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

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

Copyright © 2020 Manuel J. García Rodríguez et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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].

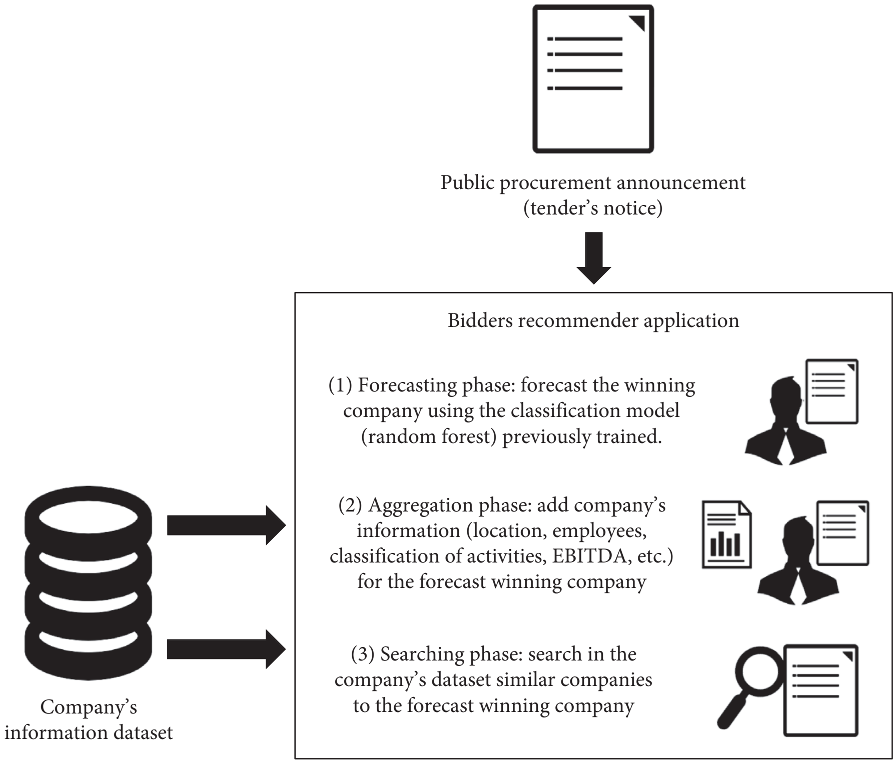

Public procurement announcement
(tender's notice)
Bidders recommender application
(1) Forecasting phase: forecast the winning company using the classification model (random forest) previously trained.
(2) Aggregation phase: add company's information (location, employees, classification of activities, EBITDA, etc.) for the forecast winning company
(3) Searching phase: search in the company's dataset similar companies to the forecast winning company
Company's
information dataset

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:

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

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 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 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 $(1 \leq i \leq n)$, y_i be the corresponding true value, ϖ_i be the corresponding sample weight, and L be the set of classes $(1 \leq l \leq L)$. Accuracy (2) is the proportion of correct predictions over n samples:

$$\text{accuracy} = \frac{1}{\text{n}} \sum_{i=1}^{n} 1\left(\hat{y}_i = y_i\right), \qquad (2)$$

where $1\left(\hat{y}_i\right)$ 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:

$$\text{balanced accuracy} = \frac{1}{\sum_{i=1}^{n} \hat{\varpi}_i} \sum_{i=1}^{n} 1\left(\hat{y}_i = y_i\right).\varpi_i, \qquad (3)$$

where $1\left(\hat{y}_i\right)$ is the indicator function and $\hat{\varpi}_i = \varpi_i / \sum_{j=1}^{n} 1 \left(\hat{y}_j = y_i\right).\varpi_j$.

