estimator is made.

Looking at Table 3, the following issues are observed:

(i) There are a lot of winning companies and bidding organisations. On average, each public procurement agency makes 16.46 tenders and each company wins 3.37 tenders.

(ii) There is a great dispersion of prices (for both Tender_Price and Award_Price) looking at the median, the mean, and the maximum.

(iii) There is a big difference between Tender_Price and Award_Price looking at the differences between both medians (€14,897) and means (€135,812.48). Therefore, it makes sense to propose a predictor of Award_Price because Tender_Price is not an accurate estimator.

(iv) The 5 types of CPV with greater weight add up to 48.55% of the total number of tenders.

Table 1: Laws about public procurement and the reuse of public sector information.

Law	Description	Level	Permanent link
Directive 2003/98/EC	Reuse of public sector information	Europe	http://data.europa.eu/eli/dir/2003/98/oj
Directive 2013/37/EU	Modifying previous directive 2003/98/EC	Europe	http://data.europa.eu/eli/dir/2013/37/oj
Directive 2007/2/EC	Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)	Europe	http://data.europa.eu/eli/dir/2007/2/oj
Law 37/2007	Transposing into Spanish law the European directive 2003/98/EC	Spain	https://boe.es/eli/es/l/2007/11/16/37
Royal Decree-Law 1495/2011	Developing the Spanish law 37/2007	Spain	https://boe.es/eli/es/rd/2011/10/24/1495
Commission Decision 2011/833/EU	On the reuse of commission documents	Europe	http://data.europa.eu/eli/dec/2011/833
Law 19/2013	Transparency, access to public sector information and good governance	Spain	https://boe.es/eli/es/l/2013/12/09/19
Law 20/2013	Market unit guarantee	Spain	https://boe.es/eli/es/l/2013/12/09/20
Law 18/2015	Transposing into Spanish law the European directive 2013/37/EU	Spain	https://boe.es/eli/es/l/2015/07/09/18
Directive 2014/23/EU	Award of concession contracts	Europe	http://data.europa.eu/eli/dir/2014/23/oj
Directive 2014/24/EU	Public procurement	Europe	http://data.europa.eu/eli/dir/2014/24/oj
Law 9/2017	Transposing into Spanish law the previous European directives 2014/23/UE and 2014/24/UE	Spain	https://boe.es/eli/es/l/2017/11/08/9

Table 2: Most relevant data fields in the public procurement notices (tenders) used in the dataset

Name	Description	Name column dataset
Tender status	Status of the tender during the development of the procedure: prior notice, in time, pending adjudication, awarded, resolved or cancelled	Not used (similar to Result_code)
Contract file number	Unique identifier for a contract file	Not used
Object of the contract	Summary description of the contract	Not used (unstructured textual information)
Public procurement agency	Public procurement agency that made the tender: name, identifier (NIF or DIR3), website, address, postal code, city, country, contact name, telephone, fax, e-mail, etc	**Name_Organisation Postalzone Postalzone_CCAA Postalzone_Province Postalzone_Municipality**
Tender price	Amount of bidding budgeted (taxes included)	**Tender_Price**
Duration	Time (days) to execute the contract	**Duration**
CPV classification	CPV (common procurement vocabulary) is a European system for classifying the type of work in public contracts defined in the Commission Regulation (EC) No 213/2008: http://data.europa.eu/eli/reg/2008/213/oj The numerical code consists of 8 digits, subdivided into divisions (first 2 digits of the code), groups (first 3 digits), classes (first 4 digits), and categories (first 5 digits)	**CPV** **CPV_Aggregated** (first 2 digits of the code)
Contract type	Type of contract defined by legislation (Law 9/2017): works, services, supplies, public works concession, works concession, public services management, services concession, public sector and private sector collaboration, special administrative, private, patrimonial, or others	**Type_code**
Contract subtype	Code to indicate a subtype of contract. If it is a type of service contract: based upon the 2004/18/CE Directive, Annex II. If it is a type of work contract: works contract codes defined by the Spanish DGPE	**Subtype_code**

Table 3: Quantitative description of the dataset.

Topic	Description	Value
General values	Total number of tenders in the dataset	58,337
	Temporal range of tenders	2012/01/01–2018/12/28
	Total number of tendering organisations	3,544
	Total number of winning/award companies	17,305
	Mean number of offers received per tender	4.55
	Mean duration of tender's works	382.21 days
Dataset's variables	Input variables of tender's notice: Procedure_code, Urgency_code, Type_code, Subtype_code, Result_code, Name_Organisation, Postalzone, Postalzone_CCAA, Postalzone_Province, Postalzone_Municipality, Tender_Price, CPV, CPV_Aggregated, Duration, and Date	15 input variables (description in Table 2)
	Output variables of tender's resolution: Award_Price, Winner_Province, CIF_Winner, and Received_Offers	4 output variables (description in Table 2)
Tender price (taxes included)	Mean tender price	€538,707.39
	Median tender price	€86,715.00
	Maximum tender price	€3,196,970,000
	Aggregated tender price of all tenders	€31,426,572,936
Award price (taxes included)	Mean award price	€402,894.91
	Median award price	€71,818.00
	Maximum award price	€786,472,000
	Aggregated award price of all tenders	€23,503,680,419
Number of tenders by CPV	Tenders with CPV 45: construction work	12,166 (20.85%)
	Tenders with CPV 50: repair and maintenance services	5,174 (8.87%)
	Tenders with CPV 79: business services (law, marketing, consulting, recruitment, printing, and security)	3,992 (6.84%)
	Tenders with CPV 72: IT services (consulting, software development, Internet, and support)	3,725 (6.39%)
	Tenders with CPV 34: transport equipment and auxiliary products to transportation	3,264 (5.60%)
Number of tenders by type code	Tenders with Type_code 1: goods/supplies	17,876 (30.64%)
	Tenders with Type_code 2: services	28,363 (48.62%)
	Tenders with Type_code 3: works	12,008 (20.58%)

To obtain new relevant information through the variables, the Spearman correlation method was used; Figure 2 shows the Spearman correlation matrix (a symmetric matrix with respect to the diagonal). Among the three typical correlation methods (Pearson, Kendall, and Spearman), the Spearman correlation method is chosen because it evaluates the strength of a monotonic relationship between two variables. A monotonic function preserves order (increasing or decreasing). The Spearman correlation coefficient (r_s) is defined for a sample of size n, and the n raw scores X_i, Y_i are converted to ranks rg_{X_i}, rg_{Y_i}:

$$r_s \quad \frac{\operatorname{cov}(rg_X \; rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \tag{1}$$

where $\operatorname{cov}(rg_X \; rg_Y)$ is the covariance of the rank variables and σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.

Looking at Figure 2, the greatest correlations are the following:

(i) Tender_Price vs. Award_Price (0.97): this high correlation is in accordance with common sense since high bids are associated with high awards and low bids with low awards.

(ii) Type_code vs. Subtype_code (0.74): each type of contract has its associated subtypes of contract. This is the reason for the high correlation.

(iii) Name_Organisation vs. Postalzone_Municipality (0.42): each public procurement agency has a location associated with a postal code.

(iv) Type_code vs. CPV (0.38): each type of contract is usually used for certain types of works.

(v) Procedure_code vs. Tender_Price (−0.38) and Award_Price (−0.36): each type of contract procedure tends to correspond to a range of bidding and adjudication amounts.

(vi) CPV vs. Duration (0.34): each type of work is usually associated with a temporal range (duration) for its realisation.

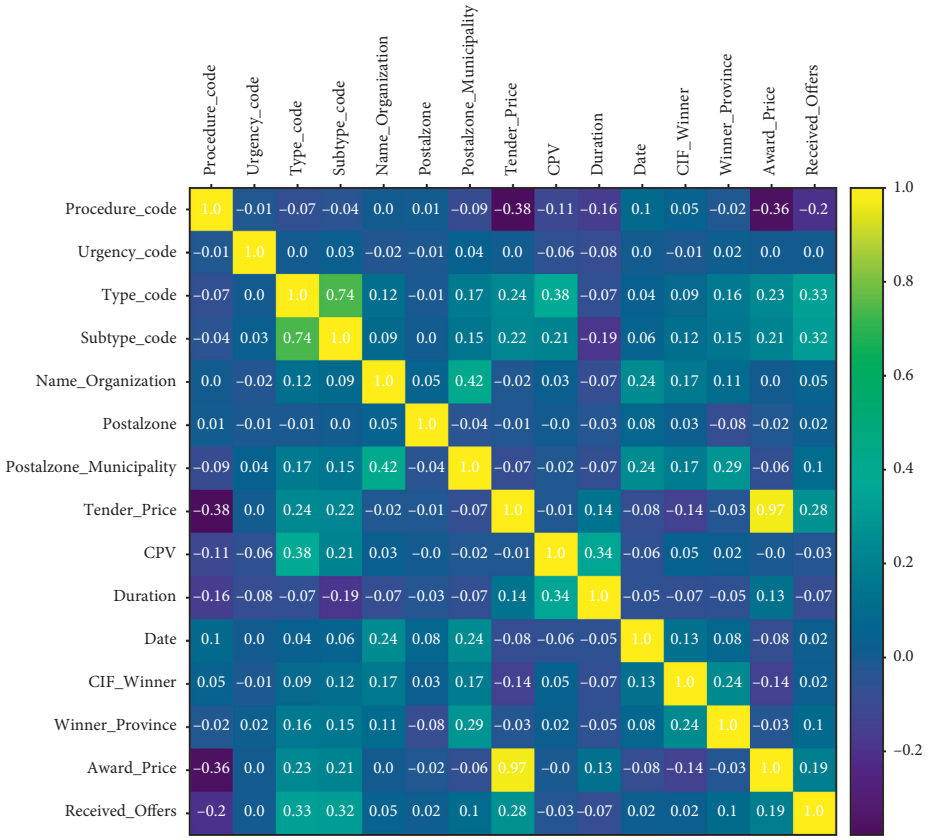


Figure 2: Correlation matrix between the variables of the dataset. Spearman's rank correlation coefficient is the method applied.

(vii) Received_Offers vs. Type_code (0.33) and Subtype_code (0.32): the number of received offers by tender has a correlation with the type and subtype of the contract.

(viii) Winner_Province vs. Postalzone_Municipality (0.29): there is a correlation between the origin (province) of the winning company and the location (municipality) of the tender. In general, tenders from a specific geographical region are won by companies from the same region. There are different socioeconomic reasons for this.

Higher correlation values have not been obtained due to the numerical form of expressing the information and the limitations of the correlation method (all methods have disadvantages). For example, Name_Organisation and Postalzone_Municipality have a direct relation: an organisation usually has a unique assigned postal code. However, this relation can follow any mathematical pattern or function.

Another way to analyse the data is through the scatter matrix (see Figure 3) where the variables are plotted two by two and the matrix's diagonal is the probability density function of the corresponding variable. Although it cannot be appreciated in detail by the large amount of data and variables, the following relations are seen:

(i) Procedure_code, Urgency_code, Type_code, and Subtype_code generate straight lines because they are variables with few values (they are codes) but have great dispersion when they are confronted with the rest of the variables.

(ii) Name_Organisation, Postalzone, and Postalzone_Municipality have a large dispersion. In the probability density function of Postalzone, a great maximum is seen in Madrid's postal codes. This is because many tenders in Spain have been put forward by agencies located in the capital (Madrid).

(iii) The CPVs show that some codes have high tender and award prices, a longer duration, and more received offers. This is true because each type of work has certain characteristics such as price, duration, or competence in the sector.

(iv) The relation between Tender_Price and Award_Price will be analysed in detail later, but a certain relation can be seen. It had already appeared in the correlation matrix.

3.2. Evaluation Metrics. To compare the variables and calculate the errors or deviations of the prediction algorithms, first it is necessary to define some error metrics. The use of

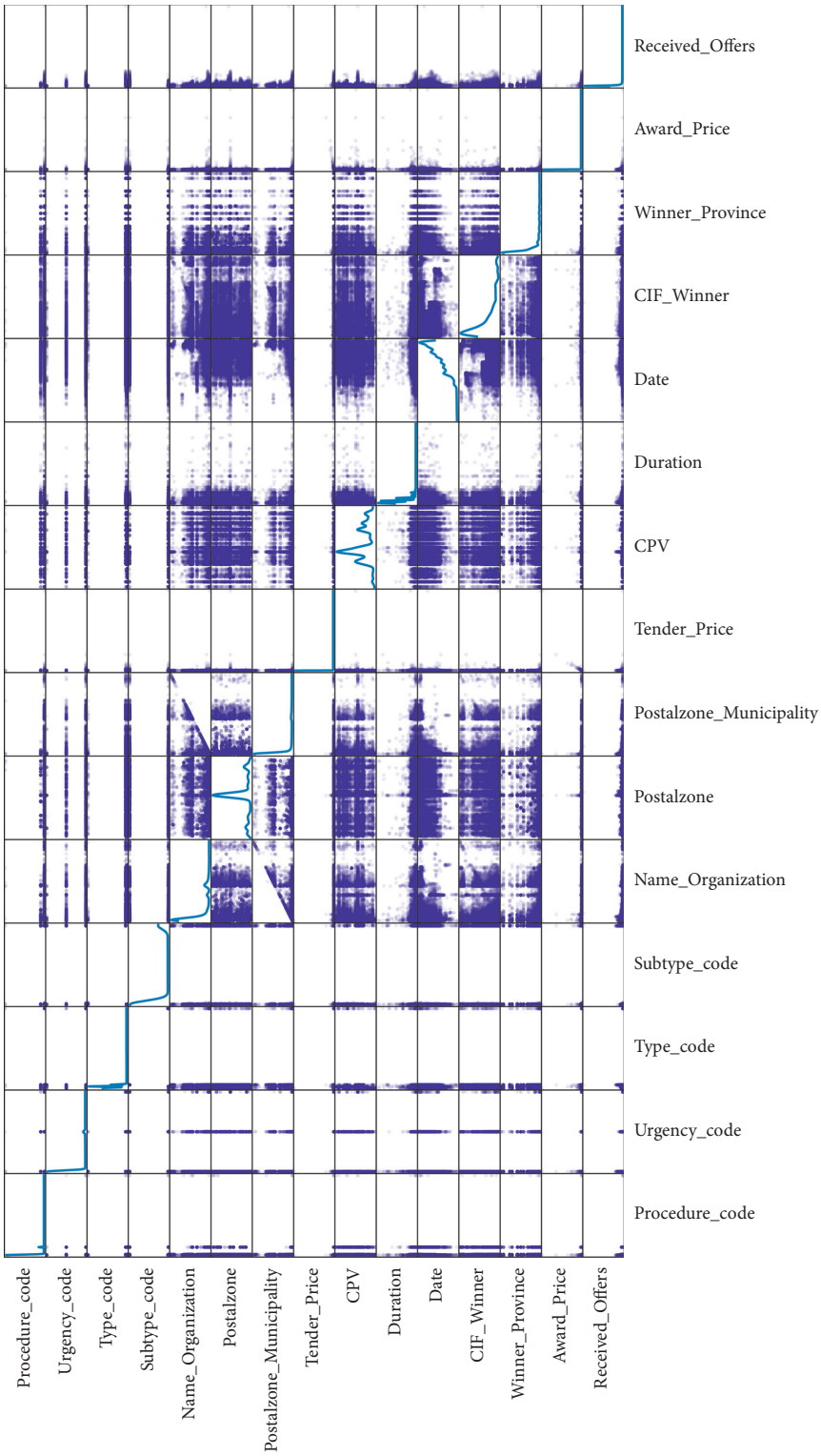


Figure 3: Scatter matrix between the variables of the dataset.

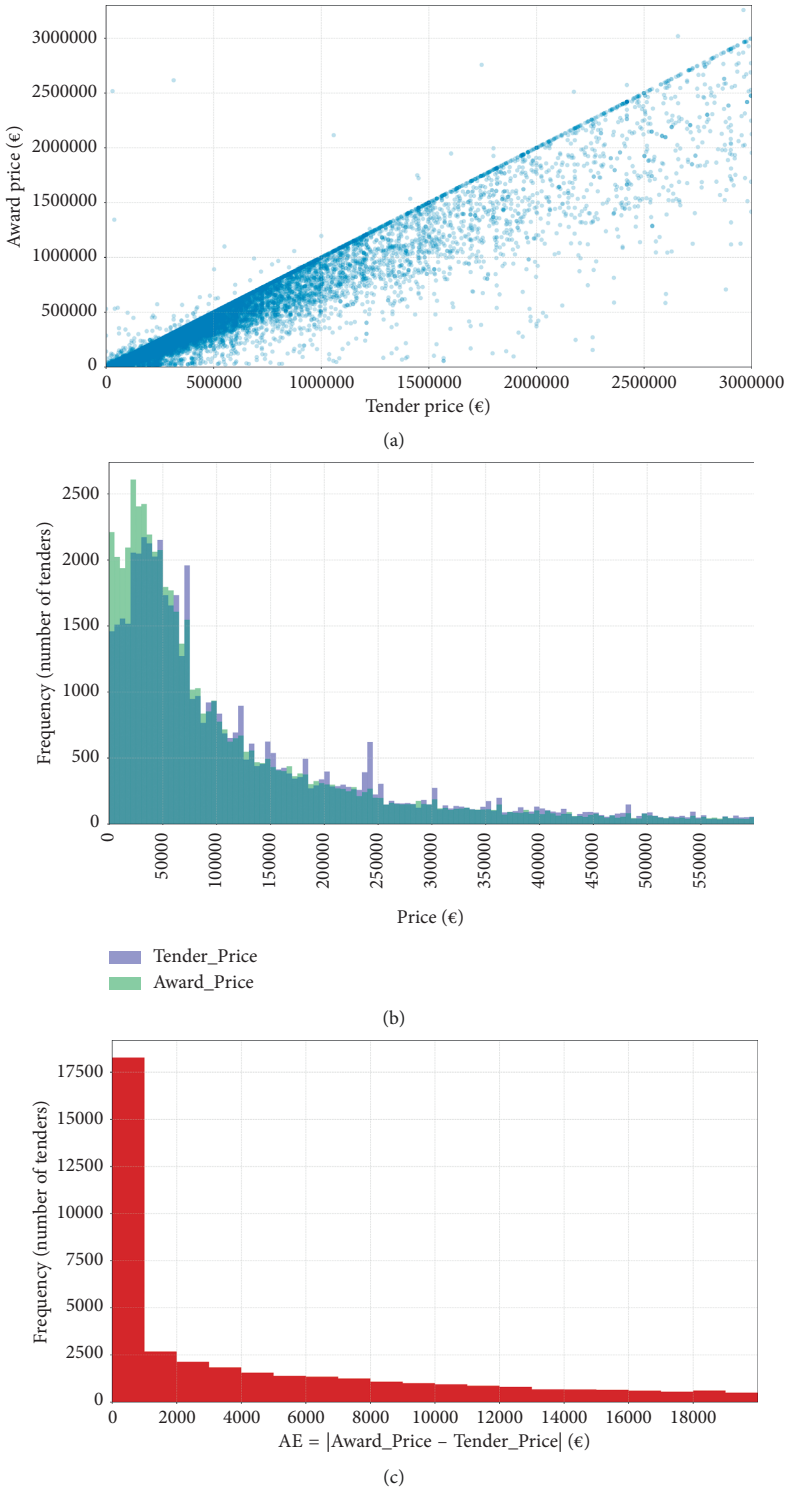


Figure 4: Relation between tender price and award price. (a) Scatter plot. (b) Histograms of frequency (number of tenders). (c) Absolute error (AE) histogram.

metrics based on medians and relative percentage is useful in this survey because the dataset has outliers of great weight, and the use of such metrics helps us to counteract the effect of these outliers.

Absolute error (AE), absolute percentage error (APE), mean absolute error (MAE), mean absolute percentage error (MAPE), median absolute error (MdAE), median absolute percentage error (MdAPE), root mean square error (RMSE), normalised root mean square error (NRMSE), and coefficient of determination (R^2) were selected as evaluation criteria (2)–(10): A_t is the actual value for period t, F_t is the expected or estimated value for period t, and n is the number of periods.

$$\mathrm{AE}_t \quad |A_t - F_t| \tag{2}$$

$$\mathrm{APE}_t\,(\%) \quad 100\left|\frac{\mathrm{AE}_t}{A_t}\right| \quad 100\left|\frac{A_t - F_t}{A_t}\right| \tag{3}$$

$$\mathrm{MAE} \quad \frac{1}{n}\sum_{t\ 1}^{n}\mathrm{AE}_t \quad \frac{1}{n}\sum_{t\ 1}^{n}|A_t - F_t| \tag{4}$$

$$\mathrm{MAPE}\,(\%) \quad \frac{100}{n}\sum_{t\ 1}^{n}\mathrm{APE}_t \quad \frac{100}{n}\sum_{t\ 1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{5}$$

$$\mathrm{MdAE} \quad \frac{1}{n}\mathrm{median}\left(|A_1 - F_1|\ |A_2 - F_2|\ \ldots\ |A_n - F_n|\right) \tag{6}$$

$$\mathrm{MdAPE}\,(\%) \quad \frac{100}{n}\mathrm{median} \cdot\left(\left|\frac{A_1 - F_1}{A_1}\right|\ \left|\frac{A_2 - F_2}{A_2}\right|\ \cdots\ \left|\frac{A_n - F_n}{A_n}\right|\right) \tag{7}$$

$$\mathrm{RMSE} \quad \sqrt{\frac{1}{n}\sum_{t\ 1}^{n}|A_t - F_t|^2} \tag{8}$$

$$\mathrm{NRMSE} \quad \frac{\mathrm{RMSE}}{\max(A_t) - \min(A_t)} \quad \frac{\sqrt{(1/n)\sum_{t\ 1}^{n}|A_t - F_t|^2}}{\max(A_t) - \min(A_t)} \tag{9}$$

$$R^2 \quad 1 - \frac{\sum_{t\ 1}^{n}|A_t - F_t|^2}{\sum_{t\ 1}^{n}|A_t - \overline{A}|^2} \tag{10}$$

where $\overline{A}$ is the mean: $\overline{A} \quad (1/n)\sum_{t\ 1}^{n}A_t$.

3.3. Tender Price vs. Award Price. Figure 4(a) shows graphically the variable tender price versus award price for all tenders when tender price is less than €3,000,000. This threshold is 3.5 times the median of tender price. A line at 45 degrees can be seen; its points satisfy the condition that tender price is equal to award price. Therefore, in this line there, is no error between the two variables, and so the tender price would be a perfect estimator. Below this line, there is a large dispersion of points. When the distance between a point and the line is high, the error is also high. Finally, there are few points above the line. This is because only rarely is the award price higher than the tender price. This can happen due to special conditions of the contract or, alternatively, it can be wrong data. There is no information about how the public procurement agencies calculate the tender price or if it is validated before entering the dataset.

Table 4: Error metrics between tender price and award price.

Error	Value
Absolute error (AE)	See Figure 4(c)
Absolute percentage error (APE)	See Figure 5
Median absolute error (MdAE)	€6,955.00
Median absolute percentage error (MdAPE)	11.84%
Mean absolute error (MAE)	€137,778.64
Mean absolute percentage error (MAPE)	39.79%
Root mean square error (RMSE)	101,451,609,620,714
Coefficient of determination	−3.10

Figure 4(b) shows the frequency histogram of both variables. The frequency is the number of tenders for each bar of €5,000. For example, the most frequent range for the tender price is €30,000–€35,000; for the award price, it is €20,000–€25,000. Figure 4(c) shows the frequency histogram of the AE between both variables by ranges of €1,000. It can be observed that approximately 18,000 tenders (30% of the total) have less than €1,000 error. There is a big difference with the rest of the bars.

Table 4 presents the error metrics (or evaluation metrics) calculated between the variables tender price and award price for the entire dataset. An error between tender price and award price, in terms of project management, means that there is a budget deviation between the tender price and the price finally awarded.

An interesting analysis is how the award price is affected by the competitiveness of the companies (see Table 5). It is necessary to group the tenders according to the number of offers received. For this purpose, 4 groups have been created: no competitiveness (1 offer), low competitiveness (2–4 offers), medium competitiveness (5–10 offers), and high competitiveness (more than 10 offers). As competitiveness increases, the difference between the award price and tender price is greater because MdAE, MdAPE, MAE, and MAPE are greater. This shows that companies are more aggressive (bid lower prices) to win the tender. Consequently, the award price is lower in a scenario with less competitiveness or, in other words, public procurement agencies save money.

Figure 5 shows the APE boxplot grouped by CPV. Box diagrams are a standard method to graphically represent numerical data through their quartiles. The outliers of the dataset have not been represented because they are values very far out, which would make it difficult to scale the axes. MAPE (red colour) and MdAPE (green colour) for each CPV group are marked. The great differences of APE,

Table 5: Description of the dataset and the errors between tender price and award price by number of received offers.

Description	Groups by competitiveness			
	No competitiveness	Low Received offers (2–4)	Medium Received offers (5–10)	High Received offers >10
Total number of tenders in the dataset	18,790	22,714	11,553	5,271
Total number of tendering organisations	1,956	2,553	2,135	1,053
Total number of winning/award companies	7,550	9,555	5,222	2,402
Mean received offers by tender	1.0	2.80	6.73	20.01
Mean duration of tender's works	401.07 days	396.65 days	370.95 days	277.50 days
Mean tender price	€354,882.49	€388,526.27	€785,455.49	€1,301,031.70
Median tender price	€60,500.00	€75,000.00	€121,000.00	€254,376.00
Mean award price	€341,874.79	€323,611.87	€460,548.68	€836,188.79
Median award price	€58,984.50	€64,833.00	€90,689.00	€174,986.00
Median absolute error (MdAE)	€93.50	€7,661.50	€22,854.00	€76,420.00
Median absolute percentage error (MdAPE)	0.12%	13.39%	29.63%	45.94%
Mean absolute error (MAE)	€13,966.65	€68,244.60	€326,698.33	€464,907.75
Mean absolute percentage error (MAPE)	10.02%	25.65%	54.48%	77.98%

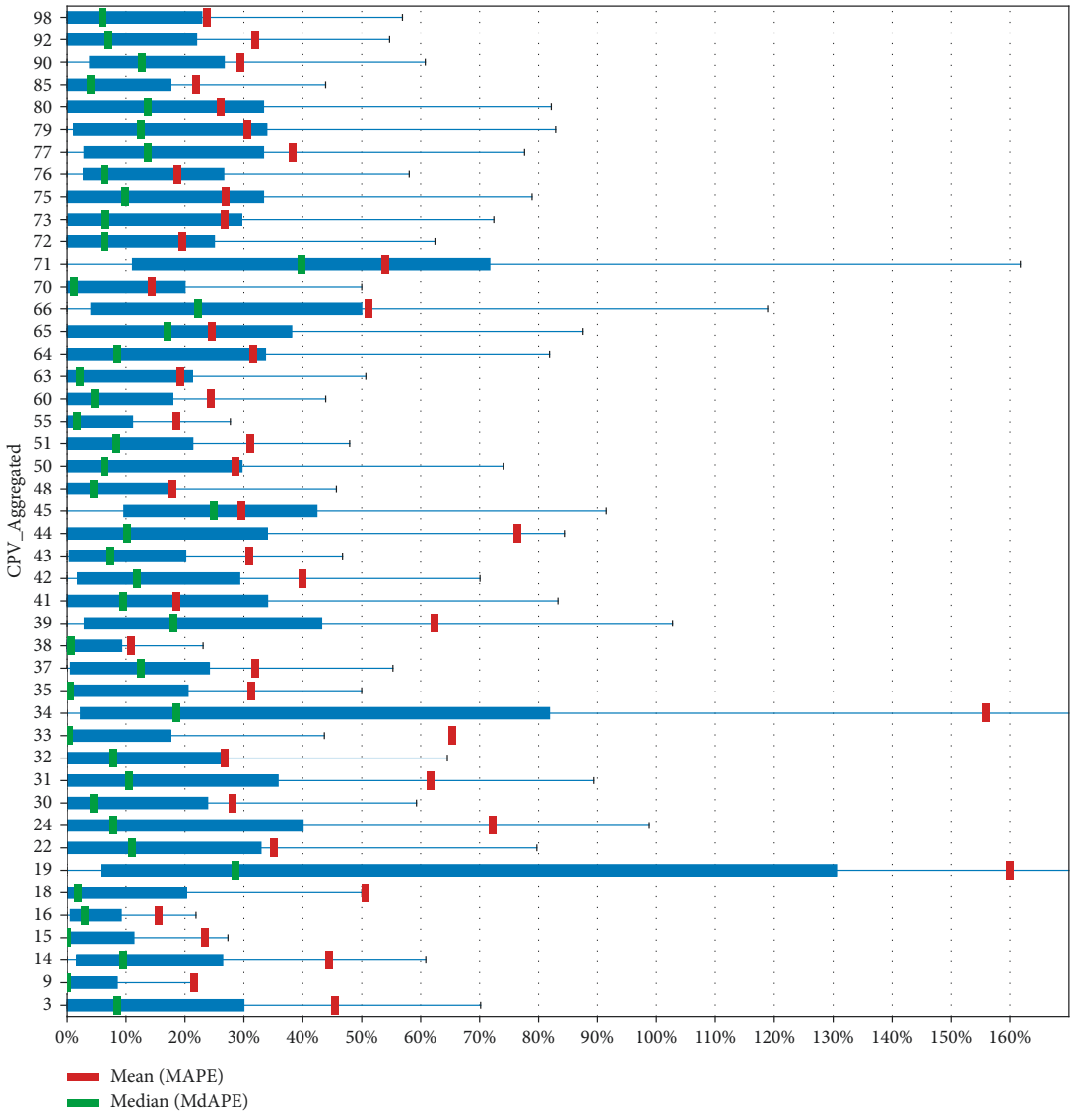


Figure 5: Boxplot of absolute percentage error (APE) between award price and tender price grouped by CPV.

MAPE, and MdAPE according to the CPV can be clearly seen. In general, MdAPE is between 20% and 40% and MAPE is higher than 40%. The total value of MAPE and MdAPE (without dividing by CPV) has already been calculated, as shown in Table 4.

In conclusion, in view of the graphical and quantitative results, it can be affirmed that tender price is a bad estimator of award price. Perhaps it is not excessively bad in median (11.84%) but it is so in mean (39.79%). This is certainly due to the high dispersion between both prices (as seen in Figure 4(a)). This is the reason to create an award price estimator in the following section.

4. Award Price Estimator

A good award price estimator would be very useful and valuable for companies and public procurement agencies. It would be a key tool in their project management decision making because it reduces the economic risks. Due to the complexity involved, machine learning techniques have been chosen to create the estimator, in particular, random forest. In Section 4.1, random forest for regression is presented, from the theoretical framework to its application to the Spanish tenders' dataset. In Section 4.2, the empirical results and analysis are presented, for example, the error metrics of the award price estimator created. In Section 4.3 a similar analysis is presented using a dataset from other countries, creating a new award price estimator.

4.1. Random Forest for Regression. Random forests (RF), introduced by Breiman [24] in 2001, is an ensemble learning method for regression or classification that operates by constructing a multitude of decision trees at training time and outputting the class which is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is a popular learning algorithm that offers excellent performance [25, 26], no overfitting [27], and a versatility of applicability to large-scale problems and in handling different types of data [25, 28]. It provides its own internal generalisation error estimate, called out-of-bag (OOB) error.

Simplified algorithm of RF for regression [29]:

(1) For b 1 to B:

  (a) Draw a bootstrap sample $\mathbb{Z}^*$ of size N from the training data.
  (b) Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{\min}$ is reached.
    (i) Select m variables at random from the p variables.
    (ii) Pick the best variable/split-point among the m.
    (iii) Split the node into two daughter nodes.

(2) Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x, $\widehat{f}_{rf}^{B}(x)$ $(1/B)\sum_{b\ 1}^{B} T_b(x)$.

At each split in each tree, the improvement in the split criterion is the measure of the importance attributed to the splitting variable and is accumulated over all the trees in the forest separately for each variable. It is called variable importance [24].

There are other implementations of RF algorithms, such as Boruta [30], regularised random forest (RRF) [31], conditional forest [32], quantile regression forest (QRF) [33], or extremely randomised regression trees (extraTrees) [34]. The last one was tested with this dataset, but it has a worse accuracy than random forest, so finally it was discarded. The reason is because the function to measure the quality of a split is the Gini index, which is worse than MAE (mean absolute error) or MSE (mean squared error). A comparison between the use of MAE and MSE is shown in Figure 6 for 30 to 1000 trees generated in RF. MAE used as the quality function has clearly better values for the error metrics (especially MAPE and NRMSE) than the MSE quality function for this dataset. Therefore, the function selected is MAE.

The random forest method has been used for multiple and different real-world applications [25], such as the estimation of traffic car issues [35–37], wind speed prediction [38], classification of protein sequences [39], discrimination between seismic events and nuclear explosions [40], pedestrian detection [41], aggregated recommender systems [42], bed occupancy predictor in hospitals [43], classification of phishing e-mail [44], network intrusion detection [45], and employee turnover prediction [46].

Figure 7 shows different ratios between the training and testing subsets (train : test in percentage): 65 : 35, 70 : 30, 75 : 25, 80 : 20, 85 : 15, and 90 : 10. The most important errors for this study, MdAPE and MAPE, are constantly in the order of 9% and 30%, respectively. OBB and NMRSE do not change significantly. Hence, the train : test ratio is not relevant. The typical ratio 80 : 20 will be used in this article.

RandomForestRegressor from *Scikit-learn*, which is a machine learning library for the Python programming language, with 400 trees is the function used in this article. The 14 input variables used in RF are Tender_Price, Date, Duration, Name_Organisation, CPV, CPV_Aggregated, Procedure_code, Type_code, Subtype_code, Urgency_code, Postalzone, Postalzone_CCAA, Postalzone_Province, and Postalzone_Municipality. The variable to perform the regression is Award_Price, and the output generated by RF (prediction) will be called Forecast_Price.

This article does not use the other 3 variables of the tender's resolution (Winner_Province, CIF_Winner, and Received_Offers; Table 3) because they are not variables of the tender's notice. In a real scenario, the award price estimator only can use the variables of the tender's notice. However, if these 3 output variables are used in RF plus 14 input variables, the errors would decrease logically. This is demonstrated as shown in Figure 8: MdAPE is about 5% and MAPE 25%. MdAPE and MAPE are, respectively, 4% and 5% lower than the real scenario with only variables of the tender's notice (see Figure 7). The variable importances (RF

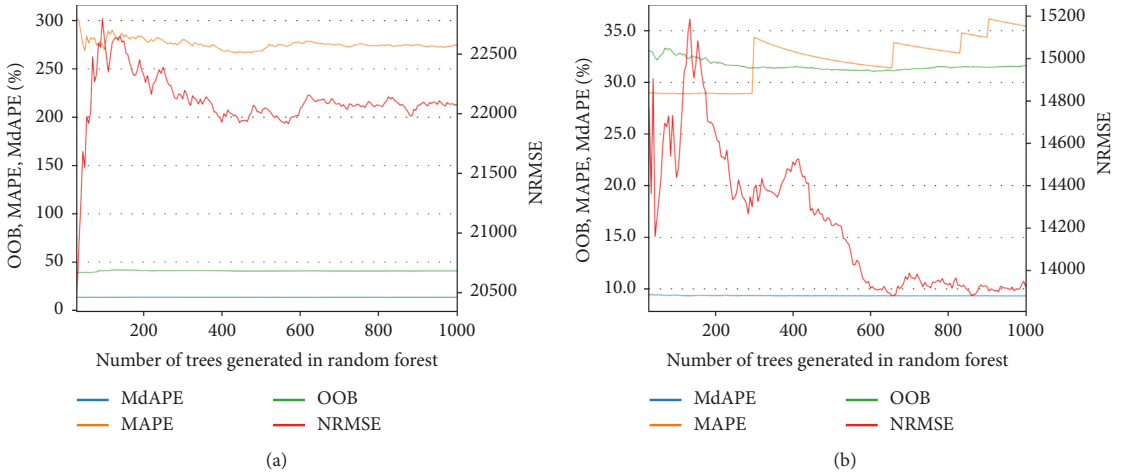


Figure 6: Relationship between trees in random forests (number of estimators) and error metrics (MdAPE, MAPE, OBB, and NRMSE) for two functions to measure the quality of a split. (a) The quality function is mean squared error (MSE). (b) The quality function is mean absolute error (MAE).

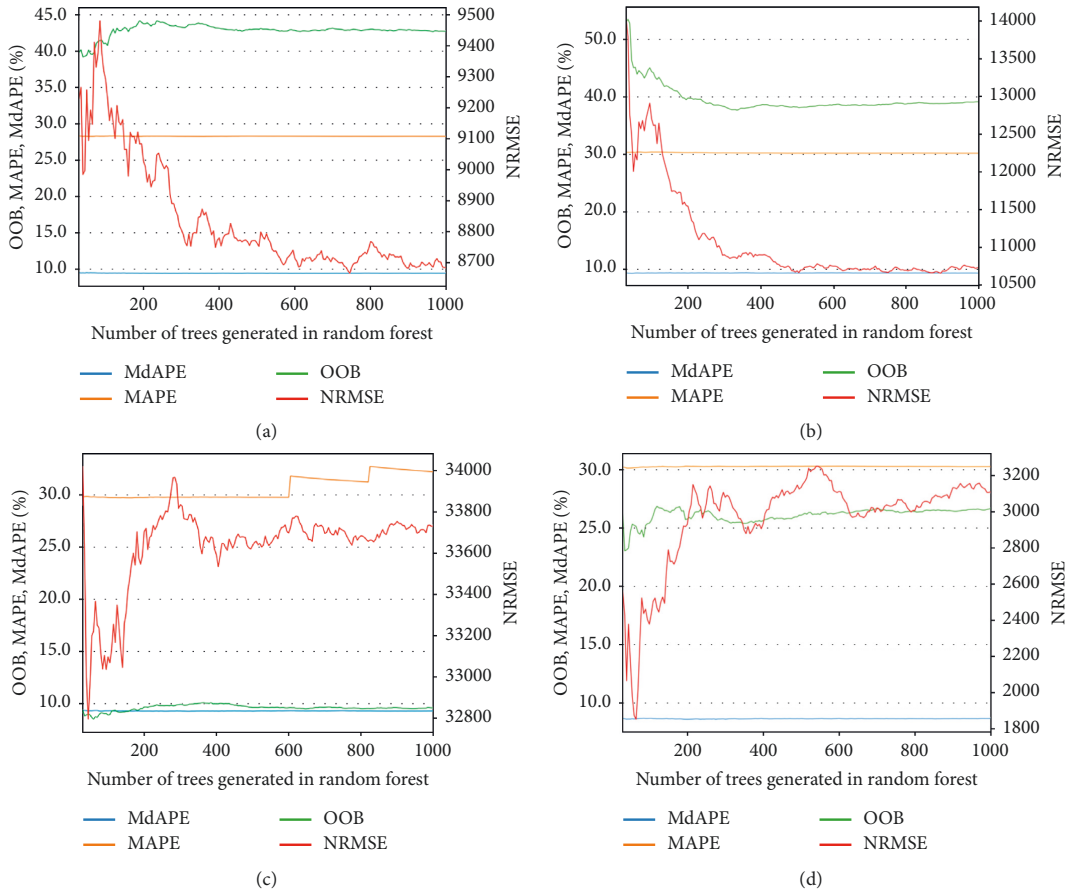


Figure 7: Continued.

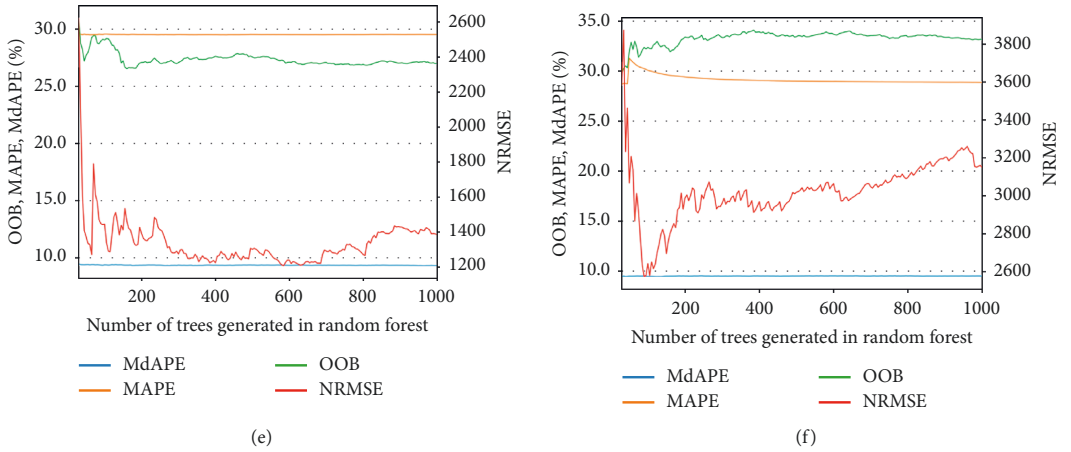


Figure 7: Relationship between trees in random forests and error metrics (MdAPE, MAPE, OBB, and NRMSE) for different ratios of training and testing subsets. (a) 65 : 35. (b) 70 : 30. (c) 75 : 25. (d) 80 : 20. (e) 85 : 15. (f) 90 : 10.

