UNIVERSIDAD DE OVIEDO

DEPARTAMENTO DE EXPLOTACIÓN Y PROSPECCIÓN DE MINAS

MÁSTER INTERUNIVERSITARIO EN "DIRECCIÓN DE PROYECTOS"

TRABAJO FIN DE MÁSTER

(Curso 2016 - 2017)

ADAPTACIÓN DE ESTÁNDARES DE DIRECCIÓN DE PROYECTOS PARTICULARIZADOS PARA LA MINERÍA DE DATOS

AUTOR: Julio César Mérida Sánchez

DIRECTOR: Joaquín M. Villanueva Balsera

CODIRECTOR: Jose Manuel Mesa

Índice de contenidos

1 INTRODUCCIÓN

Hoy en día el campo de la minería y el análisis de datos se encuentra en auge. Cada vez más organizaciones recurren a la minería de datos para obtener ventajas sobre sus competidores, situación que se ve reflejada por el aumento de la contratación de personal especializado por parte de las empresas, independientemente de su área de negocio, así como por el caudal de dinero movido por la industria especializada.

Este auge del sector contrasta con el estancamiento de las metodología de gestión de proyectos de minería de datos, las cuales más de 20 años sin sufrir cambios significativos, a pesar de los cambios tecnológicos o los últimos movimientos del sector, el cual, en los últimos años, ha dejado de estar restringido a unos pocos profesionales del ramo de las "Tecnologías de la Información", siendo adoptado por empresas de los sectores más diversos, extendiéndose por todas las áreas departamentales de las organizaciones y gracias a los cambios tecnológicos, estas técnicas de gestión de datos han pasado incluso a tener aceptación por profesionales no específicos del sector.

Teniendo en cuenta que todos los datos indican que esta tendencia está en aumento parece, por tanto, conveniente revisar la vigencia de ***las metodologías tradicionales de gestión de proyectos de minería y análisis de datos***, con el fin de identificar sus debilidades y fortalezas, así como de verificar su vigencia en el marco actual.

1.1 MINERÍA DE DATOS, ANÁLISIS DE DATOS Y SUS CONDICIONANTES

Se cree necesario para el correcto entendimiento de este Trabajo de Final de Máster establecer la diferencia entre minería de datos y análisis de datos:

Minería de datos es el proceso mediante el que se extrae información de una fuente u origen de datos para posteriormente transformar los datos extraídos en un conjunto entendible y estructurado de forma que se permita el uso posterior de los mismos.

Análisis de datos es el proceso mediante el que se inspecciona, limpia, transforma y finalmente modela un conjunto de datos con el fin de "descubrir" información útil, extraer conclusiones o apoyar la toma de decisiones de forma informada.

Si bien hoy en día tiende a hablarse de ambos procesos en su conjunto, integrando ambos bajo el término "Minería de Datos" cabe destacar que las metodologías más empleadas en gestión de proyectos de Minería y Análisis de Datos tienden a diferenciar ambos conceptos, separándolos en ámbitos o pasos bien diferenciados dentro del proceso global de "Minería de Datos". Es por ello que, si bien los condicionantes que aceptan a este tipo de proyectos suelen ser agrupados, parece adecuado detallar que condicionantes afectan a cada uno de estos procesos, es decir separar los condicionantes que influyen sobre el proceso de minería de datos de aquellos que influyen sobre el análisis de los mismos.

Así tendremos que:

Los condicionantes que afectan al proceso de "minería de datos" son:

- **El tamaño del conjunto de datos.** Generalmente afecta a la capacidad de computación requerida, así como al tiempo necesario para extraer la información. Este condicionante desemboca en la siguiente clasificación de los conjuntos de datos:
  - **Small Data:** Conjuntos de tamaño inferior a 10 GB
  - **Medium Data**: Conjuntos de tamaño situado entre 10 GB y 1TB
  - **Big Data:** Conjuntos de tamaño superior a 1TB.
- **El origen de los datos:** Se refiere al medio en el que se encuentran presentes los datos a extraer. Esto incluye por un lado la localización de los mismos (internet, datos internos empresa, …) como el contenedor (archivos de texto, tablas de datos, archivos de audio, videos …)
- **La estructura en que se encuentran los datos:** Los datos pueden seguir algún tipo de ordenación (datos estructurados) o no tener ordenación aparente (datos no estructurados)
- **La frecuencia de actualización de los datos,** así como el volumen de los mismos: Los datos pueden proceder de orígenes fijos o cambiantes, como pudieran ser los datos meteorológicos, por ejemplo. Así mismo es también importante el volumen de datos obtenido en cada actualización de la fuente de origen.

Algunos de los condicionantes que afectan al proceso de "análisis de datos" son:

- **La "naturaleza" de los datos contenidos por el conjunto**, se refiere al grado de comprensión técnica necesaria para el entendimiento del conjunto de datos a analizar.
- **La homogeneidad de los datos** se refiere al grado de variación de los datos obtenidos.
- **Fiabilidad de los datos:** Los datos pueden ser cuantitativos o cualitativos y no siempre tienen porqué proceder de una fuente contrastada. El grado de fiabilidad de los datos puede por tanto influir sobre el análisis de los mismos.

1.2 EL CRECIENTE INTERÉS EN LA MINERÍA DE DATOS

Durante los últimos años el campo de la minería de datos ha adquirido gran importancia en numerosas áreas de negocio, industrias y corporaciones, dada su capacidad para tratar ingentes cantidades de datos cuyo volumen los hacía no aprovechables hace años.

Existe una cantidad creciente de datos procedentes de organizaciones e individuos. Las empresas tienen cada vez mayor a almacenar los datos de sus operaciones diarias, lo cual unidos a los datos de acceso público, los generados en las interacciones con sus clientes y los ofrecidos por entidades especializadas en la recopilación de información hacen que un flujo de información cada vez mayor esté disponible para las empresas, lo cual aumenta de forma

creciente el interés de estas en filtrar y analizar información relevante que les permita obtener ventajas competitivas sobre el mercado.

Algunos hechos que ponen en relieve el creciente interés en las ciencias de análisis de datos son los siguientes:

1. **Criterios económicos**

Los beneficios generados por el sector de Big Data se estimaron en 27,3 billones durante el ejercicio de 2015, con una previsión de crecimiento que alcanzará un beneficio de 92,2 billones de dólares en 2026. [1]

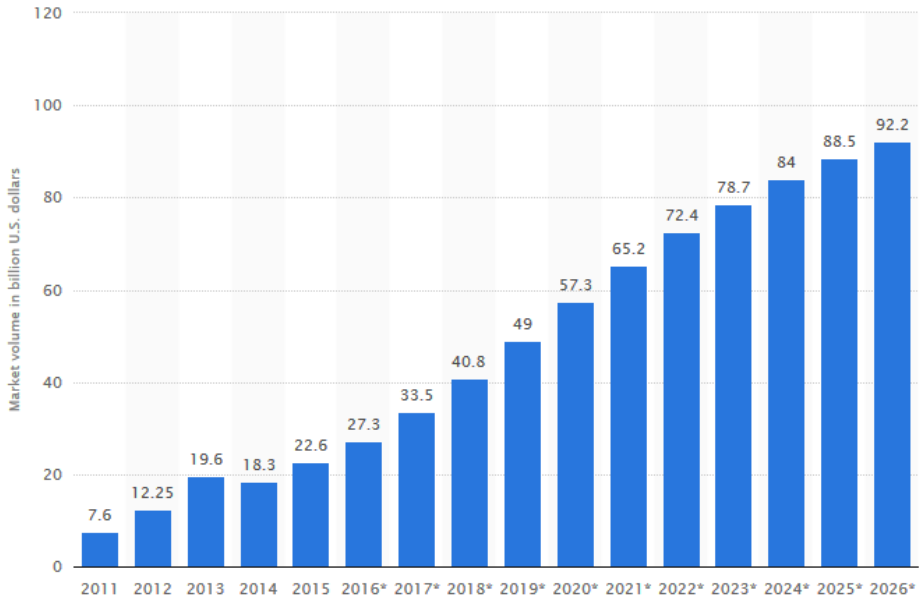

Figura 1. Beneficios generados por el mercado de "Big Data", en billones de dólares (statista.com; datos recogidos por Dr. Ralph Finos, Wikibon – 2017)

2. **Demanda de empleo intersectorial**

De acuerdo a datos publicados por Forbes en 2014 se produjo un incremento del 89,9% de puestos de trabajo en el campo de Big Data.

Si bien los sectores que más empleo producen en el campo siguen siendo sectores relacionados tradicionalmente con las ciencias de la información, las estadísticas de empleo por sectores muestran un interés creciente en áreas no tradicionales, con un llamativo tercer puesto del sector manufacturero (12,35% de nuevos empleos generados) [2]

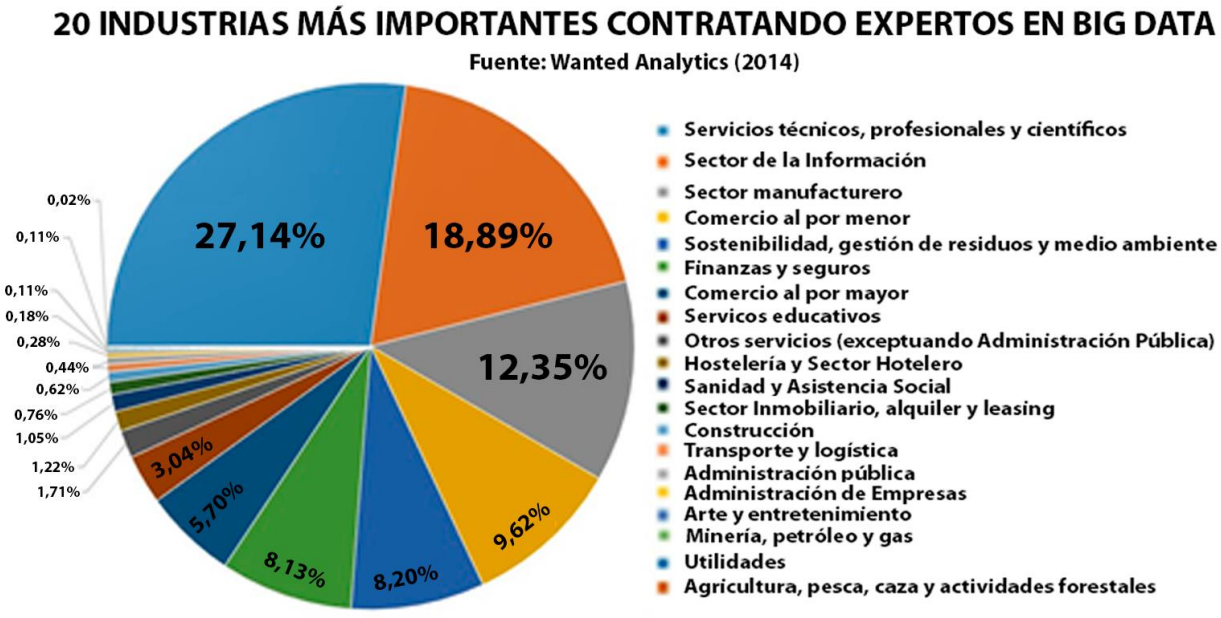

Figura 2. Distribución de empleos generados en el campo de "Big Data" distribuidos por sectores (Forbes – 2014)

3. Implantación interdepartamental

Una encuesta realizada a profesionales del sector en 2015 por la Evans Data Corporation revela que el empleo de técnicas de Big Data tiene una distribución uniforme entre varios departamentos de las empresas participes del estudio. [3]

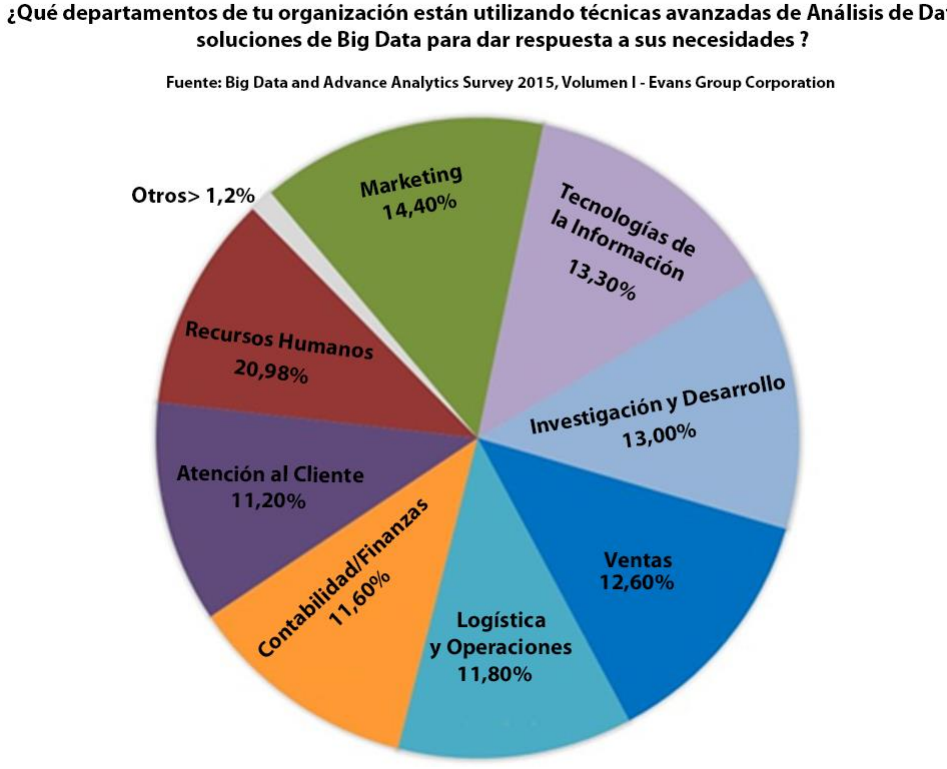

Figura 3. Distribución de departamentos que emplean Big Data entre las empresas encuestadas (Big Data and Advanced Analytics Survey 2015, Volume I - Evans Data Corporation)

4. **Aumento del número de datos almacenados en servidores y velocidad de creación de los mismos**

De acuerdo con un artículo publicado por Forbes entre 2014 y 2015 se crearon más datos que en el resto de la historia anterior del ser humano [4]. En 2013 existía una cantidad almacenada de 4,4 Zetabytes, cantidad equivalente a mil millones de Terabytes (10^{21}), con la previsión de alcanzar los 44 Zetabytes en 2020 [5] [6], fecha para la cual se espera alcanzar una velocidad de creación de datos de 1,7 MB por segundo.