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

$$\text{precision} = \frac{1}{L} \sum_{l=1}^{L} \frac{\left|y_l \cap \hat{y}_l\right|}{\left|y_l\right|}. \qquad (4)$$

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, $\widehat{C}_{rf}^B(x) = \text{majority vote}\left\{\widehat{C}_b(x)\right\}_1^B$.
```

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 $z_i = (y_i, \widehat{y}_i)$. 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) $\text{OperatingIncome}_{\text{co.}} \geq F_{\text{OI}} \cdot \text{OperatingIncome}_{\text{forecastco.}} \cdot$

(b) $\text{EBIT}_{\text{co.}} \geq F_{\text{EBIT}} \cdot \text{EBIT}_{\text{forecast co.}} \cdot$

(c) $\text{EBITDA}_{\text{co.}} \geq F_{\text{EBITDA}} \cdot \text{EBITDA}_{\text{forecast co.}} \cdot$

(d) $\text{Employees}_{\text{co.}} \geq F_E \cdot \text{Employees}_{\text{forecast co.}} \cdot$

(e) $\sum_{i=1}^{C} 1\left[\{\text{Code}\}_{\text{co.}} = \{\text{Code}\}_{\text{forecast co.}}\right] \geq F_{CEA} \cdot C$ 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:

$$\text{Code} = \{\text{NACE2, IAE, SIC and NAICS}\}.$$

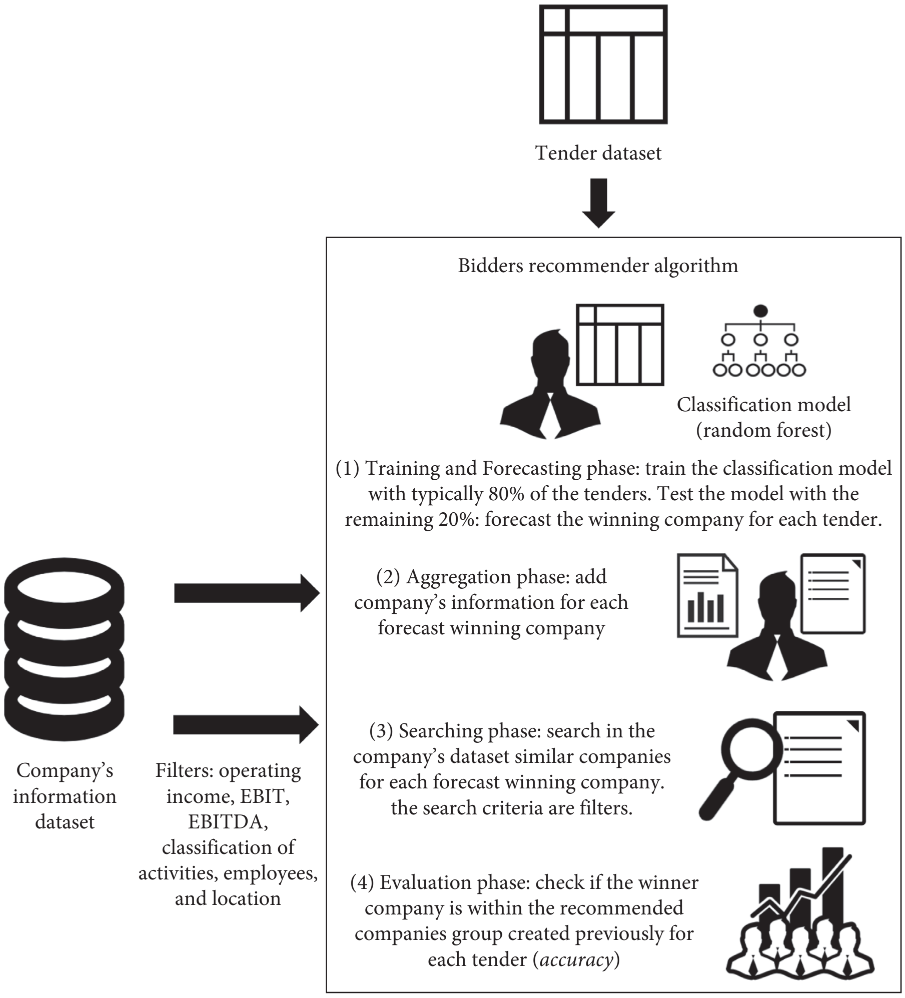

Tender dataset
Bidders recommender algorithm
Classification model
(random forest)
(1) Training and Forecasting phase: train the classification model with typically 80% of the tenders. Test the model with the remaining 20%: forecast the winning company for each tender.
(2) Aggregation phase: add company's information for each forecast winning company
(3) Searching phase: search in the company's dataset similar companies for each forecast winning company. the search criteria are filters.
(4) Evaluation phase: check if the winner company is within the recommended companies group created previously for each tender (*accuracy*)
Company's information dataset