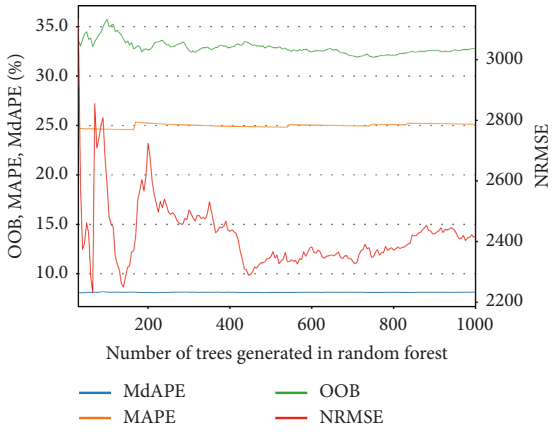


Figure 8: Relationship between trees in random forests and error metrics (MdAPE, MAPE, OBB, and NRMSE) using the 14 input variables plus 3 variables of the tender's resolution.

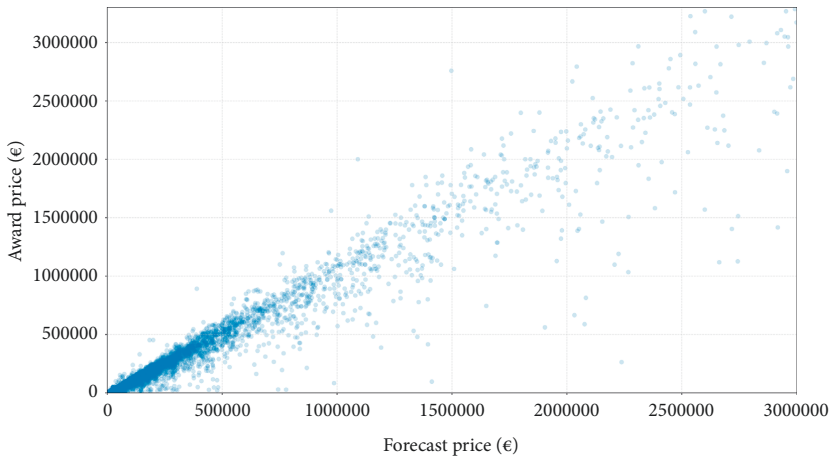


Figure 9: Scatter plot between forecast price and award price.

Table 6: Error metrics between award price and forecast price.

Error	Value	Difference with respect to Tender_Price
Absolute percentage error (APE)	See Figure 8	See Figure 8
Median absolute error (MdAE)	€7,575.45	+€620.45
Median absolute percentage error (MdAPE)	9.26%	−2.58%
Mean absolute error (MAE)	€67,241.34	+€70,537.3
Mean absolute percentage error (MAPE)	28.60%	−11.19%
Root mean square error (RMSE)	364,901,707,583	−101,086,707,913,131
Coefficient of determination (R^2)	0.92	—

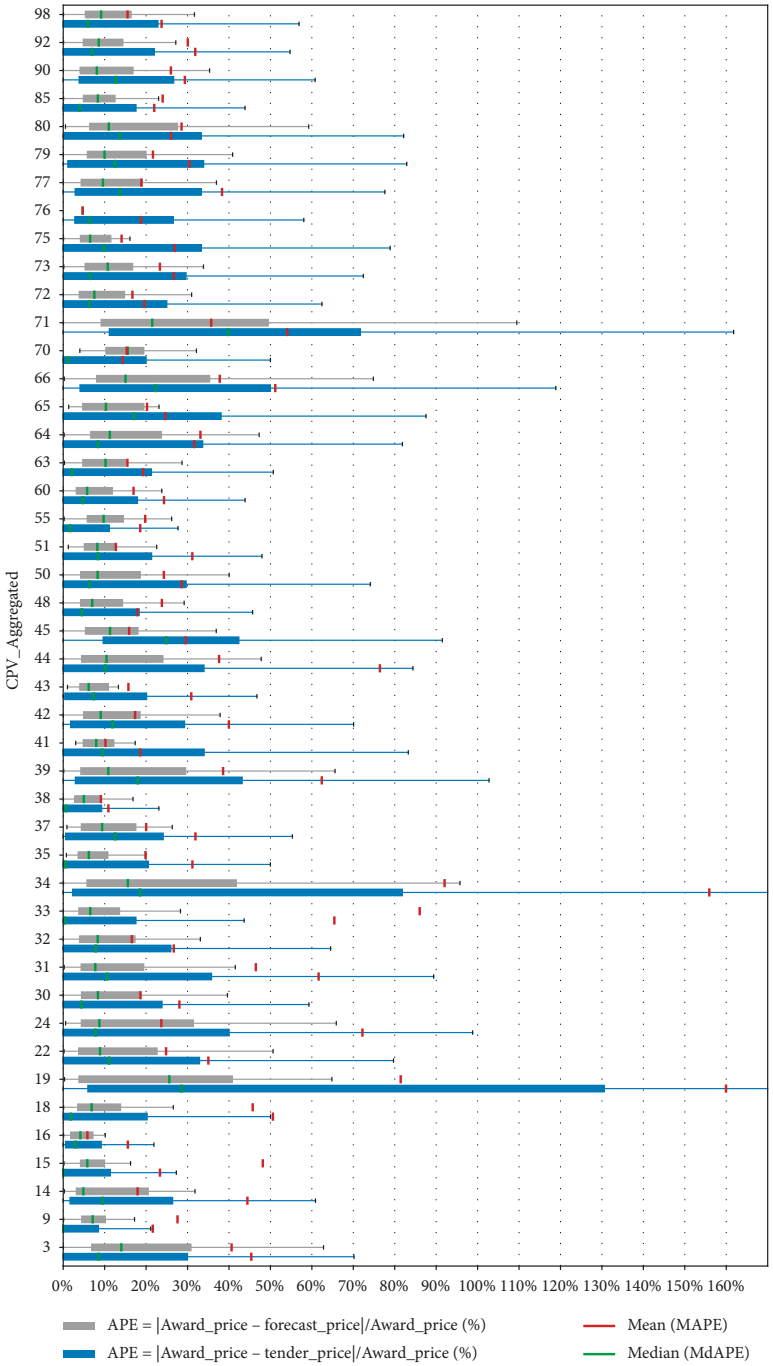


Figure 10: Boxplot of absolute percentage error (APE, grey colour) between award price and forecast price, grouped by CPV. The APE reference (blue colour) is the award price and tender price shown in Figure 5.

Table 7: European countries' dataset: quantitative description.

Topic	Description	Value
General values	Total number of tenders in the dataset	41,556
	Number of tenders by country: France (FR), Croatia (HR), Slovenia (SI), Bulgaria (BG), Germany (DE), Italy (IT), Hungary (HU), and Latvia (LV)	12,449 (FR); 7,910 (HR); 6,473 (SI); 6,096 (BG); 3,918 (DE); 3,782 (IT); 3,724 (HU); 1,736 (LV)
	Temporal range of tenders	2016/12/22–2017/12/29
	Total number of tendering organisations	6,163
	Total number of winning/award companies	19,100
	Mean received offers by tender	5.02
Dataset's variables	Input variables of tender's notice: Date, Name_Organisation, Postalzone, ISO_country_code, Main_activity, Type_code, CPV, CPV_Aggregated, Tender_Price, and Procedure_code	10 input variables
	Output variables of tender's resolution: Award_Price	1 output variable
Prices (without taxes)	Median tender price	€425,000.00
	Median award price	€394,951.26
Number of tenders by CPV	Tenders with CPV 33: medical equipments, pharmaceuticals, and personal care products	10,927 (26.29%)
	Tenders with CPV 15: food, beverages, tobacco, and related products	4,363 (10.50%)
	Tenders with CPV 45: construction work	4,053 (9.75%)
	Tenders with CPV 71: architectural, construction, engineering, and inspection services	1,973 (4.75%)
	Tenders with CPV 34: transport equipment and auxiliary products to transportation	1,893 (4.56%)
Number of tenders by type code	Tenders with Type_code 1: goods/supplies	24,593 (59.18%)
	Tenders with Type_code 2: services	12,849 (30.92%)
	Tenders with Type_code 3: works	4,114 (9.90%)

Table 8: European countries' dataset: errors between award price vs. tender price and award price vs forecast price and their differences.

Error	Award price vs. tender price	Award price vs. forecast price	Difference
Median absolute error (MdAE)	€4,514.50	€20,982.94	+€16,468.44
Median absolute percentage error (MdAPE)	4.17%	6.48%	+2.31%
Mean absolute percentage error (MAPE)	27.49%	23.57%	−3.92%
Normalised root mean square error (NRMSE)	99,018.04	2,816,245.06	+2,717,227.02
Coefficient of determination (R^2)	0.9680	0.7303	−0.2377

output parameter) ordered from highest to lowest are Tender_Price (0.870%), Received_Offers (0.035%), Duration (0.017%), Date (0.013%), Name_Organisation (0.012%), CIF_Winner (0.010%), CPV (0.009%), Postalzone (0.007%), Subtype_code (0.006%), CPV_Aggregated (0.005%), Winner_Province (0.004%), Type_code (0.004%), Procedure_code (0.003%), Postalzone_Municipality (0.002%), Postalzone_Province (0.001%), Postalzone_CCAA (0.001%), and Urgency_code (0.0001%). It is clear that the 3 output variables are important in the previous ranking.

4.2. Empirical Results and Analysis. RF has been trained with 80% of tenders (46,670). The remaining 20% (11,667) have been used as the test group. Figure 9 shows the scatter plot between forecast price and award price for the test group. As has already been mentioned, if the estimator were perfect, all points would have to be on the line at 45 degrees.

The prediction's errors are presented in Table 6. Furthermore, in the third column, it is compared with the error made by Tender_Price (see Table 4) to check if the proposed estimator is better or worse. It makes no sense to compare the absolute errors because the sizes of the datasets are different. It is best to compare the percentage errors, such as MdAPE and MAPE; they are significantly lower, MdAPE—2.58% and MAPE—11.19%.

Figure 10 shows the boxplot of APE (grey colour) between award price and forecast price grouped by CPV. It is also plotted the APE reference (blue colour) which has been presented previously in Figure 5. It is clearly visible how the APE of the estimator has boxplots with a smaller interquartile range (IQR). In general, MdAPE and MAPE are lower than the APE reference. In conclusion, the proposed estimator reduces significantly the error with respect to tender price (analysed in Section 3.3).

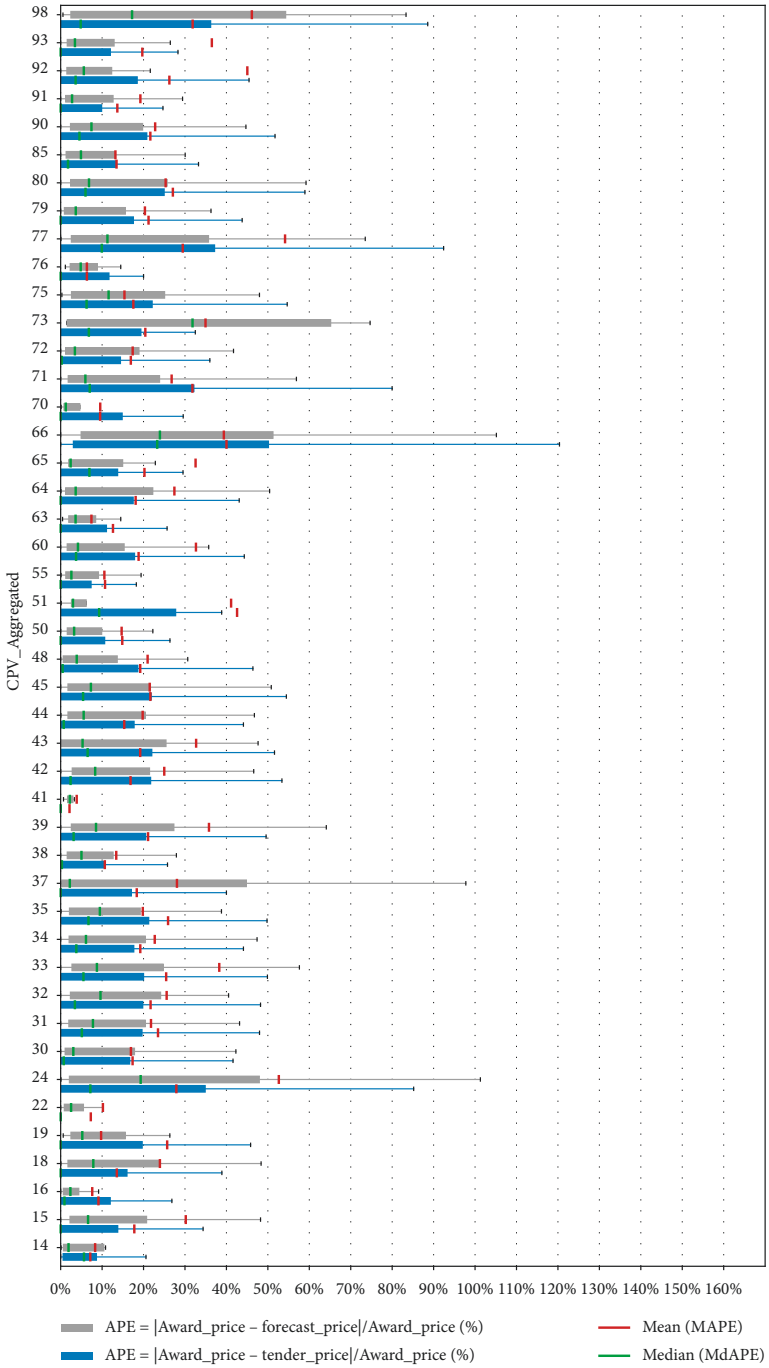


Figure 11: European countries' boxplot: absolute percentage error (APE) between award price and tender price (blue colour) and award price and forecast price (grey colour), grouped by CPV.

The variable importances (RF output parameter) ordered from highest to lowest are Tender_Price (88.34%), Date (1.84%), Duration (1.83%), Name_Organisation (1.56%), Subtype_code (1.52%), CPV (1.10%), Postalzone (1.09%), Type_code (0.97%), Procedure_code (0.66%), CPV_Aggregated (0.49%), Postalzone_Municipality (0.24%), Postalzone_Province (0.18%), Postalzone_CCAA (0.17%), and Urgency_code (0.03%).

4.3. Empirical Results and Analysis for Other Countries. In this section, a study is made with tenders from other countries, similar to the previous one for Spanish tenders. The

purpose is to evaluate the award price estimator with a different dataset using the same machine learning technique (random forest). The countries selected are from the European Union because they have almost the same characteristics associated with public procurement announcements: legislation, tender's regulation, public administrations, purchase procedures, etc. The raw data have been downloaded from the European Open Data Portal [21], in particular the tenders' database of the year 2017 (link in the Data Availability section). However, the quality of the data is not good: fields without data, errors in tender and award prices, the winning company does not have its tax identification number, tender and award prices have the same value, etc. It is an official dataset provided by the European Union, but it does not have as good a quality as the Spanish dataset. In the beginning, there were 706,104 tenders. After data preprocessing, there were only 41,556 tenders.

Table 7 shows the quantitative description of the dataset for the following 8 European countries: France, Germany, Italy, Croatia, Slovenia, Bulgaria, Hungary, and Latvia. They have been selected because they have the highest number of tenders after data preprocessing.

This dataset has been trained with 80% of the tenders (33,244). The remaining 20% (8,312) have been used as the test group. The random forest process is analogous to the Spanish one. The 10 input variables used in RF are Date, Name_Organisation, Postalzone, ISO_country_code, Main_activity, Type_code, CPV, CPV_Aggregated, Tender_Price, and Procedure_code. The variable to perform the regression is Award_Price, and the output generated by RF (prediction) will be called Forecast_Price.

The errors MdAE, MdAPE, MAPE, and NRMSE and R^2 are shown in Table 8. The second column shows award price vs. tender price (the reference), and the third column shows award price vs. forecast price (the estimator created with RF). MdAPE between award price and tender price is very low (4.17%) if it is compared to the Spanish MdAPE (11.84%, see Table 4). This means that award price is very close to tender price or, in other words, a lot of tenders have the same price for both and, consequently, without error. MAPE is also lower (27.49%) than the Spanish MAPE (39.79%). The estimator is better in MAPE (−3.92%) but it is worse in MdAPE (+2.31%) (see fourth column in Table 8).

Figure 11 shows the boxplot of APE (grey colour) between award price and forecast price grouped by CPV. The APE reference (blue colour) between award price and tender price is also plotted. It is not clearly visible how the APE of the estimator has boxplots with a smaller interquartile range (IQR). In general, MdAPE and MAPE are similar to the APE reference.

In conclusion, the estimator created for this dataset has similar error metrics with respect to tender price. Why do a lot of tender notices in the European countries have the same value of tender price and award price? Why not in the Spanish case? This could be due to the bad quality of the European dataset (tender's notices with mistakes) or, a less likely hypothesis, the fact that the Spanish public procurement agencies fail to estimate the tender price and the European agencies never fail in anything. The proposed method can be useful and generalisable to other countries with a large dataset without mistakes.

5. Conclusions and Future Research

The European and Spanish public procurement legislation has been presented. A dataset of 58,337 Spanish public tenders from 2012 to 2018 has been analysed. The relations between the main fields of the public procurement notices have been studied mathematically. Error metrics between the tender price and the award price have been calculated (MdAPE 11.84% and MAPE 39.79%). An award price estimator, which reduces the previous errors (MdAPE 9.26% and MAPE 28.60%), has been proposed by using a machine learning algorithm (random forest). The estimator has 14 fields as input variables, of which the most important are the tender price, date, duration, public procurement agency name, subtype code, CPV classification, and postal zone.

A good award price estimator would be useful for companies and public procurement agencies. It would be useful for companies because it can be a key tool in their project management decision making: it would reduce the economic risks, thus winning tenders more easily. For public procurement agencies, it would be useful because, for example, in the Spanish dataset, the tender price could have been reduced by 2.24% (MdAPE reduction), equivalent to approximately 811 million euros. This is a significant error reduction that, consequently, would improve the accuracy of the budget for public procurement.

An analogous analysis has been made with 8 European countries (France, Germany, Italy, Croatia, Slovenia, Bulgaria, Hungary, and Latvia) to generalise the award price estimator to other real situations and check the results. The dataset used has 41,556 tenders, but the quality of the data is worse than the Spanish dataset. The new award price estimator obtains predictions with error metrics that are similar to those between the tender price and award price. The estimator is better in MAPE (−3.92%) but it is worse in MdAPE (+2.31%).

An accurate estimate is impossible to achieve because the market is theoretically open and free and, therefore, unpredictable. Furthermore, the award price is not always the final price paid by the public procurement agency because the contract may be modified during its execution. However, this article illustrates how a machine learning algorithm can be useful. Particularly, random forest predicts the award prices with less uncertainty, adapting to the real market. This market reality is gathered implicitly through the public procurement notices. Therefore, this estimator is interesting for the public procurement agencies and for the companies because their risk is reduced.

Thanks to the open data sources of public procurement, it is possible to avoid depending on government statistics offices such as the Spanish (INE [47]) or the European (Eurostat [23]). Therefore, there is independence, and there are resources to perform low-level analysis or cross data with

other databases or external services to extract more valuable information.

This article opens the doors to future research related to the analysis of massive data on public procurement, in particular:

(i) It achieves a more accurate estimator by incorporating business data of the winning bidder: location, core business, annual turnover, number of employees, financial situation, etc. With the new data, the estimator has more input variables that could be relevant to predicting the award price.

(ii) It compares other machine learning algorithms to estimate award prices, number of received offers, and other interesting fields.

(iii) It performs data business analysis such as companies with a higher success rate in public procurement or the characterisation of the winning bidder: type of company, size, national origin or foreign, etc.

Data Availability

The processed data used to support the findings of this study are available from the corresponding author upon request. The raw data from Spain are available at the Ministry of Finance, Spain. Open data of Spanish tenders are hosted in http://www.hacienda.gob.es/es-ES/GobiernoAbierto/Datos%20Abiertos/Paginas/licitaciones_plataforma_contratacion.aspx. The raw data from other countries are available in the European Union Open Data Portal (Publications Office of the European Union) hosted in https://data.europa.eu/euodp/en/data/dataset/ted-csv.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Plan of Science, Technology and Innovation of the Principality of Asturias (Ref: FC-GRUPIN-IDI/2018/000225).

References

[1] European Commission, "European semester thematic factsheet public procurement," 2017, https://ec.europa.eu/info/sites/info/files/file_import/european-semester_thematic-factsheet_public-procurement_en_0.pdf.

[2] The National Securities Market Commission (CNMV) from Spain, "Radiography of public procurement procedures in Spain," 2019, https://www.cnmc.es/sites/default/files/2314114_5.pdf.

[3] N. Obwegeser and S. D. Müller, "Innovation and public procurement: terminology, concepts, and applications," *Technovation*, vol. 74-75, pp. 1–17, 2018.

[4] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," *Government Information Quarterly*, vol. 32, no. 4, pp. 399–418, 2015.

[5] H. R. Varian, "Big data: new tricks for econometrics," *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3–28, 2014.

[6] S. Mullainathan and J. Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 87–106, 2017.

[7] J. M. Alvarez-Rodríguez, J. E. Labra-Gayo, and P. O. De Pablos, "New trends on e-procurement applying semantic technologies: current status and future challenges," *Computers in Industry*, vol. 65, no. 5, pp. 800–820, 2014.

[8] M. Nečaský, J. Klímek, J. Mynarz, T. Knap, V. Svátek, and J. Stárka, "Linked data support for filing public contracts," *Computers in Industry*, vol. 65, no. 5, pp. 862–877, 2014.

[9] J. D. Twizeyimana and A. Andersson, "The public value of e-government—a literature review," *Government Information Quarterly*, vol. 36, no. 2, pp. 167–178, 2019.

[10] M. A. Bergman and S. Lundberg, "Tender evaluation and supplier selection methods in public procurement," *Journal of Purchasing and Supply Management*, vol. 19, no. 2, pp. 73–83, 2013.

[11] T. D. Fry, R. A. Leitch, P. R. Philipoom, and Y. Tian, "Empirical analysis of cost estimation accuracy in procurement auctions," *International Journal of Business and Management*, vol. 11, no. 3, p. 1, 2016.

[12] H. Jung, G. Kosmopoulou, C. Lamarche, and R. Sicotte, "Strategic bidding and contract renegotiation," *International Economic Review*, vol. 60, no. 2, pp. 801–820, 2019.

[13] K. Bloomfield, T. Williams, C. Bovis, and Y. Merali, "Systemic risk in major public contracts," *International Journal of Forecasting*, vol. 35, no. 2, pp. 667–676, 2019.

[14] P. Ballesteros-Pérez, M. Skitmore, E. Pellicer, and J. H. Gutiérrez-Bahamondes, "Improving the estimation of probability of bidder participation in procurement auctions," *International Journal of Project Management*, vol. 34, no. 2, pp. 158–172, 2016.

[15] T. Hanák and P. Muchová, "Impact of competition on prices in public sector procurement," *Procedia Computer Science*, vol. 64, pp. 729–735, 2015.

[16] V. Titl and B. Geys, "Political donations and the allocation of public procurement contracts," *European Economic Review*, vol. 111, pp. 443–458, 2019.

[17] S. Tadelis, "Public procurement design: lessons from the private sector," *International Journal of Industrial Organization*, vol. 30, no. 3, pp. 297–302, 2012.

[18] T. Hanák and C. Serrat, "Analysis of construction auctions data in Slovak public procurement," *Advances in Civil Engineering*, vol. 2018, Article ID 9036340, 13 pages, 2018.

[19] J.-M. Kim and H. Jung, "Predicting bid prices by using machine learning methods," *Applied Economics*, vol. 51, no. 19, pp. 2011–2018, 2019.

[20] Publications Office of the European Union, *The Official Journal of the European Union*, Publications Office of the European Union, Brussels, Belgium, 2019, https://eur-lex.europa.eu/oj/direct-access.html.

[21] Publications Office of the European Union, *European Union Open Data Portal*, Publications Office of the European Union, Brussels, Belgium, 2019, http://data.europa.eu/euodp.

[22] Tenders Electronic Daily (TED), "Online version of the supplement to the official journal of the EU," 2019, https://ted.europa.eu.

[23] European Commission, Eurostat, https://ec.europa.eu/eurostat.

[24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: a survey and results of new tests," *Pattern Recognition*, vol. 44, no. 2, pp. 330–349, 2011.

[26] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande, "An extensive experimental survey of regression methods," *Neural Networks*, vol. 111, pp. 11–34, 2019.

[27] M. R. Segal, *Machine Learning Benchmarks and Random Forest Regression*, UCSF: Center for Bioinformatics and Molecular Biostatistics, San Francisco, CA, USA, 2004, https://escholarship.org/uc/item/35x3v9t4.

[28] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, Berlin, Germany, 2nd edition, 2008.

[30] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *Journal Of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010.

[31] H. Deng and G. Runger, "Gene selection with guided regularized random forest," *Pattern Recognition*, vol. 46, no. 12, pp. 3483–3489, 2013.

[32] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: a conditional inference framework," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, 2006.

[33] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, pp. 983–999, 2006.

[34] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[35] Y. Cheng, X. Chen, X. Ding, and L. Zeng, "Optimizing location of car-sharing stations based on potential travel demand and present operation characteristics: the case of Chengdu," *Journal of Advanced Transportation*, vol. 2019, Article ID 7546303, 13 pages, 2019.

[36] Q. Shang, D. Tan, S. Gao, and L. Feng, "A hybrid method for traffic incident duration prediction using BOA-optimized random forest combined with neighborhood components analysis," *Journal of Advanced Transportation*, vol. 2019, Article ID 4202735, 11 pages, 2019.

[37] J. Xing and G. Zheng, "Stress field gradient analysis technique using lower-order C^0 elements," *Mathematical Problems in Engineering*, vol. 2015, Article ID 457046, 12 pages, 2015.

[38] Z. Sun, H. Sun, and J. Zhang, "Multistep wind speed and wind power prediction based on a predictive deep belief network and an optimized random forest," *Mathematical Problems in Engineering*, vol. 2018, no. 4, Article ID 6231745, 15 pages, 2018.

[39] Z. Liao, Y. Ju, and Q. Zou, "Prediction of G protein-coupled receptors with SVM-prot features and random forest," *Scientifica*, vol. 2016, Article ID 8309253, 10 pages, 2016.

[40] L. Dong, X. Li, and G. Xie, "Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive bayes classification," *Abstract and Applied Analysis*, vol. 2014, Article ID 459137, 8 pages, 2014.

[41] T. Xiang, T. Li, M. Ye, and Z. Liu, "Random forest with adaptive local template for pedestrian detection," *Mathematical Problems in Engineering*, vol. 2015, Article ID 767423, 11 pages, 2015.

[42] H. R. Zhang, F. Min, and X. He, "Aggregated recommendation through random forests," *The Scientific World Journal*, vol. 2014, Article ID 649596, 11 pages, 2014.

[43] J. Ruyssinck, J. van der Herten, R. Houthooft et al., "Random survival forests for predicting the bed occupancy in the intensive care unit," *Computational and Mathematical Methods in Medicine*, vol. 2016, Article ID 7087053, 7 pages, 2016.

[44] A. A. Akinyelu and A. O. Adewumi, "Classification of phishing email using random forest machine learning technique," *Journal of Applied Mathematics*, vol. 2014, Article ID 425731, 6 pages, 2014.

[45] E. Min, J. Long, Q. Liu, J. Cui, and W. Chen, "Tr-ids: anomaly-based intrusion detection through text-convolutional neural network and random forest," *Security and Communication Networks*, vol. 2018, Article ID 4943509, 9 pages, 2018.

[46] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," *Mathematical Problems in Engineering*, vol. 2019, Article ID 4140707, 12 pages, 2019.

[47] National Statistics Institute (INE), Spain, https://www.ine.es.

Publicación B

Estimador del importe de adjudicación de licitaciones usando ML: caso de estudio con licitaciones de España

Studies in Informatics and Control, 30(4) 67-76, December 2021 ISSN: 1220-1766 eISSN: 1841-429X 67

Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain

Manuel J. GARCIA RODRIGUEZ[1], Vicente RODRIGUEZ MONTEQUIN[1]*, Andoni ARANGUREN UBIERNA[2], Roberto SANTANA HERMIDA[2], Basilio SIERRA ARAUJO[2], Ana ZELAIA JAUREGI[2]

[1] Project Engineering Area, University of Oviedo, Oviedo, 33004, Spain
manueljgarciar@gmail.com, montequi@uniovi.es (**Corresponding author*)

[2] Department of Computer Sciences and Artificial Intelligence, University of the Basque Country, San Sebastián, 20018, Spain
andoni.aranguren@gmail.com, roberto.santana@ehu.eus, b.sierra@ehu.eus, ana.zelaia@ehu.eus

Abstract: The public procurement process plays an important role in the efficient use of public resources. In this context, the evaluation of machine learning techniques that are able to predict the award price is a relevant research topic. In this paper, the suitability of a representative set of machine learning algorithms is evaluated for this problem. The traditional regression methods, such as linear regression and random forest, are compared with the less investigated paradigms, such as isotonic regression and popular artificial neural network models. Extensive experiments are conducted based on the Spanish public procurement announcements (tenders) dataset and employ diverse error metrics and implementations in WEKA and Tensorflow 2.

Keywords: Machine learning, Neural networks, Public procurement, Spanish tender.

1. Introduction

The importance of the public procurement is well known. In terms of projects and cost, the largest adjudicators of a country are the public procurement agencies. For example, the public authorities of the European Union (EU) spent around 14% of their GDP (around €2 trillion) on public procurement (purchase of services, works and supplies) in 2017 (European Commission, 2017). Therefore, improving public procurement can yield enormous savings: even a 1% efficiency gain could save €20 billion per year. It is crucial to analyse the public procurement notice (also called auctions, requests for tender or simply tenders) in order to understand its behaviour in terms of prices. Through the use of new technologies, like machine learning (ML), among others, new tools can be created to improve these public procurement processes.

ML involves computer algorithms that are used for knowledge discovery from large amounts of data. It is considered to be a type of artificial intelligence (AI), and it is regarded as one of the most disruptive innovations and a strong enabler of competitive advantages. While ML has been around for more than 60 years, it has only recently showed significant potential for disrupting economies and societies (Lee & Shin, 2020). Mirroring a trend that has increased pace in the last 5 years in the private sector worldwide, the adoption of AI within public administration processes has the potential to provide enormous benefits. It improves the efficiency and effectiveness of policy making and service delivery to businesses and citizens, ultimately enhancing their level of satisfaction and trust in the quality of public service (Kuziemski & Misuraca, 2020).

The award price estimator is a regression problem. The tender has x known input features (e.g., date, tender price, type of contract, and public procurement agency) and a y unknown output feature (award price). The tender price, which is calculated by the public procurement agency, is the key input parameter to the award price estimator. The tender price is the theoretical price and the estimator adjusts it regarding the real and changing market conditions to predict the award price by the winning bidder.

The aim of this article is to improve the accuracy of the award price estimator studied previously in (García Rodríguez et al., 2019a). That article applied only one algorithm (random forest) to predict the award price, and it was validated over two tender datasets from Spain and Europe. Further, this article increases the prediction accuracy, and it compares four algorithms: linear

https://doi.org/10.24846/v30i4y202106

regression, isotonic regression, random forest and artificial neural network. The last two algorithms are ML methods, particularly supervised learning.

An award price estimator would produce significant benefits. It would be an excellent tool for the cost planning of public tendering agencies by allowing them to have more realistic budgets. Additionally, such a price estimator would provide support to small- and medium-sized enterprises (SMEs) that play a crucial role in most economies. For example, SMEs represent 50% of the GDP in the EU (European Commission, 2020). However, they have difficulty when competing on equal terms with big suppliers in the public procurement space. Other benefits could be the reduction of fraud between bidders, which would improve the transparency of the process and lead to better quantification of the product quality.

The paper begins with reviewing the literature and identifying the research gap to be examined (Section 2). Then, the dataset of public procurement auctions, the ML algorithms being compared (random forest, linear regression, isotonic regression and artificial neural networks (ANNs)) and the error metrics that are used are described (Section 3). Next, the major quantitative results of the experimental analysis are summarized for identifying the best ML algorithm to predict the award price (Section 4). Finally, some concluding remarks, limitations, and avenues for future research are presented (Section 5).

2. Literature Review

While an increasing number of studies in public procurement is being published every year, an overview of the field is missing. In the literature on public procurement, an ambiguous wording is usually used, and a consensus on the terminology and concepts involved has not been reached yet (Obwegeser & Müller, 2018). Technological and organizational challenges faced during public electronic procurement processes are not well understood despite past studies focusing on these topics (Mohungoo, Brown & Kabanda, 2020). The data analysis of public tenders can provide valuable information for different stakeholders: public tendering agencies, public procurement managers, project managers, executives, politicians and, indirectly, citizens. In the particular case of Spain, an initial analysis published in (García Rodríguez et al., 2019b) explains the Spanish public tendering system and the potential applications and benefits of employing massive data processing.

The past decades have seen the rapid development of the computer hardware, communication technologies and computer sciences (artificial intelligence and big data). These new technologies make it possible to implement the informatization of conventional public procurement tendering processes. Public procurement has the typical objectives of the private sector: to acquire the right goods or services from the right supplier, at the right price, at the highest service level, and considering laws and norms requirements. But it also requires strict compliance with the principles of non-discrimination, free competition, and transparency of the awarding procedures (Dotoli, Epicoco & Falagario, 2020).

There is extensive literature about prediction techniques (forecasting) and data analysis in public tendering. There are mainly two approaches: statistical models (e.g., mathematical algorithms) and statistical learning (e.g., ML algorithms). There is not a clear demarcation or boundary between both approaches because research on ML also covers the conception of mathematical algorithms. Thus, statistics and ML are closely related fields in terms of methods, but distinct with regard to their principal goal: statistics draws population inferences from a sample, while ML finds generalizable predictive patterns (Bzdok, Altman & Krzywinski, 2018).

Statistical models are the traditional or conventional approach used to analyse and validate hypotheses. For example, there are models for statistical relationships for tender forecasting in capped tender (Ballesteros-Pérez, González-Cruz & Cañavate-Grimal, 2012), scoring probability graphs (Ballesteros-Pérez, González-Cruz & Cañavate-Grimal, 2013), multicriteria decision making (Dotoli, Epicoco & Falagario, 2020), the probability of bidder participation (Ballesteros-Pérez et al., 2015; Ballesteros-Pérez et al., 2016), and the optimal

bidder participation to achieve the lowest procurement prices (Onur & Tas, 2019). There is also a mathematical model where the bidders are evaluated on the basis of price and quality through a score function (Lorentziadis, 2020), the detection of groups of bidders in collusive auctions (also called not competitive tenders or bid-rigging cartels) (Conley & Decarolis, 2016) or discriminatory competitive procedures in public procurement with unverifiable quality (Albano, Cesi & Iozzi, 2017).

On the other hand, a variety of ML techniques has also been successfully applied to public procurement and created empirical models. For example, among the particular problems addressed by this type of algorithm are those related to the behaviour of bidders: the estimation of the number of bidders in tenders (KNN) (Gorgun, Kutlu & Onur Tas, 2020), the identification of the optimal bidder (fuzzy logic) (Wang et al., 2014), creating a search engine of suppliers to recommend potential bidders for a characterized tender (random forest) (García Rodríguez et al., 2020) the detection of collusive auctions (ensemble method) (Huber & Imhof, 2019), or the proposal of an objective system (key performance indicators) for supporting the estimators (benchmarking) during the tender evaluation process (ANNs) (Bilal & Oyedele, 2020).

However, there are almost no studies about award price forecasting, so there is a research gap. The first holistic approach that considers all kinds of tenders (multi-sectorial) and a large volume of tenders is (García Rodríguez et al., 2019a) whose dataset is used in this article. Previously, two articles created award price estimators with ML algorithms, but they were applied only to construction auctions: bridge projects (Chou et al., 2015) and highway procurement (Kim & Jung, 2019). It is typical to find literature focused only on public procurement for construction or civil engineering projects; this is mainly because they are the biggest and most important projects in public procurement (García Rodríguez et al., 2019a). This paper is the first attempt to compare different algorithms in order to improve the accuracy of award price forecasting in multi-sectorial tenders.

In conclusion, this article is a true reflection of the applicability of ML in public procurement. The fundamental insight behind this breakthrough is as much statistical as computational. Artificial intelligence became possible once researchers stopped approaching intelligence tasks procedurally and began tackling them empirically (Mullainathan & Spiess, 2017). ML algorithms produce a powerful, flexible way of making quality predictions, but they have a weakness: they do not contain strong assumptions and instead contain mostly unverifiable assumptions due to the fact that ML approaches do not generally produce stable estimates of the underlying parameters (Mullainathan & Spiess, 2017).

3. Experimental Procedures

The main objective of this work is to analyze different ML paradigms for predicting the award (winning) price of Spanish public tenders. In this section, the dataset and the learning models are presented, and details about the error metrics and validation method are given.

3.1 Dataset

The original data were extracted from the information files published by the Spanish Ministry of Finance (see Data Availability). It contained information about tenders published between 2012 and 2018. The data were preprocessed for a preliminary study published in (García Rodríguez et al., 2019a) and a dataset of 58,337 Spanish tenders was obtained. To compare the results, the same dataset was used in the experiments presented in this article.

Tenders in the dataset were defined by 14 input variables that provided the following information: the name of the public procurement agency that made the tender, geographical information about the agency (municipality, province, region, and wider region code), the tender price (the amount of budgeted bidding), the duration (days to execute the contract), the type of work according to the common procurement vocabulary (CPV) in two levels of detail, the type of contract defined by legislation (in two levels of detail), the procedure by which the contract was awarded, the urgency level

and the date of agreement in the award of the contract. Note that during preprocessing, all this information was converted to integer values to make it suitable for the learning methods being evaluated. The output variable was the award price, which is the amount offered by the winning bidder of the contract.

3.2 Machine Learning Algorithms

The random forest for regression ML model was selected to create an award price predictor in the preliminary study (García Rodríguez et al., 2019b). The research presented here aimed to investigate a wider range of ML paradigms, compare them and select the most suitable one for the task. Models for regression need to be selected, since the output variable to predict is the award price. Very widely used supervised ML algorithms were considered: random forest, linear regression, isotonic regression and artificial neural networks (ANNs). A brief description of them is presented here.

A random forest algorithm (Breiman, 2001) is a combination of tree predictors, where each tree depends on the values of a random vector independently sampled and with the same distribution for all the trees in the forest. The prediction of the ensemble is computed by averaging the predictions of the individual models. It is a typical example of an ensemble method that reduces the bias of individual models and provides a more flexible predictor that is less prone to overfitting.

While the random forest model is robust, there are situations where simpler algorithms, like linear regression, could produce better results. This explains the convenience of evaluating the performance of linear regression for award price estimation.

Linear regression (equation 1) is a machine learning technique used to model the linear relationship between the input variables x_i and the output variable y:

$$y = \sum_{i=1}^{n} (\beta_i x_i + \varepsilon) \quad (1)$$

where β_i are the parameters that measure the influence of the input variables, and ε is a constant value.

Another technique that is increasingly applied to regression problems is isotonic regression (equation 2). This method tries to find a line as close to the observations as possible:

$$\min_{g} \sum_{i=1}^{m} w_i (g(x_i) - f(x_i))^2 \quad (2)$$

where x_i are the input variables, g is the isotonic estimator, f is a function, w_i are the weights and m is the number of observations. This method produced a series of predictions for the training data that were the closest to the targets in terms of the mean square error (MSE). These predictions were interpolated to predict unseen data. The predictions from the isotonic regression thus formed a function that was piecewise linear (Chakravarti, 1989).

For the three ML algorithms presented in this section, the implementations available in WEKA (Hall et al., 2009; Witten et al., 2011) were used in the experiments. WEKA is a machine learning platform developed by Waikako University, that supports a large number of learning algorithms (Waikako University, 2021).

3.3 ANNs

Recently, ANNs have re-emerged as a powerful tool to deal with a variety of ML problems. In particular, they have been applied to regression problems where the input data can be noisy or not fully observed. An ANN is a computational model inspired by biological neural networks. It consists of a collection of units or nodes (artificial neurons) organized in connected layers. The parameters of the model are the weights and biases associated to the connections. Information is processed from the input layer to the output layer.

The learning process is based on minimizing a cost function (also known as loss function) that evaluates the performance of the network for the given task. Backpropagation is used to learn the weights associated to the connections. One of the ANNs used in this work is a multi-layer perceptron (MLP) (Hastie, Tibshirani & Friedman, 2009) implemented in WEKA (Hall et al., 2009; Witten et al., 2011).

3.4 ANN Optimization (Deep Learning)

In addition to using the MLP implementation in WEKA, a set of ANN architectures that represented a different number of layers (to evaluate the impact of the depth) and a different number of neurons in each layer was selected. The particular choice of the number of neurons is arbitrary and was intended to keep a balance between the goals of increasing the capacity of the model and keeping a manageable complexity.

The selected ANN architectures were the following: two architectures of one hidden layer with 16 nodes and 32 nodes; five architectures of three hidden layers of 16 nodes in each (16-16-16), 32 nodes in each (32-32-32), and a different number of nodes in each (16-8-16, 32-8-32, 32-16-32); and an architecture of five hidden layers with 32-16-8-16-32 nodes (see Figure 1). For each of these ANN designs, different activation functions, loss functions and gradient descent optimization algorithms were evaluated.

The activation function determines the type of non-linear transformation made to the linear combination of the weights and input neurons. In most cases, the rectified linear unit (ReLU) general activation function is used. Recently, the scaled exponential linear unit (SeLU) activation function (Klambauer et al., 2017) has been reported to produce promising results. This is an activation function that induces self-normalizing properties.