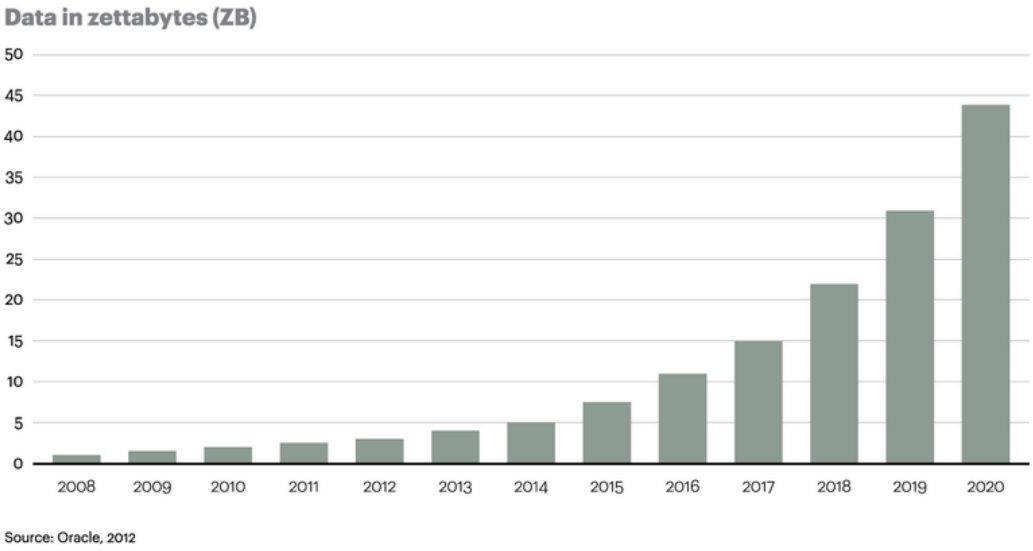

Figura 4. Evolución de los datos almacenados a nivel mundial (Oracle – 2012) *[7]*

5. **Aumento creciente del número de búsquedas, artículos científicos y libros registrados por Google sobre la temática**

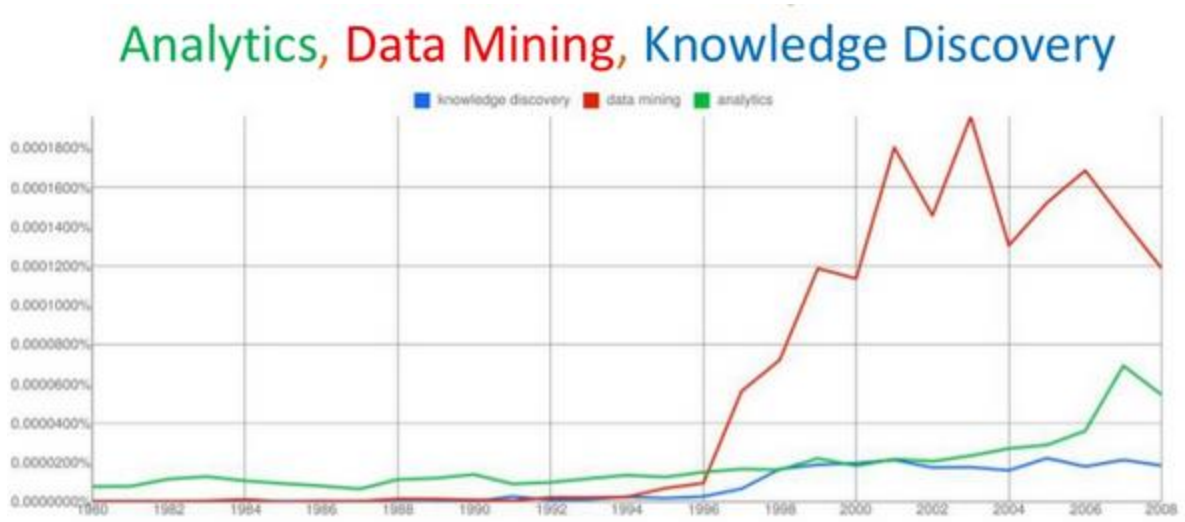

Figura 5. Evolución en la literatura científica de los términos relacionados con el análisis y minería de datos (Google Ngram Viewer)

Estos datos son consecuencia de la creciente cantidad de datos generados por organizaciones e individuos. El aumento de la disponibilidad de datos almacenados ha llevado en la última

década a la mejora de los procesos de "Minería de Datos". A su vez el éxito en la gestión y análisis de los mismos ha alimentado la necesidad de empresas y gobiernos de almacenar cada vez un número mayor y más variado de datos con el fin de obtener información relevante que se traduzca en ventajas competitivas sobre sus competidores.

Como respuesta a esta situación en la última década hemos visto un refinamiento cada vez mayor de las herramientas que facilitan el análisis de los datos obtenidos, aportando una gran cierta facilidad al proceso (lenguajes accesibles, librerías de programación especializadas, aplicaciones Ready To Use). Esto junto a la disponibilidad de soluciones de computación cada vez más potentes y económicas han facilitado la adopción de las técnicas de minería y análisis de datos de un número creciente de técnicos y organizaciones fuera del sector de tecnologías de la información, facilitando la incorporación de estas técnicas en ámbitos y proyectos cada vez más extensos.

Si bien parte de estos proyectos se realizan mediante su externalización empresas especializadas otros muchos se realizan dentro de la misma organización, integrándose dentro de proyectos más amplios y usando recursos internos para la consecución de estas tareas, lo cual unido a la propia naturaleza de los datos obtenidos, los cuáles muchas veces requieren conocimientos específicos para su discretización y análisis, conllevan una presencia creciente de equipos multidisciplinares en estas organizaciones, ya que la comprensión de la información obtenida puede caer fuera del área de conocimiento de parte de los miembros involucrados en el proyecto.

Es por estos motivos, anteriormente expuestos, que pudiera ser conveniente revisar las metodologías que actualmente se consideran "buenas prácticas" en la gestión de proyectos de minería y análisis de datos con el fin de verificar la vigencia y adaptación de los conceptos establecidos por los mismos a la realidad creciente hacia la que se dirige parte de la industria especializada.

2 OBJETO Y ALCANCE

El objetivo del presente "Trabajo de Final de Master", en adelante TFM, es ***proponer una técnica de gestión de proyectos capaz de abordar proyectos multidisciplinares***.

Para ello se establecen una serie de metas que faciliten la consecución de dicho objetivo, siendo estas:

1. Analizar las técnicas de gestión de proyectos de minería y análisis de datos empleadas en el sector, enumerando sus virtudes y debilidades.
2. Determinar los puntos que se cree debe de satisfacer una metodología de minería de datos para abordar con éxito proyectos multidisciplinares.
3. Dar respuesta a los puntos críticos inferidos, estableciendo un método adecuado que permita abordar con éxito proyectos multidisciplinares.

3 ESTADO DEL ARTE

3.1 METODOLOGÍAS TRADICIONALES DE MINERÍA DE DATOS

Cuando hablamos de "Data Mining" existen tres metodologías ampliamente utilizadas en el sector, estas son CRISP-DM, SEMMA y KDD. Estas tres metodologías han sido comparadas en numerosos artículos científicos y suelen ser las metodologías de uso genérico más utilizadas por los expertos, tal como suelen revelar las encuestas publicadas en el sitio web kdnuggets, uno de los mayores puntos de encuentro en internet entre expertos en la materia.

Las encuestas realizadas por kdnuggets entre 2002 y 2014 tienden a reflejar el dominio de CRISP-DM como metodología más empleada en la gestión de proyectos de Data Mining, seguida de SEMMA y KDD, no obstante, existe un porcentaje importante de expertos que emplea su propia metodología, una establecida por la organización en la que trabaja u otras metodologías de menor calado.

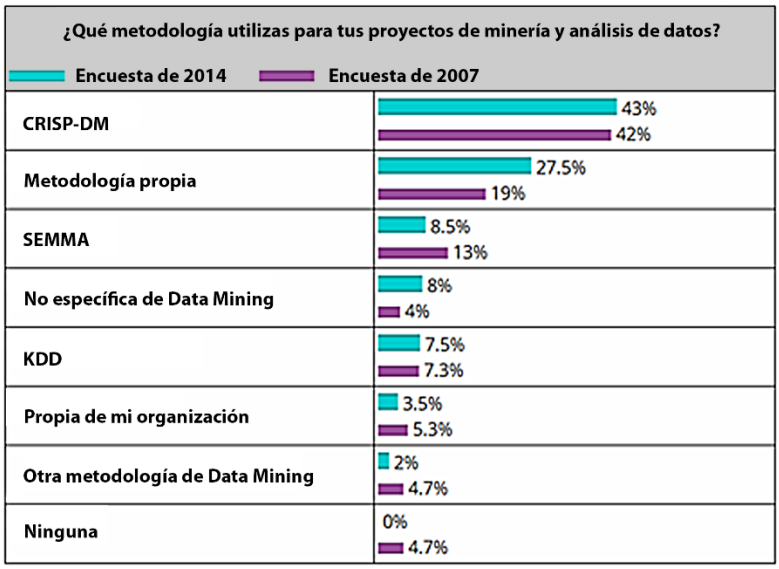

Figura 6. Evolución de los datos almacenados a nivel mundial (Oracle – 2012) *[7]*

En apartados sucesivos se describirán estas tres metodologías de unos tradicional en proyectos de "Data Mining".

3.1.1 KDD (KNOWLEDGE DISCOVERY IN DATABASES)

KDD es un método que empezó a gestarse en 1989, mediante la organización del "Grupo de Trabajo" denominado "Knowledge Discovery in Databases" (descubrimiento de conocimiento en bases de datos) durante la Conferencia Internacional de Inteligencia Artificial celebrada en ese mismo año en Detroit. Este grupo de trabajo, habitualmente referido como KDD-89, fue organizado por Gregory Piatetsky-Shapiro, quien decidió utilizar este término para diferenciar el objetivo del grupo de los métodos que se empleaban hasta la fecha en minería de datos, ampliamente basados en árboles de conocimiento. [8]

El grupo de trabajo se siguió estableciendo con carácter anual, derivándose los primeros resultados interesantes de estas reuniones en 1994, mediante la publicación del artículo *"The Process of Knowledge Discovery in Databases: A First Sketch"* [9] por Ronald J. Brachman y Tej

Anand. En este artículo se establece un borrador inicial del método de descubrimiento de conocimiento en bases de datos, esquema del cual se muestra a continuación.

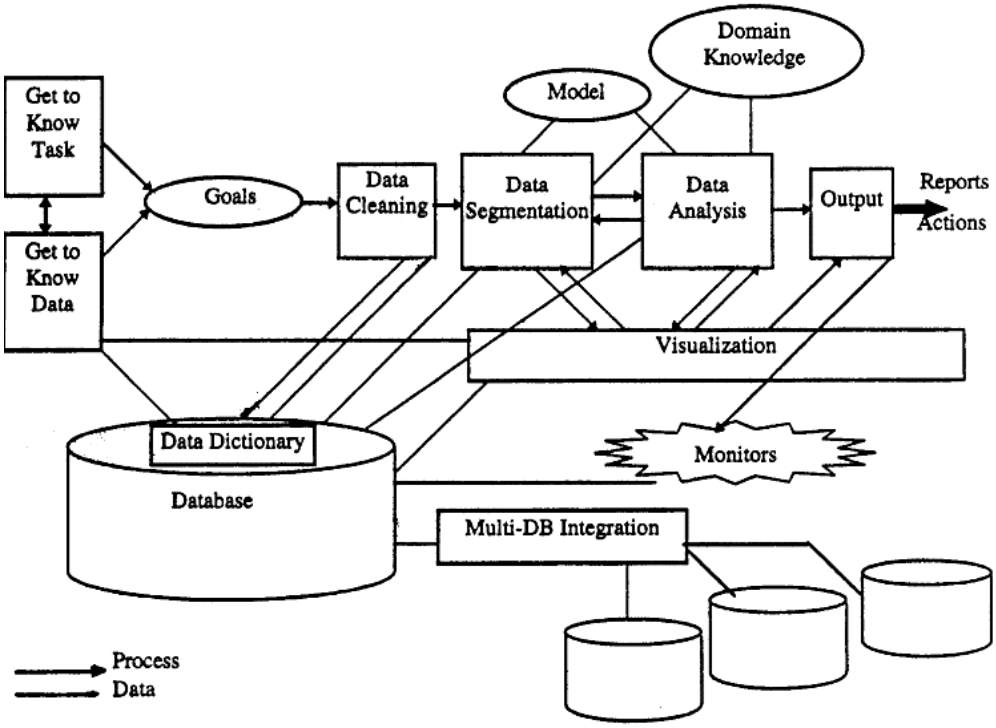

Figura 7. Esbozo inicial de KDD. Brachman et al., 1994

A partir de la descripción del procedimiento dada por ***Brachman et al.,*** pueden extraerse los siguientes procesos y esquema simplificado, con el fin de facilitar el entendimiento del mismo.

1. **Descubrimiento.** Es la etapa inicial, consiste en un análisis previo de la base de datos para hacerse una idea del conocimiento que puede ser extraído de la misma
2. **Determinación de "Objetivos".** Es un proceso donde el analista se sienta con el cliente, una vez se hace una idea previa del "conocimiento" que puede ser extraído de la base datos, con el fin de marcar los objetivos del proyecto.
3. **Modelado de datos.** El modelado la extracción, modelado y limpieza de los datos en conjuntos entendibles.
4. **Análisis de datos.** El análisis de datos consiste en el entendimiento de los datos extraídos, así como la comparación entre diversos grupos de datos con el fin de extraer
5. **Visualización de resultados**. Brachman et al consideran importante la visualización de resultados como parte del entendimiento de los datos. Como tal esa sirve para replantearse el conocimiento extraíble de los datos, así como los objetivos del proyecto. Así mismo es también parte de los entregables finales del proyecto.
6. **Resultados.** Consiste en la toma de conclusiones y elaboración de entregables finales del proyecto.
7. **Integración.** Consiste en la implementación de medidas en la organización en base a los resultados extraídos. (esto será aplicable o no dependiendo del tipo de proyecto)
8. **Monitorización.** Consiste en la observación de cambios de los datos en el origen de datos (en caso de que estos sean dinámicos), y en el control de las medidas implementadas en la

organización, en caso de que estas se produzcan. Este proceso realimenta al proceso de descubrimiento pudiendo derivar en cambios en el proyecto o en nuevos proyectos.

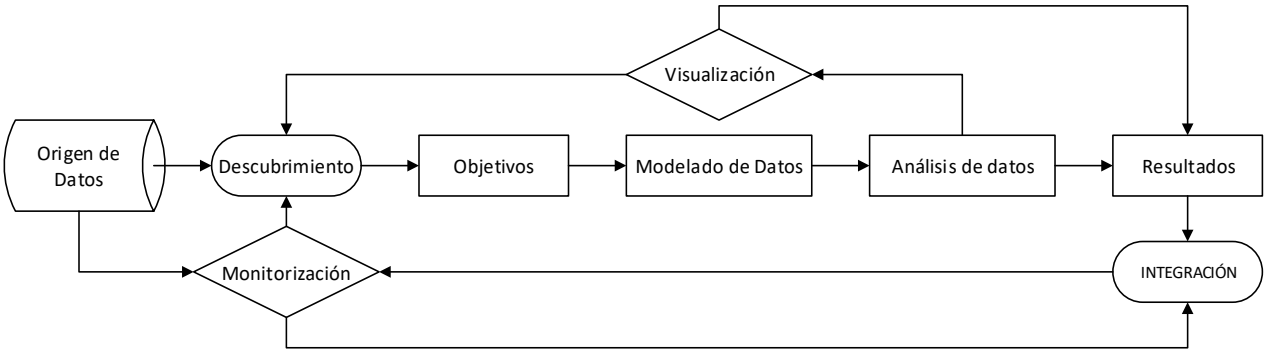

Figura 8. Esquema simplificado del procedimiento propuesto por Brachman et al.

Posteriormente, en 1996 Usama Fayyad, con la colaboración de Gregory Piatetsky-Shapiro y Padhraic Smyth le dieron la forma al procedimiento de KDD que se usa de forma extendida en la actualidad, en el artículo "From Data Mining to Knowledge Discovery in Databases" [10]. El diagrama de procesos propuesto en este artículo, de uso en la actualidad se muestra a continuación.