Filters: operating income, EBIT, EBITDA, classification of activities, employees, and location

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

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

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 $\text{Operating\_Income}_{\text{co.}}$, $\text{EBIT}_{\text{co.}}$, 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: $\text{Operating\_Income}_{\text{co.}}$, $\text{EBIT}_{\text{co.}}$, 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: $\text{Distance}_{\text{tender−co.}}$. 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.

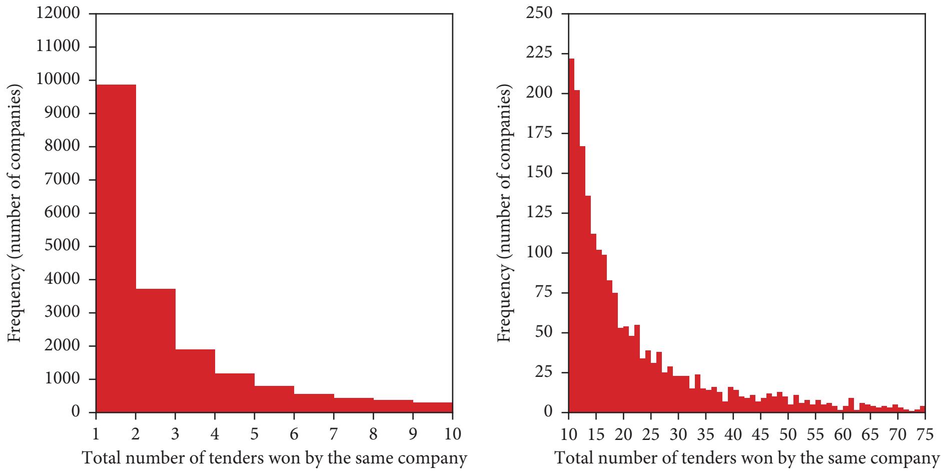
Topic	Description	Value
General values	Total number of tenders in the dataset	102,087
	Temporal range of tenders	2014/01/02–2020/03/31
	Total number of tendering organisations	5,761
	Total number of winning companies	21,268
	Mean number of offers received per tender	4.38