Regarding the choice of ReLU and SeLU, preliminary experiments were made with other activation functions including a sigmoid. Due to the page number restrictions and the poor results achieved with these functions, it was decided to include only the results for the ReLU and SeLU. It is emphasized that both functions are theoretically more sound since they address the vanishing and exploding gradient problems experienced by the sigmoid and hyperbolic tangent functions. They can be used for all the main neural network paradigms (i.e., MLPs, CNNs, and RNNs). In particular, SeLU, one of the newest activation functions proposed in the literature, was introduced with an eye on standard feed-forward neural networks and not envisioning CNNs.

Regarding the selection of the regression loss-functions, the common ones were used: MSE or the sum of squared distances, mean absolute error (MAE) or the sum of absolute differences (see subsection 3.5). Among the available gradient descent optimization algorithms commonly used, Adagrad (Duchi, Bartlett & Wainwright, 2012) and Adam and Adamax (Kingma & Ba, 2014) were selected for the experiments (Keras, 2021). For the optimization process, the maximum number of epochs (times the learning algorithm iterated through the training dataset) was set to 50,000.

The eight ANN structures combined with two activation functions, two loss functions and three optimizers provided 96 different ANN designs. Only the training dataset (46,670 tenders) was used for the optimization process. Two different validation frameworks were evaluated: a train/test

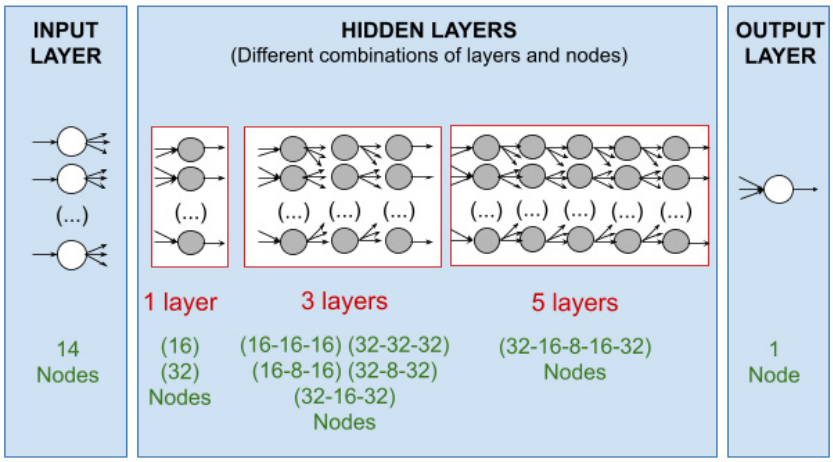


Figure 1. ANN structures

division (Hold-out 80/20) and a K-fold cross-validation with K=10. Figure 2 and Figure 3 show the results obtained for the four error metrics (MAE, root mean square error (RMSE), relative absolute error (RAE) and root relative square error (RRSE)). The rows of the tables correspond to the eight different ANN architectures, the columns correspond to the two activation functions (RELU, SELU), two loss functions (MAE, MSE) and three optimizers (Adam, Adamax, Adagrad). The best results (minimum error) are coloured in green, the intermediate ones are in orange and the worst are in red.

A set of ANN configurations that performed well during the optimization phase was selected for the final test:

- ANN1: Three hidden layers with 16-8-16 nodes, SeLU activation function, MAE loss function and Adam optimizer (see Figure 2);
- ANN2: Three hidden layers with 32-8-32 nodes, SeLU activation function, MAE loss function and Adagrad optimizer (see Figure 3);
- ANN3: Three hidden layers with 16-8-16 nodes, ReLU activation function, MAE loss function and Adagrad optimizer (see Figure 2);
- ANN4: Three hidden layers with 16 nodes in each, ReLU activation function, MSE loss function and Adam optimizer (see Figure 3).

The optimization for the ANN architectures was performed using Tensorflow (Abadi et al., 2016; Tensorflow, 2021), which is an open source software library for ML. It is a very appropriate platform to evaluate with different ANN architectures. Keras is an API designed to simplify the use of Tensorflow.

3.5 Error Metrics

In this subsection, the error metrics used to measure the deviation of the predicted values compared to the real ones are presented.

The MAE (equation 3) and RMSE (equation 4) are two of the most common metrics used to measure accuracy in absolute terms for continuous variables. The MAE measures the average magnitude of the errors in a set of predictions without considering their direction. It was also used as the loss function for the ANNs in the present experiments. The RMSE is a quadratic scoring rule that also measures the average magnitude of the error:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|r_i - p_i| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(r_i - p_i)^2} \tag{4}$$

where r_i are the actual observations (the true values), P_i are the predicted values and m is the number of observations.

In relative terms, the RAE (equation 5) and RRSE (equation 6) calculate the error values as a ratio (percentage):

$$RAE = \frac{\sum_{i=1}^{m}|r_i - p_i|}{\sum_{i=1}^{m}|r_i - \overline{r}|} \tag{5}$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^{m}(r_i - p_i)^2}{\sum_{i=1}^{m}(r_i - \overline{r})^2}} \tag{6}$$

where $\overline{r}$ is the mean of the actual observations. Values over 100% appear when the absolute or quadratic difference between the predicted values and the actual observations are bigger than the differences between the actual observations and their means.

Finally, the MSE (equation 7) is used as a loss function in the present experiments with the ANNs:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(r_i - p_i)^2 \tag{7}$$

3.6 Validation

To evaluate the performance of the different ML paradigms, models were trained with 80% of the tenders (46,670). The remaining 20% (11,667) were used as a test. The same validation framework and train/test division of the dataset were used in the preliminary study.

4. Experimental Results

The validation was performed for the 11,667 tenders in the test dataset. The random forest model was considered to be the baseline because it was used in the preliminary study and therefore enabled the comparison of the behaviour of different ML paradigms.

Error metrics	ANN configuration	Experimental results													
	Activation function	RELU						SELU							
	Regression loss function	Mean Absolute Error			Mean Squared Error			Mean Absolute Error			Mean Squared Error			Colour legend	
	Optimizer / Layer structure	Adam	Adamax	Adagrad	Adam	Adamax	Adagrad	Adam	Adamax	Adagrad	Adam	Adamax	Adagrad	Percentile value	Percentile
MAE (M€)	16x1	0.94M€	0.77M€	0.88M€	1.04M€	0.89M€	0.73M€	0.80M€	0.75M€	0.86M€	0.89M€	0.83M€	0.75M€	1.27M€	100
	32x1	0.91M€	0.76M€	0.94M€	0.92M€	0.86M€	0.75M€	0.82M€	0.75M€	1.03M€	0.94M€	0.88M€	0.74M€	0.87M€	75
	16x3	0.78M€	0.75M€	0.76M€	0.73M€	0.80M€	0.80M€	0.77M€	0.75M€	0.79M€	0.80M€	0.86M€	0.80M€	0.82M€	59
	32x3	0.78M€	0.74M€	0.76M€	0.94M€	0.77M€	0.74M€	0.78M€	0.74M€	0.79M€	0.80M€	0.82M€	0.73M€	0.79M€	41
	16-8-16	0.78M€	0.76M€	**0.67M€**	0.73M€	0.79M€	0.75M€	**0.55M€**	0.76M€	0.80M€	0.81M€	0.77M€	0.73M€	0.76M€	25
	32-8-32	1.11M€	1.03M€	0.86M€	0.80M€	0.79M€	0.74M€	0.95M€	1.25M€	0.86M€	0.77M€	0.77M€	0.85M€	0.75M€	16
	32-16-32	0.90M€	1.01M€	0.82M€	0.79M€	0.85M€	0.73M€	1.06M€	1.27M€	0.80M€	0.83M€	0.83M€	0.82M€	0.73M€	8
	32-16-8-16-32	1.00M€	1.23M€	0.84M€	0.83M€	0.83M€	0.78M€	0.98M€	1.22M€	0.81M€	0.76M€	0.78M€	0.76M€	0.55M€	0
RMSE (M€)	16x1	10.80M€	10.86M€	10.78M€	10.62M€	10.62M€	10.74M€	10.84M€	10.68M€	10.45M€	10.59M€	10.63M€	10.64M€	10.99M€	100
	32x1	10.68M€	10.76M€	10.63M€	10.69M€	10.61M€	10.65M€	10.71M€	10.71M€	10.68M€	10.70M€	10.62M€	10.56M€	10.78M€	75
	16x3	10.72M€	10.73M€	10.61M€	10.68M€	10.66M€	10.57M€	10.81M€	10.47M€	10.54M€	10.66M€	10.71M€	10.56M€	10.71M€	59
	32x3	10.95M€	10.53M€	10.60M€	10.60M€	10.51M€	10.66M€	10.82M€	10.50M€	10.63M€	10.62M€	10.62M€	10.62M€	10.66M€	41
	16-8-16	10.77M€	10.59M€	**10.36M€**	10.64M€	10.57M€	10.58M€	**10.11M€**	10.68M€	10.60M€	10.68M€	10.56M€	10.58M€	10.62M€	25
	32-8-32	10.91M€	10.79M€	10.86M€	10.70M€	10.71M€	10.67M€	10.86M€	10.86M€	10.93M€	10.73M€	10.64M€	10.64M€	10.58M€	16
	32-16-32	10.91M€	10.99M€	10.92M€	10.70M€	10.65M€	10.58M€	10.89M€	10.92M€	10.91M€	10.71M€	10.71M€	10.71M€	10.55M€	8
	32-16-8-16-32	10.95M€	10.84M€	10.86M€	10.70M€	10.73M€	10.68M€	10.92M€	10.84M€	10.83M€	10.71M€	10.72M€	10.71M€	10.11M€	0
RAE (%)	16x1	145%	119%	137%	160%	137%	112%	123%	116%	133%	137%	127%	115%	195%	100
	32x1	140%	118%	145%	142%	133%	115%	127%	116%	159%	145%	136%	114%	134%	75
	16x3	120%	116%	118%	113%	123%	123%	119%	115%	122%	123%	132%	124%	127%	59
	32x3	121%	115%	118%	145%	118%	114%	120%	114%	122%	124%	127%	112%	121%	41
	16-8-16	120%	117%	**104%**	113%	121%	115%	**85%**	117%	123%	125%	119%	113%	118%	25
	32-8-32	171%	159%	133%	124%	122%	115%	147%	192%	133%	119%	118%	131%	115%	16
	32-16-32	139%	156%	126%	122%	131%	112%	164%	195%	124%	128%	128%	127%	113%	8
	32-16-8-16-32	155%	189%	129%	128%	129%	120%	151%	188%	125%	117%	120%	117%	85%	0
RRSE (%)	16x1	109.4%	110.0%	109.2%	107.5%	107.6%	108.8%	109.8%	108.2%	105.8%	107.2%	107.6%	107.8%	111.3%	100
	32x1	108.2%	109.0%	107.7%	108.3%	107.5%	107.8%	108.5%	108.5%	108.1%	108.4%	107.5%	107.0%	109.2%	75
	16x3	**108.5%**	108.6%	107.5%	108.2%	108.0%	107.1%	109.5%	106.1%	106.7%	108.0%	108.5%	107.0%	108.5%	59
	32x3	110.9%	106.7%	107.4%	107.4%	106.5%	107.9%	109.6%	106.3%	107.6%	107.5%	107.5%	107.6%	108.0%	41
	16-8-16	109.1%	107.3%	**104.9%**	107.8%	107.0%	107.1%	**102.4%**	108.2%	107.4%	108.2%	106.9%	107.2%	107.5%	25
	32-8-32	110.5%	109.2%	110.0%	108.4%	108.5%	108.1%	110.0%	110.0%	110.7%	108.6%	107.7%	107.8%	107.2%	16
	32-16-32	110.5%	111.3%	110.6%	108.4%	107.9%	107.2%	110.3%	110.6%	110.5%	108.5%	108.5%	108.4%	106.8%	8
	32-16-8-16-32	110.9%	109.8%	110.0%	108.3%	108.6%	108.2%	110.6%	109.8%	109.7%	108.4%	108.5%	108.5%	102.4%	0

Figure 2. Error metrics (MAE, RMSE, RAE and RRSE) for different ANN configurations with validation framework train/test division (hold-out 80/20)

Error metrics	ANN configuration	Experimental results													
	Activation function	RELU						SELU							
	Regression loss function	Mean Absolute Error			Mean Squared Error			Mean Absolute Error			Mean Squared Error			Colour legend	
	Optimizer / Layer structure	Adam	Adamax	Adagrad	Adam	Adamax	Adagrad	Adam	Adamax	Adagrad	Adam	Adamax	Adagrad	Percentile value	Percentile
MAE (M€)	16x1	0.99M€	0.90M€	0.96M€	0.97M€	1.01M€	0.89M€	0.97M€	0.89M€	0.94M€	1.08M€	1.02M€	0.88M€	1.25M€	100
	32x1	1.15M€	0.91M€	0.95M€	1.01M€	1.14M€	0.92M€	1.01M€	0.90M€	1.02M€	0.97M€	1.01M€	0.91M€	0.97M€	75
	16x3	0.89M€	0.89M€	0.88M€	**0.81M€**	0.88M€	0.89M€	0.90M€	0.87M€	0.91M€	0.91M€	0.99M€	0.89M€	0.91M€	59
	32x3	0.89M€	0.88M€	0.86M€	0.91M€	0.83M€	0.98M€	0.90M€	0.90M€	0.86M€	0.82M€	0.83M€	0.98M€	0.89M€	41
	16-8-16	0.84M€	0.90M€	0.91M€	0.95M€	1.04M€	0.89M€	**0.81M€**	0.90M€	0.91M€	0.87M€	1.02M€	0.91M€	0.88M€	25
	32-8-32	0.89M€	0.91M€	0.85M€	0.86M€	0.90M€	0.84M€	0.82M€	0.83M€	**0.77M€**	0.92M€	0.95M€	0.87M€	0.86M€	16
	32-16-32	0.89M€	0.90M€	0.89M€	0.88M€	0.99M€	1.09M€	0.88M€	0.89M€	0.89M€	0.96M€	1.15M€	1.01M€	0.83M€	8
	32-16-8-16-32	0.83M€	0.87M€	0.87M€	0.98M€	0.91M€	0.88M€	0.85M€	0.89M€	**0.80M€**	0.86M€	1.25M€	0.87M€	0.77M€	0
RMSE (M€)	16x1	14.30M€	14.29M€	14.25M€	14.26M€	14.29M€	14.24M€	14.26M€	14.26M€	14.26M€	14.30M€	14.27M€	14.25M€	14.32M€	100
	32x1	14.30M€	14.31M€	14.26M€	14.25M€	14.29M€	14.23M€	14.31M€	14.28M€	14.24M€	14.32M€	14.24M€	14.27M€	14.27M€	75
	16x3	14.26M€	14.29M€	14.27M€	**14.18M€**	14.22M€	14.23M€	14.32M€	14.24M€	14.27M€	14.19M€	14.21M€	14.23M€	14.26M€	59
	32x3	14.30M€	14.28M€	14.26M€	14.25M€	14.21M€	14.28M€	14.32M€	14.28M€	14.27M€	14.25M€	14.19M€	14.22M€	14.24M€	41
	16-8-16	14.24M€	14.24M€	14.25M€	14.23M€	14.24M€	14.23M€	**14.21M€**	14.27M€	14.27M€	14.18M€	14.23M€	14.22M€	14.23M€	25
	32-8-32	14.27M€	14.29M€	14.18M€	14.27M€	14.25M€	14.18M€	14.30M€	14.19M€	**14.08M€**	14.24M€	14.25M€	14.25M€	14.22M€	16
	32-16-32	14.25M€	14.28M€	14.24M€	14.29M€	14.21M€	14.25M€	14.25M€	14.23M€	14.24M€	14.23M€	14.21M€	14.23M€	14.19M€	8
	32-16-8-16-32	14.21M€	14.26M€	14.23M€	14.26M€	14.26M€	14.23M€	14.25M€	14.25M€	**14.16M€**	14.23M€	14.30M€	14.23M€	14.08M€	0
RAE (%)	16x1	124%	112%	121%	121%	127%	112%	122%	112%	118%	136%	128%	110%	156%	100
	32x1	144%	114%	119%	127%	143%	115%	127%	113%	128%	122%	126%	114%	121%	75
	16x3	112%	111%	110%	**102%**	111%	112%	113%	109%	113%	115%	124%	112%	114%	59
	32x3	112%	111%	108%	114%	104%	123%	112%	113%	108%	103%	103%	123%	112%	41
	16-8-16	106%	113%	114%	119%	130%	112%	**102%**	112%	114%	109%	128%	115%	110%	25
	32-8-32	111%	114%	107%	108%	113%	105%	103%	104%	**96%**	116%	119%	109%	108%	16
	32-16-32	111%	112%	112%	110%	124%	137%	111%	112%	111%	121%	144%	127%	104%	8
	32-16-8-16-32	104%	109%	108%	123%	115%	110%	107%	112%	**101%**	108%	156%	110%	96%	0
RRSE (%)	16x1	103.0%	102.9%	102.6%	102.7%	103.0%	102.6%	102.7%	102.7%	102.7%	103.0%	102.8%	102.7%	103.2%	100
	32x1	103.0%	103.1%	102.7%	102.6%	103.0%	102.5%	103.1%	102.9%	102.6%	103.2%	102.6%	102.8%	102.8%	75
	16x3	102.7%	102.9%	102.8%	**102.1%**	102.4%	102.5%	103.1%	102.6%	102.8%	102.3%	102.4%	102.5%	102.7%	59
	32x3	103.0%	102.8%	102.8%	102.6%	102.4%	102.8%	103.1%	102.9%	102.8%	102.7%	102.2%	102.4%	102.6%	41
	16-8-16	102.6%	102.6%	102.7%	102.5%	102.6%	102.5%	**102.3%**	102.8%	102.8%	102.1%	102.5%	102.5%	102.5%	25
	32-8-32	102.8%	103.0%	**102.1%**	102.8%	102.6%	102.2%	103.0%	102.2%	**101.4%**	102.6%	102.7%	102.6%	102.4%	16
	32-16-32	102.6%	102.9%	102.6%	102.9%	102.4%	102.6%	102.7%	102.5%	102.6%	102.5%	102.3%	102.5%	102.2%	8
	32-16-8-16-32	102.4%	102.7%	102.5%	102.8%	102.8%	102.5%	102.6%	102.7%	**102.0%**	102.5%	103.0%	102.5%	101.4%	0

Figure 3. Error metrics (MAE, RMSE, RAE and RRSE) for different ANN configurations with validation framework K-fold cross-validation (K=10)

Table 1 shows that the results obtained for the random forest model were improved for all the error metrics (the lowest errors are in bold). The linear regression model did not perform well because the results obtained are worse than the ones obtained with the random forest model for all the error metrics. Therefore, it was concluded that the model is not appropriate for the problem at hand. Isotonic regression and MLP performed better. In fact, both improved the results obtained with the random forest model for some of the error metrics. Isotonic regression improved all the error metrics. MLP substantially improved the results for the RMSE and RRSE error metrics (the values in bold) and are the best compared to the results obtained from the other models.

For all the error metrics, the ANNs are the models that obtained the best results (values in bold). The ANN2 architecture improved the simple MLP and had the best MAE and RAE errors. This comprised a network structure of only 3 hidden layers with 32-8-32 nodes, SeLU activation function, MAE loss function. It was trained using the Adagrad optimizer and appears to be a very promising configuration in terms of the MAE. The simplicity of this network design makes it very suitable in terms of generalization to other data.

Similarly, when considering the RMSE metrics, the MLP with parameters by default outperformed all the other configurations. Relative errors for the previous two ANN configurations were also very good. The ANN2 model had the best RAE, and the MLP model had the best RRSE. These results confirmed that these are the best ANN designs among the ones evaluated herein. Experts may select ANN or ANN2 depending on the risk they are taking: ANN2 minimizes the absolute error value, while ANN (MLP) obtains the minimum value for the square of the errors, which could be considered as a riskier bidding.

Summarizing, ANNs are very promising models for award price prediction. The quality of the final predictions is very good considering that only 96 ANN designs were tested.

5. Conclusion and Future Work

While the importance of using public datasets to make a more efficient use of public resources is generally acknowledged, the choice of the particular type of ML technique to apply to each problem is not straightforward. For award price prediction in public procurement auctions, it was previously reported that the random forest model is an efficient algorithm. The present paper investigates this question considering a larger set of ML models. Extensive experiments were conducted aiming to predict the award price of Spanish tenders.

The contributions of this study are the following. Using different metrics, it was demonstrated that ANNs and isotonic regression can improve the performance of random forests for the award price estimation of public procurement auctions. Furthermore, the influence of the neural network hyperparameters and gradient optimizers on the performance of the ANN was evaluated in detail and it was concluded that a careful choice of hyperparameters can further improve the predictions of the model.

These experiments used different error metrics, and the performance of different ML paradigms was evaluated. Upon analysing the obtained results, it was concluded that among the methods that are not based on ANNs, isotonic regression is the model that gives the best results. Using its implementation in WEKA, it was corroborated that it is a fast and efficient method for training and testing. However, according to all the error metrics considered, the ANN models can

Table 1. Results compared with the baseline model (random forest)

	MAE	RMSE	RAE	RRSE
Random Forest (baseline model)	179,247.80€	6,621,784.24€	31.11%	74.86%
Linear regression	228,491.36€	15,535,231.61€	39.66%	175.63%
Isotonic regression	136,971.39€	5,648,693.54€	23.76%	63.86%
ANN (MLP)	270,953.50€	**1,974,981.24€**	47.03%	**22.33%**
ANN1	140,763.27€	7,416,004.50€	23.03%	83.84%
ANN2	**123,570.91€**	5,110,687.50€	**20.22%**	57.78%
ANN3	157,181.00€	9,543,883.00€	25.71%	107.90%
ANN4	124,035.82€	3,304,259.20€	20.29%	37.36%

outperform the results from isotonic regression. It was proved that a hyperparameter optimization phase can contribute to improving the predictions made by the ANNs.

There are a number of ways in which this work could be extended. Procurement datasets are updated daily, so we can increase the size of the dataset. An update of the dataset in order to include tender information up to 2021 and a revaluation of the performance of the ML algorithms are planned. On the other hand, three interesting input variables that have not yet been used and that could improve the award price estimator in terms of accuracy were discovered during the analysis. These variables include the price criteria weighing variable and the number of bidders for each tender and their economic offers. Unfortunately, this information has not been consistently collected in the Spanish public procurement datasets until now. When these values become available, they will be added to the input variables of this study.

Data Availability

The processed data used to support the findings of this study are available from the corresponding author upon request. The raw data from Spain are available from the Ministry of Finance, Spain. Open data of Spanish tenders are hosted in:

https://www.hacienda.gob.es/es-ES/GobiernoAbierto/Datos%20Abiertos/Paginas/licitaciones_plataforma_contratacion.aspx

REFERENCES

Abadi, M. et al. (2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. Available at: <http://arxiv.org/abs/1603.04467>, last accessed: 20 May, 2021.

Albano, G. L., Cesi, B. & Iozzi, A. (2017). Public procurement with unverifiable quality: The case for discriminatory competitive procedures, *Journal of Public Economics*, *145*, 14-26. DOI: 10.1016/j.jpubeco.2016.11.004

Ballesteros-Pérez, P., Campo-Hitschfeld, M., Mora-Meliá, D. & Domínguez, D. (2015). Modeling bidding competitiveness and position performance in multi-attribute construction auctions, *Operations Research Perspectives*, *2*, 24–35. DOI: 10.1016/j.orp.2015.02.001

Ballesteros-Pérez, P., González-Cruz, M. C. & Cañavate-Grimal, A. (2012). Mathematical relationships between scoring parameters in capped tendering, *International Journal of Project Management*, *30*(7), 850-862. DOI: 10.1016/j.ijproman.2012.01.008

Ballesteros-Pérez, P., González-Cruz, M. C. & Cañavate-Grimal, A. (2013). On competitive bidding: Scoring and position probability graphs, *International Journal of Project Management*, *31*(3), 434-448. DOI: 10.1016/j.ijproman.2012.09.012

Ballesteros-Pérez, P., Skitmore, M., Pellicer, E. & Gutierrez, J. H. (2016). Improving the estimation of probability of bidder participation in procurement auctions, *International Journal of Project Management*, *34*(2), 158-172. DOI: 10.1016/j.ijproman.2015.11.001

Bilal, M. & Oyedele, L. O. (2020). Big Data with deep learning for benchmarking profitability performance in project tendering, *Expert Systems with Applications*, *147*, 113194. DOI: 10.1016/j.eswa.2020.113194

Breiman, L. (2001). Random forests, *Machine Learning*, *45*(1), 5-32. DOI: 10.1023/A:1010933404324

Bzdok, D., Altman, N. & Krzywinski, M. (2018). Statistics versus machine learning, *Nature Methods*, *15*(4), 233-234. DOI: 10.1038/nmeth.4642

Chakravarti, N. (1989). Isotonic Median Regression: A Linear Programming Approach, *Mathematics of Operations Research*, *14*(2), 303-308. DOI: 10.1287/moor.14.2.303

Chou, J. S., Lin, C.-W., Pham, A.-D. & Shao, J.-Y. (2015). Optimized artificial intelligence models for predicting project award price, *Automation in Construction*, *54*, 106-115. DOI: 10.1016/j.autcon.2015.02.006

Conley, T. G. & Decarolis, F. (2016). Detecting Bidders Groups in Collusive Auctions, *American Economic Journal: Microeconomics*, *8*(2), 1-38. DOI: 10.1257/mic.20130254

Dotoli, M., Epicoco, N. & Falagario, M. (2020). Multi-Criteria Decision Making techniques for the management of public procurement tenders: A case study, *Applied Soft Computing*, *88*, 106064. DOI: 10.1016/j.asoc.2020.106064

Duchi, J. C., Bartlett, P. L. & Wainwright, M. J. (2012). Randomized smoothing for (parallel) stochastic optimization. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)* (pp. 5442-5444). IEEE. DOI: 10.1109/CDC.2012.6426698

European Commission (2017). *Public procurement*, *European semester thematic factsheet*. Available at:

<https://ec.europa.eu/info/sites/info/files/file_import/european-semester_thematic-factsheet_public-procurement_en_0.pdf>, last accessed: 20 May, 2021.

European Commission (2020). *Unleashing the Full Potential of SMEs*, p. 3. DOI: 10.2775/296379

García Rodríguez, M. J., Montequín, V. R., Ortega-Fernández, F. & Balsera, J. (2019a). Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning, *Complexity*, *2019*(v), 1-20. DOI: 10.1155/2019/2360610

García Rodríguez, M. J., Montequín, V. R., Ortega-Fernández, F. & Balsera, J. (2019b). Spanish Public Procurement: Legislation, open data source and extracting valuable information of procurement announcements, *Procedia Computer Science*, *164*, 441-448. DOI: 10.1016/j.procs.2019.12.204

García Rodríguez, M. J., Montequín, V. R., Ortega-Fernández, F. & Balsera, J. (2020). Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain, *Complexity*, *1*, 1-20. DOI: 10.1155/2020/8858258

Gorgun, M. K., Kutlu, M. & Onur Tas, B. K. (2020). Predicting The Number of Bidders in Public Procurement. In *2020 5th International Conference on Computer Science and Engineering (UBMK)*. (pp. 360-365). IEEE. DOI: 10.1109/UBMK50275.2020.9219404

Hall, M. Frank, E., Holmes, G., Bernhard Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). The WEKA data mining software, *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18. DOI: 10.1145/1656274.1656278

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*, second edition. New York, NY: Springer New York. Part of the Springer Series in Statistics book series. DOI: 10.1007/978-0-387-84858-7

Huber, M. & Imhof, D. (2019). Machine learning with screens for detecting bid-rigging cartels, *International Journal of Industrial Organization*, *65*, 277-301. DOI: 10.1016/j.ijindorg.2019.04.002

Keras. (2021). *Optimizers*. Available at: <https://keras.io/api/optimizers/>, last accessed: 20 May, 2021.

Kim, J. M. & Jung, H. (2019). Predicting bid prices by using machine learning methods, *Applied Economics*, *51*(19), 2011-2018. DOI: 10.1080/00036846.2018.1537477

Kingma, D. P. & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (pp. 1–15). Available at: <http://arxiv.org/abs/1412.6980>, last accessed: 20 May, 2021.

Klambauer, G., Unterthiner, T., Mayr, A. & Hochreiter, S. (2017). Self-Normalizing Neural Networks, *Advances in Neural Information Processing Systems*, *2017*(Decem), 972-981. Available at: <http://arxiv.org/abs/1706.02515>, last accessed: 20 May, 2021.

Kuziemski, M. & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings, *Telecommunications Policy*, *44*(6), 101976. DOI: 10.1016/j.telpol.2020.101976.

Lee, I. & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges, *Business Horizons*, 63(2), 157-170. DOI: 10.1016/j.bushor.2019.10.005.

Lorentziadis, P. L. (2020). Competitive bidding in asymmetric multidimensional public procurement, *European Journal of Operational Research*, *282*(1), 211-220. DOI: 10.1016/j.ejor.2019.09.005

Mohungoo, I., Brown, I. & Kabanda, S. (2020) A Systematic Review of Implementation Challenges in Public E-Procurement, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2020) 12067 LNCS*, 46-58. Springer International Publishing. DOI: 10.1007/978-3-030-45002-1_5

Mullainathan, S. & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach, *Journal of Economic Perspectives*, *31*(2), 87-106. DOI: 10.1257/jep.31.2.87

Obwegeser, N. & Müller, S. D. (2018). Innovation and public procurement: Terminology, concepts, and applications, *Technovation*, *74–75*(April 2016), 1-17. DOI: 10.1016/j.technovation.2018.02.015

Onur, I. & Tas, B. K. O. (2019). Optimal bidder participation in public procurement auctions, *International Tax and Public Finance*, *26*(3), 595-617. DOI: 10.1007/s10797-018-9515-2

TensorFlow (2021). *An end-to-end open source machine learning* platform. Available at:<https://www.tensorflow.org>, last accessed: 20 May, 2021.

Waikako University (2021). *Weka 3: Machine Learning Software in Java*. Available at: <http://old-www.cms.waikato.ac.nz/ml/weka/>, last accessed: 20 May, 2021.

Wang, Y. Xi, C., Zhang, S., Yu, D., Zhang, W. & Li, Y. (2014). A combination of extended fuzzy AHP and Fuzzy GRA for government e-tendering in hybrid fuzzy environment, *Scientific World Journal*, *2014*(1), 123675. DOI: 10.1155/2014/123675

Witten, I., Frank, E. & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Third. Boston: Elsevier. DOI: 10.1016/C2009-0-19715-5

Publicación C

Recomendador de licitadores usando ML: análisis de datos, algoritmo y caso de estudio con licitaciones de España

Hindawi
Complexity
Volume 2020, Article ID 8858258, 20 pages
https://doi.org/10.1155/2020/8858258

WILEY | Hindawi

Research Article

Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain

Manuel J. García Rodríguez, Vicente Rodríguez Montequín, Francisco Ortega Fernández, and Joaquín M. Villanueva Balsera

Project Engineering Area, University of Oviedo, Oviedo 33012, Spain

Correspondence should be addressed to Vicente Rodríguez Montequín; montequi@uniovi.es

Received 14 September 2020; Revised 9 November 2020; Accepted 11 November 2020; Published 25 November 2020

Academic Editor: Thiago Christiano Silva

Recommending the identity of bidders in public procurement auctions (tenders) has a significant impact in many areas of public procurement, but it has not yet been studied in depth. A bidders recommender would be a very beneficial tool because a supplier (company) can search appropriate tenders and, vice versa, a public procurement agency can discover automatically unknown companies which are suitable for its tender. This paper develops a pioneering algorithm to recommend potential bidders using a machine learning method, particularly a random forest classifier. The bidders recommender is described theoretically, so it can be implemented or adapted to any particular situation. It has been successfully validated with a case study: an actual Spanish tender dataset (free public information) which has 102,087 tenders from 2014 to 2020 and a company dataset (nonfree public information) which has 1,353,213 Spanish companies. Quantitative, graphical, and statistical descriptions of both datasets are presented. The results of the case study were satisfactory: the winning bidding company is within the recommended companies group, from 24% to 38% of the tenders, according to different test conditions and scenarios.

1. Introduction

The largest adjudicators of a country, by number of projects and by cost, are public procurement agencies. For example, public authorities in the European Union spend around 14% of GDP (around €2 trillion) on public procurement [1] every year. The definition of public procurement is the purchase of goods, works, or services by a public agency. Public procurement is clearly important to politicians, citizens, researchers, and companies because of its size. On the other hand, the European open data market size (products and services enabled by open data) was €184.45 billion in 2019, according to the official European Data Portal [2]. High growth is expected in the near future. The availability of open data in public procurement announcements (also known as tenders) enables the building of a bidders recommender.

The bidders recommender may be a strategic tool for improving the efficiency and competitiveness of organisations and is particularly suitable for the two main stakeholders: suppliers and public procurement agencies. On the one hand, it is useful to the supplier because it assists in identifying the most suited tenders, i.e., those that they should prioritise. On the other hand, the contracting agency could automatically search companies with a compatible profile for the tender's announcement, e.g., selective tendering where suppliers are only allowed by invitation. Thus, it could be called a "bidders search engine" or a "bidders recommender."

Many public agencies do not easily obtain competitive offers when they publish public procurement announcements. It is a serious problem with negative consequences for the project in terms of cost, quality, lifetime,

sustainability, etc. A bidders recommender would produce significant benefits as follows:

(i) Tenders with more bidders have lower award prices and, consequently, the public agencies will reduce costs. This relationship is quantitatively demonstrated for Spanish tenders in this paper, but there are more empirical studies, e.g., in Italy [3] and the Czech Republic [4, 5].

(ii) This new tool will provide support to small- and medium-sized enterprises (SMEs), which play a crucial role in most economies. It will make it easier and more efficient for SMEs to access procurement auctions, promote inclusive growth, and support principles such as equal treatment, open access, and effective competition [6].

(iii) In scenarios of high participation, it is more difficult to generate corruption or collusive tendering (where the bidders do not compete honestly).

The main objective of this paper is to propose an algorithm to search for suppliers (companies) to invite to tender. Discovering the number and identity of bidders is challenging, since there does not exist a suitable quantitative model to forecast the identities of a single or a group of specific key competitors likely to submit a future tender [7]. So, the input parameters of the bidders recommender have to have the tender's announcement but also be a generic algorithm that can be implemented or adapted to any particular situation. The main issue is to get information about bidders and the rest of the companies in the market because in many countries, the information is not public or free.

Some papers have proposed similar tools, but only the tenders are characterised or analysed, not the bidders, e.g., a product search service [8] or a similar tenders engine search (comparison of one tender to all other tenders according to specific criteria) [9]. Our work is based on the profile of the winning companies rather than the characteristics of the tender. Thus, this paper is a novel study which brings a new and modern perspective to gathering tenders and bidders. The bidders recommender has used tenders that have been published in Spain. In particular, the tender dataset has 102,087 Spanish tenders from 2014 to 2020. All types of works are included, not only construction auctions (which are the favourite subjects in the public procurement literature, for several reasons). The company dataset has 1,353,213 Spanish companies to search suitable bidders. In [10, 11], the Spanish public procurement system as well as the European and national legislation is described, and they have also analysed Spanish tenders for other purposes.

The application of this pioneering bidders recommender by public procurement agencies or potential bidders is summarised in Figure 1. It has three sequential steps or phases, and the input is obviously a new public procurement announcement, also known as a tender notice. Initially, it is based on forecasting the winning company of the tender thanks to a machine learning method called a random forest classifier model. This classification model has previously been trained with lots of tenders and their respective winning companies. The second phase is to add the business information of the forecast winning company for creating a profile of a winning company. The business information is in the company dataset (data from the Business Register). Finally, similar or compatible companies are searched, according to their profile, where the search criteria are filters or fixed rules.

The paper is structured as follows. Section 2 summarises the literature review associated with the bidders recommender in public procurement auctions. Section 3 presents the fields of the dataset and the machine learning algorithm (called random forest classifier) which will be used in the recommender. Furthermore, the bidders recommender is explained in detail (Section 3.5) and some evaluation metrics are defined to measure the accuracy of detecting the winning company of the tender within the group of bidders. Section 4 quantitatively describes the datasets for the real case study from Spain to test the bidders recommender. It is tested under different scenarios, and the results are presented in Section 4.3. In Section 5, the recommender is discussed from a general perspective to be applied to other countries or datasets. Finally, some concluding remarks, limitations, and avenues for future research are presented in Section 6.

2. Literature Review

This paper involves (either directly or indirectly) diverse topics such as open government data, public procurement and its regulation, machine learning, tender evaluation, prediction techniques, business registers, and so on. The bidders recommender has a multidisciplinary nature which fills a gap in the literature. Nevertheless, the key components have an extensive literature which will be summarised in the following paragraphs.

In this article, we used open data and, especially, Open Government Data (OGD). The OGD initiatives have grown very strongly in the academic field [12–14]. That is to say, open data are produced by governmental entities in order to promote government transparency and accountability. Hence, there are different stakeholders, user groups, and perspectives [15, 16]. The OGD is a part of the public value of e-government [17], and it is a new and important resource with economic value [18, 19]. For example, data.europe.au and *data.gov* are online portals that provide open access datasets in a machine-readable format [20] and are generated by the European Union and the United States of America public agencies, respectively. However, there are challenges and risks in dealing with the data quality of open datasets (quality over quantity) [21] and this article suffers from these too. It is very important to measure the transparency and the metadata quality in the open government data portals [22–24].

Other public procurement fields that have recently sparked the interest of governments, policy makers, and researchers are Big and Open Linked Data (BOLD) [25], the growing awareness of public procurement as an innovation policy tool [26], and the role of e-government in sustainable public procurement [27].

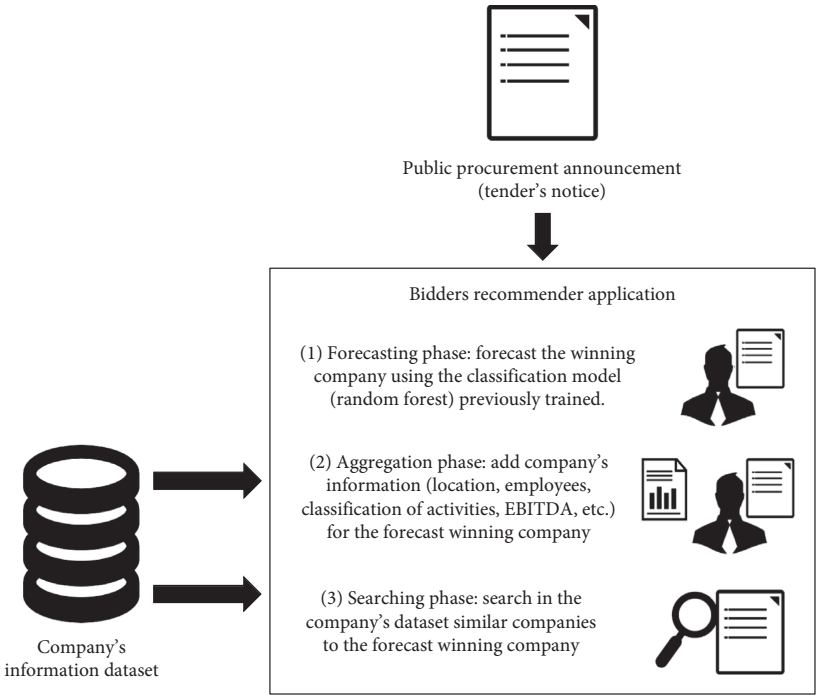


Figure 1: Flowchart of the application of the bidders recommender for a new tender.

This article uses a machine learning algorithm. The big data and machine learning technologies can be used for econometrics [28, 29], enterprises [30], tender evaluation [31], or analysis of public procurement notices [32]. Therefore, this paper follows the trends in the literature.

There is extensive literature about tender evaluation (also called bidding selection methods) for the selection of the optimal supplier in public procurement [33] with different techniques such as the economic scoring formulas [34], data envelopment analysis [35] or multicriteria decision making [36, 37], and where multiple bidders are evaluated on the basis of price and quality [38]. In particular, the most studied public procurement auctions are related to construction, i.e., distribution of bids [39], bidding competitiveness and position performance [40], strategic bidding [41], tender evaluation and contractor selection [42, 43], and empirical analysis in countries such as Slovakia [44]. There are almost no studies which include all kinds of business sectors and a large volume of tenders. However, this article has a holistic approach due to the large tender dataset of all sectors.

Another relevant subject in the public procurement literature is the detection of collusive tendering or bid rigging [45] with case studies in Spain [46], India [47], and Hungary [48]. This occurs when businesses that would otherwise be expected to compete secretly conspire to raise prices or lower the quality of goods or services for purchasers in a public procurement auction (this is called a cartel). In addition, public procurement contracts have other issues such as optimal quality [49], too many regulations [50], systemic risk [51], or corruption [52–54]. Corruption is a form of dishonesty undertaken by a person or organisation with the authority to acquire illicit benefit. There are empirical studies to detect corruption by analysing public tenders in many countries, for example, in China [55], Russia [56], the Czech Republic [57], and Hungary [58]. The application of algorithms by governments or enterprises to detect collusion or corruption [59], especially using machine learning methods [60–62], has become an almost inevitable topic and the subject of numerous studies. Indirectly, this article could create a useful tool for these topics since it is able to forecast the most probable winning bidders and, therefore, the detection of unlikely winners too.

Forecasting and prediction techniques are widely studied and applied in the academic field of public procurement auctions. In [63], the mathematical relationship between scoring parameters in tendering is studied because, among other reasons, it is useful for the bid tender forecasting model [64]. There are some notable key parameters which have been analysed in the forecasting literature, especially for construction auctions, from traditional techniques to new machine learning methods, for example, the probability of bidder participation [7], an award price estimator [10, 65, 66], or cost estimator [67, 68]. However, as far as we know, this article is the first attempt to forecast the winning company for all tenders in a country.