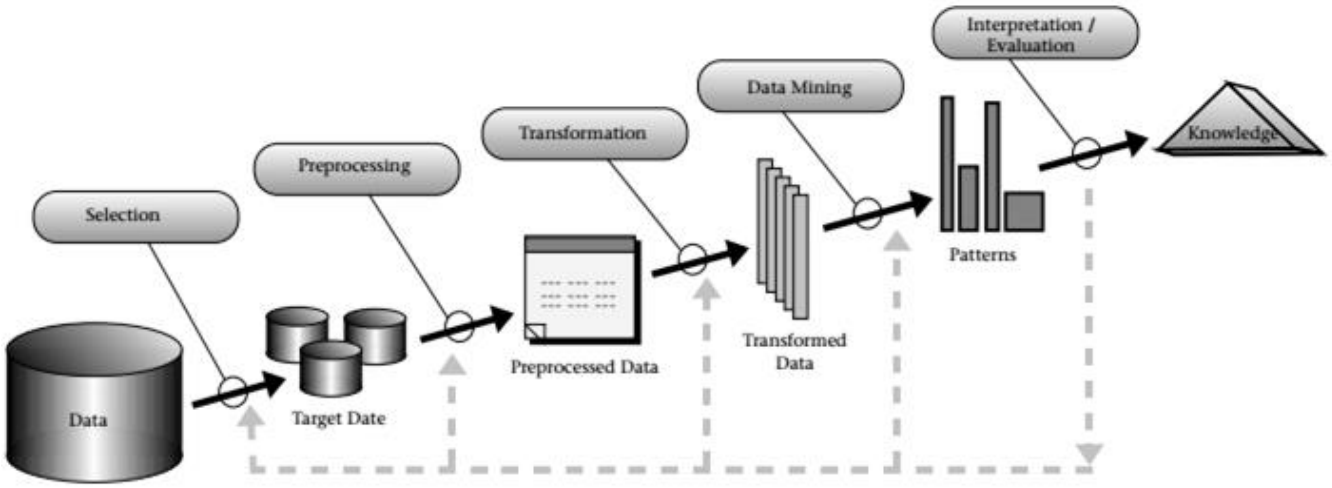

Figura 9. Diagrama de procesos de KDD, implementación actual. (Fayyad et al., 1996)

Para Fayad et al. KDD es el procedimiento que permite extraer conocimientos útiles de las bases de datos estudiadas, mientras que la "minería de datos" debe de referirse simplemente a una parte del procedimiento, dándole por tanto categoría de proceso. Estos autores inciden en que la aplicación de técnicas de minería de datos de forma ciega, sin el correcto entendimiento de los mismos puede llevar a la extracción de "patrones de conocimiento" irrelevantes o incorrectos.

Otro punto interesante es el reconocimiento de la interdisciplinaridad del procedimiento de "extracción de conocimiento en bases de datos", aunque se solo tienen en cuenta distintas ramas de las ciencias de la información (Inteligencia Artificial, Machine Learning, Reconocimiento de Patrones, Métodos Estadísticos, …)

Los pasos propuestos por Fayyad et al. son los siguientes:

0. **Análisis previo:** Aunque Fayyad et al. no incluyen este proceso en el diagrama propuesto, definen la necesidad de una etapa inicial de aproximación al problema. Durante esta etapa debe de adquirirse el conocimiento relevante a los datos estudiados, así como establecer los objetivos en colaboración con el cliente.
1. **Selección:** El segundo paso es a partir de la base de datos disponibles crear una o varias series de datos a analizar. Esta etapa termina con el entregable **"Datos a Analizar" (Target Data)**.
2. **Pre-procesado:** Este proceso consiste en la limpieza y preprocesado de las series de datos a analizar (eliminación de datos no-válidos, Manejo de series incompletas, ordenación de datos…) Esta etapa termina con el entregable **"Datos Preprocesados" (Preprocessed Data).**
3. **Transformación:** Este proceso consiste en la reducción de los datos contenidos en la serie de datos, uso de técnicas de representación dependiendo del objetivo final, … Esta etapa termina con el entregable **"Datos Transformados" (Transformed Data)**.
4. **Minería de Datos:** Este proceso consiste en coger los datos transformados y someterlos a las técnicas de minería de datos que sean relevantes respecto al objetivo final del proyecto y a la tipología de los datos. Esta etapa termina con el entregable **"Patrones" (Patterns)**.
5. **Interpretación/Evaluación:** Una vez obtenidos los patrones existentes en los datos comienza el proceso de análisis mediante el cual deben extraerse conclusiones de los mismos. Dependiendo de este proceso puede ser necesario volver a realizar uno o más de los procesos anteriores (1 a 4). Una vez que estemos satisfechos con las conclusiones extraídas finalizaremos el procedimiento de extracción de conocimiento con el entregable **"Conocimiento".**
6. **Actuar sobre el conocimiento: Aunque tampoco se incluye en el diagrama de proceso** Fayyad et al. mencionan que una vez extraído el conocimiento debe existir una etapa posterior en el que se busca un uso para el conocimiento extraído. Este proceso debe incluir el contraste de resultados con el conocimiento previamente extraído o las creencias anteriores sobre la materia.

Por último, cabe destacar que Fayyad et al. reconocen que dependiendo del tipo de proyecto a abordar puede ser necesario establecer iteraciones o bucles entre distintos procesos, modificando así el esquema propuesto. De este modo una de las representaciones habituales de KDD pasa de 6 pasos a 9 pasos, consolidándose la fase de análisis previo (paso 0) y dividiendo la fase de data mining (paso 4) en tres pasos (elección de la tarea de minería de datos, elección del algoritmo de minería de datos, minería de datos).
Si valoramos el diagrama propuesto por Fayyad et al., de uso normalizado en la actualidad, con el propuesto por Brachman et al. si bien son ampliamente similares, existen unas pequeñas diferencias en los objetivos y concepto de los mismo.

- Lo primero en que hemos de fijarnos es en que Fayyad et al. establecen entregable ente los procesos propuesto, mientras que Brachman et al. no entran en este detalle.
- El segundo punto observable es que Brachman et al. le dan importancia al proceso de "Descubrimiento", el cual dividen en dos sub-procesos ("get to know the data", "get to

know the task"). Fayyad et al. lo mencionan como importante pero no lo añaden a su diagrama y por tanto al proceso en sí.

Esto es importante ya que Brachman et al. establecen el destino de sus realimentaciones hacia el proceso de descubrimiento.

- Brachman et al. le dan especial importancia a la monitorización y visualización de datos. Fayyad et al. no describen un proceso de control (monitorización) e incluyen la visualización como una herramienta más de la transformación de datos.

En general el procedimiento de Brachman et al. parece más enfocado a un entorno dinámico donde los datos se actualizan constantemente y es necesario hacer reajustes en el método de obtención de los mismos, mientras que el propuesto por Fayyad et al., estándar de la metodología KDD actual, parece enfocado a la obtención de datos en bases estáticas. Por lo general el primero sería un proceso continuo, mientras que el segundo tendría un final, estableciéndose un nuevo proyecto en caso de que cambien los datos a analizar.

3.1.2 CRISP-DM

CISP-DM, siglas de Cross-Industry Standard Process for Data Mining, (Proceso Estandarizado Interindustrial para Minería de Datos) fue concebido en 1997 como un proyecto con financiación por parte de la Unión Europea en el cual participó un consorcio formado por 5 empresas: la automovilística Daimler-Benz (actualmente DaimlerChrysler AG), las empresas de análisis de datos SPSS, Teradata, la financiera NCR Corporation y la aseguradora Ohra.

El proyecto concluyó en 1999 mediante la publicación del documento CRISP-DM 1.0 [11], el cuál describe ampliamente la metodología paso a paso. Aunque han existido intentos de actualizar de forma oficial esta metodología de gestión, llegándose a formar un nuevo grupo de trabajo entre 2006 y 2008 con el fin de desarrollar una versión 2.0 de la guía, la falta de acuerdos derivó en la disolución del proyecto. Actualmente las webs oficiales del proyecto ya están activas.

La metodología CRISP – DM consiste en un modelo jerárquico, compuesto de 6 fases, las cuáles se dividen en tareas (genéricas y especializadas), y estas a su vez en instancias de proceso. (ver figura 8)

Las 6 fases que componen a esta metodología son las siguientes:

1. **Entendimiento del área de negocio (Bussiness Understanding):** Esta fase se centra en el entendimiento del negocio a analizar centrándose en las variables que tradicionalmente originan el éxito en el negocio, la situación actual de la empresa en el mercado, los criterios técnicos específicos del área de negocio… Esta fase termina con la elaboración de un plan inicial del proyecto.
2. **Entendimiento de los datos a analizar (Data Understanding):** En esta fase se recogen los datos a analizar y se estudia la forma en que se entregan los datos, así como la calidad de los mismos.
3. **Preparación de los datos (Data Preparation):** En esta fase se realiza una selección de los datos relevantes, así como limpiar, ordenar, dar formato a los datos y todos aquellos procesos necesarios para preparar los datos para fases sucesivas.

4. **Modelado de los datos (Modeling):** En esta fase se manipula los datos con el fin de encontrar relaciones entre los mismos y poder extraer futuras conclusiones.
5. **Evaluación de los datos (Evaluation):** En esta fase se evalúa las relaciones y representaciones obtenidas en la fase anterior, extrayendo conclusiones de los mismos.
6. **Implantación de medidas (Deployment):** En esta fase se coge las conclusiones obtenidas, se evalúa la relevancia de las mismas sobre el área de negocio y se busca las posibles conclusiones a implantar sobre el mismo.

A continuación, se muestra un esquema de las fases que componen la metodología CRISP-DM junto a los procesos correspondientes a cada una de las fases. [12]

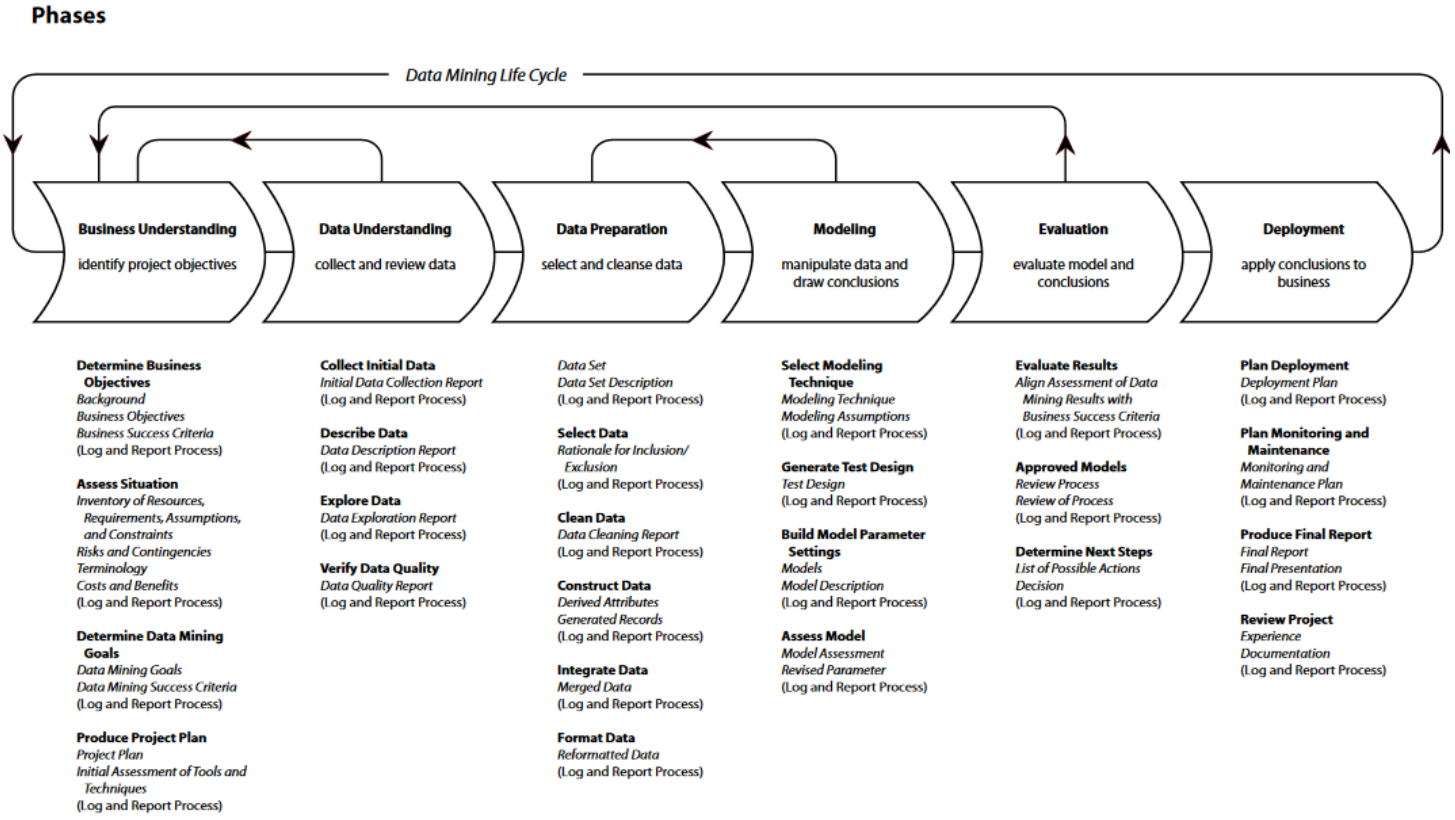

Figura 10. A visual guide to CRISP-DM methodology. Nicole Leaper (2009)

3.1.3 SEMMA

SEMMA es una metodología elaborada por el SAS Institute, una multinacional americana dedicada al desarrollo de software de análisis de datos. SEMMA fue desarrollada por SAS como apoyo a su herramienta de análisis de datos "SAS Enterprise Miner". [13]

SEMMA se centra principalmente en las tareas de modelado, minado y análisis de datos, dejando a un lado el entendimiento del área de negocio. Así mismo el enfoque de SEMMA como modelo de soporte de la aplicación "SAS Enterprise Miner" hace que su aplicación fuera de esta aplicación necesite de retoques con el fin de adaptar el modelo, pudiendo producirse diferencias en la aplicación de la misma de una entidad a otra.

La metodología SEMMA está compuesta de 5 fases, las cuales forman el acrónimo que da nombre a esta metodólogía. Estas fases son Sample (Muestreo), Explore (Explorar), Modify (Modificar), Model (Modelar) y Assess (Evaluar).