	Mean duration of tender's works	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%)



(a)

(b)

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.

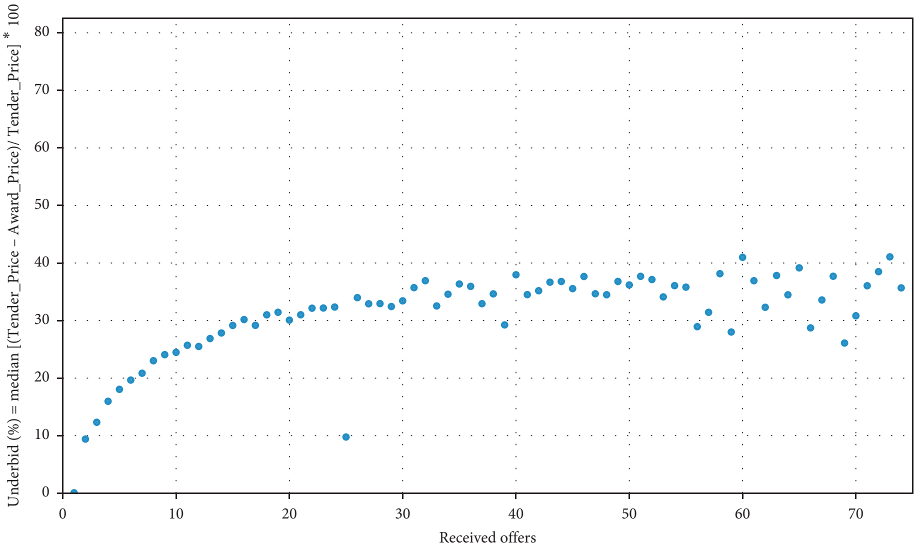


Figure 4: Relation between the received offers of bidders and the underbid (median absolute percentage error between tender price and award price).

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

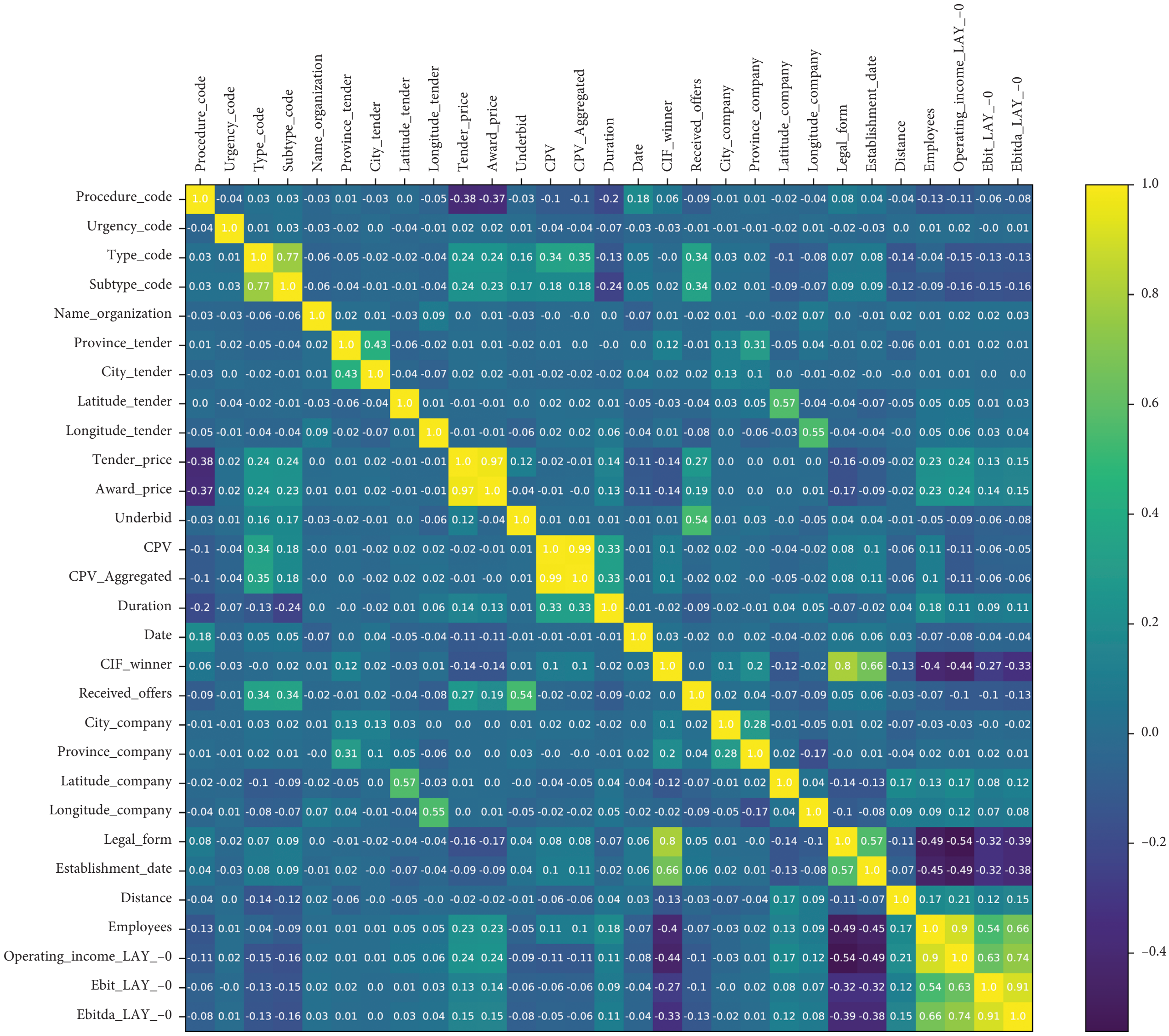
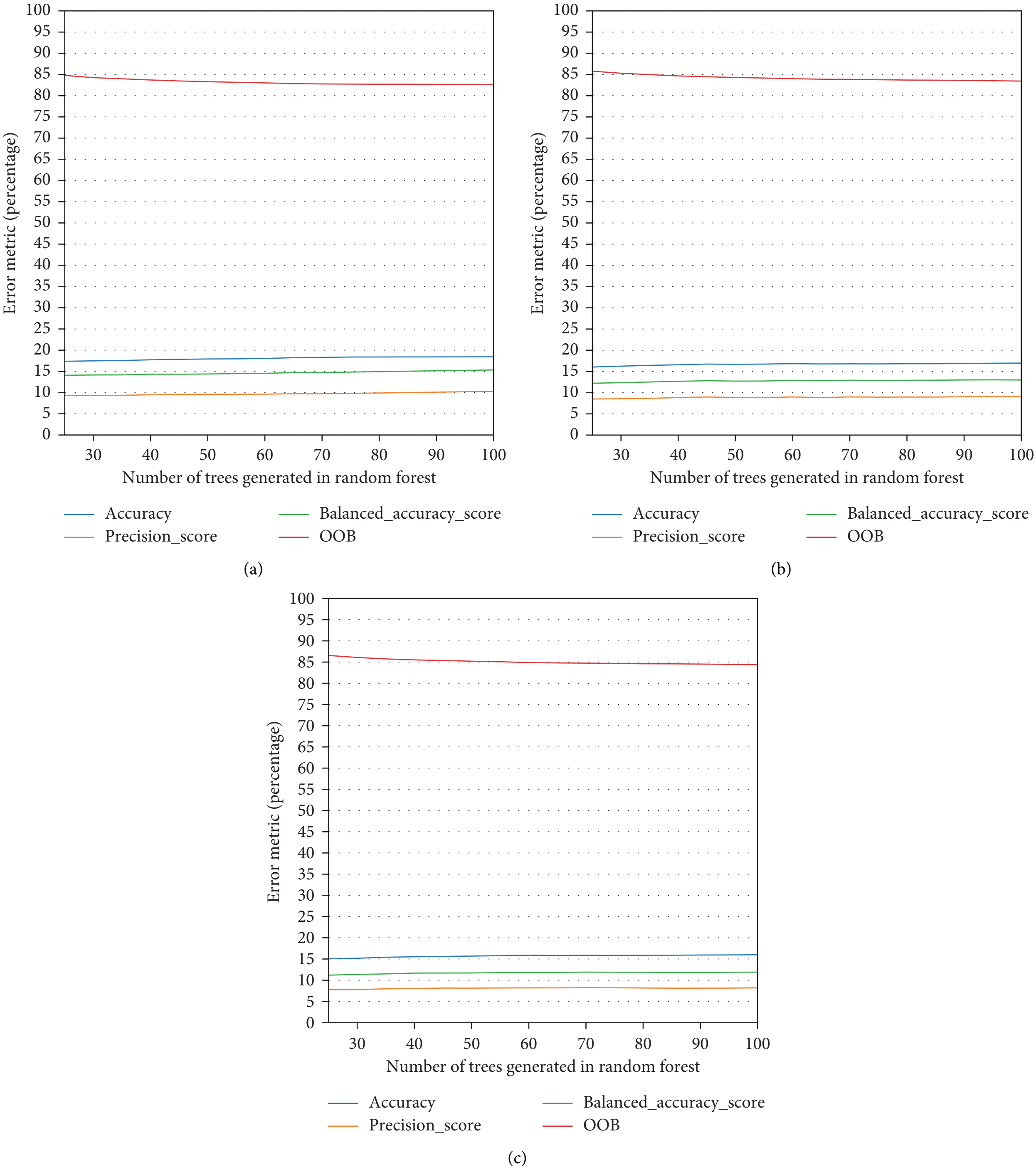


Figure 5: Correlation matrix between the variables of the two datasets (tenders and companies). Spearman's rank correlation coefficient is the method applied.

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



(a)

(b)

(c)

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	Accuracy$_{n=1}$: winner company is the forecast company	17.07%				
	Accuracy$_{n=5}$: winner company is within the top 5 forecast companies	31.58%				