In conclusion, this paper creates a smart search engine to recommend a group of companies for each tender, according to the forecast winning company. This means they have a similar business, technical, and economic profiles. Therefore, it is necessary to find these profiles in the Business Registers [69, 70] or other databases where the company's annual accounts are available. For instance, it is even possible to forecast

the corporate distress using machine learning in such reports [71]. The analysis of a company's profile has the same basis as the academic topic called bankruptcy prediction. This is the measurement of corporate solvency and the creation of prediction models [72] to forecast the company failure or distress. It has been intensively discussed over the past decades [73], using traditional statistical techniques [74–76] or machine learning methods, such as gradient boosting [72], neural networks [77], support vector machine [78], or the comparison of different methods [79, 80].

3. Materials and Methods

This section describes the necessary components to create the bidders recommender proposed in this article. It is described theoretically so that it can be implemented in any country, not only in Spain. Section 3.1 presents the origin of the tender dataset and describes its fields, and, analogously, the company dataset is presented in Section 3.2. Section 3.3 explains the random forest classifier which is used in the first phase of the bidders recommender method. In Section 3.4, the evaluation metrics are defined to measure the recommender's accuracy. Finally, the bidders recommender algorithm is described in detail in Section 3.5.

3.1. Tender Dataset. The European and Spanish legislation on public procurement and on the reuse of public information is extensively detailed in [11]. The official website of the Public Sector Contracting Platform (P.S.C.P.) of Spain publishes the public procurement notices and their resolutions of all contracting agencies belonging to the Spanish Public Sector.

The P.S.C.P. has an open data section for the reuse of this information which will be used in this article to generate the tender dataset. The information is provided by the Ministry of Finance (the link is given in the Data Availability section) and has been published as open data since 2012. The fields, their descriptions, and the process to obtain the dataset are the same as discussed in [10]. However, these fields are shown in Table 1 for the convenience of the reader. A remarkable limitation is that only the identity of the winning company is known, not the rest of the bidders, and this will be a constraint for the recommendation system.

3.2. Company Dataset. In general, to obtain business information (companies' annual accounts) over several years is not easy or free. In Europe, Business Registers offer a range of services, which may vary from one country to another. However, the core services provided by all registers are to examine and store company information and to make this information available to the public [69]. *European Regulation 2015/884* [81] interconnects the Business Registers of the EU countries. The *European Business Registry Association* [82] has a list of Business Registers from around the world, for more information.

The authors have collected a dataset of annual accounts from Spanish companies, based on the information available in the Spanish Business Register. It is a public institution, but access is not free of charge. It is the main legal instrument for recording business activity: the company documents and submission of the annual accounts. The companies become a legal entity through their registration on the Business Register.

The fields of the company dataset are explained in Table 2. They can be divided into 5 headings: general information, human resources, location, accounting measures (operating income, EBIT, and EBITDA), and different systems for classifying industries or economic activities (CNAE, NACE2, IAE, US SIC, and NAICS). It should be noted that the company's annual accounts have more fields, but the authors have not been able to access and collect them. The fields of this dataset try to characterise the company from different points of view: main business activities (CNAE, NACE2, IAE, US SIC, and NAICS), nearby market (location), work capacity (employees), size (operating income), financial performance (EBITDA), etc. Not all of the fields have been used because they are not relevant to the analysis in this paper.

3.3. Random Forest Classifier. Random forest (RF), introduced by Breiman [83] in 2001, is an ensemble learning method for classification or regression that operates by constructing a multitude of decision trees at training times and outputting the class, which is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is a popular learning algorithm that offers excellent performance [84], no overfitting [85, 86], a versatility of applicability to large-scale problems and in handling different types of data [85, 87]. Particularly, Random Forest has been applied with remarkable success in tender datasets, for example in [10]. It provides its own internal generalisation error estimate, called the out-of-bag (OOB) error. Simplified algorithm of RF for classification [88] is summarized in Algorithm 1.

At each split in each tree, the improvement in the split criterion is the measure of the importance attributed to the splitting variable and is accumulated over all the trees in the forest separately for each variable. This is called "variable importance" [83].

3.4. Evaluation Metrics. It is necessary to define some error metrics to compare similar variables of the datasets and calculate the prediction error of the bidders recommender. The use of metrics based on medians and relative percentage is useful because the dataset has outliers of great weight, and the use of such metrics helps us to counteract the effect of these outliers. To compare variables of the dataset, the median absolute percentage error (MdAPE) was used, as defined in the following equation:

$$\mathrm{MdAPE}\,(\%) \quad \frac{100}{\mathrm{n}}\,\mathrm{median}\left(\left|\frac{A_1 - F_1}{A_1}\right| \left|\frac{A_2 - F_2}{A_2}\right| \quad \left|\frac{A_\mathrm{n} - F_\mathrm{n}}{A_\mathrm{n}}\right|\right) \tag{1}$$

where A_t is the actual value for period t, F_t is the expected value for period t, and n is the number of periods.

The following error metrics are to measure the prediction error of the RF classifier method for multiclass classification on imbalanced datasets [89]. Multiclass

Table 1: Most relevant data fields in the Spanish Public Procurement Notices (tenders) used in the dataset.

Name	Description	Name column dataset
Tender status	Status of the tender during the development of the procedure: prior notice, in time, pending adjudication, awarded, resolved, or cancelled	Not used (similar to Result_code)
Contract file number	Unique identifier for a contract file	Not used
Object of the contract	Summary description of the contract	Not used (unstructured textual information)
Public procurement agency	Public procurement agency that made the tender: name, identifier (NIF or DIR3), website, address, postal code, city, country, contact name, telephone, fax, e-mail, etc. CCAA is the Autonomous Community which is a first-level division in Spain. Latitude and longitude have been calculated from postal code, and they are not official fields in the notice.	Name_Organisation Postalzone CCAA Province Municipality Latitude Longitude
Tender price	Amount of bidding budgeted (taxes included)	Tender_Price
Duration	Time (days) to execute the contract	Duration
CPV classification	CPV (Common Procurement Vocabulary) is a European system for classifying the type of work in public contracts defined in the Commission Regulation (EC) No 213/2008: http://data.europa.eu/eli/reg/2008/213/oj	CPV
	The numerical code consists of 8 digits, subdivided into divisions (first 2 digits of the code), groups (first 3 digits), classes (first 4 digits), and categories (first 5 digits)	CPV_Aggregated (first 2 digits of the CPV number)
Contract type	Type of contract defined by legislation (Law 9/2017): works, services, supplies, public works concession, works concession, public services management, services concession, public sector and private sector collaboration, special administrative, private, patrimonial, or others	Type_code
Contract subtype	Code to indicate a subtype of contract. If it is a type of service contract: based upon the 2004/18/CE Directive, Annex II. If it is a type of works contract: works contract codes defined by the Spanish DGPE.	Subtype_code
Contract execution place	Contract's execution has a place through the Nomenclature of Statistical Territorial Units (NUTS), created by Eurostat [47]	Not used (assumed equal to postalzone)
Type of procedure	Procedure by which the contracts was awarded: open, restricted, negotiated with advertising, negotiated without publicity, competitive dialogue, internal rules, derived from framework agreement, project contest, simplified open, association for innovation, derivative of association for innovation, based on a system dynamic acquisition, bidding with negotiation, or others	Procedure_code
Contracting system	The contracting system indicates whether it is a contract itself or a framework agreement or dynamic acquisition system	Not used
Type of processing	Type of processing: ordinary, urgent, or emergency	Urgency_code
Award result	Type of results: awarded, formalised, desert, resignation, and withdrawal	Result_code
Winner identifier (CIF)	Identifier of the winning bidder (called CIF in Spain) and its province (region)	CIF_Winner Winner_Province
Award price	Amount offered by the winning bidder of the contract (taxes included)	Award_Price
Date	Date of agreement in the award of the contract	Date
Number of received offers	Number of received offers (bidders participating) in each tender	Received_Offers

classification occurs when the input is classified into one, and only one, nonoverlapping class. An imbalanced dataset occurs when there is a disproportionate ratio of observations in each class.

Let $\hat{y}_i$ be the predicted value of the $i-$th sample ($1 \le i \le n$), y_i be the corresponding true value, ω_i be the corresponding sample weight, and L be the set of classes ($1 \le l \le L$). Accuracy (2) is the proportion of correct predictions over n samples:

$$\text{accuracy} \quad \frac{1}{n}\sum_{i=1}^{n} 1(\hat{y}_i \quad y_i) \tag{2}$$

where $1(\hat{y}_i)$ is the indicator function. The equation returns a 1 if the classes match and 0 otherwise.

Balanced accuracy (3) avoids inflated performance estimates on imbalanced datasets:

$$\text{balanced accuracy} \quad \frac{1}{\sum_{i=1}^{n} \hat{\omega}_i}\sum_{i=1}^{n} 1(\hat{y}_i \quad y_i)\,\omega_i \tag{3}$$

where $1(\hat{y}_i)$ is the indicator function and $\hat{\omega}_i \quad \omega_i / \sum_{j=1}^{n} 1(\hat{y}_j \quad y_i)\,\omega_j$.

Let y_l be the subset of true values with class l. The precision (average macro) is calculated as follows:

$$\text{precision} \quad \frac{1}{L}\sum_{l=1}^{L} \frac{|y_l \cap \hat{y}_l|}{|y_l|} \tag{4}$$

Finally, the out-of-bag (OOB) is a method of measuring the prediction error in RF and other machine learning

Table 2: Data fields in the company's information database.

Name	Description	Name column dataset
Name company	Name of the company	Not used
CIF	CIF (for the Spanish term Certificado de Identificación Fiscal) is the company registration number. This identifier provides formal registration on the company tax system in Spain. In many countries, a company would be issued with a separate VAT number, while in Spain, the CIF also forms the VAT number.	CIF
Establishment date	It is the date on which the company starts its activities	Establishment_Date
Legal form	It is the entity type of company defined in the Spanish legal system. Mainly, there are two types: public limited company (PLC) and private company limited by shares (Ltd.)	Legal_Form
Last available year info	Last available year with economic information (operating income, EBIT, and EBITDA) of the company	Last_Available_Year_Info
Social capital	Minimum capital required to register the company in the legal system	Not used
Status company	Opened company (active) or closed company (inactive)	Status_Company
City, province, and country	City, province, and country of the company	City_Company Province_Company
Latitude and longitude	It represents the coordinates at geographic coordinate system of the company's location	Latitude_Company Longitude_Company
Web	Website of the company	Not used
President and CEO	President and Chief Executive Officer (CEO) of the company	Not used
Employees	Number of employees	Employees
Number group companies	Number of companies controlled (owned) by the company	Not used
Number investee companies	Number of companies in which the investor (company) makes a direct investment	Not used
Operating income	It measures the amount of profit realised from a business's operations, after deducting operating expenses (cost of goods sold, wages, depreciation, etc.). Value per year. Operating income gross income − operating expenses net profit + interest + taxes	Operating_Income
EBIT	Earnings before interest and taxes (EBIT) is a company's net income before interest and income tax expenses have been deducted. It is an indicator of a company's profitability. EBIT can be calculated as revenue minus expenses excluding tax and interest. The most important difference between operating income and EBIT is that EBIT includes any nonoperating income the company generates. Value per year. EBIT net income + interest + tax	EBIT
EBITDA	Earnings before interest, taxes, depreciation, and amortization (EBITDA) is a measure of a company's overall financial performance. Value per year. EBITDA net income + interest + taxes + depreciation + amortization operating income + depreciation + amortization	EBITDA
Activity description	Textual description of the main business activities of the company	Not used
CNAE	CNAE (for the Spanish term Clasificación Nacional de Actividades Económicas) is the national classification of economic activities from Spain for statistical purposes. The last version of the CNAE has been adopted in 2009 (Royal Decree-Law 475/2007). It is equivalent to the European classification NACE2. It has primary and secondary codes.	CNAE_Primary CNAE_Secondary
NACE2	NACE2 (for the French term Nomenclature statistique des Activités Économiques dans la Communauté Européenne) is the statistical classification of economic activities in the European Community. The current version is revision 2 and was established by Regulation (EC) No 1893/2006. It is the European implementation of the United Nations (UN) classification ISIC (revision 4). There is a correspondence between NACE and ISIC. It has primary and secondary codes.	NACE2_Primary NACE2_Secondary
IAE	IAE (for the Spanish term Impuestos de Actividades Económicas) is the classification of economic activities in the Spanish Tax Agency for tax purposes. It has primary and secondary codes.	IAE_Primary IAE_Secondary
US SIC	The Standard Industrial Classification (SIC) is a system for classifying industries established in the United States (US) but also used by agencies in other countries. In the US, the SIC has been replaced by NAICS but some US government departments and agencies continued to use SIC codes. It has primary and secondary codes.	SIC_Primary SIC_Secondary
NAICS	The North American Industry Classification System (NAICS2017) is a classification of business establishments by type of economic activity (process of production). It has largely replaced the older SIC. It has primary and secondary codes.	NAICS_Primary NAICS_Secondary

(1) For b 1 to B (number of trees):
(a) Draw a bootstrap sample $\mathbb{Z}^*$ of size N from the training data.
(b) Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{\min}$ is reached.
(i) Select m variables at random from the p variables.
(ii) Pick the best variable/split point among the m.
(iii) Split the node into two daughter nodes.
(2) Output the ensemble of trees $\{T_b\}_1^B$.
To make a prediction at a new point x, let $\widehat{C}_b(x)$ be the class prediction of the $b-$th random forest tree. Then, $\widehat{C}_{rf}^B(x)$ majority vote $\left\{\widehat{C}_b(x)\right\}_1^B$.

Algorithm 1: Simplified algorithm of random forest for classification.

models. The RF classifier is trained using bootstrap aggregation, where each new tree is fitted from a bootstrap sample of the training observations z_i $(y_i\ \hat{y}_i)$. The OOB error is the average error for each z_i calculated using predictions from the trees that do not contain z_i in their respective bootstrap sample. This allows the RF classifier to be fitted and validated while being trained [88].

3.5. Bidders Recommender Algorithm

3.5.1. Creation of the Bidders Recommender Algorithm. The flowchart for the creation of the bidders recommender is summarised in Figure 2. The two data sources and the steps for its development are illustrated. It is important to note that the application of the bidders recommender is one thing (see Figure 1), but its creation and setting is another. The steps are quite similar, but they are not the same.

The creation of the bidders recommender has the following four sequential steps. It is based on initially training the classification model, then forecasting the winning company, and aggregating its business information. Finally, it requires searching for similar companies, according to the profile where the search criteria are filters or fixed rules.

(1) Training and Forecasting Phase. Train the classification model (random forest classifier) over the tender dataset. Typically, 80% of the data is for the training subset and 20% is for the testing subset. Then, forecast the winning company for each tender of the testing subset by applying the previous classification model. The following input and output variables (described in Table 1) have been used by the random forest classifier:

(1) Input variables: Procedure_code, Subtype_code, Name_Organisation, Date, CCAA, Province, Municipality, Latitude, Longitude, Tender_Price, CPV, and Duration.
(2) Output variables (forecast): N winning companies (variable called CIF_Winner) for each tender. Typically, N 1 but it is also possible to predict the N most probable companies to win the tender.

At this point, the accuracy$_{n\ N}$ of the testing subset can be calculated. It will be the minimum accuracy of the bidders recommender because these N forecast winning companies will be inserted into the recommended companies group.

(2) Aggregation Phase. Add the business fields from the company dataset (described in Table 2) to the forecast winning company estimated in the previous step. The business fields are

(1) General information: *CIF*, *Last_Available_Year_Info*, *Status_Company*, and *Employees*.
(2) Location: *Latitude* and *Longitude*.
(3) Economic indicators per year: *Operating_Income*, *EBIT*, and *EBITDA*.
(4) Systems of classification of economic activities: NACE2, IAE, SIC, and NAICS.

(3) Searching Phase. In the company dataset, search for similar companies to the forecast winning company. Hence, it will create a recommended companies group for each tender. The search criteria (filters) are a basic mechanism to modulate the number of recommended companies, and they are described below. Each filter has a constant factor (numeric value from 0 to infinite) to increase or decrease the size of the search.

(a) $\text{OperatingIncome}_{\text{co}} \geq F_{\text{OI}} \cdot \text{OperatingIncome}_{\text{forecastco}}$.
(b) $\text{EBIT}_{\text{co}} \geq F_{\text{EBIT}} \cdot \text{EBIT}_{\text{forecast co}}$.
(c) $\text{EBITDA}_{\text{co}} \geq F_{\text{EBITDA}} \cdot \text{EBITDA}_{\text{forecast co}}$.
(d) $\text{Employees}_{\text{co}} \geq F_E \cdot \text{Employees}_{\text{forecast co}}$.
(e) $\sum_{i\ 1}^{C} 1[\ \text{Code}_{\ \text{co}}\ \ \text{Code}_{\ \text{forecast co}}\ ] \geq F_{CEA} \cdot C$ where $1[\text{Code}]$ is the indicator function (returns 1 if the codes match and 0 otherwise), C is the total number of codes of the forecast company, and Code is the identification number of the different systems of classifications of economic activities registered by the forecast company:

Code NACE2 IAE SIC and NAICS .

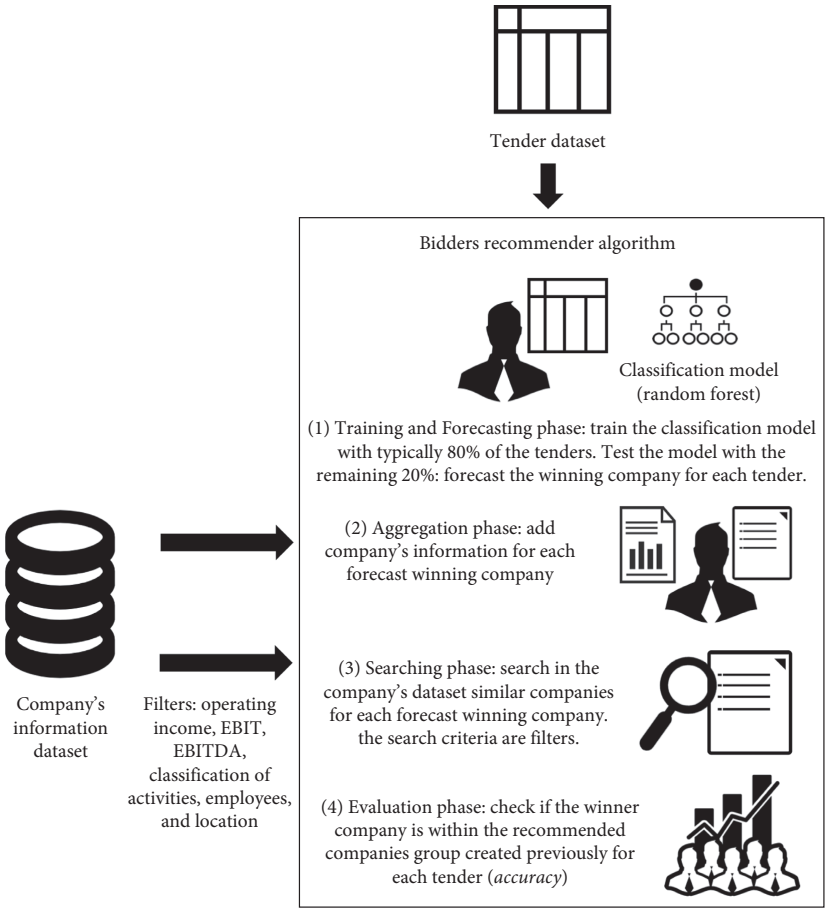


Figure 2: Flowchart of the creation of the bidders recommender.

(f) $\text{Distance}_{\text{tender}-\text{co}} \leq F_{\text{D}} \cdot \text{Distance}_{\text{tender}-\text{forecast co}}$.

Therefore, it is necessary to set up the bidders recommender by assigning numeric values to the previous six factors: F_{OI}, F_{EBIT}, F_{EBITDA}, F_E, F_{CEA}, and F_D. The three economic filters (operating income, EBIT, and EBITDA) are annual values. The minimum annual value for $\text{Operating\_Income}_{\text{forecast co}}$, $\text{EBIT}_{\text{forecast co}}$, and $\text{EBITDA}_{\text{forecast co}}$ for the last available 5 years were selected. For searching companies, the $\text{Operating\_Income}_{\text{co}}$, EBIT_{co}, and $\text{EBITDA}_{\text{co}}$ of the tender's year date were selected.

(4) Evaluation Phase. Check if the real winner company is within the recommended companies group created for each tender (phase 3). This evaluation metric is called $\text{accuracy}_{n\ M}$. Logically, $\text{accuracy}_{n\ M} \geq \text{accuracy}_{n\ N}$ because the N forecast winning companies (phase 1) are automatically within the recommended companies group. Furthermore, the mean and median number of the recommended companies of each tender is calculated. Large groups are more likely to contain the real winner company but, obviously, the smart search engine is less useful because it recommends too many companies.

Therefore, the bidders recommender selects winning companies from the tender dataset but also incorporates new companies available in the market (company dataset) that have a similar profile to the forecast winning company. Creating this profile to search similar companies is a very complex issue, which has been simplified. For this reason, the searching phase (3) has basic filters or rules. Moreover, it is possible to modify or add other filters according to the available company dataset used in the aggregation phase. The fields available in the company dataset (filters) will strongly depend on the country. In our case study, the filters are the following:

(i) Economic resources to finance the project: $\text{Operating\_Income}_{\text{co}}$, EBIT_{co}, and $\text{EBITDA}_{\text{co}}$.

(ii) Human resources to do the work: $\text{Employees}_{\text{co}}$.

(iii) Kind of specialised work which the company can do: NACE2 IAE SIC and NAICS.

(iv) Geographical distance between the company's location and the tender's location: $\text{Distance}_{\text{tender}-\text{co}}$. It will be shown that it is a fundamental parameter. Intuitively, the proximity has business benefits such as lower costs.

3.5.2. Application of the Bidders Recommender. The application of the bidders recommender (see Figure 1) by public agencies or potential bidders for a new tender was summarised in Section 1. It has three phases, which is very similar to its creation. The first phase (forecasting) is to predict the most probable company to win the tender using the model, already trained by the random forest classifier. The second phase (aggregation) is exactly the same: add the business fields from the company to the forecast winning company. Finally, the third phase (searching) is simply applying the filters (numeric factors) that were previously fixed in the creation, in order to search the recommended companies.

4. Experimental Analysis

A real case study from Spain is presented to evaluate the bidders recommender. Section 4.1 summarises the preprocessing of the two data sources: tender dataset and company dataset. Section 4.2 provides a quantitative description of both datasets and their relationship such as the correlation. In Section 4.3, the bidders recommender is applied under two different scenarios with five different settings in each one. Finally, the results are presented and analysed for these ten different tests.

4.1. Data Preprocessing. Data preprocessing of the tender dataset is necessary due to the fact that information has not been verified automatically to correct human errors, such as incorrect formatting, wrong values, empty fields, and so on. Data preprocessing can be divided into the following 5 consecutive tasks: extraction, reduction, cleaning, transformation, and filtering. They are described in detail in [10] because the data source and the data preprocessing are the same in both articles. At first, there were 612,090 tenders. After data preprocessing, there were 110,987 tenders.

Data preprocessing of the company dataset is a simple task since the data source is already a database. Therefore, it is not necessary to verify or check the data. The company dataset has 1,353,213 Spanish companies listed.

Finally, the tender dataset has been merged with the company dataset. This relationship is possible thanks to the CIF field (ID company number) which both datasets have. The merged dataset has 102,087 tenders and their respective winner companies. About 8,900 tenders have been lost because the winning company's CIF has not been found for some reason. The possible reasons include foreign company, wrong CIF value, winning company's CIF not stored in the database, etc.

4.2. Statistical Analysis of the Datasets. Firstly, the most relevant information of the tender dataset will be explained, quantitatively. Secondly, the company dataset will also be explained, and, finally, the correlations between both datasets will be analysed.

Table 3 shows the quantitative description of the tender dataset: total numbers, means, medians, maximum, percentages, etc. The dataset has 19 fields or variables: 15 announcement fields and 4 award fields. There are 102,087 tenders from 2014 to 2020 spread across Spain, and any CPV code is possible. Therefore, there are a wide number of heterogeneous tenders which will be used in the bidders recommender.

Looking at Table 3, the following issues are observed:

(i) There are a lot of winning companies and tendering organisations. On average, each public procurement agency creates 17.72 tenders and each company wins 4.80 tenders.

(ii) There is a great dispersion of prices (for both Tender_Price and Award_Price) considering the median, the mean, and the maximum. Furthermore, there is a remarkable difference between Tender_Price and Award_Price, looking at the differences between their medians (€12,535.60) and their means (€93,177.42).

(iii) The 5 types of CPV with greater weight add up to 51.16% of the total number of tenders.

(iv) With every passing year, a greater number of tenders are recorded in the Spanish Public Procurement System without wrong values or incomplete data.

(v) The Spanish capital (Madrid) accounts for 37.50% of the tenders. The 5 Provinces with greater weight add up to 56.21% of the total number of tenders (Spain has 50 provinces).

(vi) 32.43% of Spanish auctions have only one bidder. A large number of tenders with only one bidder could be a sign of anomaly (collusion, corruption, economical disorder, or others). However, according to the European public reports [90], this ratio is similar to other countries, like, for example, Poland (37.5%), Romania (34%), or Czech Republic (26.6%).

Table 4 shows the quantitative description of the company dataset. There are 1,353,213 companies, and 61.44% of them are active. The dataset has 23 fields (see the description in Table 2): general information of the company, location, employees, 3 economic indicators (operating income, EBIT, and EBITDA), and 5 systems of classification of economic activities (CNAE, NACE2, IAE, SIC, and NAICS).

Looking at Table 4, the following issues are discussed:

(1) The Spanish companies have a small size for 3 reasons. First of all, 91.58% are limited companies (private companies limited by shares). Secondly, the mean number of employees is 11.51 employees per company. Thirdly, in the year 2018, the median operating income was only €299,130, the median EBIT was only €10,472.40, and the median EBITDA was only €18,733.35.

(2) The highest number of economic fields (operating income, EBIT, and EBITDA) were recorded in the year 2016 (about 700,000 companies), followed by 2015 and then 2017.

(3) The 5 Provinces with greater weight add up to 45.38% of the total number of companies. So, the companies are concentrated in certain locations.

Table 3: Quantitative description of the tender dataset.

Topic	Description	Value
General values	Total number of tenders in the dataset	102,087
	Temporal range of tenders	2014/01/02–2020/03/31
	Total number of tendering organisations	5,761
	Total number of winning companies	21,268
	Mean number of offers received per tender	4.38
	Mean duration of tender's works	376.30 days
Dataset's variables	Input variables of tender's notice: Procedure_code, Urgency_code, Type_code, Subtype_code, Result_code, Name_Organisation, Postalzone, Postalzone_CCAA, Postalzone_Province, Postalzone_Municipality, Tender_Price, CPV, CPV_Aggregated, Duration, and Date	15 input variables (description in Table 1)
	Output variables of tender's resolution: Award_Price, Winner_Province, CIF_Winner, and Received_Offers	4 output variables (description in Table 1)
Tender price (taxes included)	Mean tender price	€422,293.27
	Median tender price	€78,650.00
	Maximum tender price	€3,196,970,000
	Aggregated tender price of all tenders	€43,110,653,361
Award price (taxes included)	Mean award price	€329,115.85
	Median award price	€66,114.40
	Maximum award price	€786,472,000
	Aggregated award price of all tenders	€33,598,449,589
Number of tenders by received offers (bidders)	Tenders with Received_Offers 1 (one bidder)	33,112 (32.43%)
	Tenders with Received_Offers 2 (two bidders)	16,302 (15.97%)
	Tenders with Received_Offers 3 (three bidders)	13,583 (13.31%)
	Tenders with Received_Offers ≥ 4 (four or more bidders)	39,090 (38.29%)
Number of tenders by CPV	Tenders with CPV 45 : Construction work	24,699 (24.19%)
	Tenders with CPV 50 : Repair and maintenance services	8,692 (8.51%)
	Tenders with CPV 79 : Business services (law, marketing, consulting, recruitment, printing and security)	6,900 (6.76%)
	Tenders with CPV 72 : IT services (consulting, software development, internet and support)	6,444 (6.31%)
	Tenders with CPV 34 : Transport equipment and auxiliary products to transportation	5,506 (5.39%)
Number of tenders by type code	Tenders with Type_code 1: Goods/Supplies	31,065 (30.43%)
	Tenders with Type_code 2: Services	46,377 (45.43%)
	Tenders with Type_code 3: Works	24,480 (23.98%)
Number of tenders by year	Number of tenders in 2014	1,002 (0.98%)
	Number of tenders in 2015	5,165 (5.06%)
	Number of tenders in 2016	9,746 (9.55%)
	Number of tenders in 2017	15,081 (14.77%)
	Number of tenders in 2018	25,879 (25.35%)
	Number of tenders in 2019	38,571 (37.78%)
	Number of tenders in 2020 (until March inclusive)	6,643 (6.51%)
Number of tenders by location (province)	Top 1: number of tenders from Madrid	38,285 (37.50%)
	Top 2: number of tenders from Valencia	7,616 (7.46%)
	Top 3: number of tenders from Alicante	4,097 (4.01%)
	Top 4: number of tenders from Baleares	3,866 (3.79%)
	Top 5: number of tenders from Sevilla	3,526 (3.45%)

Figure 3 shows the frequency histogram of the number of tenders won by the same company. The reader must not confuse this histogram with the number of tenders by received offers (bidders) which is described in Table 3. The most frequent number of tenders won by the same company is 1. This means that about 10,000 companies have won only 1 tender. It is more or less 47% of the total number of winning companies. About 3,800 companies (18%) have won 2 tenders and so on (the trend is decreasing). Therefore, only 53% of companies have won 2 or more tenders. This distribution is important for the bidders recommender. It is more difficult to forecast the winning company successfully if a lot of companies have won only 1 tender because there are no patterns, trends, or relationships between tenders.

Figure 4 shows the relationship between the received offers of bidders for each tender and the underbid (also called discount). Actually, the underbid is the evaluation metric called MdAPE (median absolute percentage error) between

Table 4: Quantitative description of the company dataset.

Topic	Description	Value
General values	Total number of companies in the dataset	1,353,213
	Total number of opened companies (actives)	831,356 (61.44%)
	Total number of closed companies (inactives)	521,857 (38.56%)
	Temporal range of the opened companies' establishment date	1842/03/17–2019/03/25
	Mean of the opened companies' establishment date (seniority date)	2002/12/18
	Mean employees of opened companies (actives)	11.51
	Total number of the opened companies of legal entity type: private company limited by shares (Ltd.) (SL in Spanish)	761,358 (91.58%)
	Total number of the opened companies of legal entity type: public limited company (PLC) (SA in Spanish)	60,633 (7.29%)
Dataset's variables	CIF, Establishment_Date, Legal_Form, Last_Available_Year_Info, Status_Company, City_Company, Province_Company, Latitude_Company, Longitude_Company, Employees, Operating_Income, EBIT, EBITDA, CNAE_Primary, CNAE_Secondary, NACE2_Primary, NACE2_Secondary, IAE_Primary, IAE_Secondary, SIC_Primary, SIC_Secondary, NAICS_Primary, and NAICS_Secondary	23 variables (description in Table 2)
Operating income, EBIT, and EBITDA	Total number of opened companies with annual operating income available information (data from 2006 to 2018)	14,695 (2006); 22,080 (2007); 31,067; 38,120; 46,762; 85,210; 460,751; 589,239; 621,926; 659,266; 694,059; 648,598; 124,514 (2018)
	Total number of opened companies with annual EBIT available information (data from 2006 to 2018)	16,642 (2006); 24,618 (2007); 35,441; 41,558; 50,253; 89,890; 476,655; 608,397; 640,520; 677,366; 711,972; 663,761; 127,267 (2018);
	Total number of opened companies with annual EBITDA available information (data from 2006 to 2018)	16,654 (2006); 24,637 (2007); 35,452; 41,571; 50,266; 89,917; 476,719; 608,482; 640,623; 677,468; 712,085; 663,880; 127,295 (2018)
	Mean operating income of the year 2018	€4,122,727.11
	Median operating income of the year 2018	€299,130.00
	Mean EBIT of the year 2018	€397,964.64
	Median EBIT of the year 2018	€10,472.40
	Mean EBITDA of the year 2018	€542,772.79
	Median EBITDA of the year 2018	€18,733.35
Number of opened companies by location (province)	Top 1: number of opened companies from Madrid	157,705 (18.97%)
	Top 2: number of opened companies from Barcelona	114,207 (13.74%)
	Top 3: number of opened companies from Valencia	45,590 (5.48%)
	Top 4: number of opened companies from Alicante	33,386 (4.02%)
	Top 5: number of opened companies from Sevilla	26,368 (3.17%)

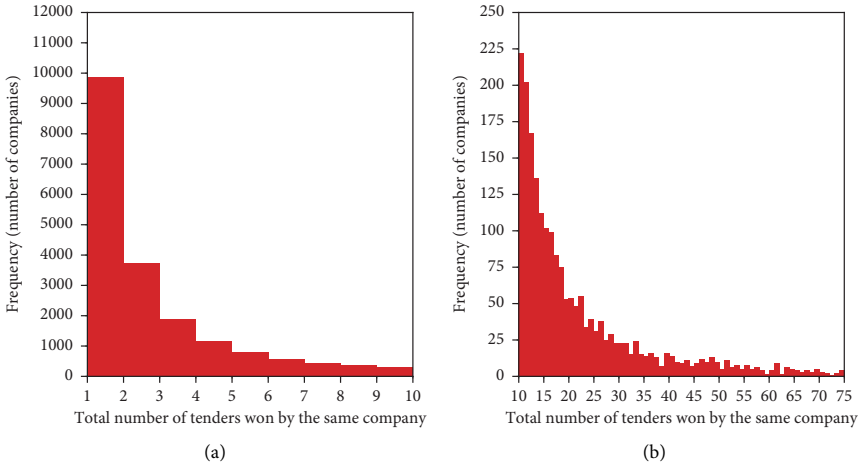


Figure 3: Histogram of frequency (number of companies) based on the total number of tenders in the dataset won by the same company (bidder). The graph is divided into two for better visualisation.

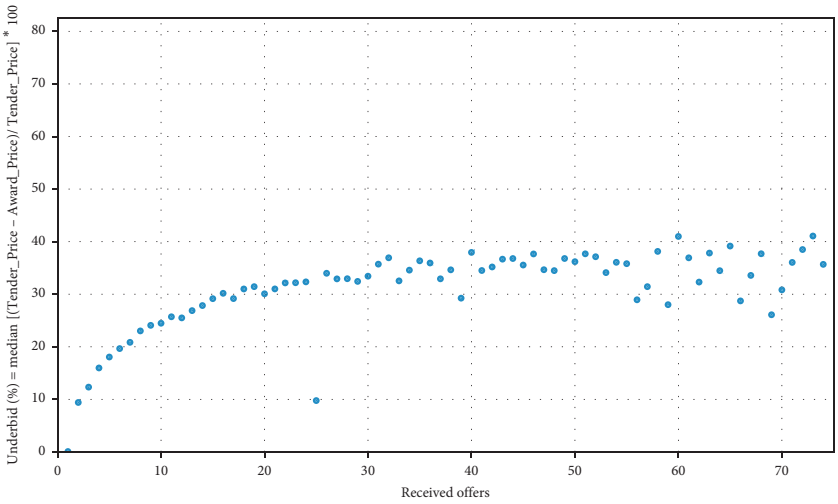


Figure 4: Relation between the received offers of bidders and the underbid (median absolute percentage error between tender price and award price).

the tender price and the award price, which is explained in Section 3.4. The trend is clear: the underbid increases until stabilising at around 35%. Hence, we have quantitatively demonstrated how the tenders with more bidders have lower award prices. In other words, the award price is lower in a tender with more competitiveness and the public procurement agencies will save money. So, the objective of the agencies should be to encourage the participation of companies to receive more offers. For this reason, the bidders recommender is a very useful tool for these agencies because they can effectively increase the number of participants in each tender.

To obtain new, relevant information through the variables in the merged dataset (the tender variables plus company variables), the Spearman correlation method was used. Figure 5 shows the Spearman correlation matrix (a symmetric matrix with respect to the diagonal). It is mathematically described in [10], and it is also used for the same purpose.

Looking at Figure 5, the most important correlations are the following:

(1) Tender_Price vs. Award_Price (0.97): this high correlation is in accordance with common sense since high bids are associated with high awards and low bids with low awards.

(2) Type_code vs. Subtype_code (0.77): each type of contract has its associated subtypes of contract.

(3) City_Tender vs. Province_Tender (0.43): the public procurement agency is in a city which belongs to a Province. So, the relationship city-province is always the same.

(4) Underbid vs. Received_Offers (0.54): the underbid (or discount) is the absolute percentage error (APE %) between Tender_Price and Award_Price. When the public procurement agency receives more offers from bidding companies, the underbid is bigger. This important correlation will be explained in detail in the following section.

(5) CPV vs. Duration (0.33): each type of work is usually associated with a temporal range (duration) for its realisation.

(6) CPV vs. CPV_Aggregated (0.99) has an obvious correlation: CPV_Aggregated is the first 2 digits of the CPV number.

(7) Latitude_Tender vs. Latitude_Company (0.57) and Longitude_Tender vs. Longitude_Company (0.55): this means that both locations (tender and company) are close and therefore the distance tender-company will be an input parameter for the bidders recommender.

(8) Employees, Operating_Income_LAY_-0, EBIT_LAY_-0, and EBITDA_LAY_-0 are strongly correlated with each other. Big companies have a lot of employees, and these companies can earn more profits.

4.3. Bidders Recommender Validation. There are two related validations: firstly, to validate the classification model (random forest) applied in phase 1 (train and forecast) of the bidders recommender and secondly, and more importantly, the validation of the bidders recommender results which is phase 4 (evaluation). This checks if the real winner company is within the recommended companies group.

For validating the classification model, Figure 6 shows three different ratios between the training and testing subsets (train : test in percentage) randomly chosen: 90 : 10, 80 : 20, and 70 : 30. Furthermore, it shows the behaviour of the error metrics (accuracy, precision, balanced accuracy, and OOB) for a different number of trees generated in the random

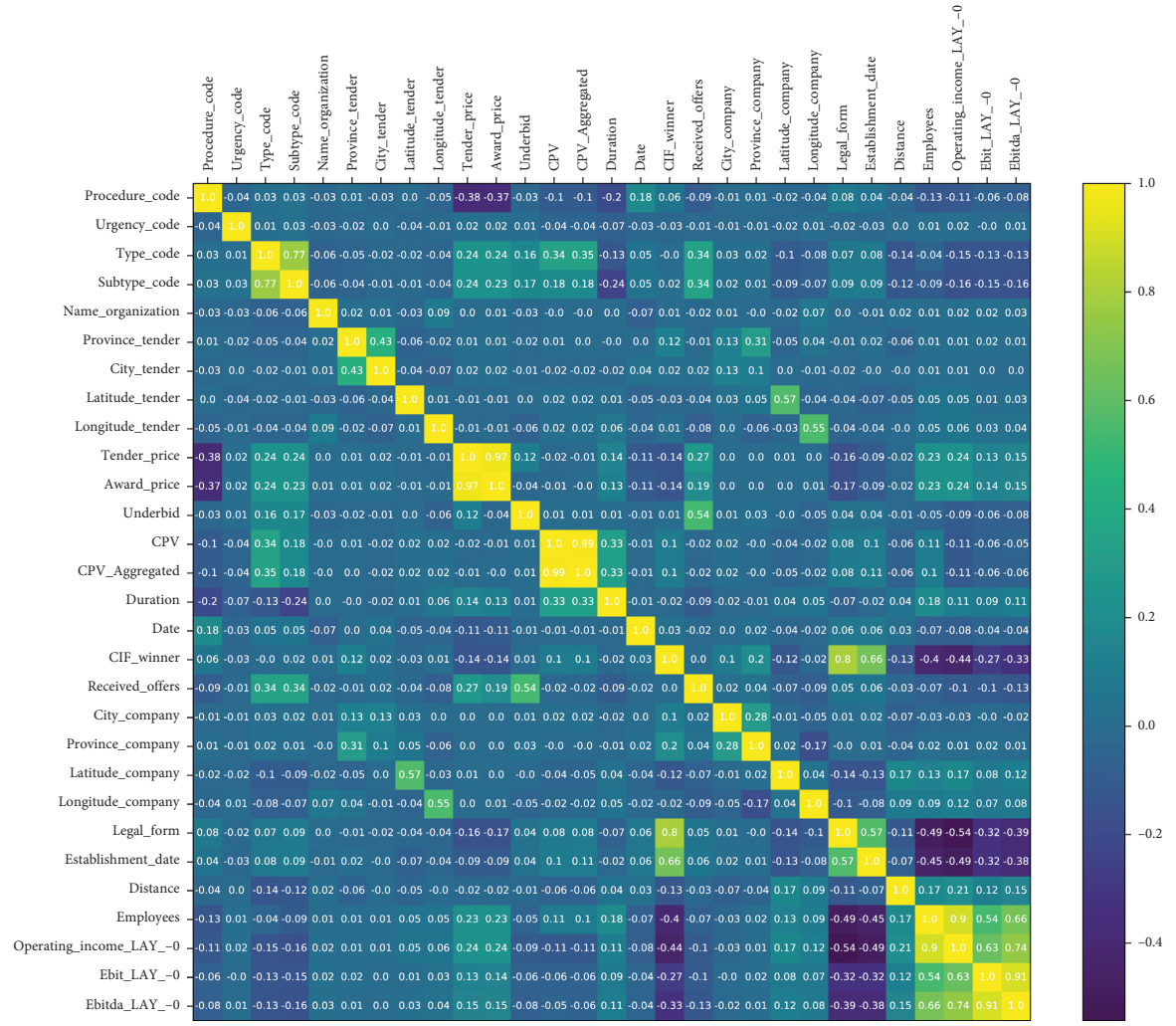


Figure 5: Correlation matrix between the variables of the two datasets (tenders and companies). Spearman's rank correlation coefficient is the method applied.

forest classifier. The accuracy$_{n\ 1}$ is the most important error for this study, and, in each graph, it is constantly of the order of 18%, 17%, and 15%, respectively. Logically, when decreasing the training data percentage, the accuracy is lower. Hence, the number of trees is not relevant and the election of the ratio also has a minimal impact. *RandomForestClassifier* from *Scikit-learn*, which is a machine learning library for the Python programming language, has 75 trees and a ratio of 80 : 20 and is the function used in this article.