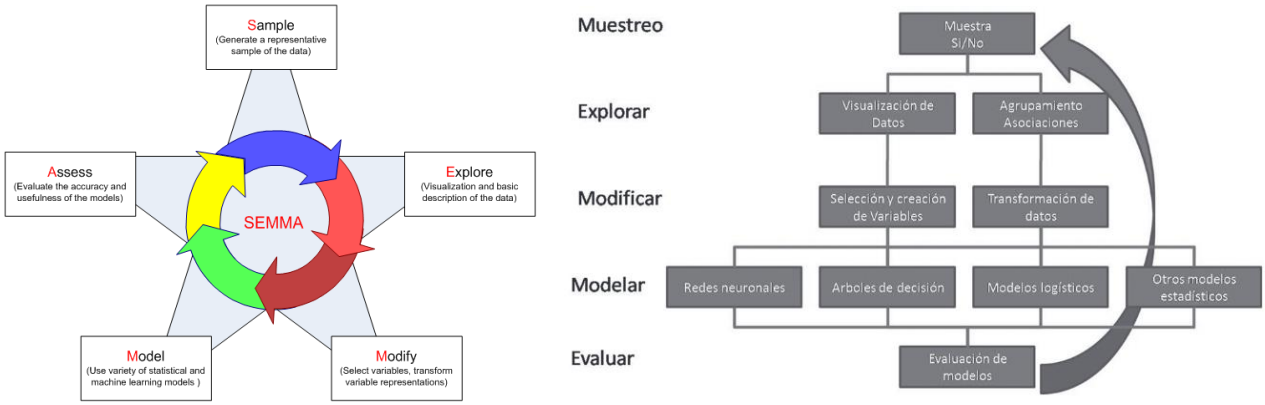

Figura 11. Representación de la composición de fases y tareas de la metodología SEMMA

La descripción de fases de la metodología SEMMA es la siguiente:

1. **Muestreo**
   El objetivo de la fase de muestreo es extraer una muestra representativa del conjunto de datos. Esta muestra debe de ser lo más pequeña posible y a su vez contener información relevante.
   La idea es que si existen patrones en el conjunto de datos estos mismos deben de ser reconocibles en la muestra.
   SEMMA define 3 procesos:_Entrenamiento, validación y prueba con el fin de facilitar la obtención de muestras representativas.
2. **Explorar**
   Una vez que se determina la presencia de patrones mediante la fase de muestreo, se trata todo el conjunto de datos con el fin de buscar tendencias y anomalías en el mismo. La idea es lograr un entendimiento de todo el conjunto de datos.
3. **Modificar**
   En la fase de modificación se altera los datos en base a los descubrimientos hallados en la fase anterior. El objetivo es preparar los datos para su modelización.
4. **Modelar**
   Se modelan los datos permitiendo que el software busque automáticamente una combinación de datos que prediga con cierta certeza un resultado deseado.
   En este proceso se emplean técnicas tradicionales de modelado en minería de datos (redes neuronales, árboles de decisión, modelos lógicos, …)
5. **Evaluar**
   En esta fase se califican los datos mediante la evaluación de la utilidad y fiabilidad de los resultados obtenidos en el proceso de minería de datos.

3.2 NECESIDAD DE UNA NUEVA METODOLOGÍA

Entre los autores que consideran necesaria la creación de una nueva metodología para la gestión de proyectos de minería de datos está Jeffrey S. Saltz. [14]. De acuerdo con el señor Saltz el incremento de la velocidad, volumen y variedad de datos disponibles para las organizaciones ha aumentado el interés y el número de proyectos relacionados con el aprovechamiento de estos datos. Esto ha derivado en equipos cada vez más grandes, los cuáles involucran cada vez aun mayor número de profesionales, en muchas ocasiones de diferentes disciplinas.

Saltz cree que las metodologías tradicionales no se han refinado con el objetivo de facilitar equipos multidisciplinares o de gran amplitud si no que los profesionales del sector se han centrado únicamente en la mejora de las técnicas de extracción y análisis de los datos. Un dato que muestra esta falta de interés es que de los 296 artículos e itinerarios de conferencias distribuídos durante la "Conferencia sobre Big Data 2014" promovida por IEEE (Institute of Electrical and Electronics Engineers) ninguno se centraba en mejoras sobre las metodologías de gestión de proyectos en el área y solo un 8% mencionaba los desafíos técnicos y sociales que puede llegar a implicar un proyecto de minería de datos.

Otro aspecto relevante mencionado por Saltz es que generalmente las metodologías de minería de datos suelen tomar una aproximación de "tareas" hacia el problema, conformado por una serie de etapas que pueden repetirse de forma iterativa. Este es un punto de vista que no ha evolucionado, en su opinión, en al menos 20 años, siendo la mayoría de las metodologías adaptaciones de KDD y solo diferenciándose algo CRISP-DM, la cual valora como un primer paso hacia el establecimiento de una metodología completa para la gestión de estos proyectos.

Sus motivos para considerar necesaria la definición de una nueva metodología son los siguientes:

1. La descripción paso por paso empleada en las metodologías tradicionales no está pensada para grandes equipos. Por ello se acaban realizando soluciones específicas para cada tipo de proyecto no siendo generalizables y debiendo llegar a ellas por el método prueba-error. Esto denota una **"baja madurez" del sector**.
2. Existe cierta dificultad de prever si el proyecto va a ser exitoso, se va a terminar su ejecución en tiempo y dentro de presupuesto.
3. Un 55% de los proyectos de "Big Data" no llegan a completarse y muchos otros no llegan a alcanzar el 100% de sus objetivos. [15]
4. Hace falta una metodología que defina con claridad la organización del grupo de trabajo, las funciones de cada individuo o el control de costes y plazos.

De acuerdo con Saltz la metodología propuesta debe de abordar al menos los siguientes problemas:

- Coordinación del equipo
- Gestión de la calidad de los resultados
- Propiedad intelectual de los datos, seguridad y privacidad
- Análisis de requisitos
- Priorización de requisitos
- Implantación

Por otro lado, Javier Segovia establece también un punto muy válido. En su opinión CRISP-DM intenta ser al mismo tiempo un modelo de proceso, una metodología y un ciclo de trabajo, esto hace que acabe careciendo de definición y detalle. En la opinión de Segovia esto hace que el proceso de minería de datos se encuentre aislado del resto del proceso de ingeniería de software pese a que tiene una implicación muy importante en el mismo.

Si bien Segovia piensa en un proyecto puramente de software al hacer esta reflexión esta misma es bien aplicable a otros tipos de proyectos. Las metodologías tradicionales de minería de datos entienden, tal y como mencionó Saltz los proyectos como una serie de pasos sucesivos los cuáles llevan o bien a la consecución de un objetivo o deben de ser repetidos hasta hallar el resultado deseado. Esto produce de algún modo una abstracción del resto de variables y entorno que rodean el proyecto (competidores, involucrados, mercado, ambiente dentro de la organización …) [16]

3.3 ALTERNATIVAS PROPUESTAS

En los últimos años ha habido un buen número de autores que han propuesto metodologías alternativas a SCRUM-DM y el resto de metodologías de gestión de proyectos de minería de datos con el fin de mejorar la eficiencia en la aplicación de estas técnicas a algún campo o área en particular.

No obstante, la mayoría de estas metodologías difieren poco de las metodologías tradicionales, limitándose a una mera reinterpretación de alguna de las fases que componen las mismas con el fin de facilitar su adaptación a un campo u organización. Ejemplos de estas son:

- La reinterpretación de Robin Way de la metodología SEMMA, ligada altamente al sector bancario y de inversiones y con un fuerte énfasis en la fase de implantación. [17]
- La adaptación de CRISP-DM al sector médico de Olegas Niakšu, quien establece una descomposición de las fases que componen la metodología CRISP en procesos específicos para el sector médico. [18]

- Lee and Kerschberg quienes establecen un procedimiento de control para la metodología KDD mediante un panel de expertos [19]
- Hofmann and Tierney quienes establecen responsables para cada una de las etapas del KDD de 9 pasos. [20]

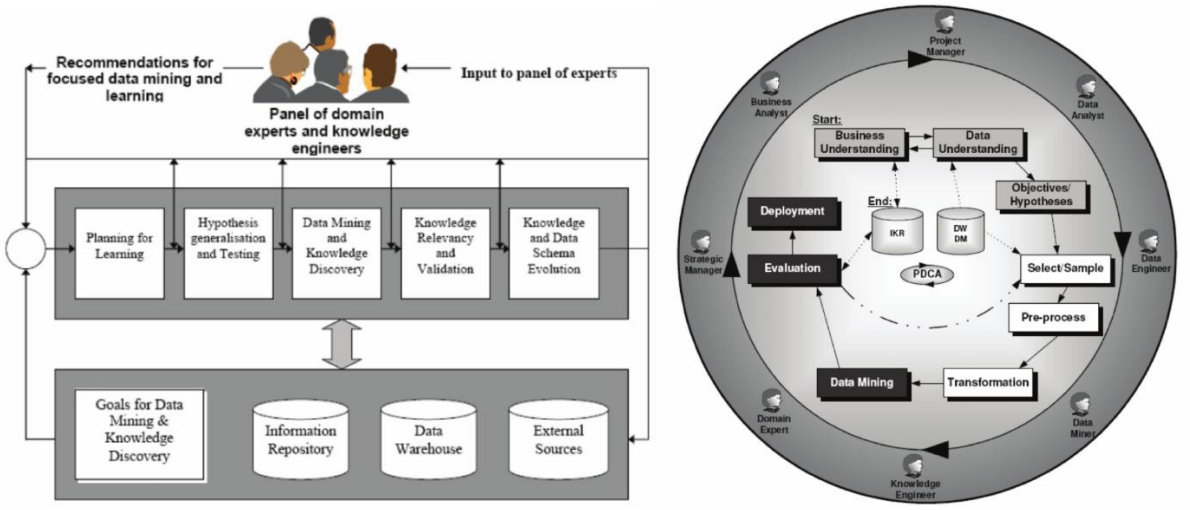

Figura 12. Izquierda: Modelo propuesto por Lee y Kerschberg (1998). Derecha: Modelo propuesto por Hofmann and Tierney (2007)

Estas metodologías, aunque interesantes no aportan ideas novedosas y se limitan a ser adaptaciones de las metodologías existentes para un fin en particular. No obstante, durante el estudio del "Estado del Arte" se ha detectado unos pocos autores que se considera aportan ideas novedosas en la materia y cuyas propuestas se cree se diferencian los suficiente de las metodologías tradicionales de gestión de proyectos de minería de datos.

Estas metodologías pasarán a detallarse en apartados sucesivos:

3.3.1 METODOLOGÍAS ÁGILES

3.3.1.1 ¿QUÉ SON LAS METODOLOGÍAS ÁGILES?

La metodología ágil (agile) surgió a finales de los años 90, tomando forma mediante la publicación del "manifiesto por el desarrollo Ágil de Software en el año 2001. Este manifiesto fue redactado inicialmente por 14 autores y ha recibido numerosas firmas de apoyo desde su publicación hasta la actualidad.

Agile surge a partir de la creencia de sus autores de la necesidad de priorizar las **relaciones personales entre individuos** sobre las herramientas y procesos, la **operatividad del software** sobre la documentación del mismo, la **colaboración con el cliente** sobre la negociación del contrato y la **respuesta al cambio** sobre un plan prefijado. Esta necesidad se ve representada por los **"12 principios del manifiesto "Ágil"** [21]**:**

1. **Satisfacción del cliente** *"Nuestra mayor prioridad es satisfacer al cliente mediante la entrega temprana y continua de software con valor".*

2. **Aceptación del cambio.** *"Aceptamos que los requisitos cambien, incluso en etapas tardías del desarrollo. Los procesos Ágiles aprovechan el cambio para proporcionar ventaja competitiva al cliente."*
3. **Entregas graduales. "**Entregamos software funcional frecuentemente, entre dos semanas y dos meses, con preferencia al periodo de tiempo más corto posible".
4. **Colaboración con el cliente.** *"Los responsables de negocio y los desarrolladores trabajamos juntos de forma cotidiana durante todo el proyecto."*
5. **Motivación y confianza.** *"Los proyectos se desarrollan en torno a individuos motivados. Hay que darles el entorno y el apoyo que necesitan, y confiarles la ejecución del trabajo."*
6. **Comunicación cara a cara.** "El método más eficiente y efectivo de comunicar información al equipo de desarrollo y entre sus miembros es la conversación cara a cara."
7. **Funcionabilidad.** *"El software funcionando es la medida principal de progreso."*
8. **Desarrollo sostenible.** *"Los procesos Ágiles promueven el desarrollo sostenible. Los promotores, desarrolladores y usuarios debemos ser capaces de mantener un ritmo constante de forma indefinida."*
9. **Atención a los detalles.** La atención continua a la excelencia técnica y al buen diseño mejora la Agilidad.
10. **Simplicidad**. *"La simplicidad, o el arte de maximizar la cantidad de trabajo no realizado, es esencial."*
11. **Autogestión de los equipos.** *"Las mejores arquitecturas, requisitos y diseños emergen de equipos auto-organizados."*
12. **Adaptación circunstancias cambiantes.** *"A intervalos regulares el equipo reflexiona sobre cómo ser más efectivo para a continuación ajustar y perfeccionar su comportamiento en consecuencia. "*

Las **ventajas** [22] generalmente atribuidas a agile son:

- **Gestión del cambio.** Al existir ciclos de planificación corta es fácil realizar cambios en cualquier punto del proyecto.
- **Es posible no conocer la finalidad del proyecto al inicio.** Generalmente se asocia Agile a proyectos en los cuáles el objetivo final no está completamente definido.
- **Permite entregas más rápidas, continuadas en el tiempo y de mayor calidad.**
- **Interacción fuerte entre los involucrados en el proyecto.**
- **El cliente siempre es escuchado**
- **La mejora del producto es siempre continua.**

A su vez las **desventajas** [22] que se atribuyen a esta metodología son las siguientes:

- **La planificación del proyecto es menos concreta**
- **El equipo de trabajo debe tener conocimientos multidisciplinares.** Los equipos de trabajo ágiles suelen ser pequeños, lo cual hace que sus individuos deban de abarcar varias áreas de conocimiento, entender la metodología ágil y sentirse cómodos usándola.

- **Entrega total al proyecto.** La metodología ágil es más efectiva cuando los involucrados pueden dedicarle todo su tiempo al proyecto, dada la naturaleza de esta metodología, la cual requiere una mayor implicación de los involucrados.
- **La documentación puede ser dejada a un lado.** La preferencia del producto sobre la documentación del mismo puede hacer que se deje a un lado el desarrollo de esta.
- **El producto final puede variar significativamente respecto de las especificaciones iniciales.** Debido a la flexibilidad al cambio de esta metodología puede darse el caso de que un número incremental de cambios deriven en un producto totalmente distinto del requerido al inicio del proyecto.

Las metodologías ágiles se caracterizan por una serie de fases que no tienen por qué producirse en sucesión, pudiendo producirse en paralelo, así como alterar el orden entre las mismas.

Figura 13. Comparación entre Agile y las metodologías tradicionales (Santy Abreu)

3.3.1.2 Relación entre las metodologías ágiles y la Minería de Datos

Existe cierta similitud entre el concepto de Agile y el concepto empleado en las metodologías clásicas de minería de proyectos, la cual exploraremos en apartados posteriores. Es probable que debido a este hecho ciertos autores hayan promovido la adopción de los principios de la metodología ágil en los proyectos de minería de datos. Algunos de los hitos que se han considerado de mayor relevancia en esta tendencia se presentan a continuación.