	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	Accuracy$_{n=1}$: winner company is the forecast company	10.25%				
	Accuracy$_{n=5}$: winner company is within the top 5 forecast companies	23.12%				
	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 Accuracy$_{n=5}$). This means that the 5 most probable companies to win a tender can be incorporated to the recommender companies group (called Accuracy$_{n=M}$). For scenario 1, Accuracy$_{n=5}$ = 31.58% and Accuracy$_{n=M}$ = [33.25% − 38.52%]. For scenario 2, Accuracy$_{n=5}$ = 23.12% and 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

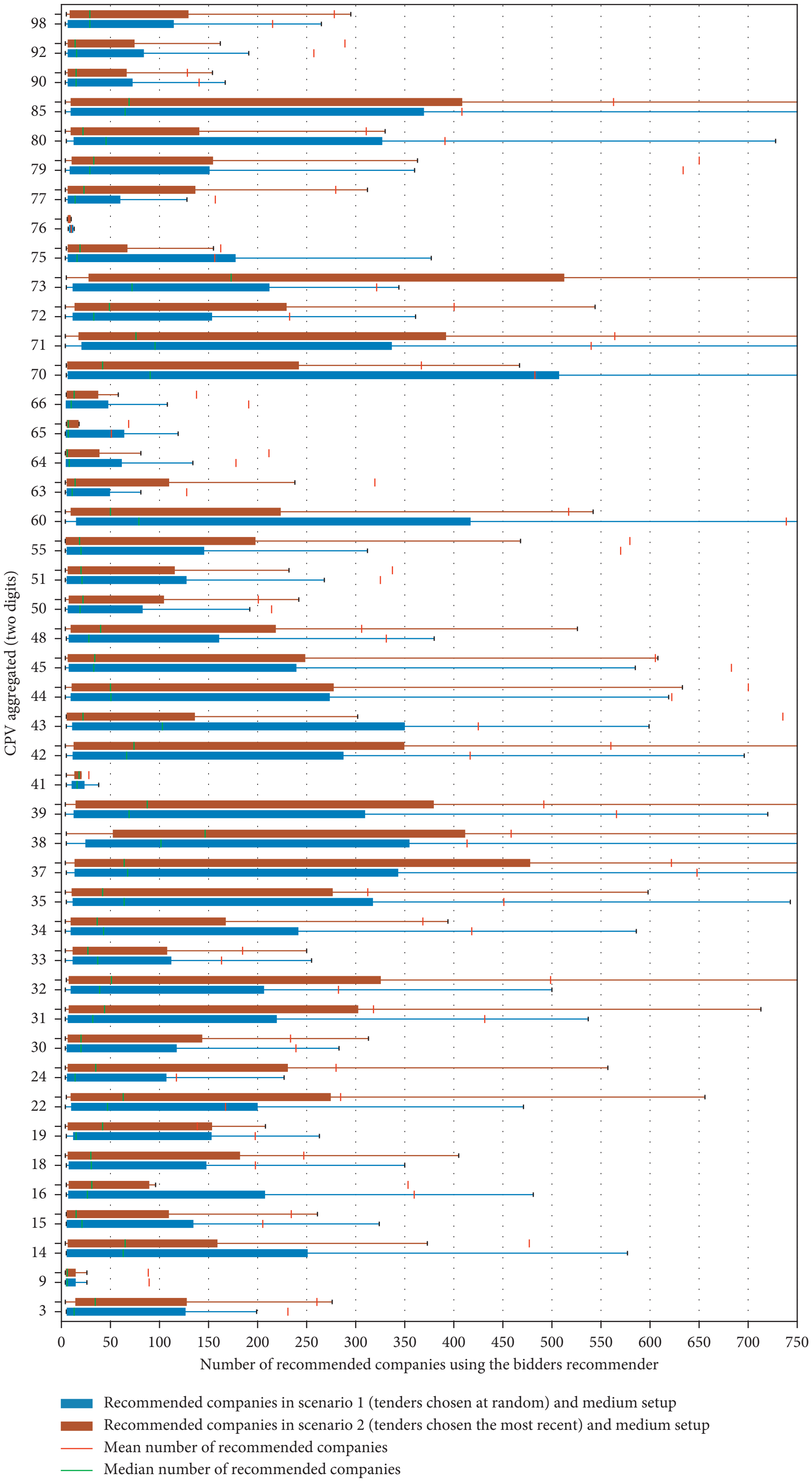


FIGURE 7: Boxplots for the size of the recommended companies group generated by the bidders recommender, disaggregated by CPV. Scenario 1 (blue colour) and scenario 2 (brown colour) both have a medium setup.

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.