Validation of the bidders recommender results was tested over two scenarios with five different setups. In the first scenario, the testing subset is 20% and is chosen randomly. In the second scenario, the dataset is ordered by tender date and the testing subset is the latest 20%, i.e., the most recent tenders. So, the second scenario is more appropriate to test a real engine search. Each scenario has the same five setups (filter settings), from very low (restrictive) filters to very high. The filters are described in detail in Section 3.5. Basically, there are six factors (F_{OI}, F_{EBIT}, F_{EBITDA}, F_E, F_{CEA}, and F_D), and it is necessary to assign numeric values. Hence, there are 10 combinations to test the bidders recommender.

The validation of the bidders recommender is shown in Table 5. The evaluation metric to measure the success of the recommender is the accuracy: the percentage of tenders where the winning company is within the recommended companies group. For scenario 1, when N 1 (it is predicted that the most probable company will win the tender), the accuracy is 17.07%. When N 5 (the 5 most probable companies to win the tender), the accuracy rises to 31.58%. Finally, the bidders recommender searches a group of compatible companies, automatically including the previous 5 companies, for each tender. The range of the accuracy is from 33.25% to 38.52% according to the settings applied. The

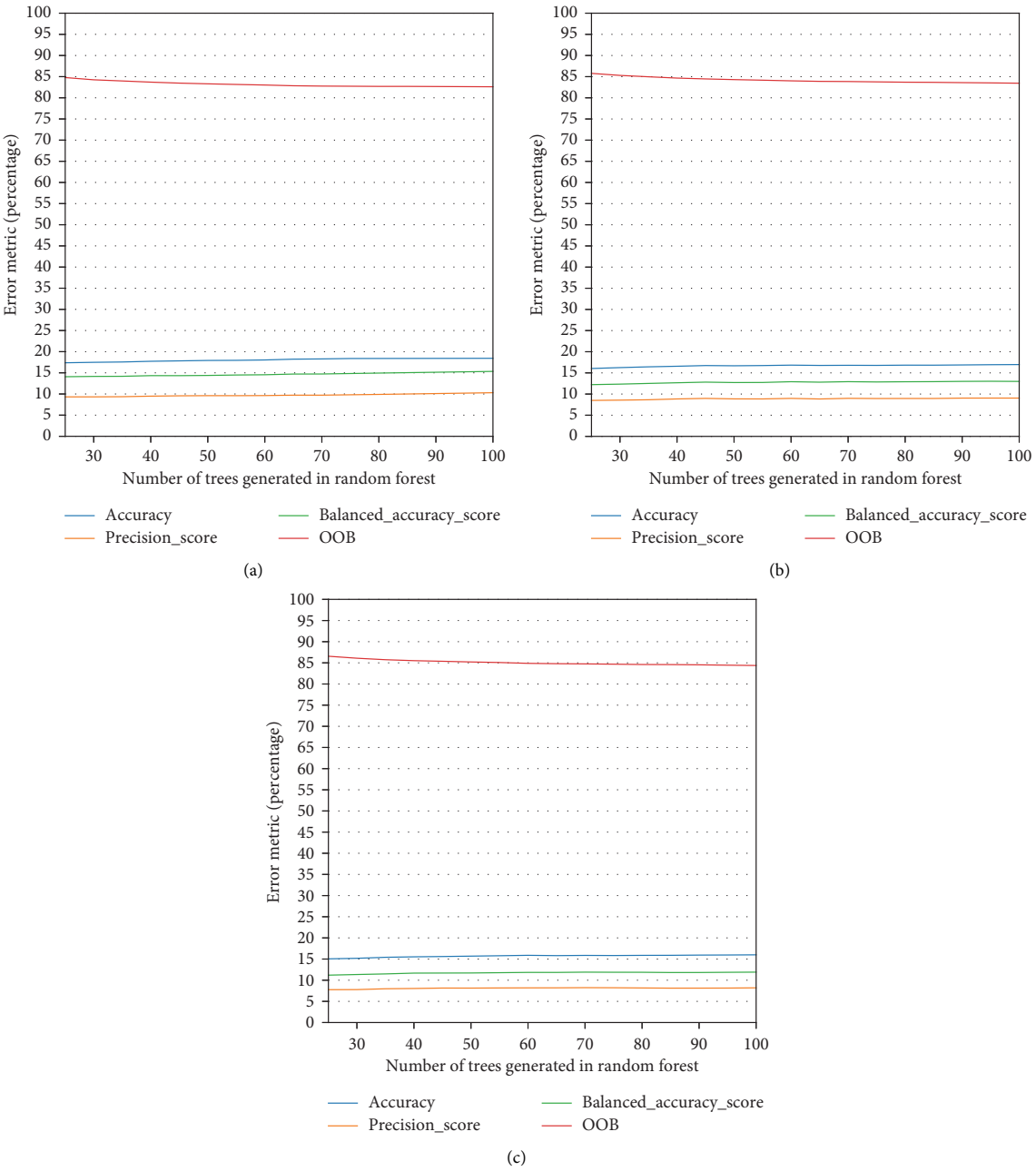


Figure 6: Relationship between trees in random forests and error metrics (accuracy, precision, balanced accuracy, and OOB) for different ratios of training and testing subsets. (a) 90 : 10. (b) 80 : 20. (c) 70 : 30.

reason to the increasing accuracy is simple: there are more recommended companies. Consequently, the mean (and median) number of recommended companies is higher.

Analogously for scenario 2, $\text{Accuracy}_{n\ 1}$ 10 25%, $\text{Accuracy}_{n\ 5}$ 23 12%, and $\text{Accuracy}_{n\ M}$ [24 79% − 30 52%]. This accuracy is significantly lower than that in scenario 1, and it could be for multiple reasons. For example, recent tenders have less business information because the annual accounts of the winner company are published the following year. In particular, the company dataset does not have information about operating income, EBIT, and EBITDA in 2019 and 2020 (see Table 4). However, there are a lot of tenders in 2019 and 2020 (see Table 3).

One area of interesting analysis is the size of the companies group generated by the bidders recommender. This recommender will be more efficient if the group is small and

Table 5: Testing the bidders recommender for two scenarios: results of the accuracy and number of recommended companies per tender for five different setups.

Description		Different bidders recommender settings				
		Very low	Low	Medium	High	Very high
Bidders recommender factors for the settings	F_{OI}: operating income factor	0.25	0.5	0.65	0.75	1.0
	F_{EBIT}: EBIT factor	0.25	0.5	0.65	0.75	1.0
	F_{EBITDA}: EBITDA factor	0.25	0.5	0.65	0.75	1.0
	F_E: employees factor	0.15	0.25	0.25	0.35	0.45
	F_{CEA}: classification economic activities factor	0.125	0.15	0.14	0.175	0.2
	F_D: distance tender-company factor	1.6	1.4	1.4	1.2	1
Results of scenario 1: testing subset is the 20% of the dataset randomly chosen	$\text{Accuracy}_{n\ 1}$: winner company is the forecast company			17.07%		
	$\text{Accuracy}_{n\ 5}$: winner company is within the top 5 forecast companies			31.58%		
	$\text{Accuracy}_{n\ M}$: winner company is within the recommended companies group	38.52%	36.20%	35.92%	34.04%	33.25%
	Mean and median number of the recommended companies of each tender	877.43; 86	469.69; 35	430.48; 31	226.07; 11	145.97; 9
Results of scenario 2: testing subset is the last 20% of the dataset ordered by tender's date	$\text{Accuracy}_{n\ 1}$: winner company is the forecast company			10.25%		
	$\text{Accuracy}_{n\ 5}$: winner company is within the top 5 forecast companies			23.12%		
	$\text{Accuracy}_{n\ M}$: winner company is within the recommended companies group	30.52%	28.00%	27.73%	25.55%	24.79%
	Mean and median number of the recommended companies of each tender	900.64; 95	470.41; 37	430.33; 33	210.92; 11	132.10; 9

the accuracy is high. Figure 7 shows the boxplots, disaggregated by CPV, for scenarios 1 and 2 (medium setup). CPV is the system for classifying the type of work in public contracts. The total mean is very similar in both scenarios: 430.48 potential bidders (median is 31) and 430.33 potential bidders (median is 33), respectively. The median value, disaggregated by CPV, is usually below 50 companies. However, the mean value of each CPV has great variability.

5. Discussion

The main objective is to find out and recommend companies for a new tender announcement. However, it is not easy to measure the performance of the bidders recommender; each company is unique and different from the rest, so the searching, comparison, and recommendation of companies is relative (subjective evaluation). Accuracy has been selected as the evaluation metric to measure the performance: the percentage of tenders where the winning company is within the recommended companies group.

Table 5 shows the results of the bidders recommender: the accuracy, mean, and median number of recommended companies over two scenarios with five different set ups (very low, low, medium, high, and very high). The main determining factor to get a good performance is due to the top 5 forecast companies (called $\text{Accuracy}_{n\ 5}$). This means that the 5 most probable companies to win a tender can be incorporated to the recommender companies group (called $\text{Accuracy}_{n\ M}$). For scenario 1, $\text{Accuracy}_{n\ 5}$ 31 58% and $\text{Accuracy}_{n\ M}$ [33 25% − 38 52%]. For scenario 2, $\text{Accuracy}_{n\ 5}$ 23 12% and $\text{Accuracy}_{n\ M}$ [24 79% − 30 52%]. The range is governed by the bidders recommender settings. Hence, the user can configure the factors for the settings (F_{OI}, F_{EBIT}, F_{EBITDA}, F_E, F_{CEA}, and F_D) to search more or less companies.

Figure 7 shows the boxplots for the size of the recommended companies group, disaggregated by the type of tender's work (CPV). There are considerable differences in the size, mean, and median values for each CPV. Other interesting analyses would be to disaggregate by geographic regions, business sectors, or markets.

As seen in this article, the bidders recommender depends strongly on the fields of public procurement announcements and the information available to characterise the bidders. Therefore, the recommender cannot be the same for each country since their public procurement systems are not unified or standardised for several reasons: regulations, laws, diverse information systems, different tender criteria, distinct levels of technological maturity in public administration, etc. However, this paper establishes the basis to create a bidders recommender which can be adapted to each country according to the two basic data sources: tender information and company information. This is because the recommender is an open frame which can easily add or modify other

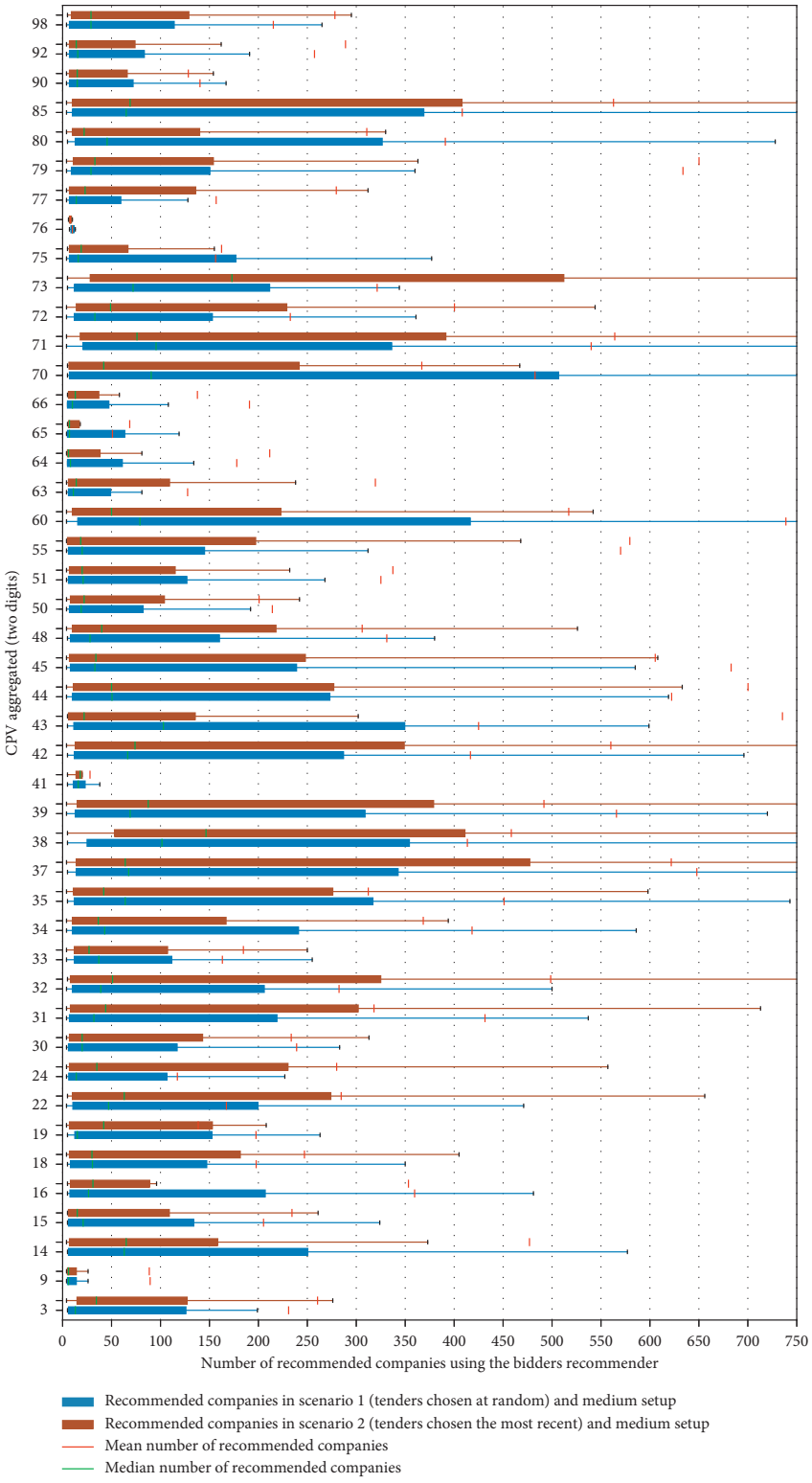


Figure 7: Boxplots for the size of the recommended companies group generated by the bidders recommender, disaggregated by CPV. Scenario 1 (blue colour) and scenario 2 (brown colour) both have a medium setup.

available fields or data sources. The selection and optimisation of the recommender's parameters can significantly improve it. It is a laborious task and particular to each country.

In summary, the recommender is an effective tool for society because it enables and increases the bidders participation in tenders with less effort and resources. Furthermore, this will serve to modernise the public procurement systems with a new approach based on machine learning methods and data analysis. Thus, the beneficiaries are the government, the citizens, and the two main users:

(1) *Public Contracting Agencies*. When they publish a tender notice, the algorithm automatically recommends suppliers which have a suitable profile for the tender. The agencies could contact these suppliers directly and invite them to participate if they are really interested in the tender.

(2) *Potential Bidders*. They will be able to search suitable tenders effortlessly, according to the type of tender and the profile of previous winning companies.

6. Conclusions and Future Research

The public procurement systems of many countries continue to use the inefficient mechanisms and tools of the 20th century for the publication of tenders and the attraction the offers and bidders. However, more and more new technologies (open data, big data, machine learning, etc.) are emerging in the public administration sector to improve their systems, proceedings, and services. This article clearly demonstrates how it is possible to create new tools using these technologies.

Especially, this paper develops a pioneering algorithm to recommend potential bidders. It is a multidisciplinary system which fills a gap in the literature. The bidders recommender proposed here is a promising and strategic instrument for improving the efficiency of public procurement agencies and should also facilitate access to the tenders for the suppliers. The recommender brings a trendy new perspective to gathering tenders and bidders.

The bidders recommender is described theoretically and also validated experimentally, using a case study from Spain. Two datasets have been used: tender dataset (102,087 Spanish tenders from 2014 to 2020) and company dataset (1,353,213 Spanish companies). The company dataset is difficult to collect because it is nonfree public information in Spain, so it is a valuable dataset. Quantitative, graphical, and statistical descriptions of both datasets have been presented.

The results of the case study have been successful because of the accuracy; it means that the winning bidding company is within the recommended companies group (from 24% to 38% of the tenders). The accuracy range is due to the two test scenarios (either being chosen from the most recent tenders or chosen at random), and each scenario has five different settings for the bidders recommender. Hence, the recommender has been validated for over 10 combinations of testing and the results are quite successful and promising, opening the research up to other countries and datasets.

The main limitation of this research is inherent to the design of the recommender's algorithm because it necessarily assumes that winning companies will behave as they behaved in the past. Companies and the market are living entities which are continuously changing. On the other hand, only the identity of the winning company is known in the Spanish tender dataset, not the rest of the bidders. Moreover, the fields of the company's dataset are very limited. Therefore, there is little knowledge about the profile of other companies which applied for the tender. Maybe in other countries the rest of the bidders are known. It would be easy to adapt the bidder recommender to this more favourable situation.

This paper opens the door to future research for creating bidder recommendation systems. In particular, for this recommender, some research can be done to improve it, as follows:

(i) The training and forecasting phase of the algorithm (step 1) to predict the winning company is based on the random forest classifier. Alternative methods of machine learning can be studied to increase the accuracy.

(ii) The aggregation phase (step 2) can use other fields of business information to create the profile of the winning company for the tender.

(iii) The searching phase (step 3) implements basic rules or filters to search similar companies. It would be interesting to explore more sophisticated methods, for example: clustering to group similar companies.

(iv) There is no ranking of recommended companies. This means that the algorithm only recommends companies without any associated probabilities, so the user cannot choose the companies that are most likely to be recommended to win the tender. This can be solved by applying a voting system or some kind of distance in the searching phase (step 3) of the algorithm.

Data Availability

The processed data used to support the findings of this study are available from the corresponding author upon request. The raw data from Spain are available at the Ministry of Finance, Spain (open data of Spanish tenders are hosted in http://www.hacienda.gob.es/es-ES/GobiernoAbierto/Datos%20Abiertos/Paginas/licitaciones_plataforma_contratacion.aspx).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors are grateful to Pablo Ballesteros-Pérez (PhD in Project Management and expert researcher in procurement auctions from the University of Cádiz (Spain)), for his very

valuable comments and suggestions to improve this article. This study was supported by the Plan of Science, Technology and Innovation of the Principality of Asturias (Ref: FC-GRUPIN-IDI/2018/000225).

References

[1] European Commission, "Public procurement," 2017.
[2] E. Huyer and L. van Knippenberg, "The economic impact of open data opportunities for value creation in europe," 2020.
[3] S. Curto, S. Ghislandi, K. Van de Vooren, S. Duranti, and L. Garattini, "Regional tenders on biosimilars in Italy: an empirical analysis of awarded prices," *Health Policy*, vol. 116, no. 2-3, pp. 182–187, 2014.
[4] T. Hanák and P. Muchová, "Impact of competition on prices in public sector procurement," *Procedia Computer Science*, vol. 64, pp. 729–735, 2015.
[5] J. Soudek and J. Skuhrovec, "Procurement procedure, competition and final unit price: the case of commodities," *Journal of Public Procedure*, vol. 16, no. 1, pp. 1–21, 2016.
[6] OECD Public Governance Reviews, *SMEs in Public Procurement: Practices and Strategies for Shared Benefits*, OECD Publishing, Paris, 2018.
[7] P. Ballesteros-Pérez, M. Skitmore, E. Pellicer, and J. H. Gutiérrez-Bahamondes, "Improving the estimation of probability of bidder participation in procurement auctions," *International Journal of Project Management*, vol. 34, no. 2, pp. 158–172, 2016.
[8] A. Mehrbod and A. Grilo, "Advanced Engineering Informatics Tender calls search using a procurement product named entity recogniser," *Advanced Engineering Informatics*, vol. 36, 2018.
[9] M. Nečaský, J. Klímek, J. Mynarz, T. Knap, V. Svátek, and J. Stárka, "Linked data support for filing public contracts," *Complexity*, vol. 65, no. 5, pp. 862–877, 2014.
[10] M. J. García Rodríguez, V. Rodríguez Montequín, F. Ortega Fernández, and J. M. Villanueva Balsera, "Public procurement announcements in spain: regulations, data analysis, and award price estimator using machine learning," *Complexity*, vol. 2019, 2019.
[11] M. J. García Rodríguez, V. R. Montequín, F. O. Fernández, and J. V. Balsera, "Spanish Public Procurement: legislation, open data source and extracting valuable information of procurement announcements," *Procedia Computer Science*, vol. 164, pp. 441–448, 2019.
[12] D. Corrales-Garay, M. Ortiz-de-Urbina-Criado, and E. M. Mora-Valentín, "Knowledge areas, themes and future research on open data: a co-word analysis," *Government Information Quarterly*, vol. 36, no. 1, pp. 77–87, 2018.
[13] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," *Government Information Quarterly*, vol. 32, no. 4, pp. 399–418, 2015.
[14] E. Afful-Dadzie and A. Afful-Dadzie, "Liberation of public data: exploring central themes in open government data and freedom of information research," *International Journal of Information Management*, vol. 37, no. 6, pp. 664–672, 2017.
[15] J. Lassinantti, A. Ståhlbröst, and M. Runardotter, "Relevant social groups for open data use and engagement," *Government Information Quarterly*, vol. 36, no. 1, pp. 98–111, 2018.
[16] F. Gonzalez-Zapata and R. Heeks, "The multiple meanings of open government data: understanding different stakeholders and their perspectives," *Government Information Quarterly*, vol. 32, no. 4, pp. 441–452, 2015.
[17] J. D. Twizeyimana and A. Andersson, "The public value of E-Government-a literature review," *Government Information Quarterly*, vol. 36, no. 2, pp. 167–178, 2019.
[18] F. Ahmadi Zeleti, A. Ojo, and E. Curry, "Exploring the economic value of open government data," *Government Information Quarterly*, vol. 33, no. 3, pp. 535–551, 2016.
[19] G. Magalhaes and C. Roseira, "Open government data and the private sector: an empirical view on business models and value creation," *Government Information Quarterly*, vol. 23, pp. 1–10, 2017.
[20] R. Krishnamurthy and Y. Awazu, "Liberating data for public value: the case of Data.gov," *International Journal of Information Management*, vol. 36, no. 4, pp. 668–672, 2016.
[21] S. Sadiq and M. Indulska, "Open data: quality over quantity," *International Journal of Information Management*, vol. 37, no. 3, pp. 150–154, 2017.
[22] S. Kubler, J. Robert, S. Neumaier, J. Umbrich, and Y. Le Traon, "Comparison of metadata quality in open data portals using the Analytic Hierarchy Process," *Government Information Quarterly*, vol. 35, no. 1, pp. 13–29, 2018.
[23] R. P. Lourenço, "An analysis of open government portals: a perspective of transparency for accountability," *Government Information Quarterly*, vol. 32, no. 3, pp. 323–332, 2015.
[24] N. Veljković, S. Bogdanović-Dinić, and L. Stoimenov, "Benchmarking open government: an open data perspective," *Government Information Quarterly*, vol. 31, no. 2, pp. 278–290, 2014.
[25] M. Lnenicka and J. Komarkova, "Big and open linked data analytics ecosystem: theoretical background and essential elements," *Government Information Quarterly*, vol. 36, no. 1, pp. 129–144, 2018.
[26] N. Obwegeser and S. D. Müller, "Innovation and public procurement: terminology, concepts, and applications," *Technovation*, vol. 74, 2018.
[27] P. Adjei-bamfo, T. Maloreh-nyamekye, and A. Ahenkan, "The role of e-government in sustainable public procurement in developing countries : a systematic literature review," *Government Information Quarterly*, vol. 142, 2018.
[28] S. Mullainathan and J. Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 87–106, 2017.
[29] H. R. Varian, "Big data: new tricks for econometrics," *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3–28, 2014.
[30] I. Lee and Y. J. Shin, "Machine learning for enterprises: applications, algorithm selection, and challenges," *Business Horizons*, vol. 63, no. 2, pp. 157–170, 2020.
[31] M. Bilal and L. O. Oyedele, "Big Data with deep learning for benchmarking profitability performance in project tendering," *Expert Systems with Applications*, vol. 147, 2020.
[32] J. J. Grandia, "Assessing the implementation of sustainable public procurement using quantitative text-analysis tools: a large-scale analysis of Belgian public procurement notices," *Journal of Purchasing and Supply Management*, vol. 19, 2020.
[33] M. A. Bergman and S. Lundberg, "Tender evaluation and supplier selection methods in public procurement," *Journal of Purchasing and Supply Management*, vol. 19, no. 2, pp. 73–83, 2013.
[34] P. Ballesteros-Pérez, M. C. González-Cruz, and A. Cañavate-Grimal, "On competitive bidding: scoring and position probability graphs," *International Journal of Project Management*, vol. 31, no. 3, pp. 434–448, 2013.
[35] M. Falagario, F. Sciancalepore, N. Costantino, and R. Pietroforte, "Using a DEA-cross efficiency approach in

public procurement tenders," *European Journal of Operational Research*, vol. 218, no. 2, pp. 523–529, 2012.
[36] M. Dotoli, N. Epicoco, and M. Falagario, "Multi-Criteria Decision Making techniques for the management of public procurement tenders: a case study," *European Journal of Operational Research*, vol. 88, 2020.
[37] Y. Wang, C. Xi, S. Zhang, D. Yu, W. Zhang, and Y. Li, "A combination of extended fuzzy AHP and Fuzzy GRA for government e-tendering in hybrid fuzzy environment," *European Journal of Operational Research*, vol. 2014, 2014.
[38] P. L. Lorentziadis, "Competitive bidding in asymmetric multidimensional public procurement," *European Journal of Operational Research*, vol. 282, no. 1, pp. 211–220, 2020.
[39] P. Ballesteros-Pérez and M. Skitmore, "On the distribution of bids for construction contract auctions," *Construction Management and Economics*, vol. 35, no. 3, pp. 106–121, 2017.
[40] P. Ballesteros-Pérez, M. L. del Campo-Hitschfeld, D. Mora-Melià, and D. Domínguez, "Modeling bidding competitiveness and position performance in multi-attribute construction auctions," *Operations Research Perspectives*, vol. 2, pp. 24–35, 2015.
[41] H. Jung, G. Kosmopoulou, C. Lamarche, and R. Sicotte, "Strategic bidding and contract renegotiation," *International Economic Review*, vol. 60, no. 2, pp. 801–820, 2019.
[42] A. Cheaitou, R. Larbi, and B. Al Housani, "Decision making framework for tender evaluation and contractor selection in public organisations with risk considerations," *International Economic Review*, vol. 68, 2019.
[43] J. Bochenek, "The contractor selection criteria in open and restricted procedures in public sector in selected EU countries," *Procedia Engineering*, vol. 85, pp. 69–74, 2014.
[44] T. Hanák and C. Serrat, "Analysis of construction auctions data in Slovak public procurement," *Advances in Civil Engineering*, vol. 2018, 2018.
[45] D. Imhof, *Empirical Methods for Detecting Bid-Rigging Cartels*, Université Bourgogne Franche-Comté, London, UK, 2018.
[46] P. Ballesteros-Pérez, M. C. González-Cruz, A. Cañavate-Grimal, and E. Pellicer, "Detecting abnormal and collusive bids in capped tendering," *Automation in Construction*, vol. 31, pp. 215–229, 2013.
[47] S. S. Padhi, S. M. Wagner, and P. K. J. Mohapatra, "Design of auction parameters to reduce the effect of collusion," *Decision Sciences*, vol. 47, no. 6, pp. 1016–1047, 2016.
[48] B. Tóth, M. Fazekas, and T. István János, "Toolkit for detecting collusive bidding in public procurement with examples from hungary," 2015.
[49] G. L. Albano, B. Cesi, and A. Iozzi, "Public procurement with unverifiable quality: the case for discriminatory competitive procedures," *Journal of Public Economics*, vol. 145, pp. 14–26, 2017.
[50] S. Tadelis, "Public procurement design: lessons from the private sector," *International Journal of Industrial Organization*, vol. 30, no. 3, pp. 297–302, 2012.
[51] K. Bloomfield, T. Williams, C. Bovis, and Y. Merali, "Systemic risk in major public contracts," *International Journal of Forecasting*, vol. 35, no. 2, pp. 667–676, 2019.
[52] G. Locatelli, G. Mariani, T. Sainati, and M. Greco, "Corruption in public projects and megaprojects: there is an elephant in the room!," *International Journal of Project Management*, vol. 35, no. 3, pp. 252–268, 2017.
[53] K. G. Dastidar and D. Mukherjee, "Corruption in delegated public procurement auctions," *European Journal of Political Economy*, vol. 35, pp. 122–127, 2014.
[54] A. Estache and R. Foucart, "The scope and limits of accounting and judicial courts intervention in inefficient public procurement," *European Journal of Political Economy*, vol. 157, 2018.
[55] Y. Huang, "An empirical study of scoring auctions and quality manipulation corruption," *European Economic Review*, vol. 120, 2019.
[56] P. Detkova, E. Podkolzina, and A. Tkachenko, "Corruption, centralization and competition: evidence from Russian public procurement," *International Journal of Public Administration*, vol. 41, no. 5-6, pp. 414–434, 2018.
[57] V. Titl and B. Geys, "Political donations and the allocation of public procurement contracts," *European Economic Review*, vol. 111, pp. 443–458, 2019.
[58] I. J. Tóth and M. Hajdu, "Cronyism in Hungary An empirical analysis of public tenders 2010-2016," 2018.
[59] OCDE, "Algorithms and collusion," 2017.
[60] M. Huber and D. Imhof, "Machine learning with screens for detecting bid-rigging cartels," *International Journal of Industrial Organization*, vol. 65, pp. 277–301, Jul. 2019.
[61] K. Rabuzin and N. Modrušan, *Prediction of Public Procurement Corruption Indices Using Machine Learning Methods*, Knowledge Engineering and Knowledge Management, New York, NY, USA, 2019.
[62] T. Sun and L. J. Sales, "Predicting public procurement irregularity: an application of neural networks," *Journal of Emerging Technologies in Accounting*, vol. 15, no. 1, pp. 141–154, 2018.
[63] P. Ballesteros-Pérez, M. C. González-Cruz, and A. Cañavate-Grimal, "Mathematical relationships between scoring parameters in capped tendering," *International Journal of Project Management*, vol. 30, no. 7, pp. 850–862, 2012.
[64] P. Ballesteros-Pérez, M. C. González-Cruz, M. Fernández-Diego, and E. Pellicer, "Estimating future bidding performance of competitor bidders in capped tenders," *Journal of Civil Engineering and Management*, vol. 20, no. 5, pp. 702–713, 2014.
[65] J.-S. Chou, C.-W. Lin, A.-D. Pham, and J.-Y. Shao, "Optimized artificial intelligence models for predicting project award price," *Automation in Construction*, vol. 54, pp. 106–115, 2015.
[66] J.-M. Kim and H. Jung, "Predicting bid prices by using machine learning methods," *Applied Economics*, vol. 51, no. 19, p. 2011, 2018.
[67] T. D. Fry, R. A. Leitch, P. R. Philipoom, and Y. Tian, "Empirical analysis of cost estimation accuracy in procurement auctions," *International Journal of Business and Management*, vol. 11, no. 3, p. 1, 2016.
[68] R. M. Skitmore and S. T. Ng, "Forecast models for actual construction time and cost," *International Journal of Business and Management*, vol. 38, no. 8, pp. 1075–1083, 2003.
[69] Official Website of the European Union, "European e-justice portal," 2003.
[70] D. Goens, "The exploitation of Business Register data from a public sector information and data protection perspective: a case study," *Computer Law & Security Review*, vol. 26, no. 4, pp. 398–405, 2010.
[71] R. Matin, C. Hansen, C. Hansen, and P. Mølgaard, "Predicting distresses using deep learning of text segments in annual reports," *Expert Systems with Applications*, vol. 132, pp. 199–208, 2019.
[72] S. Jones and T. Wang, "Predicting private company failure: a multi-class analysis," *Journal of International Financial Markets, Institutions and Money*, vol. 61, pp. 161–188, 2019.

[73] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968.
[74] S. Chava and R. A. Jarrow, *Bankruptcy Prediction with Industry Effects*, World Scientific, Berlin, Germany, 2008.
[75] D. Duffie, L. Saita, and K. Wang, "Multi-period corporate default prediction with stochastic covariates," *The Journal of Finance*, vol. 83, no. 3, pp. 635–665, 2007.
[76] E. Altman, G. Sabato, and N. Wilson, "The value of non-financial information in small and medium-sized enterprise risk management," *The Journal of Finance*, vol. 6, no. 2, pp. 1–33, 2010.
[77] Q. Yu, Y. Miche, E. Séverin, and A. Lendasse, "Bankruptcy prediction using Extreme Learning Machine and financial expertise," *Neurocomputing*, vol. 128, pp. 296–302, 2014.
[78] J. Min and Y. Lee, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters," *Expert Systems with Applications*, vol. 28, no. 4, pp. 603–614, 2005.
[79] H. Son, C. Hyun, D. Phan, and H. J. Hwang, "Data analytic approach for bankruptcy prediction," *The Journal of Finance*, vol. 28, 2019.
[80] C.-F. Tsai, "Combining cluster analysis with classifier ensembles to predict financial distress," *Information Fusion*, vol. 16, no. 1, pp. 46–58, 2014.
[81] The European Commission, "Regulation (EU) 2015/884 establishing technical specifications and procedures required for the system of interconnection of registers established by Directive 2009/101/EC," 2015.
[82] European Business Registry Association, https://ebra.be.
[83] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[84] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: a survey and results of new tests," *Pattern Recognition*, vol. 44, no. 2, pp. 330–349, 2011.
[85] M. R. Segal, "Machine learning benchmarks and random forest regression," *Pattern Recognition*, vol. 44, 2004.
[86] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande, "An extensive experimental survey of regression methods," *Neural Networks*, vol. 111, pp. 11–34, 2019.
[87] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.
[88] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, Berlin, Germany, 2008.
[89] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
[90] M. Fazekas, "Single bidding and non- competitive tendering procedures in EU co-funded projects," 2019, https://ec.europa.eu/regional_policy/en/information/publications/reports/2019/single-bidding-and-non-competitive-tendering.

Publicación D

Detección de colusión en licitaciones aplicando algoritmos de ML

Automation in Construction 133 (2022) 104047

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Review

Collusion detection in public procurement auctions with machine learning algorithms

Manuel J. García Rodríguez [a,*], Vicente Rodríguez-Montequín [a], Pablo Ballesteros-Pérez [b], Peter E.D. Love [c], Regis Signor [d]

[a] Project Engineering Area, University of Oviedo, 33012 Oviedo, Spain
[b] Departamento de Proyectos de Ingeniería, Universitat Politècnica de València, 46022, Valencia, Spain
[c] School of Civil and Mechanical Engineering, Curtin University, GPO Box U1987, Perth, Western Australia 6845, Australia
[d] Brazilian Federal Police, Rua Paschoal Apóstolo Pítsica, 4744 Florianópolis, Brazil

ARTICLE INFO

Keywords:
Auction
Collusion
Contracting
Construction
Machine learning
Procurement

ABSTRACT

Collusion is an illegal practice by which some competing companies secretly agree on the prices (bids) they will submit to a future auction. Worldwide, collusion is a pervasive phenomenon in public sector procurement. It undermines the benefits of a competitive marketplace and wastes taxpayers' money. More often than not, contracting authorities cannot identify non-competitive bids and frequently award contracts at higher prices than they would have in collusion's absence. This paper tests the accuracy of eleven Machine Learning (ML) algorithms for detecting collusion using collusive datasets obtained from Brazil, Italy, Japan, Switzerland and the United States. While the use of ML in public procurement remains largely unexplored, its potential use to identify collusion are promising. ML algorithms are quite information-intensive (they need a substantial number of historical auctions to be calibrated), but they are also highly flexible tools, producing reasonable detection rates even with a minimal amount of information.

1. Introduction

Public procurement is a common form of public spending whose purpose is to provide works, goods or services to a purchasing entity [1]. Within the context of procuring capital works, companies compete to be awarded a contract to build, improve or maintain a capital asset. Such contracts can vary in nature and may require the construction of new civil (e.g., roads and bridges) and social (e.g., schools and hospitals) infrastructures, the modification of existing assets or require maintenance [2].

Public procurement can be an intensive and complex process and thus can consume significant resources. For example, the European Union spends around 16% of its Gross Domestic Product on public procurement [3]. Collusion in these auctions (also called bid-rigging) refers to various illegal agreements among competing firms that aim to increase their profit margins. These collusive practices usually take the form of coordinated (non-competitive) price increases that are set between the companies (commonly referred to as *cartels*) [4]. Collusion is a recurring problem confronting the public sector, particularly when procuring capital works, with some being the most expensive items to be acquired [4]. Criminal investigations are regularly initiated to combat collusive activity, but being able to prosecute and obtain a conviction is challenging [5].

A major issue that stymies public institutions (e.g. contracting authorities, police bodies, competition commissions and courts of justice) from obtaining a conviction is detecting and proving that collusion has occurred [6]. However, the secrecy surrounding illegal agreements between firms tends to be underpinned by a carefully coordinated and sophisticated strategy, which is difficult to expose. In stark contrast, procurement authorities adhere to transparent and relatively stable purchasing patterns whereby they reuse awarding procedures, purchase standard products, resort to similar service specifications and the like. The predictability of such procurement practices can facilitate illicit market sharing and coordinated action among collusive firms [7–9].

; **ML**, Machine Learning; **PTE**, Pre Tender Estimate; **ABA**, Average Bid Auction; **SV**, Screening Variables; **CV**, Coefficient of variation; **SPD**, Spread; **DIFFP**, Difference between the two lowest bids; **RD**, Relative distance; **SKEW**, Skewness statistic; **KSTEST**, Kolmogorov–Smirnov test.

* Corresponding author.
E-mail addresses: manueljgarciar@gmail.com (M.J. García Rodríguez), montequi@uniovi.es (V. Rodríguez-Montequín), pabbalpe@dpi.upv.es (P. Ballesteros-Pérez), p.love@curtin.edu.au (P.E.D. Love), regis.rs@pf.gov.br (R. Signor).

https://doi.org/10.1016/j.autcon.2021.104047
Received 15 June 2021; Received in revised form 18 October 2021; Accepted 8 November 2021
Available online 18 November 2021

Against this contextual backdrop, it can be said that a reliable method for detecting the presence of collusion in public procurement auctions would significantly help procurement authorities and other institutions mitigate the adverse economic and social effects of collusion.

A plethora of models for detecting collusion has been propagated in the normative literature. In this paper, the most relevant models, which we will review, have proven to flag long term collusive patterns among bidding cartels [10]. They have also helped in discover how these cartels dissuade companies from submitting competitive bids in markets dominated by them [11,12]. However, while the models have been able to detect collusion, their accuracy often comes into question as to the data that underpins them can contain noise or insufficient detail. It is common, for example, for developed models to rely on information from the bidder's (private) costs structures and/or pre-tender cost estimates (PTE), though such information is generally confidential (and collusive firms are obviously not willing to share it) or simply does not exist [13,14].

Machine Learning (ML), a branch of artificial intelligence that focuses on building an application that can automatically learn and improve from experience, analyze and draw patterns of inference from auction information, even when it is scant (i.e., just the bid values and winning bidder from each auction) [15–17]. Yet, ML algorithms usually require a significant amount of reliable information obtained from previous auctions to calibrate them [16].

This paper aims to examine the ability of various ML algorithms to detect collusive auctions accurately. Each algorithm is tested under different conditions (e.g., with access to more or less information and with/without the input of Screening Variables, SV). We refer to SV as those statistical indices (directly calculated from the bid values) whose preprocessing may help ML algorithms to increase their level of detection [17].

To test the performance of ML algorithms, we will analyze six procurement datasets from five different countries (i.e., Brazil, Italy, Japan, Switzerland and the United States, US). Access to such auction data is generally unavailable to researchers as it is deemed sensitive (e.g., contract cost estimates) [18], but access and permission have been given for the collusion detection research presented in this paper. Thus, our research demonstrates that ML algorithms can detect collusion and produce representative performance results by applying them to a wide variety of datasets from different countries boasting different types of data. To the best of the authors' knowledge, this is the first time a transversal study of this nature has been undertaken in the domain of collusion detection.

The paper commences reviewing the literature and identifying the research gap to be examined (Section 2). Then, the procurement datasets, the screening variables, the ML algorithms being compared and the error metrics adopted are described (Section 3). We next summarize the major quantitative results of the experimental analysis for identifying collusive auctions (Section 4). This summary is followed by identifying the significance and contribution of our study (Section 5). Finally, we conclude this paper by explicitly identifying the limitations and avenues for future research (Section 6).

2. Literature review

Many studies in auction theory have proven that bidders' cost structures strongly condition their competitive and/or collusive strategies [1,10,19–22]. McAfee and McMillan [23] were the first to analyze collusion in static bid rotation schemes when no compensation payments existed between cartel members. In McAfee and McMillan's [23] auction model, the awardee is independent of previous (past) auctions. Building on the work of McAfee and McMillian [23], Aoyagi [24] and Skrzypacz and Hopenhayn [25] extended their model by considering repeated collusion in dynamic bid rotation schemes.