1. **ASD-DM (Mouhib Alnoukari et al.; 2008):** ASD-DM (Adaptative Software Development – Data Mining) es una solución basada en los principios ágiles para proyectos predictivos de minería de datos. Se basa en la metodología ASD propuesta originalmente por el propio Mouhib Alnoukari basada fuertemente en un principio iterativo de prueba y error. Esta metodología tiene la ventaja de reducir fuertemente el tiempo de desarrollo en proyectos en los que el objetivo o resultados finales son predecibles, siendo su punto débil aquellos proyectos en los que los objetivos y soluciones no son fácilmente estimables al inicio. [23]

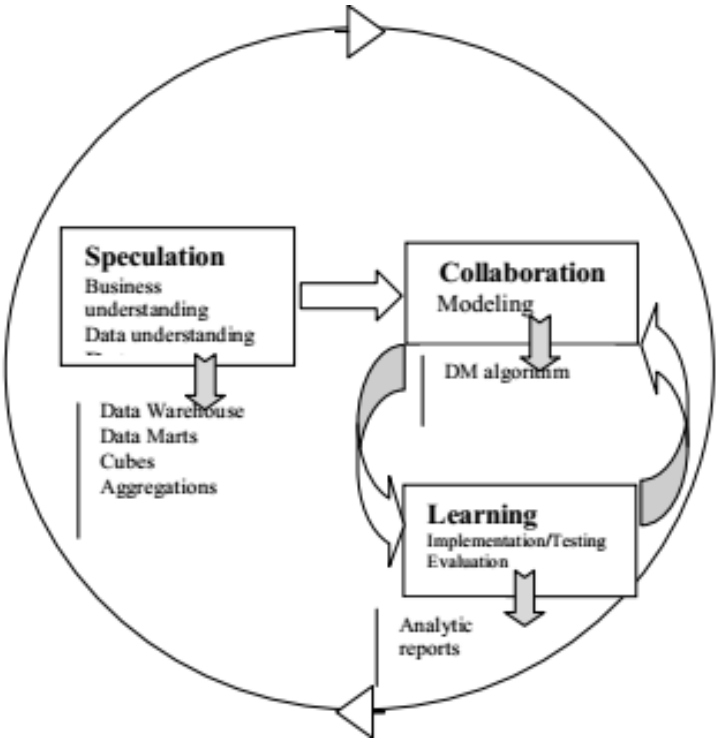

Figura 14. Modelo de proceso de la metodología ASD-DM DM *(Mouhib Alnoukari, et al.; 2008)* [25]

2. Autores como **Gonzalo Mariscal et al. (2013)** [24], **Gino Marckx (2014)** [25] y **Michał Łopuszyńsk**i **(2016)** [26] y realizaron comparaciones entre la filosofía planteada en la literatura adscrita a las metodologías tradicionales de minería de datos y los principios del manifiesto ágil resaltando las similitudes entre la filosofía de las metodologías clásicas y el manifiesto ágil, llegando a conclusiones similares y concordando en que las metodologías clásicas se beneficiarían de la adopción de los principios ágiles.

3. **Agile KDD (Givanildo Santana do Nascimento et al.; 2013).** [27] Agile KDD es una adaptación del KDD clásico que toma conceptos de CRISP-DM y del manifiesto ágil. Esta propuesta se centra en el establecimiento de metas iterativas y en ciclos más cortos entre los entregables. Esta visión de Givanildo Santana et al. pone en relieve los puntos en común entre las metodologías de minería de datos y el manifiesto ágil.

4. **Philip Guo (2013)** analizó el proceso de la minería de datos y a los desafíos técnicos que se enfrenta un "data scientist" durante un proyecto, llegando a un cilo de trabajo iterativo muy similar al descrito por el manifiesto ágil. [28]

5. **ASUM-DM (2015**). ASUM-DM (Analytics Solutions Unified Method for Data Mining) es una revisión de la metodología CRISP-DM la cuál incorpora principios ágiles, extendiendo y refinando la metodología CRISP. Esta metodología es promovida por IBM, siendo lanzada la primera versión de la misma en 2015.
   Dado el enfoque de ASSUM-DM hacia el sector empresarial, esta metodología se analizará en detalle en apartados siguientes.

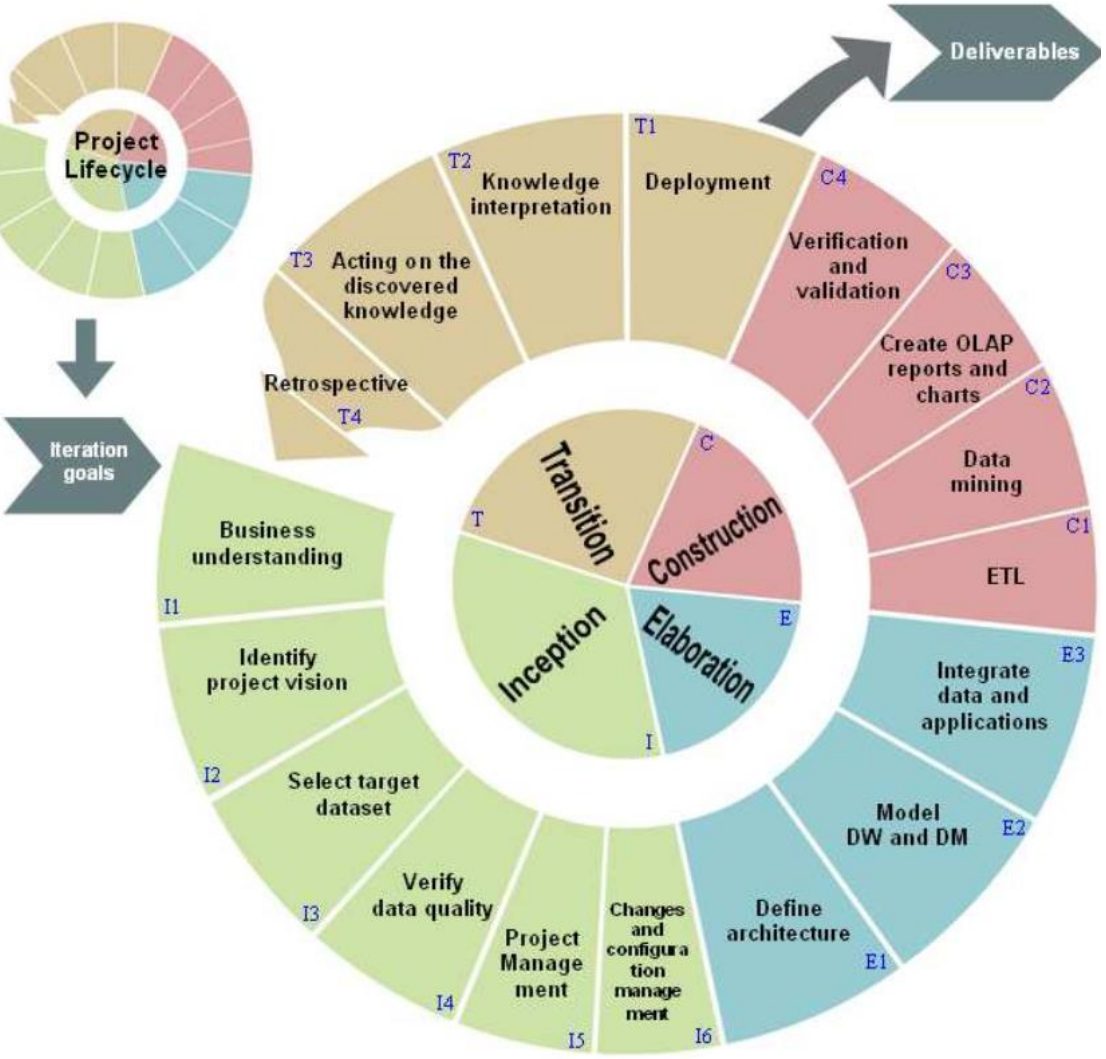

Figura 15. Fases y ciclo de vida propuestos por Agile-KDD [29]

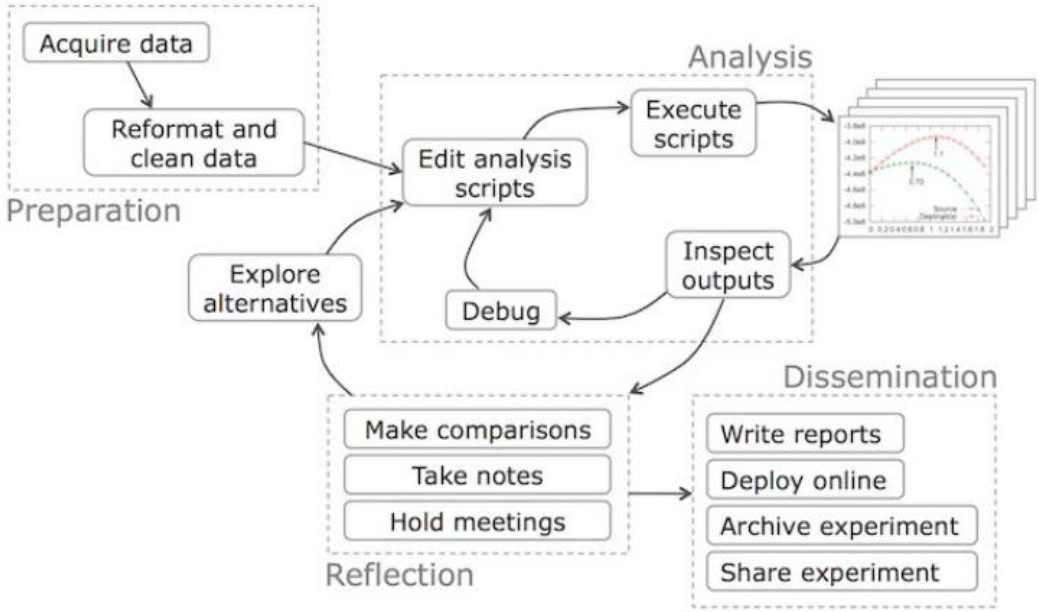

Figura 16. Ciclo de trabajo propuesto por Philip Guo (2013) *[28]*

3.3.2 METODOLOGÍAS APLICADAS AL ENTORNO EMPRESARIAL

3.3.2.1 ASUM-DM

Durante los últimos años IBM ha sido uno de los grandes promotores de la metodología CRISP-DM, en parte debido a la adquisición por parte de IBM de SPSS Inc. en 2009, una de las 5 empresas que participaron en la redacción. de la metodología. [29]

Después de varios intentos fallidos de alcanzar un acuerdo común en la redacción de la versión 2.0 de CRISP-DM por el consorcio de empresas participantes en el proyecto IBM decidió realizar una revisión de la tecnología en 2015 a la cuál denominó ASSUM-DM (Analytics Solutions Unified Method for Data Mining).

De acuerdo con IBM el punto fuerte de la metodología CRISP-DM se encuentra en el campo del análisis de datos, mientras que su punto débil radica en la falta de detalle de la metodología a la hora de cubrir las operaciones e infraestructura necesarias para llevar a cabo la fase de "implantación" del proyecto.

Algunos de los puntos clave de ASSUM-DM. Según proclama IBM son los siguientes [30]:

- ASSUM-DM está ampliamente ligado a la solución de software "IBM Analytics Solutions", no obstante, es una metodología propia que puede ser implementada de forma ajena a esta solución.
- Todas las fases son supervisadas por el "equipo de gestión del proyecto".
- Assum es una solución híbrida entre las metodologías ágiles y las metodologías tradicionales de minería de datos.
- Assum incorpora los siguientes principios ágiles:
  - El proyecto es evaluado mediante aplicación de principios ágiles.
  - Tanto el área de negocio como el área tecnológica se ven involucradas durante el proceso de implantación del proyecto.
  - Los requisitos y requerimientos se clarifican y modifican de forma continuada mediante "sprints iterativos".
  - Se adopta una implementación del proyecto por fases en base a los recursos y tiempo disponibles, así como a la prioridad de los requisitos del proyecto.
  - El proyecto se enfoca mediante un desarrollo iterativo del mismo, determinando los "hitos" alcanzados al final de cada iteración y los objetivos a alcanzar durante la iteración siguiente.
  - Los resultados de las fases implementadas se monitorizan y optimizan de forma continuada a lo largo de la vida del proyecto.
- La fase de gestión del proyecto sigue principios similares a los recomendados por la guía del PMBOK o la metodología PRINCE2.
- Usa los estándares de validación y control adoptados por la industria.

Si obviamos el carácter Ágil y enfoque iterativo del proyecto ASUM-DM no varía filosóficamente en gran medida respecto a la metodología CRISP-DM, conservando de acuerdo a IBM los puntos fuertes en de la metodología CRISP, refinándolos y expandiéndolos. Por este

motivo ASSUM-DM incorpora la definición de nuevas actividades y tareas, centrándose especialmente en la fase de "Implantación del Proyecto".

Cabe así mismo destacar que se pasa de las 6 fases en las que se dividía un proyecto en base a CRISP-DM a 4 fases:

1. **Análisis**: En esta fase se determinan los requisitos y necesidades del cliente, se definen los objetivos del proyecto, los entregables y se discute de forma inicial las primeras soluciones adoptables.
2. **Diseño**: Esta fase incluye todos aquellos procesos que atienen al desarrollo de minería y análisis de datos.
3. **Implantación:** Es el punto fuerte de ASSUM, la cual establece un flujo de trabajo para implementar los resultados dentro de la empresa de forma que no afecte a la actividad productiva normal.
4. **Operar y Optimizar:** Es la fase de control, se encarga de analizar el proyecto implementado, determinando la necesidad de aumentar el alcance del proyecto o realizar pequeños ajustes en el mismo.

Uno de los puntos notables a la hora de analizar la metodología propuesta por IBM es la falta de detalle en los procesos que conllevan la minería y análisis de datos, esto es debido a que si bien IBM define a esta metodología como una revisión de CRISP-DM si estudiamos las soluciones propuestas por la empresa podemos ver que estas 4 fases son compartidas por todas ellas lo cual denota que el objetivo de la metodología es facilitar la implantación de proyectos informáticos y que la fase de "diseño" deberá ser ocupada por los procesos y técnicas tradicionales de la minería de datos, tal y como vienen siendo definidos por otras metodologías como CRISP-DM.

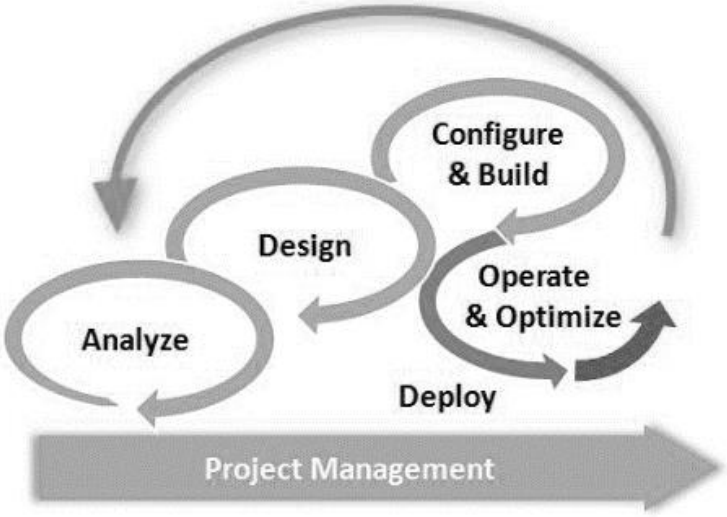

Figura 17. Ciclo de trabajo propuesto por la metodología ASSUM-DM. IBM (2015) [33]

La equivalencia entre las fases de CRISP-DM y ASSUM-DM se presenta en la siguiente tabla:

<table>
<tr><th colspan="2">CRISP -DM</th><th colspan="3">ASSUM-DM</th></tr>
<tr><td>Fases</td><td>Tareas</td><td colspan="2">Fases</td><td>Tareas</td></tr>
<tr><td>Entendimiento del área de negocio</td><td>4</td><td rowspan="8">Gestión del Proyecto</td><td>Análisis</td><td>8</td></tr>
<tr><td>Entendimiento de los datos a analizar</td><td>4</td><td rowspan="4">Diseño</td><td rowspan="4">10</td></tr>
<tr><td>Preparación de los datos</td><td>5</td></tr>
<tr><td>Modelado de los datos</td><td>4</td></tr>
<tr><td>Evaluación</td><td>3</td></tr>
<tr><td rowspan="2">Implantación</td><td rowspan="2">4</td><td>Implantación</td><td>7</td></tr>
<tr><td>Operar y Monitorizar</td><td>5</td></tr>
</table>

Tabla 1 – Comparación de metodologías tradicionales de minería de datos

Como es observable en la tabla presentada sobre este párrafo existe una mayor definición en las etapas iniciales y finales del proyecto, contrastando con la menor definición en las etapas intermedias del proyecto.