Studies have also analyzed collusion's occurrence and effect in real procurement auctions [26,27]. However, empirical-based collusion detection models are limited. One of the first attempts to develop an empirically-based model was Porter and Zona's [19], who sought to measure the probability of a bidder winning when some observable cost factors are known. However, that model aimed not to determine collusion, per se, but rather to anticipate the range of prices of future (competitive) bids. Other empirical-based models have been proposed since the propagation of Porter and Zona's [19] work. We will now summarize the four most relevant models in the remainder of this section.

The first seminal model in collusion detection is also known as *econometric screening* and was proposed by Bajari and Ye [28]. This model attempts to anticipate how a standard (competitive) distribution of bids should look based on the participating bidders' cost parameters. Unfortunately, these cost parameters constitute private data, which is generally difficult to gather and often disclosed by the bidders themselves. As a result, most data needs to be directly inferred by industry experts, resulting in a loss in accuracy. Bajari and Ye's [28] model does flag systematic deviations from a reference scenario. In this instance, the industry experts have to anticipate the reference scenario as they can assume the bidders submitting competitive bids want to be awarded the contract and will not cooperate with the cartel.

Bajari and Ye [29] model was initially tested in highway repair contract auctions in the US Midwest in 1994–1998. It was implemented as a functional reduced-form of linear regression where additional pieces of information such as bidders' past bidding history and pre-tender cost estimates (PTE) were needed (besides bidders' financial data). As a result of including this additional information, Bajari and Ye [28] could make valid comparisons with the reference scenario. However, Bajari and Ye's [28] model also has some important limitations:

- over-reliance on the functional form chosen when implementing the regression analysis;
- high sensitivity to missing information; and
- it is easy to cheat when the cartel knows 'how' it works (e.g., coordinated cover bids).

Considering the limitations above, the most important is the need for detailed data from each bidder and auction. The absence of such data precludes the model from being applicable in real bidding contexts. Fortunately, since Bajary and Ye's [28] study, more public data is available on public contracts and competitors, which can be used in the near future to improve collusion detection with ML.

The second model we examine is developed by Ballesteros-Pérez et al. [29], which focuses on analyzing possible abnormal dispersions in the distribution of bids, assuming they follow a Uniform distribution. In essence, the Ballesteros-Pérez et al. [29] model is an approximated collusion detection method used in conjunction with other approaches. It uses a simplified order statistics approach where the bids absolute order of magnitude is neglected and only the relative distances are considered. This approach, of course, leaves the possibility of cheating the method by submitting cover bids that 'emulate' a uniformly distributed pattern, no matter they are still abnormally high on average.

The third model has been proposed by Signor et al. [30], which is a *Probabilistic method* [2,34]. Signor et al.'s [30] model analyses submitted bids at two levels. Firstly, it analyses whether the bids overall distribution conform to a reference scenario (e.g., a Lognormal distribution). Additionally, the location of this distribution (i.e., absolute order of magnitude of the bids) can be closely approximated by historical auctions whenever data about their pre-tender estimates (PTE) is available. Hence, the model scrutinizes the distance of submitted bids from the PTE.

Secondly, Signor et al.'s [30] probabilistic method analyze the lowest bid's dispersion by drawing on order statistics theory. Put simply, it compares the probability of the lowest bid (i.e., the theoretical winner) being materialized as if it had been generated from the same reference distribution of the previous step. Hence, in Signor et al.'s [30]

method, the actual winning bid observed is compared against the lowest order statistic (i.e., the minimum draw of n artificially generated bids) from a calibrated reference distribution. If the statistical deviation is significant, we can be confident that such a bid is unlikely to be truly competitive. Thus, the probabilistic method is robust, but it has the limitation of being strongly dependent on the availability and reliability of a PTE for a number of previous honest auctions and the auction being tested.

Finally, the fourth model is that developed by Imhof [17,35]. This model has been the first to examine the application of ML to bidding and the detection of collusion by applying a small set of Screening Variables (SV) in a Swiss dataset of roads construction. We will use those SV and the same dataset in our study but assuming different levels of access to auction data. Additionally, Imhof [17,35] utilized two ML algorithm types: (1) the Lasso regression and an 'Ensemble method' consisting of a weighted average of several algorithms; and (2) bagged regression trees, random forests, and neural networks. In this research, we will consider a wider range of algorithmic options and various datasets to understand better the conditions leading to SV and ML algorithms performing better.

3. Materials and methods

This section describes the research methods adopted to detect collusion in auctions of public sector capital works. In Fig. 1, we present a summary of the research process used in this study.

3.1. Datasets

To assess the collusive detection capabilities of ML algorithms under different conditions (e.g., countries, types of auctions, time period, and the availability of data per auction), we acquired six public procurement datasets. These datasets are derived from five countries covering periods between 1980 and 2013.

All datasets can be found in the *Supplementary file* attached to this paper so that others can replicate our results. A quantitative description of the datasets is presented in Table 1. At this juncture, no study that has examined collusion has had access to such an extensive dataset, which enables the suitability of ML to be explored as a detection approach.

It is worth noting that all six datasets have been investigated and/or provided by public institutions [e.g. *Swiss Competition Commission* (COMCO), *Brazilian Federal Police*, *Japanese Fair-Trade Commission* (JFTC) and two courts of justice from the US and Italy]. Hence, we assume the data are reliable and trustworthy. While the datasets may contain minor contradictions, we are unable to judge the auctions' bidding consistency. Actually, the datasets' owners are also unable due to the secret nature of the agreements. For example, there are instances where an auction's winning bidder was classified as collusive while other higher (not awarded) bids were not. Clearly, in the context of capital works procurement, collusion generally involves being awarded contracts at a higher-than-usual price. In the example above, all bidders may have facilitated this outcome. However, we can only assume the awarded bidder was flagged with a consistent abnormal bidding pattern through a series of auctions. Thus, without criminal proof, other companion bidders might have avoided being flagged as collusive and consequently avoided conviction, or even being honest competitors unwittingly involved in a case of partial collusion.

Alternatively, these non-awarded bids may have been the result of estimation errors or were competitive bids with intentionally high mark-ups where evidence of coordinated action among bidders either did not exist or could not be determined. Coordinated action is a necessary condition for collusion to occur being the most difficult to prove. Despite

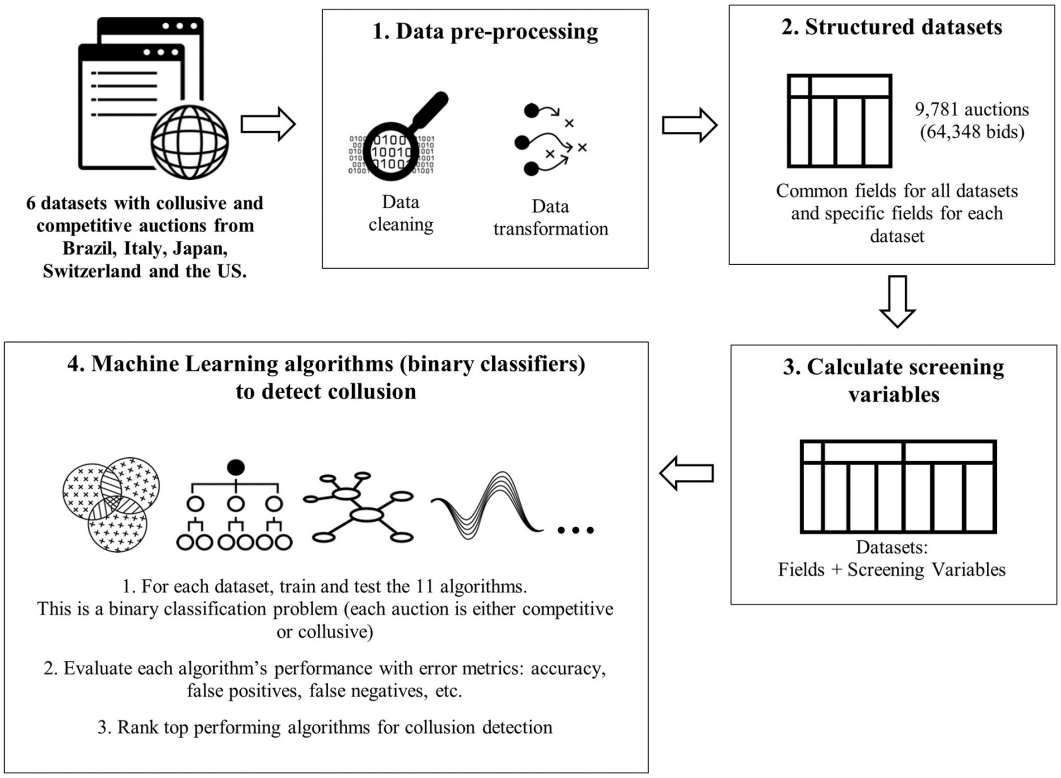


Fig. 1. Flowchart summarizing the research approach for collusion detection.

Table 1
Description of the collusive datasets.

Topic	Description	Brazil	Italy	Japan	Swiss–Ticino	Swiss–SG&GR	US
General information	**Scope**	Oil infrastructure projects	Road construction	Building constr. and civil eng.	Road construction	Road construction and civil engineering	School milk market
	Time period	2002–2013	2000–2003	2003–2007	1999–2006	14 years (over 2005)	1980–1990
	N° auctions	101	278	1080	224	4344	3754
	N° bids	683	20,286	13,515	1629	21,231	7004
	Awarding criteria	Lowest bid	Average Bid Method	Lowest bid	Lowest bid	Lowest bid	Lowest bid
	Avg. n° of bids per auction	6.76	72.97	12.51	7.27	4.89	1.91
Available information per dataset	**Common fields**	Auction code, bid values, winning bidder and number of bids per auction					
	Auction date	Yes	N/A	Yes	N/A	Yes	Yes
	Pre Tender Estimate (PTE)	Yes	Yes	Yes	N/A	N/A	N/A
	Identity of bidders	Yes. 272	Yes. 821	Yes. 1665	N/A	N/A	Yes. 120
	N° of different awardees	80 (29.41%)	19 (2.31%)	690 (41.44%)	N/A	N/A	91 (75.83%)
	Other fields (additional information)	Location and Brazilian State	Location, legal company type and economic size	Location	Consortium composition	Contract type	Inflation adjusted bid and inflation raw milk price adjusted bid
Collusive vs competitive data	**Collusive auctions**	N/A	N/A	N/A	184 (82.14%)	N/A	N/A
	Competitive auctions	N/A	N/A	N/A	40 (17.86%)	N/A	N/A
	Collusive bids	128 (18.74%)	8085 (39.86%)	1093 (8.09%)	1332 (81.77%)	12,501 (58.88%)	866 (12.36%)
	Competitive bids	555 (81.26%)	12,201 (60.14%)	12,422 (91.91%)	297 (18.23%)	8730 (41.12%)	6138 (87.64%)
	Collusive bidders	47 (17.28%)	195 (23.75%)	230 (13.81%)	N/A	N/A	11 (9.17%)
	Competitive bidders	225 (82.72%)	626 (76.25%)	1435 (86.19%)	N/A	N/A	109 (90.83%)
Bids per auction	**1 ≤ bids ≤ 4**	42 (41.58%)	0	0	29 (12.95%)	2315 (53.29%)	3727 (99.28%)
	5 ≤ bids ≤ 10	38 (37.62%)	5 (1.80%)	474 (43.89%)	171 (76.34%)	1897 (43.67%)	27 (0.72%)
	11 ≥ bids	21 (20.79%)	273 (98.20%)	606 (56.11%)	24 (10.71%)	132 (3.04%)	0
Awarding price	**Aggregated total**	€12,170,309,780	€11,520,750,772	€402,195,427	€514,972,754	€2,136,031,656	N/A (Bid values are unit price per half a pint of milk)
	Aggregated collusive	€7,918,003,543 (65.06%)	€7,911,773,729 (68.67%)	€91,405,888 (22.73%)	€458,103,059 (88.96%)	€908,666,894 (42.54%)	
	Aggregated competitive	€4,252,306,237 (34.94%)	€3,608,977,044 (31.33%)	€310,789,539 (77.27%)	€56,869,695 (11.04%)	€1,227,364,760 (57.46%)	

Note: datasets used in this paper, apart from the Italian dataset, adopt the lowest bid wins awarding criterion.

some minor inconsistencies with the data, all auctions are treated being uniform in our study. Indeed, due to differing formats for collecting data the ability to ensure its calibration poses a challenge. However, it needs to be acknowledged this is the most comprehensive study undertaken to date that examines the detection of collusion in real-life auctions. We now proceed to briefly describe the datasets, whose main features are summarized in Table 1.

3.1.1. Brazil

Between 2002 and 2013, the Brazilian Oil Company Petrobras (a publicly traded, State-controlled company) was subjected to significant bid-rigging during the procurement of infrastructure projects. The dataset has been previously analyzed and made available by Signor et al. [18,30,33,34]. In 2014, a routine investigation by the *Brazilian Federal Police* into money laundering quickly turned into a very important anticorruption operation called "Operation Car Wash". Signor et al.'s [18,30,33,34] dataset form part of an ongoing investigation where several collusive companies confessed to price-fixing and bid-rigging. It was shown that 16 of the largest Brazilian construction companies (a cartel referred to as the "*Club of 16*") colluded in many of Petrobras's auctions.

3.1.2. Italy

The Italian dataset comprises road construction auctions from the municipality of Turin [36]. The legal office of Turin collected the dataset as part of a legal case against several firms accused of bid-rigging between 2000 and 2003. This dataset employs the Average Bid Auction (ABA) method: the awardee is the bid closest to a trimmed average [36]. The ABA can be used to create incentives to coordinate bids among bidders with the intention of manipulating the bids distribution. In 2008, the *Court of Justice of Turin* convicted 95 construction firms that operated in eight cartels that had been successfully awarded contracts (<10% of the firms won >80% of the auctions).

3.1.3. Japan

The Japanese dataset comprises building construction and civil engineering contracts from Okinawa. Initially, the data was published in Ishii [37], and it was later analyzed in Imhof [38]. The dataset was obtained from the Okinawa Prefectural Government (OPG), covering the period between 2003 and 2007. The construction market in Okinawa exhibits several features facilitating collusion: (1) geographic conditions (islands); (2) restricted invitation procedure (the buyer chooses those companies allowed to bid); and (3) contracts and bidders segmented into ranks. In June 2005, the *Japanese Fair-Trade Commission* (JFTC) filed a bid-rigging investigation against many firms involved in the auctioning process. The dataset covers three periods:

1. *Pre-inspection period*: auctions before the opening of the JFTC investigation (June 2005). These auctions can be collusive or competitive, according to JFTC resolutions.
2. *Post-inspection period*: auctions between the opening of the JFTC investigation (June 2005) and the amendment of Japanese competition laws in January 2006. These auctions are not used in our analysis as it was a transition period without information from the JFTC.
3. *Post-amendment period*: auctions after the amendment of Japanese competition laws. The JFTC sentenced and sanctioned the involved cartel participants at the beginning of the post-amendment period in March 2006. Therefore, all these auctions can be considered competitive as there has not been any proof of collusion ever since.

3.1.4. Swiss – Ticino

The Swiss dataset comprises road construction projects from the Canton of Ticino in Switzerland [35,39,40]. The cartel operating in this area of Switzerland had existed since the 50s, but it was not until the mid-90s that collusion became more frequent. By then, competition pressure within cartel companies started to grow, reaching its peak in 1998. This motivated cartel members to reach a tacit agreement in 1998 to which they adhered until 2005. During this period, all cartel firms in the road construction sector rigged nearly all procurement contracts. Therefore, this is undoubtedly one of the most severe bid-rigging cartels. As a result, local politicians went to the *Swiss Competition Commission* (COMCO) to investigate how awarding prices were exaggeratedly high in Ticino compared to other country regions.

3.1.5. Swiss – St Gallen and Graubünden

The next Swiss dataset covers the period between 2004 and 2010. It comprises the operations of two cartels specialized in road construction, asphalt paving, and civil engineering works in the Swiss cantons of St. Gallen and Graubünden [40]. In the first canton, eight firms participated in bid-rigging conspiracies. They met once or twice per month until 2009, when the COMCO launched house searches in the neighbor canton. In the second canton, another cartel was made up of a local trade association for road construction and asphalt paving operated until 2010. Both cartels were well organized and were awarded a very large share of auctions. As a result, the COMCO opened an investigation after the statistical anomalies identified in the procurement data until 2010.

3.1.6. United States

The US dataset was published in Porter and Zona [19] and also used in the study of Wachs and Kertész [41]. The dataset involves school milk procurement contracts in the State of Ohio between 1980 and 1990. School district officials independently solicited bids on annual supply contracts for milk and other products to regional milk producers (dairies). Typically, the lowest bidder was selected to supply milk in half pints to the schools during the following school year. In 1993 representatives of two dairies in Ohio confessed having bid-rigged these auctions during the 1980s. Thus, all bidding data were collected by the *United States District Court of Ohio* in 1994, and 30 dairies were charged with collusion. After careful analysis of these auctions, it was concluded that the estimated average effect of collusion on this market resulted in a 6.5% price increase. The dataset is non-construction-related, but it is useful to analyze it as it serves as a frame of reference to better understanding bidding behaviors and patterns in other markets.

3.2. Screening variables

Screening Variables, or just *Screens*, are specific indices derived from each auction's bid values distribution (prices offered by bidders). These screens can help ML algorithms process auction information more efficiently to detect collusion [17]. However, there have been limited studies that have investigated the performance of different screens in collusive datasets.

Screens can be useful, not just for flagging possible collusion in a given auction but also for identifying sustained collusive patterns among specific bidders. Screens frequently consist of statistical indices calculated directly from the bid values of each auction (e.g. the bids standard deviation, skewness or kurtosis) or after removing or selecting some of the bids (e.g. the lowest and highest bid in an auction, or the lowest and second-lowest). They are generally easy to calculate and have proven to produce higher performance in ML algorithms. As a result, screens are usually beneficial when combined with ML algorithms and in our case, for detecting abnormally high bids.

The process to create a screen commences by letting t be the *t-th* auction in a dataset. We will not use an additional subscript to refer to each of the six datasets for the sake of clarity. Let sd_t be the (economic) bids standard deviation in auction t; $\overline{b}_t$ the mean (average) of all bids submitted to auction t; $b_{max,\,t}$ the maximum (most expensive) bid; $b_{min,\,t}$ the minimum (lowest, cheapest) bid; b_{2t} is the second-lowest bid; $sd_{losingbids,\,t}$ is the standard deviation of the non-awarded bids (all but the winning bid); n_t is the number of bids submitted to auction t; and b_{it} is the *i-th* bid in auction t when ordered from lowest to highest. With this notation, the following screens are initially proposed to detect collusion better:

$$CV_t = \frac{sd_t}{\overline{b}_t} \tag{1}$$

$$SPD_t = \frac{b_{max,t} - b_{min,t}}{b_{min,t}} \tag{2}$$

$$DIFFP_t = \frac{b_{2t} - b_{min,t}}{b_{min,t}} \tag{3}$$

$$RD_t = \frac{b_{2t} - b_{min,t}}{sd_{losingbids,t}} \tag{4}$$

$$SKEW_t = \frac{n_t}{(n_t-1)(n_t-2)} \sum_{i=1}^{n_t} \left(\frac{b_{it} - \overline{b}_t}{sd_t} \right)^3 \tag{5}$$

$$KURT_t = \frac{n_t(n_t+1)}{(n_t-1)(n_t-2)(n_t-3)} \sum_{i=1}^{n_t} \left(\frac{b_{it} - \overline{b}_t}{sd_t} \right)^4 - \frac{3(n_t-1)^3}{(n_t-2)(n_t-3)} \tag{6}$$

$$KSTEST_t = max\left(D_t^+, D_t^-\right) with\ D_t^+ = max_i \left(\frac{b_{it}}{sd_t} - \frac{i_t}{n_t+1} \right), D_t^- = max_i \left(\frac{i_t}{n_t+1} - \frac{b_{it}}{sd_t} \right) \tag{7}$$

All previous screening variables have been proposed by different researchers in the context of collusion detection (e.g. [35,38–40,42]). The first screen is the *Coefficient of Variation* called CV_t (Eq. (1)), a scale-invariant statistic calculated as the ratio of the bids' standard deviation divided by the average of the bids. The second screen is the *Spread* (SPD_t) represented in Eq. (2). Eq. (3) measures the relative *difference between the two lowest bids* in the auction ($DIFFP_t$). An alternative screen to the latter is the *Relative Distance* (RD_t) which replaces the term in the denominator by the losing bids standard deviation (Eq. (4)). Finally, the last three screens refer to the bid values' *Skewness* ($SKEW_t$), *Excess Kurtosis* ($KURT_t$) and *Kolmogorov-Smirnov test* ($KSTEST_t$). These three screens allow identifying possible bid distribution asymmetries (Eq. (5)), the condensation of bid values next to (or too far from) the average of the bids (Eq. (6)), and the similarity of the bid values for a uniform distribution (Eq. (7)), respectively. As the Excess Kurtosis requires at least four bids per auction to its calculation and our datasets contain a significant number of auctions with less than four bids (see Table 1), this screen will not be adopted in our study.

Other screening variables could have also been proposed, but a detailed exploration of their potential use remains outside the scope of

this investigation. The ones used are the most common in other ML applications that work with statistically distributed values. Of note, it has been observed that the statistical distribution of bids is expected to become explicit when taking the log bids instead of their natural values (i.e., a lognormal distribution) [43,44]. In our experiments, we also tested the performance of these screens with log bids besides natural bid values. However, we found no improvement in the algorithms detection rates. Thus, a bids log transformation is not to be considered in this paper.

The *Scatter matrix* of the screening variables above (Eqs. (1) to (7)) for all the datasets (64,348 bids in total) is shown in Fig. 2. This matrix is frequently generated in ML applications to identify correlations between the screening variables. It is also useful for detecting the screens that differentiate between competitive and collusive bids. However, we can see from Fig. 2 that it does not show any distinct relationship between the space dispersion of competitive (green dots) versus collusive bids (red dots). That is, we cannot find separated clusters of red versus green dots in any subgraph of Fig. 2. This finding indicates that we will need to

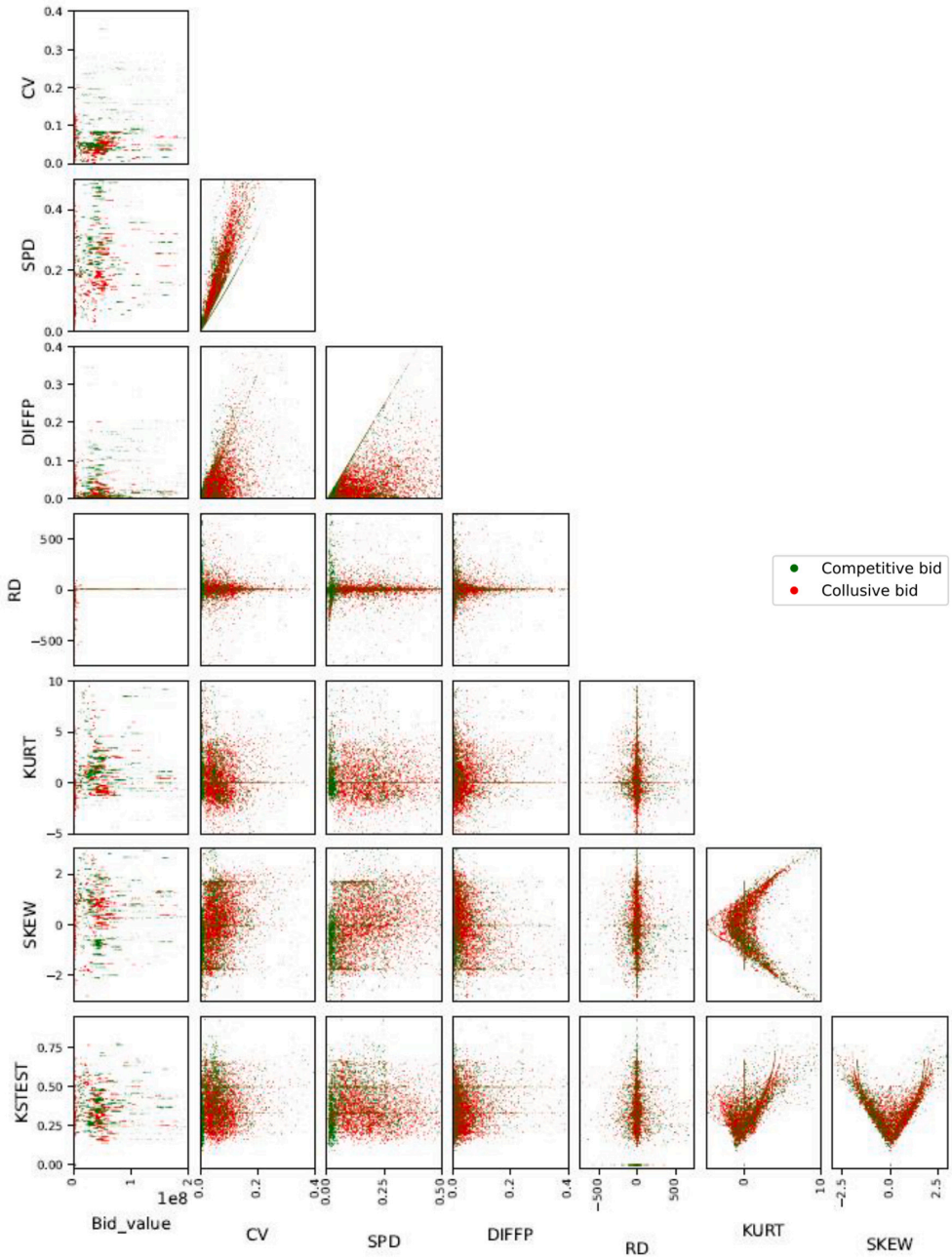


Fig. 2. Screening variables scatter matrix from all datasets.

rely on each algorithm's learning process (training) and performance (with and without the help of screens).

3.3. Machine learning algorithms settings

The collusion detection capability of 11 algorithms is tested in this paper under different scenarios of information availability. It is assumed that each auction could be classified as either 'collusive' or 'competitive'. Hence, the algorithms have to perform a binary classification for each auction *t*. The following algorithms are utilized to perform this task:

- Linear models: *SGD* (Stochastic Gradient Descent) [45];
- Ensemble methods: *Extra Trees* (Extremely Randomized Trees) [46], *Random Forest* [47], Ada Boost [48] and Gradient Boosting [49];
- Support Vector Machines: *SVC* (C-Support Vector Classification) [50];
- Nearest Neighbors: *K Neighbors* [51];
- Neural network models: *MLP* (Multi-Layer Perceptron) [52];
- Naive Bayes: *Bernoulli Naive Bayes* and *Gaussian Naive Bayes* [52]; and
- *Gaussian Process* [53].

Ensemble methods are the top-performing algorithms in our study as shown later. They combine several models (multiple learning algorithms) that produce a single optimal predictive model. This model is also generally more robust from the prediction point of view. Decision tree is usually one of those learning algorithms integrated in the Ensemble methods. This algorithm resembles a flowchart-like structure where each node implements a test on an attribute. Hence, each branch represents the outcome of a test, and each leaf node represents a class label. Two families of ensemble methods are usually distinguished:

- *Averaging methods;* they encompass several independent estimators and then average their predictions. On average, the combined estimator is usually better than any of the single base estimator because its variance is lower. Examples of averaging methods are *Extra Trees* and *Random Forest* algorithms.
- *Boosting methods;* their base estimators are implemented sequentially which reduces the bias of a combined estimator in some cases. Broadly speaking, the objective of Boosting methods is to combine several weak models to produce a single, more powerful model. *Ada Boost* and *Gradient Boosting* are some examples of Boosting methods.

These are common ML algorithms that have produced satisfactory results in many engineering applications, construction sector [54,55] and public procurement [56,57] included. All datasets and the algorithms' code can be found in the *Supplementary files* (csv format) we have provided. This will facilitate the future replicability of our results. The Python (3.0) programming language and the ML library *scikit-learn* have been used in this research [58]. Details about the eleven ML algorithms have not been provided but they are freely available from the *scikit-learn* library. However, we do provide some additional information at the end of this section about the numerical settings (parameter values) adopted for those algorithms that performed better. For those readers interested in extending their knowledge on the inner workings of each algorithm, we suggest resorting to the references provided in the list above and referring to the *Supplementary material* we have provided.

All the ML algorithms we have identified require calibration (training) before they are capable of differentiating collusive from competitive auctions. In conventional ML applications, training datasets typically comprise of thousands of entries. Algorithms generally use 80% of the data for training purposes and the remaining 20% to test their performance [59]. However, in our study and even though some of these datasets are large compared to most auction datasets reported in the construction bidding literature [60–62], many are too small to train all algorithms properly (i.e., they 'only' comprise 9781 auctions with 64,348 bids).

To avoid collusion detection results being biased by the particular choice of training and test subsets, we performed 500 iterations with each algorithm. Thus, for each algorithm and dataset, we tested their detection performance while changing the specific subset of auctions used for training and testing (random choices). Noteworthy, the bids of each auction were either all used for training or testing; that is, they were not split for different purposes. This avoids the transfer of knowledge (rendering collusion detection harder for the algorithms, as they cannot use the same auction ID to flag an auction as collusive later), but provides a realistic scenario (as the bids of the same auction are generally known at once, not in different stages). Hence, our algorithms classify an auction as collusive or competitive based on each of the specific bids it contains. Markedly, all bids from the same auction were used as a single group of analyses.

The performance of the algorithms was analyzed under four different settings (scenarios). Each setting represents access to different pieces of data per auction. We named these pieces of information as fields in Table 1. Naturally, a higher amount of data per auction should lead to better collusion detection results. However, in actual practice, some data is not always available. Yet, it is equally valuable for anticipating the detection rates of each algorithm in the absence of data. Hence, the algorithms were trained and tested individually for each dataset under the following settings:

- *Setting 1 (all fields).* In this scenario, the algorithms used all the available data with one exception: the bidders' identity (see Table 1 to identify the specific fields that were available in each dataset). The 'identity of bidders' was not used to avoid the potential risk of a bidder being easily catalogued upfront as collusive in the training process, and later classify as collusive almost all the auctions where it was involved (during the testing stage).
- *Setting 2 (all fields + screens).* Algorithms had the same data available as in setting 1 but with the assistance of the screening variables (CV, SPD, DIFFP, RD, SKEW and KSTEST). Theoretically, this should correspond to the scenario where ML algorithms perform better.
- *Setting 3 (common fields only).* In this scenario, the algorithms were only allowed to use the data shared among all datasets: that is, the auction code, bid values, winning bidder and number of bids per auction.
- *Setting 4 (common fields only + screens).* As in setting 2, this scenario assumed the data availability of setting 3 plus the aid of the screen variables described earlier.

Finally, we summarize the configuration adopted for the four ensemble methods as they were the top-performing algorithms in our study. A preliminary exploratory analysis was conducted to set the values of the algorithm parameters. Namely, we fine-tuned them based on data from related algorithm [35,38–40] and our first implementation results. With this, the best detection results were obtained for this parameters configuration:

- *Extra Trees and Random Forest*: The number of trees was 300; the function to measure the quality of a split was Gini; and the maximum depth of tree was until all leaves were pure or contained less than two samples.
- *Ada Boost*: The maximum number of estimators at which boosting terminated was 300. The base estimator was a decision tree classifier with 1 as the maximum depth of the tree with a learning rate also of 1.
- *Gradient Boosting*: The number of boosting stages to perform was 300; deviance was the loss function; and the learning rate was 0.1.

3.4. Error metrics

To compare the performance of the proposed algorithms for classification problems, it is necessary to initially define some error metrics. The most common error metrics in ML are *accuracy*, *precision*, *recall*, *balanced accuracy* and *F1 score* [63]. Each metric was calculated in our research, though all of them are reported in the manuscript.

In this study, we are dealing with a binary classification performed at the auction level. This focus on auctions rather than bids was chosen to compare previous studies, which also classify auctions as collusive or not (as a full-colluded auction is more harmful than a small percentage of collusive bids among honest ones). However, as the algorithms must first analyze every bid, every auction will be classified as collusive or competitive. This classification depends on the ratio between its collusive and competitive bids. In our study, the minimum percentage of collusive bids to classify an auction as collusive was established as follows: Brazil (≥11%), Italy (≥44%), Japan (≥11.5%), Swiss – Ticino (≥10%), Swiss – SG&GR (≥10%), and US (≥10%). As stated earlier, most of these percentages correspond to those used by the courts of justice and/or researchers who published the datasets. We only increased the Italian percentage to present good results for two reasons: the average number of bids per auction was considerably high (72.92, which is about ten times higher than the average value of the other datasets), and it has a different awarding criterion (ABA). Overall, adhering to previous percentages of collusive versus competitive bids allows us to benchmark the improvement of detection rates against previous research.

Thus, let $\widehat{y}_i$ be the predicted value of the $i-th$ sample ($1 \le i \le n$), y_i is the corresponding true value, and L is the set of classes ($1 \le l \le L$). In our case, L = 2 has two possible classes: (1) collusive or (2) competitive bid. In this instance, the *accuracy* error metric is defined as the proportion of correct predictions over *n* samples and expressed as:

$$Accuracy = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} 1\left(\widehat{y}_i = y_i\right) \tag{8}$$

where $1\left(\widehat{y}_i\right)$ is the indicator function. The equation returns 1 if the classes match and 0 otherwise.

Precision, also called positive predictive value, is intuitively the ability of the classifier not to label as positive (collusive bid) a sample that is negative (competitive bid). *Recall*, also called sensitivity or true positive rate, represents the ability of the classifier to find all positive samples. Let y_l be the subset of true values with class *l*, and $\widehat{y}_l$ the subset of true predicted values in the same class *l*:

$$Precision_l = \frac{\left|y_l \cap \widehat{y}_l\right|}{|y_l|} \tag{9}$$

$$Recall_l = \frac{\left|y_l \cap \widehat{y}_l\right|}{\left|\widehat{y}_l\right|} \tag{10}$$

The balanced accuracy avoids biased performance estimates in imbalanced datasets. Our collusion datasets are imbalanced as the number of competitive auctions in most datasets outnumber the number of collusive auctions (refer to Table 1 for the exact percentage of collusive and competitive bids in each dataset). This means, one of the two classes appears is more frequent than the other. Hence, the balanced accuracy can be defined as the average of the true positive rates (recall) of each class, that is:

$$Balanced\ Accuracy = \frac{1}{\mathrm{L}} \sum_{l=1}^{L} recall_l = \frac{1}{\mathrm{L}} \sum_{l=1}^{L} \frac{\left|y_l \cap \widehat{y}_l\right|}{\left|\widehat{y}_l\right|} \tag{11}$$

Finally, the *F1 score* can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal and expressed as:

$$F1 = 2 \bullet \frac{\mathrm{precision} \bullet \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{12}$$

The aforementioned error metrics can be adapted to our specific problem. Our study involves a binary classification (two classes), thus a *True Positive (TP)* is a correctly identified collusive bid. Additionally, a *True Negative (TN)* is a competitive bid that has also been correctly identified. A *False Positive (FP)* implies the ML algorithm flags a bid as collusive even though it was competitive. Conversely, a *False Negative (FN)* implies that the method does not classify a bid as collusive when it is so. The *FP* and *FN* have worse consequences depending on the type of public institution being involved. From the perspective of police agencies and courts of justice, *FP* is the worst type of prediction error, as it could induce an unjustified investigation in a competitive (honest) bidder. From the perspective of contracting authorities, a high percentage of *FN* is worse as there are many collusive bidders that go unnoticed. Summarizing, we have TN = Correct (not collusion), FP = Unexpected collusion, FN = Missing collusion and TP = Correct (collusion), with:

$$\mathrm{TN} + \mathrm{FP} + \mathrm{FN} + \mathrm{TP} = \mathrm{Total\ number\ of\ bids} \tag{13}$$

Hence, the previous error metrics can be expressed into our binary classification problem as:

$$Accuracy = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} 1\left(\widehat{y}_i = y_i\right) = \frac{TP + TN}{\mathrm{n}} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$Balanced\ Accuracy = \frac{1}{\mathrm{L}} \sum_{l=1}^{L} \frac{\left|y_l \cap \widehat{y}_l\right|}{\left|\widehat{y}_l\right|} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{17}$$

$$F1 = 2 \bullet \frac{\mathrm{precision} \bullet \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{18}$$

Hence, the eleven ML algorithms were trained and tested to detect collusion in six datasets from five countries. As mentioned earlier, each algorithm was run 500 times while randomly changing the training subset (80%) and the test subset (20%) from each dataset. For each repetition (run), the previous error metrics were calculated and recorded. The error metric values reported below correspond to the average values obtained from those 500 repetitions.

4. Results

Table 2 shows four of the most relevant error metrics (accuracy, FP, FN and balanced accuracy) when each dataset is used independently to detect collusion under the setting 1 (all fields) and 2 (all fields +

Table 2
Average error metrics (accuracy, FP, FN and balanced accuracy) for each dataset in settings 1 (all fields) and 2 (all fields + screens).

		Algorithm																						
		SGD		Extra Trees		Random Forest		Ada Boost		Gradient Boosting		SVC		K Neighbors		MLP		Bernoulli Naive Bayes		Gaussian Naive Bayes		Gaussian Process		Colour legend
Error metrics	Setting / Dataset	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	
Accuracy (%)	Brazil	65.2	65.1	84.9	91.2	84.9	89.8	82.4	88.1	85.2	92.4	79.3	82.7	83.2	83.5	84.0	83.5	76.9	76.5	78.6	79.6	78.5	75.9	100%
	Italy	51.3	51.1	84.4	87.4	82.5	83.1	79.5	80.8	76.1	80.2	50.8	52.4	57.2	57.5	57.4	57.1	57.4	64.8	54.4	54.4	58.1	58.5	75%
	Japan	87.8	87.8	94.7	94.5	93.1	93.0	93.5	93.1	90.5	89.2	87.8	87.9	92.5	92.5	88.7	88.8	88.7	88.6	94.6	94.6	89.5	88.9	60%
	Swiss - Ticino	74.7	74.5	79.4	90.8	77.4	86.7	73.8	91.4	77.4	87.6	60.2	55.1	75.6	76.0	81.5	81.3	81.6	80.0	81.5	81.2	19.2	16.3	45%
	Swiss - SG&GR	68.5	68.9	83.4	85.3	82.7	84.7	84.1	85.0	78.6	74.2	50.0	49.0	77.8	77.6	80.1	80.2	80.2	80.1	75.1	41.4	20.0	20.1	30%
	US	70.3	72.6	84.1	84.8	83.5	83.9	83.0	82.4	77.1	76.1	46.3	45.4	79.4	79.4	82.2	82.3	82.3	77.9	81.8	79.1	73.6	75.2	0%
False positives (FP) (%)	Brazil	23.6	23.8	4.7	2.6	6.7	5.5	8.1	6.3	4.7	4.0	12.1	10.5	6.3	5.9	4.8	5.0	1.7	6.5	14.2	13.8	0.0	0.0	100%
	Italy	23.8	24.1	9.7	6.2	9.0	8.5	11.7	9.3	13.2	11.3	37.6	39.7	17.0	17.2	0.0	0.0	2.2	8.2	34.1	32.3	0.1	0.1	75%
	Japan	5.6	5.6	0.9	0.8	2.6	2.6	2.4	2.5	4.8	5.8	9.4	9.3	3.7	3.6	0.1	0.0	0.0	0.0	0.7	0.7	0.0	0.0	60%
	Swiss - Ticino	16.0	16.0	13.9	7.5	13.1	7.7	14.6	5.8	14.3	8.5	2.9	2.8	13.6	13.5	18.5	18.7	18.4	12.8	18.5	18.8	0.0	0.1	45%
	Swiss - SG&GR	15.1	15.5	9.2	9.7	9.3	9.0	9.4	8.8	9.9	10.9	7.5	7.4	17.9	18.3	19.9	19.8	19.8	19.4	15.8	5.4	0.0	0.0	15%
	US	15.4	12.5	4.0	1.7	4.3	3.1	2.3	3.9	12.6	13.2	48.1	48.9	4.0	3.7	0.0	0.0	0.0	8.9	3.8	8.0	11.8	10.1	0%
False negatives (FN) (%)	Brazil	11.2	11.1	10.4	6.3	8.5	4.7	9.5	5.7	10.1	3.6	8.6	6.8	10.5	10.6	11.2	11.5	21.4	17.0	7.1	6.5	0.0	24.1	100%
	Italy	24.9	24.9	6.0	6.4	8.5	8.4	8.7	9.9	10.7	8.5	11.6	7.9	25.8	25.3	42.6	42.9	40.3	27.1	11.5	13.3	41.8	41.5	75%
	Japan	6.6	6.6	4.4	4.7	4.3	4.5	4.1	4.4	4.7	5.0	2.8	2.7	3.8	3.9	11.3	11.2	11.3	11.4	4.7	4.7	10.5	11.1	60%
	Swiss - Ticino	9.3	9.4	6.7	1.7	9.4	5.5	11.6	2.9	8.3	3.8	36.9	42.1	10.8	10.6	0.0	0.0	0.0	7.2	0.0	0.0	80.8	83.6	45%
	Swiss - SG&GR	16.5	15.7	7.4	5.0	7.9	6.3	6.6	6.2	11.4	14.9	42.5	43.6	4.3	4.1	0.0	0.0	0.0	0.5	9.1	53.2	80.0	79.9	15%
	US	14.3	14.9	11.9	13.6	12.2	13.0	14.6	13.6	10.4	10.7	5.6	5.7	16.7	16.9	17.8	17.7	17.7	13.1	14.4	12.9	14.7	14.7	0%
Balanced accuracy (%)	Brazil	59.5	59.8	74.0	84.6	77.0	86.0	74.3	83.7	75.5	90.6	74.2	77.5	72.9	72.7	71.3	71.5	48.9	58.7	74.0	75.3	50.0	50.0	100%
	Italy	50.5	50.4	84.7	87.2	82.3	82.7	79.5	80.3	76.4	80.3	53.5	56.4	55.5	55.6	50.0	50.0	50.8	61.0	57.2	56.7	50.7	50.4	75%
	Japan	67.6	67.9	79.8	78.7	79.3	78.6	80.4	79.2	76.3	75.3	82.9	83.1	80.7	80.7	50.0	50.1	50.0	50.1	78.3	78.7	50.0	50.0	60%
	Swiss - Ticino	50.0	50.0	57.7	78.6	58.2	75.6	52.7	82.7	61.3	76.9	69.7	66.8	56.2	56.5	50.0	50.0	50.0	61.5	50.0	50.0	50.3	50.0	45%
	Swiss - SG&GR	51.8	51.4	72.4	72.6	71.5	73.4	72.3	74.0	67.6	63.6	54.6	54.4	52.0	51.6	50.0	50.0	50.0	50.9	54.5	53.3	50.0	50.0	30%
	US	50.7	50.5	64.2	60.7	63.1	61.6	57.3	59.2	62.9	61.6	55.0	54.9	50.5	50.3	50.0	50.0	50.0	57.6	57.4	58.6	50.4	51.4	0%

Table 3
Average error metrics (accuracy, FP, FN and balanced accuracy) for each dataset in settings 3 (common fields) and 4 (common fields + screens).

		Algorithm																						
		SGD		Extra Trees		Random Forest		Ada Boost		Gradient Boosting		SVC		K Neighbors		MLP		Bernoulli Naive Bayes		Gaussian Naive Bayes		Gaussian Process		Colour legend
Error metrics	Setting / Dataset	3	4	3	4	3	4	3	4	3	4	3	4	3	4	3	4	3	4	3	4	3	4	
Accuracy (%)	Brazil	65.4	64.8	87.8	89.6	86.7	89.1	87.9	86.5	85.6	89.3	84.2	80.6	84.8	85.1	86.3	85.5	81.1	77.3	56.0	46.5	81.2	80.5	100%
	Italy	51.3	50.7	78.9	86.8	79.9	81.9	77.3	79.5	74.7	72.4	54.5	50.8	56.6	56.5	57.7	57.0	57.4	65.0	53.8	53.4	57.5	60.5	75%
	Japan	83.9	83.7	94.5	94.5	93.2	93.4	93.3	92.3	90.7	87.9	85.5	82.5	92.3	92.4	88.2	88.7	88.8	88.8	94.0	94.3	88.7	88.9	60%
	Swiss - Ticino	73.8	73.3	78.1	90.9	77.0	86.9	73.7	91.4	74.4	90.3	53.6	55.1	76.0	75.6	81.8	81.9	81.9	79.9	82.0	81.4	18.9	18.0	45%
	Swiss - SG&GR	69.3	70.0	76.6	81.1	75.8	80.3	79.4	79.2	70.5	69.4	49.8	48.3	77.8	77.7	80.2	80.2	80.1	80.2	75.5	42.2	19.3	19.8	30%
	US	70.7	70.7	83.8	83.7	82.9	83.0	82.5	81.9	77.0	74.7	47.9	47.5	79.1	79.4	82.1	82.3	82.2	78.0	82.1	79.1	72.8	74.5	15%
	All datasets	**48.7**	**48.5**	**82.0**	**86.3**	**80.5**	**84.0**	**81.6**	**81.8**	**75.6**	**72.0**	**48.1**	**47.8**	**59.2**	**59.5**	**52.5**	**52.6**	**53.7**	**58.8**	**53.6**	**53.1**	**53.3**	**52.6**	0%
False positives (FP) (%)	Brazil	23.2	23.5	2.7	3.9	4.4	4.9	4.5	6.7	4.0	4.0	10.2	12.3	3.9	5.1	3.4	3.8	0.1	3.5	39.2	48.7	0.0	0.0	100%
	Italy	22.4	23.7	11.9	6.7	10.4	9.2	12.2	10.2	14.5	14.5	38.1	40.9	16.7	16.6	0.0	0.0	0.6	7.2	27.8	28.0	0.2	0.2	75%
	Japan	8.2	8.2	1.1	0.6	2.5	2.2	2.2	3.0	5.3	7.7	11.5	14.5	3.7	3.6	0.0	0.0	0.0	0.0	0.3	0.4	0.0	0.0	60%
	Swiss - Ticino	15.7	16.5	14.9	7.7	13.1	7.6	14.8	5.7	17.6	5.9	4.3	3.0	13.5	13.7	18.2	18.1	18.1	12.4	18.0	18.6	0.0	0.2	45%
	Swiss - SG&GR	15.2	15.7	15.2	16.4	15.1	14.6	17.9	15.6	12.6	12.5	7.4	7.3	17.9	18.1	19.8	19.8	19.9	19.3	16.0	5.6	0.0	0.0	30%
	US	14.8	15.1	3.8	1.7	4.6	3.3	2.2	4.1	12.2	14.2	45.7	47.1	3.9	3.8	0.0	0.0	0.0	8.8	3.8	8.0	11.5	10.9	15%
	All datasets	**25.2**	**24.8**	**9.7**	**8.0**	**9.7**	**8.7**	**10.2**	**9.9**	**9.7**	**9.6**	**42.3**	**45.8**	**18.5**	**18.6**	**22.7**	**21.8**	**0.0**	**24.4**	**10.5**	**3.9**	**1.3**	**1.4**	0%
False negatives (FN) (%)	Brazil	11.4	11.7	9.5	6.6	8.9	6.1	7.6	6.7	10.4	6.7	5.6	7.1	11.3	9.8	10.3	10.7	18.9	19.2	4.8	4.9	18.8	19.5	100%
	Italy	26.4	25.7	9.2	6.6	9.8	8.9	10.4	10.3	10.8	13.1	7.5	8.3	26.7	26.9	42.3	43.0	42.0	27.8	18.3	18.7	42.4	39.3	75%
	Japan	7.9	8.1	4.5	4.9	4.3	4.3	4.5	4.7	4.0	4.4	3.0	3.0	4.1	3.9	11.2	11.3	11.2	11.2	5.6	5.4	11.3	11.1	60%
	Swiss - Ticino	10.5	10.1	7.0	1.5	9.9	5.6	11.5	2.9	7.9	3.8	42.0	41.9	10.5	10.8	0.0	0.0	0.0	7.7	0.0	0.0	81.1	81.9	45%
	Swiss - SG&GR	15.5	14.3	8.2	2.5	9.1	5.1	2.7	5.2	16.9	18.1	42.9	44.3	4.3	4.1	0.0	0.0	0.0	0.5	8.5	52.2	80.6	80.2	30%
	US	14.5	14.2	12.4	14.6	12.5	13.8	15.3	14.1	10.8	11.1	6.4	5.3	16.9	16.8	17.9	17.7	17.8	13.2	14.0	12.9	15.6	14.6	15%
	All datasets	**26.1**	**26.7**	**8.2**	**5.7**	**9.8**	**7.3**	**8.2**	**8.3**	**14.7**	**18.4**	**9.6**	**6.4**	**22.3**	**21.9**	**24.8**	**25.6**	**46.3**	**16.7**	**35.9**	**43.0**	**45.6**	**46.0**	0%
Balanced accuracy (%)	Brazil	57.1	58.6	73.9	83.7	74.3	83.6	78.0	81.2	71.0	83.6	78.4	73.6	69.5	74.2	71.3	72.4	50.0	54.1	64.0	58.4	50.0	50.0	100%
	Italy	50.0	50.0	78.9	86.5	79.6	81.5	77.1	79.1	74.3	72.6	58.8	55.4	54.4	54.3	50.0	50.0	50.2	61.1	55.2	55.1	51.3	51.7	75%
	Japan	60.5	59.6	79.6	78.2	79.5	79.4	78.9	77.3	77.4	75.8	80.2	78.4	80.2	80.5	50.1	50.0	50.0	50.0	74.4	75.7	50.2	50.0	60%
	Swiss - Ticino	50.0	50.0	55.4	78.3	58.1	75.8	52.6	82.5	55.9	82.3	62.1	67.8	56.7	56.7	50.0	50.0	50.0	61.3	50.0	50.0	50.3	49.8	45%
	Swiss - SG&GR	51.8	51.5	56.5	56.9	56.2	59.9	53.1	57.3	57.1	57.5	54.6	53.7	51.9	51.6	50.0	50.0	50.0	50.9	54.3	53.4	50.0	50.0	30%
	US	50.3	50.5	62.7	58.0	62.0	59.2	55.6	58.0	62.0	59.9	54.1	56.3	50.2	50.3	50.0	50.0	50.0	57.6	57.9	58.7	50.3	50.8	15%
	All datasets	**48.4**	**48.1**	**82.1**	**86.4**	**80.4**	**84.0**	**81.7**	**81.8**	**75.1**	**70.8**	**49.7**	**50.7**	**58.7**	**59.1**	**53.0**	**53.0**	**50.0**	**59.3**	**51.6**	**50.1**	**49.3**	**49.1**	0%

Note: In this table, an extra row named 'All datasets' is included as settings 3 and 4 only use fields shared among all datasets. Hence, it is possible to combine the auctions from all datasets into one.

screens). Results from the other error metrics (precision, recall and F1 score) are included later and in our *Supplementary material*. Table 3 presents the same four error metrics but applying settings 3 (common fields) and 4 (common fields + screens). Additionally, and only because settings 3 and 4 share the same input parameters, it was also possible to aggregate all datasets and analyze them as a whole. These aggregated results are presented in the bottom rows of each error metric in Table 3 (values highlighted in bold).

Tables 2 and 3 show our major results - in facilitating the process of interpreting the results presented in Tables 2 and 3, we summarize key issues in Table 4. We also would like to point out that no single algorithm performs best in all datasets. Yet, we find the ensemble methods (*Extra Trees*, *Random Forest*, *Ada Boost* and *Gradient Boosting*) are generally better among the top performers.

The screens improve the accuracy of collusion detection and decrease the rate of false positives (FP) and false negatives (FN) in almost every situation. The screens are especially effective when used with the ensemble methods. This can be readily appreciated when comparing the results of 'setting 2 versus setting 1' (Table 2) and 'setting 4 versus setting 3' (Table 3). A simple summary of this increase can be seen in the central block of Table 4. For example, Setting 2 (all fields + screens) provides evidence of the best percentages of balanced accuracy. This was expected as this is the scenario where ML algorithms have access to more auction information. For the best four algorithms (the ensemble methods) in setting 2, it is possible to see that:

- accuracy is usually higher than 80%;
- FP and FN are generally lower than 10%; and
- balanced accuracy is usually higher than 70%.

Comparing the top-performing algorithms' detection rates and results reported in the literature (bottom row of Table 4), we can see some of our algorithms have outperformed previous empirical models'. We also reveal that the US dataset was the most difficult for detecting collusion as it shows the worst percentages of balanced accuracy (about 60%) for almost all settings and algorithms. This may have arisen due to the dataset containing the lowest number of collusive bidders (11 bidders, 9.17% of the total). Similarly, the Swiss–SG&GR dataset had a low balanced accuracy (about 70%). This situation may have arisen due to the extremely high proportion of collusive versus competitive bids (59% vs 41%), rendering it difficult for the ML to differentiate between the varying bids. However, results are satisfactory when all the datasets are trained together (results in bold text in setting 4). The best algorithm, in this case, is the *Extra Trees*, which reaches a balanced accuracy of 86%. For this algorithm, the rate of FP is 8%, and the rate of FN is 6%.

The worst performing algorithms (*SGD*, *SVC*, *K Neighbors*, *MLP*, *Bernoulli* and *Gaussian Naive Bayes* and *Gaussian Processes*) hardly improve their detection results with the help of the screens. The implemented neural network algorithm (*MLP*, *Multi-Layer Perceptron*) has shown low percentages of balanced accuracy in all datasets and settings. Our *MLP* adopted four hidden layers with 240, 120, 70 and 35 neurons, respectively. However, a better combination of hidden layers

Table 4
Summary of collusion detection average results with ML algorithms.

Topic	Description	Datasets						
		Brazil	Italy	Japan	Swiss - Ticino	Swiss - SG&GR	US	All datasets
Fields	**Common fields**	Auction code, bid values, winning bid and number of bids per auction						
	All fields in the dataset	Common fields, PTE, difference Bid/PTE, location, Brazilian State and date	Common fields, PTE, difference Bid/PTE, location, type and size of bidding companies	Common fields, PTE, difference Bid/PTE, location and date	Common fields and consortium composition	Common fields, contract type and date	Common fields, bid value with and without inflation and date	Common fields only
	Num. of variables	9	9	8	5	6	7	4
	Screens	Coefficient of variation (CV), spread (SPD), percentage difference between the two lowest bids (DIFFP), relative distance (RD), skewness statistic (SKEW) and Kolmogorov–Smirnov test (KSTEST)						
Results. Best accuracy and top-performing algorithm	**Setting 1** All fields from each dataset	85.2% Gradient Boosting	84.4% Extra Trees	94.7% Extra Trees	81.6% Bernoulli Naive Bayes	84.1% Ada Boost	84.1% Extra Trees	N/A
	Setting 2 All fields from each dataset + screens	92.4% Gradient Boosting	87.4% Extra Trees	94.6% Gaussian Naive Bayes	91.4% Ada Boost	85.3% Extra Trees	84.8% Extra Trees	N/A
	Setting 3 Common fields	87.9% Ada Boost	79.9% Random Forest	94.5% Extra Trees	82.0% Gaussian Naive Bayes	80.2% MLP	83.8% Extra Trees	82.0% Extra Trees
	Setting 4 Common fields + screens	89.6% Extra Trees	86.8% Extra Trees	94.5% Extra Trees	91.4% Ada Boost	81.1% Extra Trees	83.7% Extra Trees	86.3% Extra Trees
Average accuracy increase on including screens (for the four top-performing algorithms)	**Best algorithms**	Ensemble methods: Extra Trees, Random Forest, Ada Boost and Gradient Boosting						
	Setting 2 from 1	+6.0%	+2.3%	−0.5%	+12.1%	+0.1%	−0.1%	N/A
	Setting 4 from 3	+1.6%	+2.5%	−0.9%	+14.1%	+1.9%	−0.7%	+1.1%
Detection rates reported in the literature	**Paper/s**	[30,34]	[36]	[38]	[40]	[40]	[19]	N/A
	Method	Probabilistic methods	Standard hierarchical clustering algorithm	ML methods: Random Forest & Ensemble Method	ML method: Random Forest	ML method: Random Forest	N/A	N/A
	Accuracy	81% - 96%	N/A	88% - 93%	77% - 86%	61% - 84%	N/A	N/A

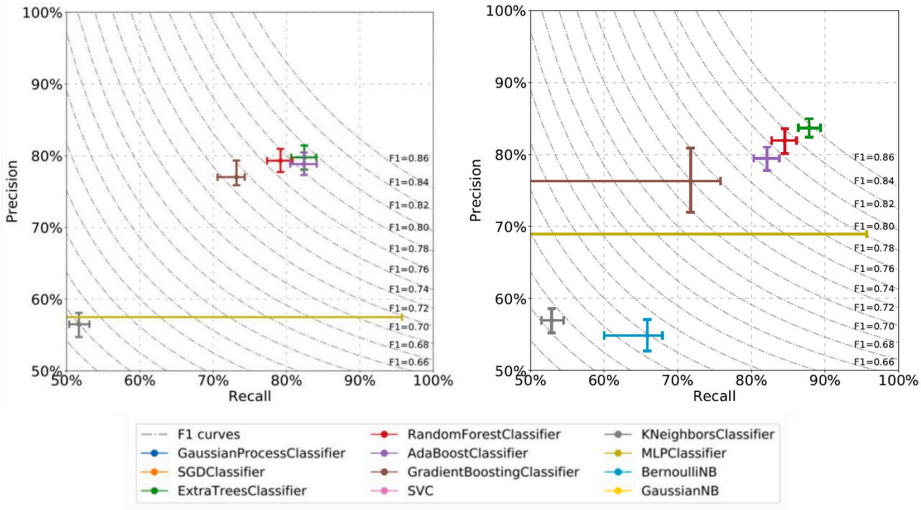


Fig. 3. Error metrics (precision, recall and F1 score) for the '*All datasets*' combination in setting 3 (common fields) on the left and setting 4 (common fields + screens) on the right.

and neurons might have reached better detection results. It should be acknowledged that combining hidden layers and neurons is an uphill task and is thus outside of the scope of this research.

Finally, Fig. 3 identifies three error metrics (precision, recall and F1 score) for settings 3 and 4 for the dataset called '*All datasets*' (auctions from all datasets merged into one). For each algorithm, the error metrics are denoted by a cross. The cutting point of the cross is the median of the precision and the recall. The endpoints of the cross are the minimum and maximum values of the precision and the recall (remember, we performed 500 iterations with each algorithm, so there are 500 values of precision and the other 500 values of recall). As a result, the precision, recall, and F1 score values remain inside the rectangle formed by the cross with a high degree of confidence. The algorithms with <50% of precision and < 50% recall are not shown in the figure. By comparing setting 3 (left graph) with setting 4 (right graph), it is seen how the screens improve the precision slightly and recall for the four top-performing algorithms (ensemble methods). Summarizing the graphical results from setting 4, we observe:

- *Extra Trees*: 83%–86% precision, 86%–89% recall and 84%–87% F1 score.
- *Random Forest*: 80%–84% precision, 82%–86% recall and 81%–85% F1 score.
- *Ada Boost*: 78%–82% precision, 80%–84% recall and 79%–83% F1 score.
- *Gradient Boosting*: 73%–81% precision, <50%–76% recall and < 78% F1 score.

For additional detail with regard to the screens boxplot and the precision, recall and F1 scores of other settings and specific datasets, we refer the readers to our *Supplementary material*.

5. Discussion

Our research demonstrates that the amount of data available per auction is positively correlated with a higher collusion detection balanced accuracy in the majority of the tested ML algorithms. Yet, even with limited access to primary data, the ML algorithms were able to achieve satisfactory collusion detection rates. To this end, the research empirically demonstrates that ML tools can be implemented and be useful even when few pieces of information are available from a large number of auctions. In this case, this basic information was the bid values and the winning bid from each auction.

The eleven ML algorithms have been tested extensively with four different settings (input data configurations). They have been analyzed with standard error metrics for binary classification problems: accuracy, false positive, false negative, balanced accuracy, precision, recall and F1 score. The results from the previous section highlight that the four ensemble methods are the top-performing algorithms for the six collusive datasets. If the field 'identity of bidders' had also been considered in settings 1 and 2, the error metrics would have also significantly improved.

Yet, we have observed some minor differences in the screen's effectiveness across datasets. In this regard, the US dataset (non-construction) and (but to a lesser extent) the Japanese dataset did not augment their average accuracy when screens were applied. Still, it is expected that screens in construction datasets will help boost collusion detection rates. Furthermore, there are no significant differences between the two awarding criteria (lowest bid versus the average bid method), at least not in accuracy for the top-performing algorithms or screens. Even though we only counted on a single dataset with different awarding criteria (the Italy dataset), hardly any differences have been found with other datasets results.

Another interesting analysis would involve training the algorithms in all but one country and then predicting collusion in the excluded country [38]. Basically, one could iteratively change the country excluded from the training data but later use it for testing purposes. This analysis would provide additional evidence on how well the methods work in terms of transferability across countries. Still, this would be a highly time-consuming, and it can only be implemented when all datasets share the same fields. Instead, we performed a similar analysis thanks to the so-called '*All datasets*' combination (combining the auctions from all datasets into one) with promising results. This combination was only possible for settings 3 and 4, though, as they were the only ones using shared information across all datasets.

6. Conclusions

Collusion has malevolent effects on public procurement, diminishes the confidence in a competitive market, and dissuades truly competitive

competitors from submitting realistic bids. Research in collusion detection in construction has focused on producing both theoretical and empirical methods. However, theoretical models have been restricted to simple applications with few bidders and under the assumption of perfect information. In contrast, the accuracy of those of an empirical nature has come into question. Our research contributes to those based on empirical models and has used a comparison of ML algorithms to demonstrate their potential for improving the accuracy of detecting collusion.

The increasing availability of public procurement information and the recent development of ML techniques has made it much easier to develop alternative empirical models to detect collusion. While ML algorithms require large amounts of data for training, they can provide robust results with fewer input variables. Recognizing the potential of ML, we have compared the performance of eleven algorithms to detect collusion. We have provided evidence that these algorithms can work with a lot of limited pieces of information. We have also shown how detection rates can be improved with the help of some screening variables. The eleven ML algorithms were tested using an extensive dataset acquired from six public procurement datasets (a total of 9781 auctions) from five countries: (1) Brazil; (2) Italy; (3) Japan; (4) Switzerland; and (5) the US.

Our analyses' three top-performing ML algorithms have been the *Extra Trees*, *Random Forest* and *Ada Boost* (ensemble methods). In the scenario where all auction information was available, these algorithms' accuracy (detection rates) ranged between 81% and 95%, with a balanced accuracy generally above 73% (excluding the US dataset). The algorithms can also be used with limited data, which poses a significant advantage over existing empirical methods. Once the algorithms are trained, they can be automatically updated with the latest auctions, and the user needs to make little effort in supervising their outcomes.

The research has limitations, which also need to be acknowledged. It is widely known that ML algorithms are akin to a black box from which it is difficult to explain the inherent complexity of the problem being analyzed (at least not in a straightforward manner). Moreover, they need a substantial amount of reliable historical data, some of which (especially the collusion-related) may not always be made available by competition commissions or law enforcement agencies – this problem is shared by other detection methods. Future research is needed to address the shortcomings of ML, specifically examining different algorithm types and fine-tuning their parameters. Access to data is critical for improving detection accuracy. A promising path for future research is to combine auction and company data (e.g., annual operating income, backlog, earnings before interest, taxes, depreciation, and amortization). By merging ML concepts with the economic theory first explored by Bajari and Ye [28] (driven by currently available data mining/scraping tools), we hope that the results will be even more accurate and their explanation better substantiated. Whereas the use of ML to detect collusion is in its infancy, we hope the research presented in this paper can foster future studies in this fertile and unexplored area.

Data availability

All auction datasets (in csv format) and algorithms code (in Python) are included as a *Supplementary file*.

Declaration of Competing Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgements

The authors are grateful to the Swiss Competition Commission (COMCO) and Dr. David Imhof for their valuable comments and sharing some of the collusive datasets used in this paper.

Appendix A. Supplementary data

Supplementary figures to this article can be found online at https://doi.org/10.1016/j.autcon.2021.104047. Furthermore, the supplementary datasets and code can be found online at (link to .zip file).

References

[1] F. Curtis, P. Maines, Closed competitive bidding, Omega. 1 (1973) 613–619, https://doi.org/10.1016/0305-0483(73)90049-2.
[2] R. Signor, P.E.D. Love, A. Oliveira, A.O. Lopes, P.S. Oliveira, Public infrastructure procurement: detecting collusion in capped first-priced auctions, J. Infrastruct. Syst. 26 (2020) 05020002, https://doi.org/10.1061/(asce)is.1943-555x.0000543.
[3] Public procurement contracts, European Parliament, Fact Sheets on the European Union, in: https://www.europarl.europa.eu/factsheets/en/sheet/34/public-procurement-contracts, 2021 (accessed May 21, 2021).
[4] Algorithms and Collusion: Competition Policy in the Digital Age, OECD, 2017. www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm (accessed May 20, 2021).
[5] OECD, Fighting Bid Rigging in IMSS Procurement: Impact of OECD Recommendations, 2018, pp. 1–256, in: http://www.oecd.org/daf/competition/IMSS-procurement-impact-OECD-recommendations2018-ENG.pdf (accessed May 20, 2021).
[6] R.C. Marshall, L.M. Marx, The vulnerability of auctions to bidder collusion, Q. J. Econ. 124 (2009) 883–910, https://doi.org/10.1162/qjec.2009.124.2.883.
[7] E.J. Anderson, T.D.H. Cau, Implicit collusion and individual market power in electricity markets, Eur. J. Oper. Res. 211 (2011) 403–414, https://doi.org/10.1016/j.ejor.2010.12.016.
[8] R. Ishii, Collusion in Repeated Procurement Auction: A Study of a Paving Market in Japan, Discussion Paper No. 710, Institute of Social and Economic Research, Osaka University, 2008, https://doi.org/10.2139/ssrn.1148064.
[9] R. Ishii, Favor exchange in collusion: empirical study of repeated procurement auctions in Japan, Int. J. Ind. Organ. 27 (2009) 137–144, https://doi.org/10.1016/j.ijindorg.2008.05.006.
[10] R.H. Porter, J.D. Zona, Detection of bid rigging in procurement auctions, J. Polit. Econ. 101 (1993) 518–538, https://doi.org/10.1086/261885.
[11] A. Blume, P. Heidhues, Modeling tacit collusion in auctions, Journal of Institutional and Theoretical Economics JITE. 164 (2008) 163–184, https://doi.org/10.1628/093245608783742101.
[12] A. Hu, T. Offerman, S. Onderstal, Fighting collusion in auctions: an experimental investigation, Int. J. Ind. Organ. 29 (2011) 84–96, https://doi.org/10.1016/j.ijindorg.2009.06.003.
[13] P. Bajari, G. Summers, Detecting collusion in procurement auctions, Antitrust Law Journal. 70 (2002) 143–170. www.jstor.org/stable/40844085.
[14] K. Hendricks, R. Porter, G. Tan, Bidding rings and the winner's curse, RAND J. Econ. 39 (2008) 1018–1041, https://doi.org/10.1111/j.1756-2171.2008.00048.x.
[15] Y. Torres Berru, V.F. López Batista, P. Torres-Carrión, M.G. Jimenez, Artificial intelligence techniques to detect and prevent corruption in procurement: a systematic literature review, in: Springer (Ed.), in: Communications in Computer and Information Science, 2020, pp. 254–268, https://doi.org/10.1007/978-3-030-42520-3_21.
[16] P. Razmi, M. Oloomi Buygi, M. Esmalifalak, A machine learning approach for collusion detection in electricity markets based on nash equilibrium theory, Journal of Modern Power Systems and Clean Energy. 9 (2021) 170–180, https://doi.org/10.35833/MPCE.2018.000566.
[17] M. Huber, D. Imhof, Machine learning with screens for detecting bid-rigging cartels, Int. J. Ind. Organ. 65 (2019) 277–301, https://doi.org/10.1016/j.ijindorg.2019.04.002.
[18] R. Signor, P.E.D. Love, L.A. Ika, White Collar Crime: Unearthing Collusion in the Procurement of Infrastructure Projects, IEEE Transactions on Engineering Management, 2020, pp. 1–12, https://doi.org/10.1109/TEM.2020.2994636.
[19] R.H. Porter, J.D. Zona, Ohio school milk markets: an analysis of bidding, RAND J. Econ. 30 (1999) 263–288. https://www.jstor.org/stable/2556080.
[20] L.H. Baldwin, R.C. Marshall, J. Richard, Bidder collusion at forest service timber sales, J. Polit. Econ. 105 (1997) 657–699, https://doi.org/10.1086/262089.
[21] E. Maskin, J. Riley, Asymmetric auctions, Review of Economic Studies. 67 (2000) 413–438, https://doi.org/10.1111/1467-937X.00137.
[22] M. Pesendorfer, A study of collusion in first-price auctions, Review of Economic Studies. 67 (2000) 381–411, https://doi.org/10.1111/1467-937X.00136.
[23] R.P. McAffe, J. McMillan, Bidding rings, Am. Econ. Rev. 82 (1992) 579–599. https://www.jstor.org/stable/2117323 (accessed May 20, 2021).
[24] M. Aoyagi, Bid rotation and collusion in repeated auctions, J. Econ. Theory 112 (2003) 79–105, https://doi.org/10.1016/S0022-0531(03)00071-1.
[25] A. Skrzypacz, H. Hopenhayn, Tacit collusion in repeated auctions, J. Econ. Theory 114 (2004) 153–169, https://doi.org/10.1016/S0022-0531(03)00128-5.
[26] J.E. Harrington, Detecting Cartels, Working Paper, No. 526, The Johns Hopkins University, Department of Economics, 2005. http://hdl.handle.net/10419/72037.
[27] J. Paha, Empirical methods in the analysis of collusion, Empirica. 38 (2011) 389–415, https://doi.org/10.1007/s10663-010-9160-1.
[28] P. Bajari, L. Ye, Deciding between competition and collusion, Rev. Econ. Stat. 85 (2003) 971–989, https://doi.org/10.1162/003465303772815871.
[29] P. Ballesteros-Pérez, M. Skitmore, R. Das, M.L. del Campo-Hitschfeld, Quick abnormal-bid-detection method for construction contract auctions, J. Constr. Eng.

Manag. 141 (2015) 04015010, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000978.
[30] R. Signor, P. Ballesteros-Perez, P.E.D. Love, Collusion Detection in Infrastructure Procurement: A Modified Order Statistic Method for Uncapped Auctions, IEEE Transactions on Engineering Management, 2021, pp. 1–14, https://doi.org/10.1109/TEM.2021.3049129.
[33] R. Signor, P.E.D. Love, J.J.C.B. Vallim, A.B. Raupp, O. Olatunji, It is not collusion unless you get caught: the case of 'operation car wash' and unearthing of a cartel, Journal of Antitrust Enforcement. 7 (2019) 177–202, https://doi.org/10.1093/jaenfo/jnz009.
[34] R. Signor, P.E.D. Love, A.T.N. Belarmino, O. Alfred Olatunji, Detection of collusive tenders in infrastructure projects: learning from operation car wash, J. Constr. Eng. Manag. 146 (2020) 05019015, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001737.
[35] D. Imhof, Y. Karagök, S. Rutz, Screening for bid rigging - does it works? Journal of Competition Law & Economics. 14 (2018) 235–261, https://doi.org/10.1093/joclec/nhy006.
[36] T.G. Conley, F. Decarolis, Detecting bidders groups in collusive auctions, American Economic Journal: Microeconomics. 8 (2016) 1–38, https://doi.org/10.1257/mic.20130254.
[37] R. Ishii, Bid roundness under collusion in Japanese procurement auctions, Rev. Ind. Organ. 44 (2014) 241–254, https://doi.org/10.1007/s11151-013-9408-6.
[38] M. Huber, D. Imhof, R. Ishii, Transnational Machine Learning with Screens for Flagging Bid-Rigging Cartels, Working Papers SES 519, Faculty of Economics and Social Sciences, University of Fribourg, 2020. https://doc.rero.ch/record/329575/files/WP_SES_519.pdf (accessed May 30, 2021).
[39] D. Imhof, Simple Statistical Screens to Detect Bid Rigging, Working Papers SES 484, Faculty of Economics and Social Sciences, University of Fribourg, 2017. http://doc.rero.ch/record/289133/files/WP_SES_484.pdf (accessed May 30, 2021).
[40] H. Wallimann, D. Imhof, M. Huber, A Machine Learning Approach for Flagging Incomplete Bid-Rigging Cartels, Working Papers SES 513, Faculty of Economics and Social Sciences, University of Fribourg, 2020. http://doc.rero.ch/record/328358/files/WP_SES_513.pdf (accessed May 30, 2021).
[41] J. Wachs, J. Kertész, A network approach to cartel detection in public auction markets, Sci. Rep. 9 (2019) 10818, https://doi.org/10.1038/s41598-019-47198-1.
[42] D. Imhof, Empirical Methods for Detecting Bid-Rigging Cartels, Université Bourgogne Franche-Comté, 2018. https://tel.archives-ouvertes.fr/tel-01963076 (accessed May 31, 2021).
[43] P. Ballesteros-Pérez, M. Skitmore, On the distribution of bids for construction contract auctions, Constr. Manag. Econ. 35 (2017) 106–121, https://doi.org/10.1080/01446193.2016.1247972.
[44] P. Ballesteros-Pérez, M.C. González-Cruz, J.L. Fuentes-Bargues, M. Skitmore, Analysis of the distribution of the number of bidders in construction contract auctions, Constr. Manag. Econ. 33 (2015) 752–770, https://doi.org/10.1080/01446193.2015.1090008.
[45] T. Zhang, Solving large scale linear prediction problems using stochastic gradient descent algorithms, in: Twenty-First International Conference on Machine Learning - ICML '04, ACM Press, New York, USA, 2004, p. 116, https://doi.org/10.1145/1015330.1015332.
[46] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42, https://doi.org/10.1007/s10994-006-6226-1.
[47] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[48] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1997) 119–139, https://doi.org/10.1006/jcss.1997.1504.
[49] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (2001) 1189–1232, https://doi.org/10.1214/aos/1013203451.
[50] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297, https://doi.org/10.1007/BF00994018.
[51] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Stat. 46 (1992) 175–185, https://doi.org/10.1080/00031305.1992.10475879.
[52] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Second, Springer, New York, NY, 2009, https://doi.org/10.1007/978-0-387-84858-7. ISBN 978-0-387-84858-7.
[53] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006. ISBN: 978-0-262-18253-9, www.GaussianProcess.org/gpml.
[54] H. Anysz, A. Foremny, J. Kulejewski, Comparison of ANN classifier to the neuro-fuzzy system for collusion detection in the tender procedures of road construction sector, in: IOP Conference Series: Materials Science and Engineering 471, 2019, p. 112064, https://doi.org/10.1088/1757-899X/471/11/112064.
[55] H. Anysz, Ł. Brzozowski, Long short-term memory (LSTM) neural networks in predicting fair price level in the road construction industry, in: IOP Conference Series: Materials Science and Engineering 1015, 2021, p. 012060, https://doi.org/10.1088/1757-899X/1015/1/012060.
[56] M.J. García Rodríguez, V. Rodríguez Montequín, F. Ortega Fernández, J. M. Villanueva Balsera, Public procurement announcements in Spain: regulations, data analysis, and award price estimator using machine learning, Complexity. 2019 (2019), https://doi.org/10.1155/2019/2360610.
[57] M.J. García Rodríguez, V. Rodríguez Montequín, F. Ortega Fernández, J. M. Villanueva Balsera, Bidders recommender for public procurement auctions using machine learning: data analysis, algorithm, and case study with tenders from Spain, Complexity. 2020 (2020) 1–20, https://doi.org/10.1155/2020/8858258.
[58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830. http://arxiv.org/abs/1201.0490 (accessed June 5, 2021).
[59] A. Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 1st ed, O'Reilly, 2017. ISBN 978-1-491-96229-9.
[60] P. Ballesteros-Pérez, M. Skitmore, E. Pellicer, X. Zhang, Scoring rules and competitive behavior in best-value construction auctions, J. Constr. Eng. Manag. 142 (2016) 04016035, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001144.
[61] P. Ballesteros-Pérez, M. Skitmore, E. Sanz-Ablanedo, P. Verhoeven, Forecasting the number and distribution of new bidders for an upcoming construction auction, J. Constr. Eng. Manag. 145 (2019) 04019056, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001694.
[62] P. Ballesteros-Pérez, M. Skitmore, E. Pellicer, M.C. González-Cruz, Scoring rules and abnormally low bids criteria in construction tenders: a taxonomic review, Constr. Manag. Econ. 33 (2015) 259–278, https://doi.org/10.1080/01446193.2015.1059951.
[63] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45 (2009) 427–437, https://doi.org/10.1016/j.ipm.2009.03.002.

Bibliografía

[1] F. Curtis y P. Maines, "Closed competitive bidding," *Omega*, vol. 1, n.º 5, págs. 613-619, oct. de 1973, issn: 03050483. doi: 10.1016/0305-0483(73)90049-2.

[2] "Public procurement contracts," *European Parliament, Fact Sheets on the European Union*, 2021. dirección: https://www.europarl.europa.eu/factsheets/en/sheet/34/public-procurement-contracts.

[3] J. VanderPlas, *Python Data Science Handbook*, 1.ª ed. O'Really, 2017, pág. 548, isbn: 978-1-491-91205-8.

[4] A. S. Patrucco, A. Moretto, S. Ronchi y D. Luzzini, "Organisational choices in public procurement: what can public management learn from the private sector?" *Local Government Studies*, vol. 45, n.º 6, págs. 977-1000, nov. de 2019, issn: 0300-3930. doi: 10.1080/03003930.2019.1608827.

[5] J. Miranzo Díaz, "Inteligencia Artificial Y Contratación Pública," en *Administración Electrónica, Transparencia y Contratación Pública*, Madrid: Iustel, 2020, págs. 105-142, isbn: 978-84-9890-385-0. dirección: https://ssrn.com/abstract=3647414.

[6] L. Cotino Hueso y A. Todolí Signes, *Explotación y regulación del uso del big data e inteligencia artificial para los servicios públicos y la ciudad inteligente*. Valencia: Tirant lo Blanch, 2022, pág. 348, isbn: 9788411133159.

[7] J. W. Tukey, "The Future of Data Analysis," *The Annals of Mathematical Statistics*, vol. 33, n.º 1, págs. 1-67, mar. de 1962, issn: 0003-4851. doi: 10.1214/aoms/1177704711.

[8] P. Naur, *Concise survey of computer methods*. Petrocelli Books, 1974, pág. 397, isbn: 9780884053149.

[9] P. del consejo de Ministros, "Real Decreto Ley de Contratos sobre Servicios Públicos," *Gaceta de Madrid*, vol. 6460, págs. 1-2, 1852. dirección: https://www.boe.es/datos/pdfs/BOE/1852/6460/A00001-00002.pdf.

[10] M. J. García Rodríguez, V. Rodríguez Montequín, F. Ortega Fernández y J. M. Villanueva Balsera, "Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning," *Complexity*, vol. 2019, n.º v, 2019, issn: 10990526. doi: 10.1155/2019/2360610.

[11] M. J. García Rodríguez, V. Rodríguez Montequín, A. Aranguen Ubierna, R. Santana Hermida, B. Sierra Araujo y A. Zelaia Jauregi, "Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain," *Studies in Informatics and Control*, vol. 30, n.º 4, págs. 67-76, dic. de 2021, issn: 12201766. doi: 10.24846/v30i4y202106.

[12] M. J. García Rodríguez, V. Rodríguez Montequín, F. Ortega Fernández y J. M. Villanueva Balsera, "Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain," *Complexity*, vol. 2020, T. C. Silva, ed., págs. 1-20, nov. de 2020, issn: 1099-0526. doi: 10.1155/2020/8858258.

[13] M. J. García Rodríguez, V. Rodríguez-Montequín, P. Ballesteros-Pérez, P. E. Love y R. Signor, "Collusion detection in public procurement auctions with machine learning algorithms," *Automation in Construction*, vol. 133, pág. 104 047, ene. de 2022, issn: 09265805. doi: 10.1016/j.autcon.2021.104047.

[14] M. J. García Rodríguez, V. R. Montequín, F. O. Fernández y J. V. Balsera, "Spanish Public Procurement: Legislation, open data source and extracting valuable information of procurement announcements," *Procedia Computer Science*, vol. 164, págs. 441-448, 2019, issn: 18770509. doi: 10.1016/j.procs.2019.12.204.

[15] D. Arosa Otero, J. C. Arvelo Flores, M. Cano Rodríguez, A. Colomer Pedrosa y M. J. García Rodríguez, "La contratación pública en España: fuentes de datos, normativa y aplicaciones tecnológicas," *Revista de la Escuela Jacobea de Posgrado*, vol. 21, págs. 87-112, 2021. dirección: https://www.jacobea.edu.mx/revista/numero21.php.

[16] M. J. García Rodríguez, "Tecnologías digitales para el control de la contratación pública," en *Auditoria pública*, ASOCEX, ed., vol. 79, Sevilla, 2022, isbn: 1136-517 X. dirección: http://auditoriapublica.com.

[17] C. Ramiró, *Inteligencia artificial y Administración pública: Robots y humanos compartiendo el servicio público*. Catarata, 2019, pág. 176, isbn: 978-84-9097-590-9.

[18] P. Valcárcel Fernández, "Strategic and Smart Public Procurement. Big Data, Open Data, public procurement 4.0 ¿Una nueva era?" En *X Seminario de Contratación Pública. Nuevos retos de la contratación pública: discusión práctica*, 2019.

[19] C. Europea, "Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones: Conseguir que la contratación pública funcione en Europa y para Europa," inf. téc., 2017, pág. 17. dirección: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=COM:2017:572:FIN.

[20] C. Europea, "Aplicación y mejores prácticas de las políticas nacionales de contratación pública en el mercado interior," Comisión Europea, Bruselas, inf. téc., 2021, pág. 11. dirección: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52021DC0245.

[21] OIReScon, "Informe Anual de Supervisión de la contratación pública en España," Oficina Independiente de Regulación y Supervisión de la Contratación (OIReScon), inf. téc., 2021. dirección: https://www.hacienda.gob.es/RSC/OIReScon/informe-anual-supervision-2021/ias-2021.pdf.

[22] M. Mendes y M. Fazekas, "DIGIWHIST Recommendations for the Implementation of Open Public Procurement Data An Implementer's Guide," págs. 1-16, 2018. dirección: https://opentender.eu/blog/2017-03-recommendations-for-implementation/.

[23] N. Modrušan, L. Mršić y K. Rabuzin, "Intelligent Public Procurement Monitoring System Powered by Text Mining and Balanced Indicators," en 2021, págs. 115-133. doi: 10.1007/978-3-030-83014-4_6.

[24] OECD, *Public Procurement for Innovation* (OECD Public Governance Reviews). OECD, jun. de 2017, isbn: 9789264265813. doi: 10.1787/9789264265820-en.