Adicionalmente ASSUM-DM incorpora una fase paralela al resto, denominada “Gestión del Proyecto” a realizar por la “Dirección del Proyecto”, encargada de que las fases del proyecto fluyan de forma fluida entre sí, esta labor es similar a la realizada por un “SCRUM MASTER”, en la metodología SCRUM, una de las metodologías derivadas de Agile.

Otro punto resaltable de la metodología ASSUM-DM es que se incorpora una definición de los interesados del proyecto con las características de cada perfil y tareas típicas, así como unas matrices de asignación de tareas.

Por último, cabe reseñar que la guía para la implantación de ASSUM-DM solo es accesible mediante registro en el sitio web específico de IBM y mediante la instalación de archivos ejecutables que dan acceso a la guía en formato html, siendo estos archivos solo accesibles mediante sistemas operativos Windows, lo cual disminuye la capacidad de acceso a la metodología por parte del público general. [31]

3.3.2.2 EL CASO DE RAMCO CEMENTS

De acuerdo con Debprotim Dutta e Indranil Bose del Indian Institute of Management Calcutta, en colaboración con la empresa Ramco Cements Limited, situada en la India, propusieron en 2014 una metodología con el fin de abordar proyectos de Big Data dentro del entorno empresarial. [32]

De acuerdo con los autores un proyecto de Big Data se caracteriza por un gran Volumen, Variedad y Velocidad en los datos a analizar (3Vs) esto conlleva 2 obstáculos característicos, la necesidad tecnológica de equipos capaces de manejar grandes fuentes de información y la necesidad social de formar equipos multidisciplinares debido a lo específico de los datos y a su esperable variedad, problemas que los autores creen que no pueden ser afrontados mediante el uso de metodologías tradicionales de gestión de proyectos de análisis de datos.

La metodología propuesta se divide en tres fases de gran amplitud, las cuáles a su vez se dividen en pasos. Estas fases son:

1. **Trabajo estratégico:** El trabajo estratégico empieza con un **análisis de negocio** en el que los "involucrados", pertenecientes a distintas áreas de la empresa dan su opinión, ayudando a determinar objetivos y expectaciones del proyecto. A esto le sigue una fase de **investigación** en donde se analiza como la competencia u otros sectores de la industria han dado solución a problemas similares.

   Uno de los puntos característico de la metodología propuesta se encuentra en el paso de **"formación del equipo multidisciplinar"**, donde a partir de los resultados anteriores se define la estructura del equipo o equipos que participarán en el proyecto y los componentes que habrán de integrarlos.

   Por último, se termina con la elaboración del **plan de proyecto**, entregable que define un "hito" entre fases.

2. **Análisis de datos:** La fase de análisis de datos sigue una estructura similar a la de las metodologías tradicionales, si bien cabe destacar que se le da relevancia a la **"visualización de datos"**.
   De acuerdo con los autores, en un proyecto de Big Data, la obtención de visualizaciones relevantes es un paso fundamental dado que la estructura tradicional de gráficas y tablas no siempre permite una visualización rápida de patrones y relaciones entre los diversos conjuntos de datos analizados.
   Esta etapa termina con el hito "percepción", el cual representa un conocimiento de la relación entre los conjuntos de datos y a partir del cual se extraen conocimientos.
3. **Implementación:** Esta etapa consiste la integración técnica del sistema propuesto en la empresa y en el entrenamiento de los individuos que deberán de utilizar este sistema.

Como la mayoría de las metodologías de uso industrial/empresarial esta metodología establece un fuerte énfasis en la implementación del sistema y entrenamiento de los individuos, no obstante, es novedoso el énfasis que se pone en apoyar la formación del equipo de proyecto en base a las ideas extraídas del análisis de negocio.

Esta es una metodología que puede ser cíclica al igual que el resto de metodologías de análisis de datos, no obstante, sigue un proceso secuencial al igual que las metodologías tradicionales.

A continuación, se presenta una representación del ciclo de trabajo propuesto por esta metodología. [32]

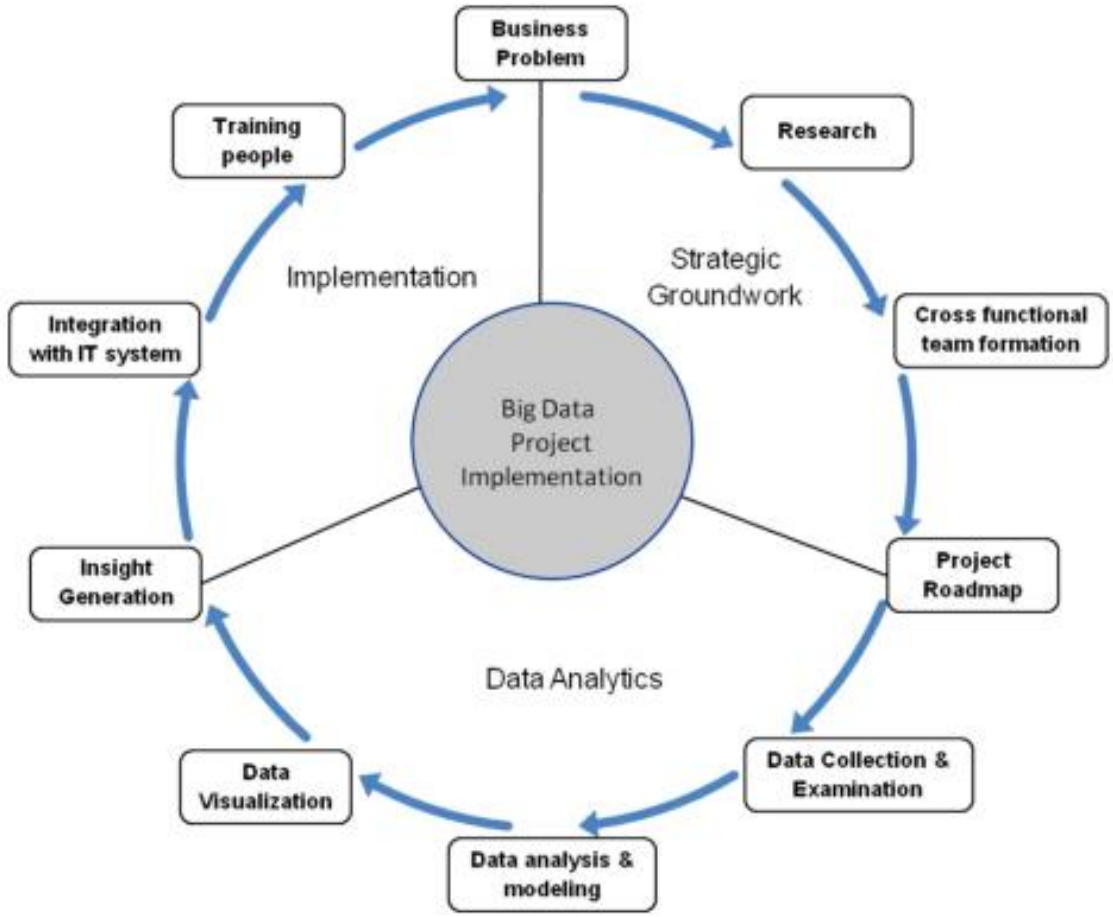

Figura 18. Ciclo de trabajo propuesto por Debprotim Dutta e Indranil Bose para proyectos multidisciplinares. *"El caso de Ramco Cements"* [34]

3.3.2.3 JOSE SOLARTE

En 2002 Jose Solarte, entonces estudiante de la Universidad de Tennessee propuso una metodología para proyectos de minería de datos con orientación a su aplicación en procesos industriales durante "Master Theses". [33]

Era opinión de Solarte que no existía en aquel entonces ninguna metodología específica que diese solución a los retos a los que se puede enfrentar un proyecto de minería de datos en dicho entorno.

La metodología propuesta por Solarte se divide en 5 fases:

1. Análisis de la organización:
2. Organización del trabajo
3. Desarrollo del modelo de análisis de datos
4. Implementación de resultados
5. Soporte técnico (Control de la implementación)

Si bien estas fases se dividen como en muchas otras metodologías en varios pasos, el resultado no difiere en gran medida de otras metodologías ya vistas con anterioridad.

La innovación de Solarte que hace diferir a su metodología de las nombradas anteriormente es su uso de matrices como apoyo a la toma de decisiones, sugiriendo como se debe de ponderar los resultados obtenidos de las mismas:

Solarte sugiere específicamente el uso de dos matrices, una para elegir la aproximación más apropiada durante la fase 3, "Desarrollo del modelo de análisis de datos" y otra para evaluar el proyecto en su cierre con el fin de realizar mejoras posteriores o cambios en el sistema en caso de que sea necesario.

A continuación, se muestra el ciclo de trabajo propuesto por Solarte, así como una muestra de estas matrices.

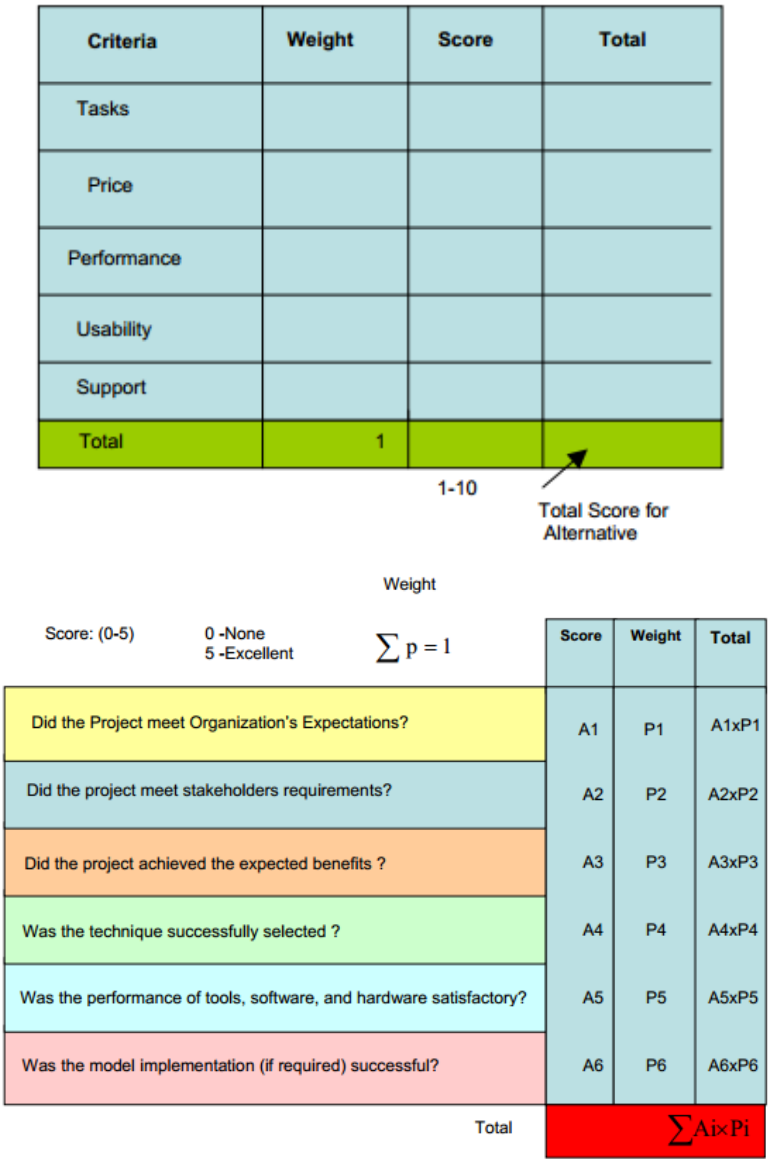

Figura 19. Ejemplos de matrices de toma de decisiones propuestas por Jose Solarte. [35]

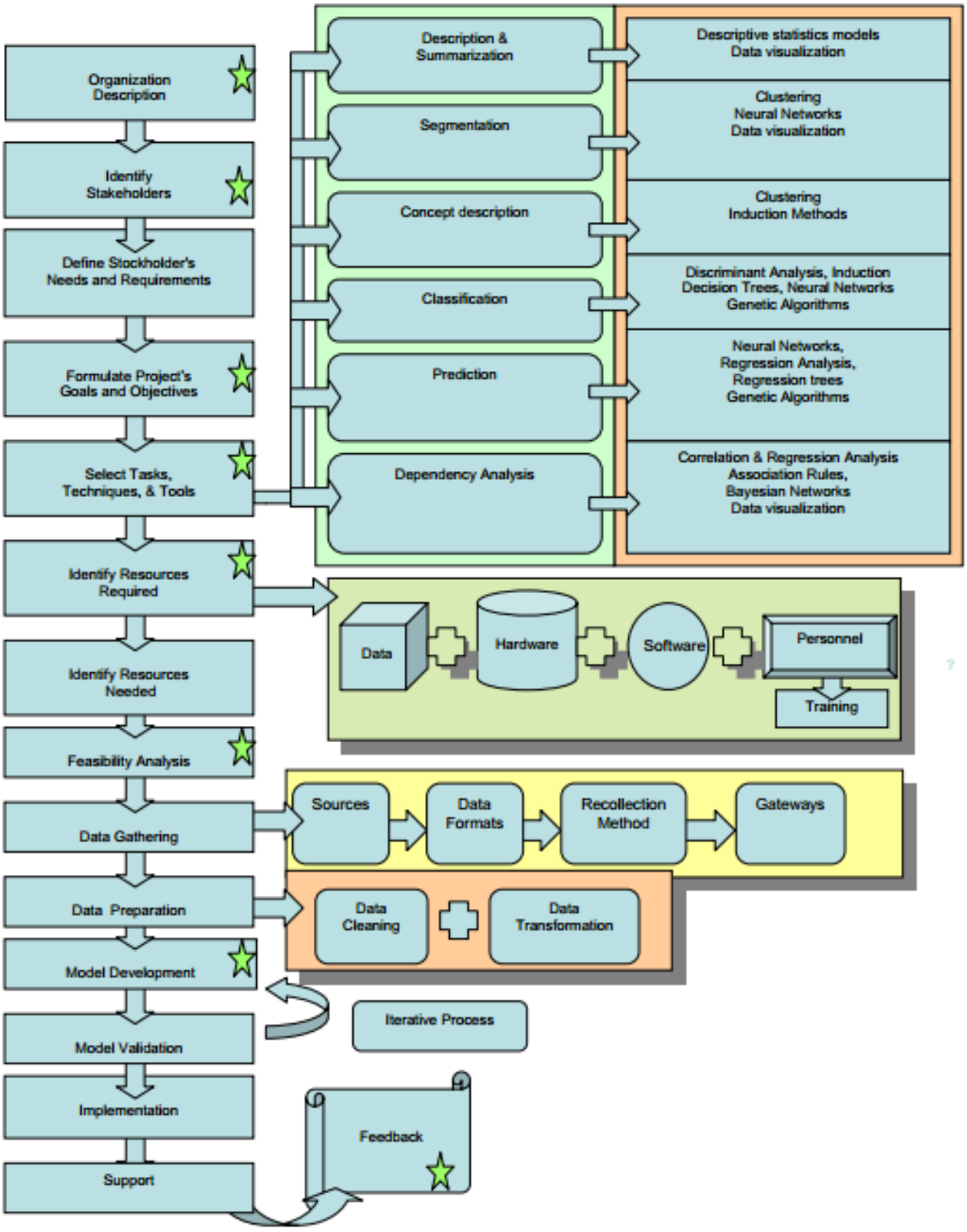

Figura 20. Ciclo de trabajo propuesto por Jose Solarte [35]

3.4 FACTORES CLAVES DE ÉXITO EN PROYECTOS DE MINERÍA DE DATOS

Antes de entrar a analizar las distintas metodologías expuestas parece conveniente revisar cuáles son los factores que determinan el éxito o fracaso de un proyecto de minería de datos.