[25] C. Bovis, "The priorities of EU public procurement regulation," *ERA Forum*, vol. 21, n.º 2, págs. 283-297, oct. de 2020, issn: 1612-3093. doi: 10.1007/s12027-020-00608-8.

[26] I. González Ríos, "La transparencia como principio vertebrador de la contratación pública: significado y problemas de articulación normativa," *Revista de Estudios de la Administración Local y Autonómica*, págs. 6-25, oct. de 2019, issn: 1989-8975. doi: 10.24965/reala.i12.10714.

[27] M. Bauhr, Á. Czibik, J. Fine Licht y M. Fazekas, "Lights on the shadows of public procurement: Transparency as an antidote to corruption," *Governance*, vol. 33, n.º 3, págs. 495-523, jul. de 2020, issn: 0952-1895. doi: 10.1111/gove.12432.

[28] M. Fernánez Salmerón y R. Martínez Gutiérrez, *Transparencia, innovación y buen gobierno en la contratación pública*. Valencia: Tirant lo Blanch, 2019, pág. 492, isbn: 9788491905943.

[29] P. Valcárcel Fernández, "Transparency in public procurement in the Spanish legal system," en *Transparency in EU Procurements. Disclosure rules within public procurement procedures and during contract period*, 4, K.-M. Halonen, R. Caranta y A. Sánchez Graells, eds., Edward Elgar Publishing, 2019, págs. 272-295, isbn: ISBN 978-1-78897-566-7. doi: 10.4337/9781788975674.00019.

[30] A. Cerrillo Martínez, "Contratación electrónica y transparencia: fundamentos necesarios de la contratación abierta," *Cuadernos de derecho local*, págs. 121-149, 2018, issn: 1696-0955.

[31] J. Valero Torrijos, "Transparencia, acceso y reutilización de la información del sector público," en *Transparencia y acceso a la información pública: de la teoría a la práctica*, Iustel, ed., 2019, págs. 225-250, isbn: 978-84-9890-373-7.

[32] P. Valcárcel Fernández, "Tres dimensiones de la transparencia en la contratación pública. Rendición de cuentas, respeto de los derechos de operadores económicos y mejora global de la gestión de este sector a través del big data," en *Observatorio de contratos públicos*, J. M. Gimeno Feliu, ed., vol. 2018, Thomson Reuters-Aranzadi, 2019, págs. 93-129, isbn: 978-84-1309-943-9.

[33] J. M. Gimeno Feliu, "Corrupción y seguridad jurídica. La necesidad de un marco normativo de las decisiones públicas anclado en los principios de Integridad y de Transparencia.," *Revista internacional de transparencia e integridad*, n.º 9, pág. 10, 2019. dirección: https://dialnet.unirioja.es/servlet/articulo?codigo=6977097.

[34] A. Zuiderwijk, N. Helbig, J. R. Gil-García y M. Janssen, "Special issue on innovation through open data - A review of the state-of-the-art and an emerging research agenda: Guest editors' introduction," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 9, n.º 2, 2014, issn: 07181876. doi: 10.4067/S0718-18762014000200001.

[35] D. Corrales-Garay, M. Ortiz-de-Urbina-Criado y E. M. Mora-Valentín, "Knowledge areas, themes and future research on open data: A co-word analysis," *Government Information Quarterly*, vol. 36, n.º 1, págs. 77-87, 2018, issn: 0740624X. doi: 10.1016/j.giq.2018.10.008.

[36] J. D. Twizeyimana y A. Andersson, "The public value of E-Government – A literature review," *Government Information Quarterly*, vol. 36, n.º 2, págs. 167-178, 2019, issn: 0740-624X. doi: 10.1016/J.GIQ.2019.01.001.

[37] G. Magalhaes y C. Roseira, "Open government data and the private sector: An empirical view on business models and value creation," *Government Information Quarterly*, n.º August, págs. 1-10, 2017, issn: 0740624X. doi: 10.1016/j.giq.2017.08.004.

[38] J. M. Alvarez-Rodríguez, J. E. Labra-Gayo y P. O. De Pablos, "New trends on e-Procurement applying semantic technologies: Current status and future challenges," *Computers in Industry*, vol. 65, n.º 5, págs. 800-820, 2014, issn: 01663615. doi: 10.1016/j.compind.2014.04.005.

[39] E. Huyer y L. van Knippenberg, "The Economic Impact of Open Data Opportunities for value creation in Europe," European Commission, inf. téc., 2020, pág. 138. doi: 10.2830/63132.

[40] W. McKinney, *Python for Data Analysis 2ed*, Second, M. Beaugureau, ed. O'Reilly, 2012, pág. 541, isbn: 978-1-491-95766-0.

[41] M. Lutz, *Learning Python*, 5.ª ed. O'Reilly, 2013, pág. 1594, isbn: 9781449355739.

[42] A. Oussous, F.-z. Benjelloun, A. Ait y S. Belfkih, "Big Data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, n.º 4, págs. 431-448, 2018, issn: 1319-1578. doi: 10.1016/j.jksuci.2017.06.001.

[43] J. Miranzo Díaz, "El nuevo Derecho de la UE: las medidas anticorrupción en la contratación pública," Tesis doct., Universidad de Castilla-La Mancha, 2018, pág. 568.

[44] E. Díaz Bravo y J. A. Moreno Molina, *Contratación Pública Global: Visiones Comparadas*. Valencia: Tirant lo Blanch, 2020, pág. 782, isbn: 9788413367279.

[45] A. Huergo Lora, "Administraciones Públicas e inteligencia artificial: ¿más o menos discrecionalidad?" *El Cronista del Estado Social y Democrático de Derecho*, vol. 96-97, págs. 78-95, 2021, issn: 1889-0016.

[46] J. D. Kelleher, B. Mac Namee y A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, First edit. The MIT Press, 2015, pág. 624, isbn: 9780262029445.

[47] I. Lee e Y. J. Shin, "Machine learning for enterprises: Applications, algorithm selection, and challenges," *Business Horizons*, vol. 63, n.º 2, págs. 157-170, 2020, issn: 00076813. doi: 10.1016/j.bushor.2019.10.005.

[48] World Bank, "Disruptive technologies in public procurement," The World Bank, Washington, inf. téc., 2021, pág. 114. dirección: https://documents1.worldbank.org/curated/en/522181612428427520/pdf/Disruptive-Technologies-in-Public-Procurement.pdf.

[49] S. Athey y G. Imbens, "Machine Learning Methods Economists Should Know About," *Working Paper,*, n.º March, mar. de 2019. arXiv: 1903.10075. dirección: https://www.gsb.stanford.edu/faculty-research/working-papers/machine-learning-methods-economists-should-know-about%20http://arxiv.org/abs/1903.10075.

[50] S. Mullainathan y J. Spiess, "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, vol. 31, n.º 2, págs. 87-106, 2017, issn: 0895-3309. doi: 10.1257/jep.31.2.87.

[51] H. R. Varian, "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, vol. 28, n.º 2, págs. 3-28, 2014, issn: 0895-3309. doi: 10.1257/jep.28.2.3.

[52] T. Hastie, R. Tibshirani y J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics), Second. New York, NY: Springer New York, 2009, pág. 764, isbn: 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7.

[53] K. P. Murphy, *Probabilistic ML - Adaptive Computation and Machine Learning*. 2022, isbn: 2021027430. dirección: https://lccn.loc.gov/2021027430.

[54] A. Soylu y col., "Data Quality Barriers for Transparency in Public Procurement," *Information*, vol. 13, n.º 2, pág. 99, feb. de 2022, issn: 2078-2489. doi: 10.3390/info13020099.

[55] M. M. Khurshid, N. H. Zakaria, A. Rashid, M. N. Ahmad, M. I. Arfeen y H. M. Faisal Shehzad, "Modeling of Open Government Data for Public Sector Organizations Using the Potential Theories and Determinants—A Systematic Review," *Informatics*, vol. 7, n.º 3, pág. 24, jul. de 2020, issn: 2227-9709. doi: 10.3390/informatics7030024.

[56] S. de Juana-Espinosa y S. Luján-Mora, "Open government data portals in the European Union: Considerations, development, and expectations," *Technological Forecasting and Social Change*, vol. 149, n.º October, pág. 119 769, 2019, issn: 00401625. doi: 10.1016/j.techfore.2019.119769.

[57] G. Vancauwenberghe, B. van Loenen y J. Crompvoets, *Open Data Exposed*. T.M.C. Asser Press, The Hague, 2018, pág. 299, isbn: 9789462652606. doi: https://doi.org/10.1007/978-94-6265-261-3.

[58] S. Sadiq y M. Indulska, "Open data: Quality over quantity," *International Journal of Information Management*, vol. 37, n.º 3, págs. 150-154, 2017, issn: 02684012. doi: 10.1016/j.ijinfomgt.2017.01.003.

[59] Federación Española de Municipios y Provincias, ed., *Datos Abiertos: Guía estratégica para su puesta en marcha Conjuntos de datos mínimos a publicar*, 1.ª ed. Spain: Wolters Kluwer España, S.A, 2017, pág. 122, isbn: 9788415651574.

[60] F. Ahmadi Zeleti, A. Ojo y E. Curry, "Exploring the economic value of open government data," *Government Information Quarterly*, vol. 33, n.º 3, págs. 535-551, 2016, issn: 0740624X. doi: 10.1016/j.giq.2016.01.008.

[61] J. Attard, F. Orlandi, S. Scerri y S. Auer, "A systematic review of open government data initiatives," *Government Information Quarterly*, vol. 32, n.º 4, págs. 399-418, 2015, issn: 0740624X. doi: 10.1016/j.giq.2015.07.006.

[62] D. Goens, "The exploitation of Business Register data from a public sector information and data protection perspective: A case study," *Computer Law & Security Review*, vol. 26, n.º 4, págs. 398-405, jul. de 2010, issn: 02673649. doi: 10.1016/j.clsr.2010.05.001.

[63] M. Kuziemski y G. Misuraca, "AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings," *Telecommunications Policy*, vol. 44, n.º 6, pág. 13, jul. de 2020, issn: 03085961. doi: 10.1016/j.telpol.2020.101976.

[64] N. Obwegeser y S. D. Müller, "Innovation and public procurement: Terminology, concepts, and applications," *Technovation*, vol. 74-75, n.º April 2016, págs. 1-17, 2018, issn: 01664972. doi: 10.1016/j.technovation.2018.02.015.

[65] M. C. B. de Araújo, L. H. Alencar y C. M. de Miranda Mota, "Project procurement management: A structured literature review," *International Journal of Project Management*, vol. 35, n.º 3, págs. 353-377, 2017. doi: 10.1016/j.ijproman.2017.01.008.

[66] K. V. Thai, ed., *Global Public Procurement Theories and Practices*. Springer International Publishing, 2017, isbn: 978-3-319-49279-7. doi: 10.1007/978-3-319-49280-3.

[67] A. S. Patrucco, D. Luzzini y S. Ronchi, "Research perspectives on public procurement: Content analysis of 14 years of publications in the journal of public procurement," *Journal of Public Procurement*, vol. 17, n.º 2, págs. 229-269, mar. de 2017, issn: 1535-0118. doi: 10.1108/JOPP-17-02-2017-B003.

[68] M. K. Gorgun, M. Kutlu y B. K. Onur Tas, "Predicting The Number of Bidders in Public Procurement," en *2020 5th International Conference on Computer Science and Engineering (UBMK)*, IEEE, sep. de 2020, págs. 360-365, isbn: 978-1-7281-7565-2. doi: 10.1109/UBMK50275.2020.9219404.

[69] M. Bilal y L. O. Oyedele, "Big Data with deep learning for benchmarking profitability performance in project tendering," *Expert Systems with Applications*, vol. 147, pág. 113 194, 2020, issn: 09574174. doi: 10.1016/j.eswa.2020.113194.

[70] J. M. Kim y H. Jung, "Predicting bid prices by using machine learning methods," *Applied Economics*, vol. 51, n.º 19, págs. 2011-2018, 2019, issn: 14664283. doi: 10.1080/00036846.2018.1537477.

[71] R. Matin, C. Hansen, C. Hansen y P. Mølgaard, "Predicting distresses using deep learning of text segments in annual reports," *Expert Systems With Applications*, vol. 132, págs. 199-208, 2019, issn: 0957-4174. doi: 10.1016/j.eswa.2019.04.071.

[72] J. S. Chou, C. W. Lin, A. D. Pham y J. Y. Shao, "Optimized artificial intelligence models for predicting project award price," *Automation in Construction*, vol. 54, págs. 106-115, 2015, issn: 09265805. doi: 10.1016/j.autcon.2015.02.006.

[73] M. Huber y D. Imhof, "Deep learning for detecting bid rigging: Flagging cartel participants based on convolutional neural networks," abr. de 2021. arXiv: 2104.11142. dirección: http://arxiv.org/abs/2104.11142.

[74] M. E. K. Niessen, J. M. Paciello y J. I. P. Fernandez, "Anomaly Detection in Public Procurements using the Open Contracting Data Standard," en *2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG)*, IEEE, abr. de 2020, págs. 127-134, isbn: 978-1-7281-5882-2. doi: 10.1109/ICEDEG48599.2020.9096674.

[75] Y. Torres Berru, V. F. López Batista, P. Torres-Carrión y M. G. Jimenez, "Artificial Intelligence Techniques to Detect and Prevent Corruption in Procurement: A Systematic Literature Review," en *Communications in Computer and Information Science*, Springer, ed., vol. 2, 2020, págs. 254-268, isbn: 9783030425203. doi: 10.1007/978-3-030-42520-3_21.

[76] M. Huber, D. Imhof y R. Ishii, "Transnational machine learning with screens for flagging bid-rigging cartels," *University of Fribourg (Switzerland)*, vol. Working Pa, n.º Faculty of Economics and Social Sciences, 2020. dirección: https://doc.rero.ch/record/329575/files/WP%7B%5C_%7DSES%7B%5C_%7D519.pdf.

[77] H. Wallimann, D. Imhof y M. Huber, "A Machine Learning Approach for Flagging Incomplete Bid-rigging Cartels," *University of Freiburg/Fribourg (Switzerland)*, abr. de 2020. arXiv: 2004.05629. dirección: http://arxiv.org/abs/2004.05629.

[78] M. Lima, R. Silva, F. Lopes de Souza Mendes, L. R. de Carvalho, A. Araujo y F. de Barros Vidal, "Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach," en *Findings of the Association for Computational Linguistics: EMNLP 2020*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, págs. 1580-1588. doi: 10.18653/v1/2020.findings-emnlp.143.

[79] C. Giorgiantonio y F. Decarolis, "Corruption Red Flags in Public Procurement: New Evidence from Italian Calls for Tenders," *SSRN Electronic Journal*, n.º February, pág. 34, 2020, issn: 1556-5068. doi: 10.2139/ssrn.3612661.

[80] J. Gallego, G. Rivero y J. Martínez, "Preventing rather than punishing: An early warning model of malfeasance in public procurement," *International Journal of Forecasting*, jul. de 2020, issn: 01692070. doi: 10.1016/j.ijforecast.2020.06.006.

[81] M. Huber y D. Imhof, "Machine learning with screens for detecting bid-rigging cartels," *International Journal of Industrial Organization*, vol. 65, págs. 277-301, jul. de 2019, issn: 01677187. doi: 10.1016/j.ijindorg.2019.04.002.

[82] K. Rabuzin y N. Modrušan, "Prediction of Public Procurement Corruption Indices using Machine Learning Methods," en *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, SCITEPRESS - Science y Technology Publications, 2019, págs. 333-340, isbn: 978-989-758-382-7. doi: 10.5220/0008353603330340.

[83] J. Wacker, R. P. Ferreira y M. Ladeira, "Detecting Fake Suppliers using Deep Image Features," en *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, IEEE, oct. de 2018, págs. 224-229, isbn: 978-1-5386-8023-0. doi: 10.1109/BRACIS.2018.00046.

[84] T. Sun y L. J. Sales, "Predicting Public Procurement Irregularity: An Application of Neural Networks," *Journal of Emerging Technologies in Accounting*, vol. 15, n.º 1, págs. 141-154, jul. de 2018, issn: 1558-7940. doi: 10.2308/jeta-52086.

[85] M. Lei, Z. Yin, S. Li y H. Li, "Detecting the collusive bidding behavior in below average bid auction," en *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, IEEE, jul. de 2017, págs. 1720-1727, isbn: 978-1-5386-2165-3. doi: 10.1109/FSKD.2017.8393026.

[86] E. Grace, A. Rai, E. Redmiles y R. Ghani, "Detecting fraud, corruption, and collusion in international development contracts: The design of a proof-of-concept automated system," en *2016 IEEE International Conference on Big Data (Big Data)*, IEEE, dic. de 2016, págs. 1444-1453, isbn: 978-1-4673-9005-7. doi: 10.1109/BigData.2016.7840752.

[87] D. Imhof, "Empirical Methods for Detecting Bid-rigging Cartels," Tesis doct., Université Bourgogne Franche-Comté, 2018. dirección: https://tel.archives-ouvertes.fr/tel-01963076.

[88] F. Pedregosa y col., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011. arXiv: 1201.0490. dirección: http://arxiv.org/abs/1201.0490.

[89] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1.ª ed. O'Reilly, 2017, isbn: 978-1-491-96229-9.

[90] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann e I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, n.º 1, págs. 10-18, nov. de 2009, issn: 1931-0145. doi: 10.1145/1656274.1656278.

[91] I. H. Witten, E. Frank y M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Third. Boston: Elsevier, 2011, isbn: 9780123748560. doi: 10.1016/C2009-0-19715-5.

[92] M. Abadi y col., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," mar. de 2016. arXiv: 1603.04467. dirección: http://arxiv.org/abs/1603.04467.

[93] L. Breiman, "Random forests," *Machine Learning*, vol. 45, n.º 1, págs. 5-32, 2001. doi: 10.1023/A:1010933404324.

[94] A. Verikas, A. Gelzinis y M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern Recognition*, vol. 44, n.º 2, págs. 330-349, 2011, issn: 00313203. doi: 10.1016/j.patcog.2010.08.011.

[95] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro y M. Febrero-Bande, "An extensive experimental survey of regression methods," *Neural Networks*, vol. 111, págs. 11-34, 2019, issn: 18792782. doi: 10.1016/j.neunet.2018.12.010.

[96] M. R. Segal, "Machine Learning Benchmarks and Random Forest Regression," *UCSF: Center for Bioinformatics and Molecular Biostatistics*, 2004. dirección: https://escholarship.org/uc/item/35x3v9t4.

[97] G. Biau y E. Scornet, "A random forest guided tour," *Test*, vol. 25, n.º 2, págs. 197-227, 2016, issn: 11330686. doi: 10.1007/s11749-016-0481-7.

[98] A. Mehrbod y A. Grilo, "Tender calls search using a procurement product named entity recogniser," *Advanced Engineering Informatics*, vol. 36, n.º June 2017, págs. 216-228, abr. de 2018, issn: 14740346. doi: 10.1016/j.aei.2018.04.005.

[99] M. Nečaský, J. Klímek, J. Mynarz, T. Knap, V. Svátek y J. Stárka, "Linked data support for filing public contracts," *Computers in Industry*, vol. 65, n.º 5, págs. 862-877, jun. de 2014, issn: 01663615. doi: 10.1016/j.compind.2013.12.006.

[100] *Algorithms and Collusion: Competition Policy in the Digital Age*, 2017. dirección: www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm.

[101] D. Imhof, Y. Karagök y S. Rutz, "Screening for Bid Rigging - Does it works?" *Journal of Competition Law & Economics*, vol. 14, n.º 2, págs. 235-261, jun. de 2018, issn: 1744-6414. doi: 10.1093/joclec/nhy006.

[102] D. Imhof, "Simple Statistical Screens to Detect Bid Rigging," *Working Papers SES. Faculty of Economics and Social Sciences. University of Fribourg*, vol. 484, 2017. dirección: http://doc.rero.ch/record/289133/files/WP_SES_484.pdf.

[103] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," en *Twenty-first international conference on Machine learning - ICML '04*, New York, New York, USA: ACM Press, 2004, pág. 116, isbn: 1581138285. doi: 10.1145/1015330.1015332.

[104] P. Geurts, D. Ernst y L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, n.º 1, págs. 3-42, 2006, issn: 08856125. doi: 10.1007/s10994-006-6226-1.

[105] Y. Freund y R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, n.º 1, págs. 119-139, ago. de 1997, issn: 00220000. doi: 10.1006/jcss.1997.1504.

[106] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *The Annals of Statistics*, vol. 29, n.º 5, págs. 1189-1232, oct. de 2001, issn: 0090-5364. doi: 10.1214/aos/1013203451.

[107] C. Cortes y V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, n.º 3, págs. 273-297, sep. de 1995, issn: 0885-6125. doi: 10.1007/BF00994018.

[108] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, n.º 3, págs. 175-185, ago. de 1992, issn: 0003-1305. doi: 10.1080/00031305.1992.10475879.

[109] C. E. Rasmussen y C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006, isbn: 9780262182539. dirección: http://gaussianprocess.org/gpml.

[110] S. Curto, S. Ghislandi, K. Van de Vooren, S. Duranti y L. Garattini, "Regional tenders on biosimilars in Italy: An empirical analysis of awarded prices," *Health Policy*, vol. 116, n.º 2-3, págs. 182-187, 2014, issn: 18726054. doi: 10.1016/j.healthpol.2014.02.011.

[111] T. Hanák y P. Muchová, "Impact of Competition on Prices in Public Sector Procurement," *Procedia Computer Science*, vol. 64, págs. 729-735, 2015, issn: 18770509. doi: 10.1016/j.procs.2015.08.601.

[112] J. Soudek y J. Skuhrovec, "Procurement procedure, competition and final unit price: The case of commodities," *Journal of Public Procurement*, vol. 16, n.º 1, págs. 1-21, mar. de 2016, issn: 1535-0118. doi: 10.1108/JOPP-16-01-2016-B001.

Enlaces a internet

- Autoridad Independiente de Responsabilidad Fiscal (AIReF): https://www.airef.es
- Autoridad Suiza de la Competencia: https://www.weko.admin.ch/weko/en/home/comco.html
- Base Nacional de Subvenciones: https://www.pap.hacienda.gob.es/bdnstrans/es/index
- Central de Información Económico-Financiera de las AA.PP.: https://www.hacienda.gob.es/es-ES/CDI/Paginas/centraldeinformacion.aspx
- Comisión Nacional de los Mercados y la Competencia (CNMC): https://www.cnmc.es
- Consejo de Transparencia y Buen Gobierno (CTBG): https://www.consejodetransparencia.es
- Diario Oficial de la Unión Europea (DOUE): https://eur-lex.europa.eu/oj/direct-access.html
- Dirección General de Racionalización y Centralización de la Contratación (DGRCC): https://contratacioncentralizada.gob.es
- DoZorro: https://dozorro.org
- Guardia Civil: https://www.guardiacivil.es
- Intervención General de la Administración del Estado (IGAE): https://www.igae.pap.hacienda.gob.es/sitios/igae
- Junta Consultiva de Contratación del Estado: https://www.hacienda.gob.es/es-ES/Areas%20Tematicas/Contratacion/Junta%20Consultiva%20de%20Contratacion%20Administrativa/Paginas/default.aspx
- Oficina Independiente de Regulación y Supervisión de la Contratación (OIReScon): https://www.hacienda.gob.es/es-ES/Oirescon/Paginas/HomeOirescon.aspx
- Open Knowledge Foundation: https://okfn.org
- Open Tender: https://www.opentender.eu
- Plataforma de Contratación del Sector Público (PLACSP): https://contrataciondelestado.es
- Policía Nacional: https://www.policia.es
- Policía Federal de Brasil: https://www.gov.br/pf/pt-br
- Portal de datos abiertos de la UE: https://data.europa.eu/
- Portal de la Transparencia de la AGE: https://transparencia.gob.es

- Red flags: https://www.redflags.eu
- Registro de Contratos: https://www.hacienda.gob.es/es-ES/Areas%20Tematicas/Contratacion/Junta%20Consultiva%20de%20Contratacion%20Administrativa/Paginas/Registro%20publico%20de%20contratos.aspx
- Registro Mercantil: https://www.registradores.org
- Tenders Electronic Daily (TED): https://ted.europa.eu/
- Transparencia Internacional: https://www.transparency.org
- Tribunal Administrativo Central de Recursos Contractuales: https://www.hacienda.gob.es/es-ES/Areas%20Tematicas/Contratacion/TACRC/Paginas/Tribunal%20Administrativo%20Central%20de%20Recursos%20Contractuales.aspx
- Tribunal de Cuentas: https://www.tcu.es

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

FORMULARIO RESUMEN DE TESIS POR COMPENDIO

1.- Datos personales solicitante	
Apellidos: García Rodríguez	Nombre: Manuel José

Curso de inicio de los estudios de doctorado	2017/18

	SI	NO
Acompaña acreditación por el Director de la Tesis de la aportación significativa del doctorando	x	
Acompaña memoria que incluye		
Introducción justificativa de la unidad temática y objetivos	x	
Copia completa de los trabajos *	x	
Resultados/discusión y conclusiones	x	
Informe con el factor de impacto de la publicaciones	x	

Se acompaña aceptación de todos y cada uno de los coautores a presentar el trabajo como tesis por compendio (Art. 32.4.b)	x	
Se acompaña renuncia de todos y cada uno de los coautores no doctores a presentar el trabajo como parte de otra tesis de compendio (Art. 32.4.c)	x	

* Ha de constar el nombre y adscripción del autor y de todos los coautores asi como la referencia completa de la revista o editorial en la que los trabajos hayan sido publicados o aceptados en cuyo caso se aportará justificante de la aceptación por parte de la revista o editorial

FOR-MAT-VOA-033

Artículos, Capítulos, Trabajos

Trabajo, Artículo 1

Titulo (o título abreviado)	Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning
Fecha de publicación	14/11/2019
Fecha de aceptación	27/9/2019
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	JCR SCIE
Factor de impacto	2.462

Coautor2 x Doctor □ No doctor . Indique nombre y apellidos	Vicente Rodríguez Montequín
Coautor3 x Doctor □ No doctor . Indique nombre y apellidos	Francisco Ortega Fernández
Coautor4 x Doctor □ No doctor . Indique nombre y apellidos	Joaquín M. Villanueva Balsera

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

Trabajo, Artículo 2

Titulo (o título abreviado)	Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain
Fecha de publicación	25/11/2020
Fecha de aceptación	11/11/2020
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	JCR SCIE
Factor de impacto	2.833

Coautor2 x Doctor □ No doctor . Indique nombre y apellidos	Vicente Rodríguez Montequín
Coautor3 x Doctor □ No doctor . Indique nombre y apellidos	Francisco Ortega Fernández
Coautor4 x Doctor □ No doctor . Indique nombre y apellidos	Joaquín M. Villanueva Balsera

Trabajo, Artículo 3

Titulo (o título abreviado)	Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain
Fecha de publicación	1/12/2021
Fecha de aceptación	10/11/2021
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	JCR SCIE
Factor de impacto	1.649

Coautor2 x Doctor □ No doctor . Indique nombre y apellidos	Vicente Rodríguez Montequín
Coautor3 □ Doctor x No doctor . Indique nombre y apellidos	Andoni Aranguren Ubierna
Coautor4 x Doctor □ No doctor . Indique nombre y apellidos	Roberto Santana Hermida
Coautor5 x Doctor □ No doctor . Indique nombre y apellidos	Ana Zelaia Jauregui

Trabajo, Artículo 4

Titulo (o título abreviado)	Collusion detection in public procurement auctions with machine learning algorithms
Fecha de publicación	18/11/2021
Fecha de aceptación	8/11/2021
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	JCR SCIE
Factor de impacto	7.700

Coautor2 x Doctor □ No doctor . Indique nombre y apellidos	Vicente Rodríguez Montequín
Coautor3 x Doctor □ No doctor . Indique nombre y apellidos	Pablo Ballesteros-Pérez
Coautor4 x Doctor □ No doctor . Indique nombre y apellidos	Peter E.D. Love
Coautor5 x Doctor □ No doctor . Indique nombre y apellidos	Regis Signord

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

INFORME PARA LA PRESENTACIÓN DE TESIS DOCTORAL COMO COMPENDIO DE PUBLICACIONES

Año Académico:__2021__/_2022__

1.- Datos personales del autor de la Tesis		
Apellidos: García Rodríguez	Nombre: Manuel José	
DNI/Pasaporte/NIE: 71664954J	Teléfono: 620773032	Correo electrónico: manueljgarciar@gmail.com

2.- Datos académicos	
Programa de Doctorado cursado: Programa de Doctorado en Ingeniería de Producción, Minero-Ambiental y de Proyectos	
Órgano responsable: CIP	
Departamento/Instituto en el que presenta la Tesis Doctoral:	
Título definitivo de la Tesis	
Español: Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning	Inglés: Public procurement announcements: data analysis and forecasting systems using machine learning methods
Rama de conocimiento: Ingeniería	

3.- Director/es de la Tesis	
D/Dª: Vicente Rodríguez Montequín	DNI/Pasaporte/NIE: 52616590L
Departamento/Instituto: Explotación y Prospección de Minas	
D/Dª: Ramiro Concepción Suárez	DNI/Pasaporte/NIE: 11049840L
Departamento/Instituto/Institución:	

4.- Informe
El doctorando cumple con los requisitos normativos establecidos para poder presentar su Tesis Doctoral bajo la modalidad de compendio de publicaciones. Derivados de esta Tesis se han publicado cuatro artículos científicos relevantes de los que es autor y primer firmante. Los artículos están publicados en revistas incluidas en el JCR SCIE. El desarrollo de la tesis se puede considerar completo con las cuatro publicaciones relacionadas.

En Oviedo, a 31 de mayo de 2022

Director de la Tesis Doctoral

Fdo.: Vicente Rodríguez Montequín / Ramiro Concepción Suárez

ACCEPTANCE OF CO-AUTHORS TO SUBMIT THEIR WORKS AS PART OF PhD THESIS AS A COMPILATION OF PUBLICATIONS

1.-Personal details of co-author		
Family name: BALLESTEROS-PÉREZ	First name: PABLO	
ID/Passport/NIE number: 33469552-k	Telephone number: +34 622 478 882	Email address: pabbalpe@dpi.upv.es

2. - Co-authored publications that will be part of the PhD Thesis

Rodríguez, M. J. G., Rodríguez-Montequín, V., **Ballesteros-Pérez, P.**, Love, P. E., & Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, *133*, 104047.

ACCEPTANCE:

I accept the above-mentioned publications to be part of the PhD Thesis entitled
Public procurement announcements: data analysis and forecasting systems using machine learning methods

and written by Manuel José García Rodríguez

26th May 2022
Signature: BALLESTEROS PEREZ PABLO - 33469552K

Firmado digitalmente por BALLESTEROS PEREZ PABLO - 33469552K
Fecha: 2022.05.26 12:32:25 +02'00'

ACCEPTANCE OF CO-AUTHORS TO SUBMIT THEIR WORKS AS PART OF PhD THESIS AS A COMPILATION OF PUBLICATIONS

1.-Personal details of co-author		
Family name: LOVE	First name: PETER	
ID/Passport/NIE number	Telephone number:	Email address: p.love@curtin.edu.au

2. - Co-authored publications that will be part of the PhD Thesis

Rodríguez, M. J. G., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E., & Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, *133*, 104047.

ACCEPTANCE:

I accept the above-mentioned publications to be part of the PhD Thesis entitled
Public procurement announcements: data analysis and forecasting systems using machine learning methods

and written by Manuel José García Rodríguez

26th May 2022

Signature:

ACCEPTANCE OF CO-AUTHORS TO SUBMIT THEIR WORKS AS PART OF PhD THESIS AS A COMPILATION OF PUBLICATIONS

1.-Personal details of co-author		
Family name: Signor	First name: Regis	
ID/Passport/NIE number: GA540059	Telephone number: +55-48-98423-3938	Email address: regis.rs@pf.gov.br

2. - Co-authored publications that will be part of the PhD Thesis

Rodríguez, M. J. G., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E., & Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, *133*, 104047.

ACCEPTANCE:

I accept the above-mentioned publications to be part of the PhD Thesis entitled
Public procurement announcements: data analysis and forecasting systems using machine learning methods

and written by Manuel José García Rodríguez

Florianópolis/SC (Brasil), 26 de maio de 2022.

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

ACEPTACIÓN COAUTORES PRESENTACIÓN TRABAJOS FORMANDO PARTE DE TESIS DOCTORAL COMO COMPENDIO DE PUBLICACIONES

1.- Datos personales del coautor		
Apellidos: Ortega Fernández	Nombre: Francisco de Asís	
DNI/Pasaporte/NIE 09380975B	Teléfono 985104272	Correo electrónico fdeasis@uniovi.es

2.- Publicaciones que formarán parte de la tesis y de las que es coautor

García Rodríguez, M. J., Rodríguez Montequín, V., Ortega Fernández, F., & Villanueva Balsera, J. M. (2019). Public procurement announcements in spain: regulations, data analysis, and award price estimator using machine learning. *Complexity*, *2019*.

García Rodríguez, M. J., Rodríguez Montequín, V., Ortega Fernández, F., & Villanueva Balsera, J. M. (2020). Bidders recommender for public procurement auctions using machine learning: data analysis, algorithm, and case study with tenders from Spain. *Complexity*, *2020*.

ACEPTACIÓN:

Acepto que las publicaciones anteriores formen parte de la tesis doctoral titulada:
Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning

Y elaborada por D. Manuel J. García Rodríguez

En Oviedo, 31 de mayo de 2022

Firma

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

ACEPTACIÓN COAUTORES PRESENTACIÓN TRABAJOS FORMANDO PARTE DE TESIS DOCTORAL COMO COMPENDIO DE PUBLICACIONES

1.- Datos personales del coautor		
Apellidos: Villanueva Balsera	Nombre: Joaquin Manuel	
DNI/Pasaporte/NIE 09384725-N	Teléfono 626747750	Correo electrónico jmvillanueva@uniovi.es

2.- Publicaciones que formarán parte de la tesis y de las que es coautor

García Rodríguez, M. J., Rodríguez Montequín, V., Ortega Fernández, F., & Villanueva Balsera, J. M. (2019). Public procurement announcements in spain: regulations, data analysis, and award price estimator using machine learning. *Complexity*, *2019*.

García Rodríguez, M. J., Rodríguez Montequín, V., Ortega Fernández, F., & Villanueva Balsera, J. M. (2020). Bidders recommender for public procurement auctions using machine learning: data analysis, algorithm, and case study with tenders from Spain. *Complexity*, *2020*.

ACEPTACIÓN:

Acepto que las publicaciones anteriores formen parte de la tesis doctoral titulada:
Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning

Y elaborada por D. Manuel J. García Rodríguez

Firma VILLANUEVA BALSERA JOAQUIN MANUEL - 09384725N

Firmado digitalmente por VILLANUEVA BALSERA JOAQUIN MANUEL - 09384725N
Fecha: 2022.05.26 11:58:36 +02'00'

Oviedo, 26/05/2022

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

ACEPTACIÓN COAUTORES PRESENTACIÓN TRABAJOS FORMANDO PARTE DE TESIS DOCTORAL COMO COMPENDIO DE PUBLICACIONES

1.- Datos personales del coautor		
Apellidos: Sierra Araujo	Nombre: Basilio	
DNI/Pasaporte/NIE 15991574L	Teléfono 943015102	Correo electrónico b.sierra@ehu.eus

2.- Publicaciones que formarán parte de la tesis y de las que es coautor

RODRIGUEZ, M. J. G., MONTEQUIN, V. R., UBIERNA, A. A., HERMIDA, R. S., ARAUJO, B. S., & JAUREGI, A. Z. (2021). Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain. *Studies in Informatics and Control*, *30*(4), 67-76.

ACEPTACIÓN:

Acepto que las publicaciones anteriores formen parte de la tesis doctoral titulada:
Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning

Y elaborada por D. Manuel J. García Rodríguez

Donostia-San Sebastián, a 26 de mayo de 2022

Firma

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

ACEPTACIÓN COAUTORES PRESENTACIÓN TRABAJOS FORMANDO PARTE DE TESIS DOCTORAL COMO COMPENDIO DE PUBLICACIONES

1.- Datos personales del coautor		
Apellidos: Zelaia Jauregi	Nombre: Ana	
DNI/Pasaporte/NIE 72443566Y	Teléfono 943 01 5033	Correo electrónico ana.zelaia@ehu.eus

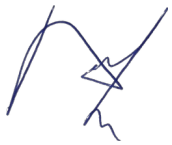
2.- Publicaciones que formarán parte de la tesis y de las que es coautor

RODRIGUEZ, M. J. G., MONTEQUIN, V. R., UBIERNA, A. A., HERMIDA, R. S., ARAUJO, B. S., & JAUREGI, A. Z. (2021). Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain. *Studies in Informatics and Control*, *30*(4), 67-76.

ACEPTACIÓN:

Acepto que las publicaciones anteriores formen parte de la tesis doctoral titulada:
Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning

Y elaborada por D. Manuel J. García Rodríguez



Donostia/San Sebastián,a 26 de mayo de 2022

Firma: Ana Zelaia Jauregi

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

ACEPTACIÓN COAUTORES PRESENTACIÓN TRABAJOS FORMANDO PARTE DE TESIS DOCTORAL COMO COMPENDIO DE PUBLICACIONES

1.- Datos personales del coautor		
Apellidos: Aranguren Ubierna	Nombre: Andoni	
DNI/Pasaporte/NIE 78994616N	Teléfono 626504049	Correo electrónico andoni.aranguren@gmail.com

2.- Publicaciones que formarán parte de la tesis y de las que es coautor

RODRIGUEZ, M. J. G., MONTEQUIN, V. R., UBIERNA, A. A., HERMIDA, R. S., ARAUJO, B. S., & JAUREGI, A. Z. (2021). Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain. *Studies in Informatics and Control*, *30*(4), 67-76.

ACEPTACIÓN:

Acepto que las publicaciones anteriores formen parte de la tesis doctoral titulada:
Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning

Y elaborada por D. Manuel J. García Rodríguez

Firma 

Sodupe, 26 de mayo de 2022

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

RENUNCIA COAUTORES PRESENTACIÓN TRABAJOS FORMANDO PARTE DE TESIS DOCTORAL

1.- Datos personales del coautor		
Apellidos: Aranguren Ubierna	Nombre: Andoni	
DNI/Pasaporte/NIE 78994616N	Teléfono 626504049	Correo electrónico andoni.aranguren@gmail.com

2.- Tesis Doctoral
Título: Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning
Autor: Manuel José García Rodríguez
Programa de doctorado: Programa de Doctorado en Ingeniería de Producción, Minero-Ambiental y de Proyectos

2.- Publicaciones que formarán parte de la tesis y de las que es coautor
RODRIGUEZ, M. J. G., MONTEQUIN, V. R., UBIERNA, A. A., HERMIDA, R. S., ARAUJO, B. S., & JAUREGI, A. Z. (2021). Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain. *Studies in Informatics and Control*, *30*(4), 67-76.

RENUNCIA:
Renuncio a que las publicaciones anteriores sean presentadas como parte de otra tesis doctoral presentada como compendio de publicaciones.
Firma  Sodupe, 26 de mayo de 2022

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

ACEPTACIÓN COAUTORES PRESENTACIÓN TRABAJOS FORMANDO PARTE DE TESIS DOCTORAL COMO COMPENDIO DE PUBLICACIONES

1.- Datos personales del coautor		
Apellidos: Santana Hermida	Nombre: Roberto	
DNI/Pasaporte/NIE 73031141T	Teléfono: 943018556	Correo electrónico: roberto.santana@ehu.es

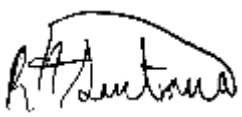
2.- Publicaciones que formarán parte de la tesis y de las que es coautor

RODRIGUEZ, M. J. G., MONTEQUIN, V. R., UBIERNA, A. A., HERMIDA, R. S., ARAUJO, B. S., & JAUREGI, A. Z. (2021). Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain. *Studies in Informatics and Control*, *30*(4), 67-76.

ACEPTACIÓN:

Acepto que las publicaciones anteriores formen parte de la tesis doctoral titulada:
Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning

Y elaborada por D. Manuel J. García Rodríguez



San Sebastian, 29-05-2022

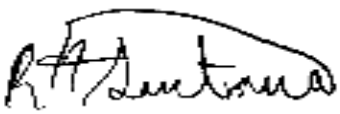
Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

RENUNCIA COAUTORES PRESENTACIÓN TRABAJOS FORMANDO PARTE DE TESIS DOCTORAL

1.- Datos personales del coautor		
Apellidos: Santana Hermida	Nombre: Roberto	
DNI/Pasaporte/NIE 73031141T	Teléfono 943018556	Correo electrónico roberto.santana@ehu.eus

2.- Tesis Doctoral
Título: Las licitaciones públicas: análisis de datos y sistemas predictores utilizando métodos de machine learning
Autor: Manuel José García Rodríguez
Programa de doctorado: Programa de Doctorado en Ingeniería de Producción, Minero-Ambiental y de Proyectos

2.- Publicaciones que formarán parte de la tesis y de las que es coautor
RODRIGUEZ, M. J. G., MONTEQUIN, V. R., UBIERNA, A. A., HERMIDA, R. S., ARAUJO, B. S., & JAUREGI, A. Z. (2021). Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain. *Studies in Informatics and Control*, *30*(4), 67-76.

RENUNCIA:	
Renuncio a que las publicaciones anteriores sean presentadas como parte de otra tesis doctoral presentada como compendio de publicaciones.	
 Firma	San Sebastián, 29-05-2022