El concepto de Factor "Clave" de Exito o Critical Succes Factor en su notación anglosajona fue propuesta por Ronald Daniel en 1961, no adquiriendo resonancia hasta 1979 año en que el concepto fue reintroducido por John F. Rockart, profesor de la escuela de la *MIT's Sloan School of Management*, fecha a partir de la que se popularizó siendo adoptado por diversas áreas de negocio y empresas.

Los Factores Clave de Éxito son aquellos factores o parámetros que influyen en mayor medida en el éxito de una empresa o proyecto, siendo por tanto los puntos prioritarios a tratar a la hora de valorar la viabilidad de un proyecto.

Después de revisar la literatura relevante se han observado dos estudios sobre factores de éxito en proyectos de minería de datos:

- Por un lado, tenemos a Jaesung Sim, quien propone una serie de factores de éxito para proyectos de minería de datos y los valida en base a encuestas realizadas a profesionales en la materia. [34]
- Por otro lado, tenemos a Jayanthi Ranjan y Vishal Bhatnagar los cuales, si bien especifican que estos factores pueden variar entre organizaciones, establecen una serie de factores de éxito en base a su experiencia. [35]

En base a sus conclusiones se determinan los siguientes "Factores de Éxito", los cuáles se organizan de acuerdo a las distintas fases que puede tener un proyecto de minería de datos en base a la literatura expuesta en apartados anteriores:

1. **Relaciones personales**
   - Poner en primer lugar los requisitos del cliente
   - El cliente debe de estar comprometido con el proyecto
   - Comunicación constante entre el cliente y el equipo de trabajo
   - Comunicación dentro del equipo de trabajo.
2. **Planificación**
   - Establecer objetivos claros antes de empezar el proyecto.
   - Selección adecuada del equipo de trabajo.
   - Planificación del proyecto adecuada.
   - Definición correcta de plazos y costes.
3. **Minería y Análisis de Datos**
   - Selección adecuada de los datos
   - Entendimiento adecuado del área de negocio
   - Uso de tecnología adecuada
   - Interpretación correcta de los resultados
4. **Implantación**
   - Feedback constante
   - Asegurar la privacidad de los datos
   - Implantación no "brusca"

5. **Control/Monitorización/Entrenamiento**
    - Paciencia. Los resultados pueden no ser inmediatos
    - Asegurar la privacidad de los datos

Cabe así mismo recordar los puntos que Saltz [14] cree que debería abordar un proyecto de minería de datos en la actualidad:

- Coordinación del equipo
- Gestión de la calidad de los resultados
- Propiedad intelectual de los datos, seguridad y privacidad
- Análisis de requisitos
- Priorización de requisitos
- Implantación

4 DISERTACIÓN SOBRE EL ESTADO DEL ARTE

En este apartado se realizará una introspección sobre diversos puntos abordados en el "Estado del Arte", con el fin de identificar aquellos puntos, en opinión del autor de este TFM, a los que debe de dar respuesta de una metodología exitosa de gestión de proyectos de minería de datos.

4.1 COMPARACIÓN E INTERRELACIÓN ENTRE LAS METODOLOGÍAS TRADICIONALES DE MINERÍA Y ANÁLISIS DE DATOS

Cuando enfrentamos dos metodologías tradicionales de gestión de proyectos de minería y análisis de datos una de las primeras cosas que puede venírsenos a la cabeza son las inherentes similitudes existentes entre ellas. Esto es provocado por la existencia de una relación claramente visible entre los pasos o fases en que se descomponen las mismas.

Para entender esta similitud cabe recordar que los conceptos básicos de la gestión de proyectos de minería de datos se encuentran en los orígenes de la metodología KDD, cuyos conceptos básicos establecen los escritos de Piatetsky-Shapiro, Fayyad, Brachman, …

Cabe recordar también que el origen de las ideas tratadas en dichos escritos surge del intercambio de conocimientos durante los grupos de trabajo de "*Descubrimiento de Conocimiento en Bases de Datos"* en los que distintos expertos sobre la materia se dan cita con carácter anual desde 1989, dando un importante punto de referencia "académico" al sector de la minería de datos.

Esta relación existente entre las metodologías tradicionales de minería de datos ha sido destacada por numerosos autores como Ana Azevedo et al. [36] quienes tienden a opinan que CRISP-DM y SEMMA derivan de KDD, siendo implementaciones particularizadas para un uso o ámbito determinado de esta metodología

A continuación, se muestra una tabla, de elaboración propia, donde se relacionan los pasos o fases de que se componen las tres metodologías anteriormente descritas, donde KDD se

muestra en sus tres variantes: la propuesta por Brachman et al., la propuesta por Fayyad et al. y la metodología de 9 pasos.

<table>
<tr><th rowspan="2">KDD (Brachman et al.)</th><th colspan="2">KDD (Fayyad et al.)</th><th rowspan="2">CRISP -DM</th><th rowspan="2">SEMMA</th></tr>
<tr><th>6 pasos</th><th>9 pasos</th></tr>
<tr><td>-</td><td>-</td><td>-</td><td>Entendimiento del área de negocio</td><td>-</td></tr>
<tr><td>Descubrimiento</td><td rowspan="2">Pre-KDD</td><td rowspan="2">Análisis Previo</td><td rowspan="2">Entendimiento de los datos a analizar</td><td>Muestreo</td></tr>
<tr><td>Determinación de objetivos</td><td>Explorar</td></tr>
<tr><td rowspan="6">Modelado de datos</td><td>Selección de datos</td><td>Selección de datos</td><td rowspan="3">Preparación de los datos</td><td rowspan="3">Modificar</td></tr>
<tr><td>Pre-Procesado de datos</td><td>Pre-Procesado de datos</td></tr>
<tr><td>Transformación</td><td>Transformación</td></tr>
<tr><td rowspan="3">Minería de datos</td><td>Elección de la tarea</td><td rowspan="3">Modelado de los datos</td><td rowspan="3">Modelar</td></tr>
<tr><td>Elección del algoritmo</td></tr>
<tr><td>Minería de datos</td></tr>
<tr><td>Análisis de datos</td><td rowspan="2">Interpretación/Evaluación</td><td rowspan="2">Interpretación/Evaluación</td><td rowspan="2">Evaluación</td><td rowspan="2">Evaluar</td></tr>
<tr><td>Visualización de resultados</td></tr>
<tr><td>Integración</td><td>Actuar sobre el conocimiento</td><td>Actuar sobre el conocimiento</td><td rowspan="2">Implantación</td><td>-</td></tr>
<tr><td>Monitorización</td><td>-</td><td>-</td><td>-</td></tr>
</table>

Tabla 2 – Comparación de metodologías tradicionales de minería de datos

De acuerdo con los datos representados en la tabla podemos extraer las siguientes conclusiones:

1. Todos los modelos son equivalentes, con pequeñas variaciones en su parte central. Es decir, la que atañe a la minería y análisis de datos.
2. Todos los modelos incorporan una fase de pre-análisis de datos, una de modelado de datos, una fase de análisis de resultados y una fase de implantación de resultados.
3. CRISP-DM es el modelo más completo ya que añade procesos de entendimiento del área de negocio y de monitorización sobre la implantación.
4. A parte de CRISP-DM solo la implementación de KDD de Brachman incorpora una fase o proceso específico de monitorización-.
5. Aunque diversos autores relacionan el entendimiento inicial del proyecto de KDD con el entendimiento de negocio estas fases iniciales de KDD se centran en la adquisición de conocimiento específico con el fin de entender los datos dejando a un lado el estado de la organización o del mercado. [36] [37]
6. SEMMA es la metodología menos completa para la gestión de proyectos ya que se centra simplemente en la obtención y análisis de los datos, dejando a un lado el entendimiento del área de negocio, la implementación de medidas a partir del resultado o el proceso de control. Así mismo es una metodología ligada a un software empresarial que requiere de adaptación para su uso fuera de dicho entorno.

4.2 ANÁLISIS DE LOS FACTORES DE ÉXITO

Los factores de éxito expuestos en el "Estado del Arte" pueden ser resumidos en los siguientes puntos, los cuales considera el autor de este TFM, que engloban aquellos aspectos que deben de ser abordados por cualquier metodología de minería de datos para la consecución de un proyecto exitoso. Estos son:

a. Conocimiento del negocio
b. Involucración del cliente
c. Control de requisitos
d. Selección adecuada del equipo de trabajo
e. Control de la comunicación entre el equipo de trabajo
f. Control de plazos y costes
g. Gestión de privacidad en los datos del cliente
h. Control de la implantación
i. Control del cambio

A continuación, en los apartados sucesivos se valorará la capacidad de las distintas metodologías presentadas durante el Estado del Arte para abordar estos puntos.

4.2.1 METODOLOGÍAS TRADICIONALES

De las metodologías tradicionales se considera que CRISP-DM es la más completa y por tanto este análisis de requisitos se hará, principalmente, con esta metodología en mente.

a. Conocimiento del negocio. Sí
CRISP-DM tiene dos fases iniciales que sirven para dar cumplimiento a este requisito. Una fase de entendimiento del área de negocio y una fase de entendimiento de los datos a analizar.
Los pasos en que se dividen estas fases sirven para dar cumplimiento a este requisito.

b. Involucración del cliente. No
En general la involucración del cliente se limita a la fase inicial de "entendimiento del área de negocio" y a la fase final de "implantación".

c. Control de requisitos. Sí
Los requisitos se establecen al inicio del proyecto y se comprueba su cumplimiento en la fase de implantación.

d. Selección adecuada del equipo de trabajo. No
Tanto CRISP como las metodologías tradicionales no entran a analizar el equipo de trabajo necesario si no que se centran en las herramientas y técnicas de minería de datos adecuadas.

e. Control de la comunicación entre el equipo de trabajo. No
No existe un procedimiento que asegure la comunicación entre el equipo de trabajo.

f. Control de plazos y costes. No
No existe un procedimiento de control de plazos y costes.

g. Gestión de privacidad en los datos del cliente. No
No existe un procedimiento que asegure la privacidad de los datos del cliente.

h. Control de la implantación. Sí

Uno de los pasos de la fase de implantación hace referencia al control de la implementación y al mantenimiento.

i. **Control del cambio.** No
En CRISP no se realiza un control de cambio de las condiciones del mercado o de los

En general las metodologías tradicionales son adecuadas para equipos pequeños en los que la interacción con el cliente se limita al principio del proyecto y el equipo se encuentra formado por profesionales de la misma área.

4.2.2 METODOLOGÍAS ÁGILES

a. **Conocimiento del negocio.** Sí
Al igual que en las metodologías tradicionales aquellas que abrazan los principios ágiles empiezan por un análisis del negocio.

b. **Involucración del cliente.** Sí
En las metodologías ágiles los proyectos se dividen en pequeñas iteraciones, realizando reuniones sobre el estado del proyecto al final de las mismas con lo que el cliente se encuentra más involucrado.
Los principios ágiles establecen también que el cliente sea siempre escuchado.

c. **Control de requisitos.** Sí
Se establece una prioridad de requisitos a cumplir al inicio del proyecto, la cual puede ser modificada en cualquier momento del mismo.
Si bien hay que tener cuidado de que por la modificación de requisitos el resultado final no derive demasiado con un control adecuado no debería de existir problema.

d. **Selección adecuada del equipo de trabajo.** No
En principio ninguna de las metodologías ágiles incorpora un proceso de selección del equipo de trabajo apoyado del análisis del negocio o de un pre-análisis de los datos

e. **Control de la comunicación entre el equipo de trabajo.** Sí
Se pueden establecer reuniones cíclicas entre los componentes del equipo. Uno de los principios ágiles es la comunicación “cara a cara”.

f. **Control de plazos y costes.** Sí
Al ser los ciclos del proyecto pequeños es más fácil controlar el plazo y coste de los mismos.

g. **Gestión de privacidad en los datos del cliente.** No
No existe un procedimiento de control de la privacidad de los datos en estas metodologías.

h. **Control de la implantación.** Sí
Se procura que la mejora sea siempre continua, parte de esto tiene que ver con el control de la implantación del proyecto en fases anteriores.

i. **Control del cambio.** Sí
Al existir ciclos de planificación corta es fácil realizar cambios en cualquier punto del proyecto

El mayor problema de la metodología ágil se encuentra en proyectos donde se implique a grupos de trabajo de elevado tamaño, donde las vías de comunicación establecidas por esta metodología puedan retrasar en exceso el proyecto.

4.2.3 ASUM-DM

a. **Conocimiento del negocio.** Sí
   Al igual que CRISP-DM ASUM-DM tiene una fase inicial de entendimiento del negocio.
b. **Involucración del cliente.** Sí
   Se involucra al cliente principalmente en la fase de "Análisis" y en la fase de "Implantación".
   La implicación del cliente en las fases intermedias viene dada a través de la fase paralela "gestión del proyecto", funcionando la dirección del proyecto como proxy entre el cliente y el equipo de proyecto.
c. **Control de requisitos.** Sí
   Sí, existe una fase en paralelo de "gestión del proyecto", encargada de gestionar los requisitos del mismo.
d. **Selección adecuada del equipo de trabajo.** No
   Si bien existen procesos de asignación de recursos esta se produce previamente a conocer los datos relevantes del proyecto.
e. **Control de la comunicación entre el equipo de trabajo.** Sí
   Sí, existe un proceso de gestión de la comunicación dentro de la fase de gestión del proyecto.
f. **Control de plazos y costes.** Sí
   Sí, existe un proceso adecuado dentro de la fase de gestión del proyecto.
g. **Gestión de privacidad en los datos del cliente.** No
   No existe un proceso de control que garantice la seguridad de los datos del cliente.
h. **Control de la implantación.** Sí
   Existe una fase destinada para tal fin con los procesos de apoyo adecuados.
i. **Control del cambio.** Sí
   Sí, existe un proceso adecuado para la gestión del cambio.

Desde el punto de vista de la gestión de proyectos la metodología propuesta por IBM "ASSUM-DM" es la más completa de las estudiadas, si bien carece de una menor definición respecto a las metodologías tradicionales en lo que respecta a la minería y análisis de datos.

4.2.4 OTRAS METODOLOGÍAS

El resto de metodologías propuestas no poseen el detalle suficiente respecto de sus procesos para hacer un análisis completo de las mismas. Entre estas destacan la empleada por Ramco Cements o la propuesta por Jose Solarte.

La metodología de Ramco destaca de otras por introducir el concepto de la elección de un equipo multidisciplinar apoyado por los datos de la investigación de mercado mediante que la metodología propuesta por Solarte destaca por su proposición de uso de matrices como apoyo a la elección o para validar resultados.

5 ABORDANDO LAS DEBILIDADES DE LAS METODOLOGÍAS TRADICIONALES

Siguiendo en la línea del análisis presentado en el capítulo anterior, en opinión del autor de este TFM las metodologías clásicas de gestión de proyectos de minería y análisis de datos tienen las siguientes debilidades una serie de debilidades, las cuales no son abordadas en el total de su conjunto por las metodologías alternativas propuestas por los distintos autores analizados durante el desarrollo del Estado del Arte.

Estas debilidades son las siguientes:

1. **Gestión de los objetivos y requisitos del proyecto:** En general en los proyectos de minería de datos los objetivos son abiertos porque no se sabe a dónde nos llevará su análisis o incluso hay ocasiones en que se analizan para buscar ventajas competitivas que no pueden intuirse preliminarmente al inicio del proyecto.
2. **Gestión de costes y plazos:** Al no conocer los objetivos finales del proyecto es difícil estimar el coste y los plazos necesarios para completar el mismo.
3. **Gestión del equipo de trabajo:** Es importante escoger de forma adecuada al equipo de trabajo de acuerdo con el proyecto planteado.
4. **Control de la comunicación.** Es importante mantener líneas de comunicación regulares entre el cliente y el equipo de trabajo, así como entre el propio equipo.
5. **Control de cambios en el área de negocio:** En las metodologías tradicionales no suele realizarse un control
6. **Gestión de la seguridad de los datos**: En los tiempos actuales parece necesario buscar metodologías que se encarguen de garantizar la seguridad de datos delicados de los clientes de forma que no se produzcan filtraciones

Para dar solución a las debilidades enumeradas se introducirá una serie de conceptos en los apartados siguientes, los cuáles se cree pueden mejorar de forma clara las metologías tradicionales. Mediante la aplicación de estos conceptos a continuación descritos se propondrá en el capítulo siguiente una metodología que dé respuesta a estas debilidades.

5.1 ABORDANDO LA GESTIÓN DE LOS OBJETIVOS DEL PROYECTO. EL CONCEPTO DE "MÍNIMO PRODUCTO VIABLE"

Uno de los problemas "raíz" de los proyectos de minería de datos es la determinación de los objetivos del proyecto. Habitualmente las organizaciones que se interesan en la realización de proyectos de minería y análisis de datos tienen objetivos simples, de concepción amplia y poco definidos. Por ejemplo, una organización puede iniciar uno de estos proyectos con la idea de reducir sus costes de operación, el tiempo que necesitan para entregar un producto, gestionar de forma más efectiva sus aprovisionamientos…

En estos casos la organización no tiene un objetivo muy amplio y no tiene una idea definida de cómo lograrlo, lo cual vendrá determinado por el análisis de los datos. Es por ello que la fase de análisis preliminar de los datos tiene una vital importancia y ha de servir para limitar el alcance y coste del proyecto lo máximo posible.

Otro problema a la hora de limitar el alcance de los proyectos de minería de datos es que como muchos otros tipos de proyectos de software se sitúan en un entorno cambiante. Los datos no siempre son estáticos si no que pueden variar, así como la aproximación inicial al tratamiento y análisis de los mismos no siempre puede ser correcta, es por ello que hay que reevaluar constantemente el alcance, el coste y plazo del proyecto.

La aplicación de los principios ágiles parece entonces adecuada, al menos para proyectos pequeños. El problema es que en proyectos de gran tamaño con implicación de una gran plantilla esto no es siempre viable y puede llegar a ser contraproducente, llevando a mayores retrasos y costes en la ejecución del proyecto.

Es en este punto donde nos gustaría proponer el concepto de "Mínimo Producto Viable" (MPV). Este término fue acuñado por Frank Robinson en 2001 y se refiere al producto capaz de satisfacer las necesidades del cliente en el menor plazo y el menor coste posible. El término es comúnmente visto el desarrollo del producto en paralelo a las necesidades del cliente.

Sí por ejemplo nuestro cliente tiene la necesidad de un medio de transporte; como se ve en la "Figura 21" el MPV será un monopatín, lo cual puede no satisfacer en alto grado al cliente, pero es el producto obtenible con menor esfuerzo que cumple sus necesidades. Este proceso suele emplearse en desarrollo de software y suele ser iterativo, así el primer programa que suele entregársele al cliente es el que cumple sus necesidades más urgentes de forma más rápida y conllevándole un menor coste, añadiendo de forma sucesiva funciones al programa según un equilibrio de necesidad del cliente, coste y tiempo.

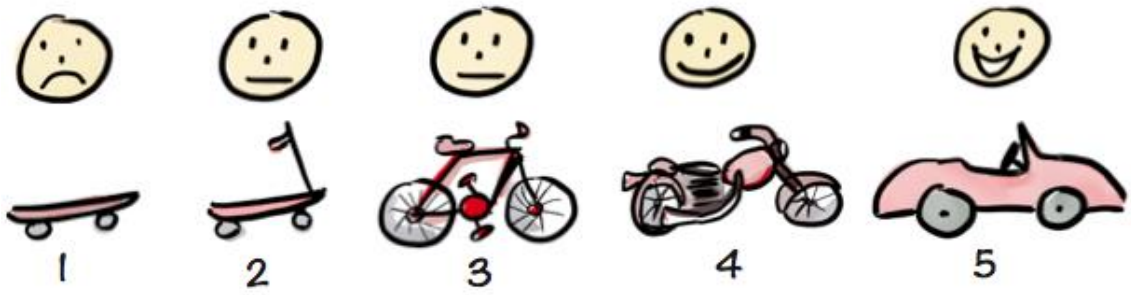

Figura 21. Desarrollo de un producto en fases según el concepto de "Mínimo Producto Viable". [35]

En este caso cuando lo aplicamos a los proyectos de minería de datos la consecuencia sería estudiar las necesidades del cliente, el tiempo y el coste que nos llevará analizar cada conjunto de datos, realizar la implantación de las soluciones extraídas de dichos datos, el entrenamiento de los individuos…, dividiendo el proyecto en distintas etapas en consecuencia del equilibrio entre las necesidades del cliente y los condicionantes expuestos.

Este es un método efectivo para controlar los objetivos del proyecto, así como limitar los riesgos adheridos al coste del mismo y reducir los plazos hasta que el cliente obtenga los primeros resultados.

La idea será determinar que fases del proyecto serán prioritarias en base a la conjunción de las necesidades del cliente, el tiempo y coste de desarrollo e implementación de la fase y la "disponibilidad". Con "disponibilidad" nos referimos a una serie de factores que afectan al plazo y coste como pueden ser si para implementar una medida hay que parar una máquina o proceso, si existen paradas programadas, si el mejor personal para realizar la tarea está disponible en ese momento…

5.2 GESTIÓN DEL EQUIPO DE TRABAJO. USO DE PRINCIPIOS ÁGILES

Una vez que determinamos las fases prioritarias del proyecto determinaremos cuales se pueden realizar en paralelo y cuales deben de aplazarse a ciclos de vida posteriores del proyecto. La idea en este caso será reducir al mínimo cada equipo implicado en cada fase del proyecto permitiendo que trabajen en paralelo, con cierta independencia, realizando tareas o "subproyectos" de tamaño suficiente para que puedan ser desarrollados con pequeñas desviaciones de coste y plazo.

Estos proyectos serán abordados mediante el uso de principios ágiles. Los proyectos de minería de datos, como se mencionó anteriormente, necesitan una capacidad de reacción rápida ante las variaciones de los datos recogidos o de las prioridades de la empresa, lo cual es beneficiado por el uso de técnicas de gestión ágil y un desarrollo del proyecto menos formal.

Estos conceptos serán empleados en las fases de minería y análisis de datos, como veremos en la metodología que describiremos en el capítulo siguiente (Metodología Propuesta).

5.3 CONTROL DEL PROYECTO. PMBOK Y LAS METODOLOGÍAS TRADICIONALES DE GESTIÓN DE PROYECTOS

Para la correcta ejecución del proyecto se propone la implantación de cuatro procesos de control. Un proceso que nos proteja sobre cambios en el entorno al que pertenece el conjunto de datos o de variaciones de patrones en los datos extraídos (**control del cambio)**, un proceso que controle que los requisitos no se desvían del objetivo y son compatibles entre las distintas fases del proyecto **(control de los requisitos),** un proceso que controle la correcta integración de los cambios propuestos en la organización **(control de la integración**) y un proceso que garantice la seguridad de los datos **("control de la seguridad").**

Para esto analizaremos lo que nos ofrecen las metodologías tradicionales de gestión de proyectos.

5.3.1 ¿QUÉ SON LAS METODOLOGÍAS TRADICIONALES DE GESTIÓN DE PROYECTOS?

Las metodologías tradicionales de gestión de proyectos, tal y como nos referimos en este Trabajo de Final de Master son aquellas metodologías que se componen de un conjunto de buenas prácticas en la gestión de proyectos que ayudan al "Director de Proyectos" a llevar a buen puerto el proyecto por aplicación de los conceptos descritos en las mismas.

Existen muchos autores que han realizado análisis comparativos de "distintas metodologías tradicionales de gestión de proyectos" No obstante, nos gustaría citar a un conjunto de autores que creemos han realizado un trabajo especialmente reseñable en este ámbito. Estos

son Sam Ghosh et al. [38] quienes realizaron un estudio entre las metodologías P2M, ICB, PRINCE2, APM y Scrum con el fin de analizar en que destaca cada una y en qué complenta al resto, llegando a la conclusión de que entre estas 6 metodologías PMBOK es la que destaca sobre el resto en la descripción de "herramientas y técnicas de gestión del proyecto"

Dentro de este conjunto de herramientas y técnicas se encuentran las de control de procesos, las cuáles cree el autor de este TFM que se corresponden con una de las debilidades de las *"metodologías de gestión de proyectos de minería y análisis de datos".*

Área de Mayor Énfasis	PMBOK	ICB	PRINCE2	P2M	APM	SCRUM
Ciclo de vida del Proyecto	✓			✓		
Programa y Portfolio				✓		
Gestión de procesos	✓			✓	✓	
Gestión del producto			✓			✓
Competencias del Director del Proyecto y habilidades interpersonales		✓			✓	
Colaboración con el cliente					✓	✓
Gestión del diseño y tecnología					✓	
Herramientas y técnicas de apoyo al proyecto	✓					
Equipo autónomo y altamente capacitado						✓
Entorno del proyecto		✓		✓	✓	
Caso de Negocio			✓		✓	
Transparencia en las comunicaciones						✓
Apoyo de la Dirección						✓
Liderazgo, Conflicto y Negociación		✓			✓	
Ingeniería de Innovación y Creación de Valor		✓			✓	
Marketing		✓			✓	
Rapidez de reacción al cambio						✓
Lecciones aprendidas				✓	✓	
Salud, Seguridad, Legislación		✓		✓	✓	
Identificación temprana de los riesgos			✓			✓
Entrega de elementos de alto valor primero						✓

Figura 22. Comparación de áreas de hincapié de las respectivas metodologías. Gosh et al. [38]

5.3.2 ¿QUÉ ES PMBOK?

PMBOK o Project Management Body of Knowledge, en castellano "Guía de los Fundamentos Para la Dirección de Proyectos) [39] es una guía publicada por el PMI (Project Management Institute) [40] que comprende un conjunto de conocimientos y prácticas aplicables a distintas situaciones formulables durante el desarrollo de un proyecto de cualquier tipo.

El PMBOK no debe entenderse como una metodología per se, sino como una guía de buenas prácticas, adaptables a proyectos de diversos campos. El PMBOK, por tanto, no ofrece un método sino un modo de obrar o proceder en distintas situaciones.

El PMBOK divide los proyectos en cuatro grupos de procesos Iniciación, Planificación, Ejecución, Seguimiento y Control y Cierre.

5.3.3 ¿EN QUÉ NOS PUEDE AYUDAR EL PMBOK EN UN PROYECTO DE MINERÍA Y ANÁLISIS DE DATOS?

PMBOK nos puede ayudar parcialmente en la "selección del equipo de proyecto", así como en tres de los procesos de control que definiremos:

- Control del Cambio
- Control de los Requisitos
- Control de la Integración

Los procesos del PMBOK en los cuáles nos apoyaremos para realizar estos procesos de control serán expuestos en el punto 8, "Caso de Uso".

Desafortunadamente PMBOK no dispone de procesos de control de la seguridad de los datos por lo que para cumplir este requerimiento necesitaremos irnos a una metodología más específicas.

5.4 EL CONTROL DE LA SEGURIDAD DE LOS DATOS

Es importante garantizar que los datos de las empresas u organizaciones participantes en los proyectos son confidenciales. En estos datos puede encontrarse el "know how" de la empresa, es decir aquella combinación de factores que hacen a la empresa competitiva.

Para el control de la "Seguridad en los Datos" se propone utilizar alguno de los conceptos propuestos en la metodología "Hermes 5". [41]

"Hermes 5" es una metodología desarrollada por la Administración Federal suiza dentro del programa de proyectos libres EGov de la eGovernment Switzerland Programme Office, siendo el proyecto nº 0054. Hermes 5 es también el standard de gestión de proyectos de la Administración Federal suiza.

Hermes se basa en módulos, los cuales define como "componentes reutilizables para la creación de escenarios". Un escenario se compone de varios módulos y a su vez un módulo puede formar parte de varios escenarios.

Cada escenario propuesto por la metodología Hermes 5 da respuesta a un "proyecto tipo" para el cual se propone el uso de una serie de módulos.

Cada módulo se compone de tareas y resultados y hace referencia a un área de conocimiento específica.

En este caso el módulo que nos interesa es el de "Seguridad de la Información y Protección de Datos" (Information Security and Data Protection) el cuál describe las tareas necesarias para establecer un proceso de control de acceso a los datos del proyecto.

6 METODOLOGÍA PROPUESTA

A continuación, se expone una propuesta de metodología que se considera apropiada para abordar proyectos multidisciplinares en entornos cambiantes, donde la variabilidad de los datos provoque numerosos cambios durante la vida del proyecto.

Así mismo esta metodología intenta abordar las necesidades que se considera pudieran surgir en un entorno empresarial como son el control de la implantación y el control de la seguridad de los datos.

6.1 DIAGRAMA DE PROCESO

Se sugiere un modelo de proceso conformado por las siguientes fases, las cuáles se describirán en apartados sucesivos.

El diagrama de proceso propuesto se compone de 7 fases obligatoria, 1 fase opcional y 4 procesos de control:

Las **fases** propuestas son:

- Visión de Negocio
- Análisis inicial de datos
- Selección de Prioridades
- Elección del Equipo de trabajo
- Minería de datos
- Análisis de datos
- Integración

La **fase opcional** es:

- Toma de datos

Los **procesos de control** son:

- Control del cambio
- Control de requisitos
- Control de la integración
- Control de seguridad de los datos

A continuación, en la “Figura 23” se muestra una propuesta de diagrama de proceso, mientras que las distintas fases del proyecto y procesos de control propuestos se describirán en apartados sucesivos.

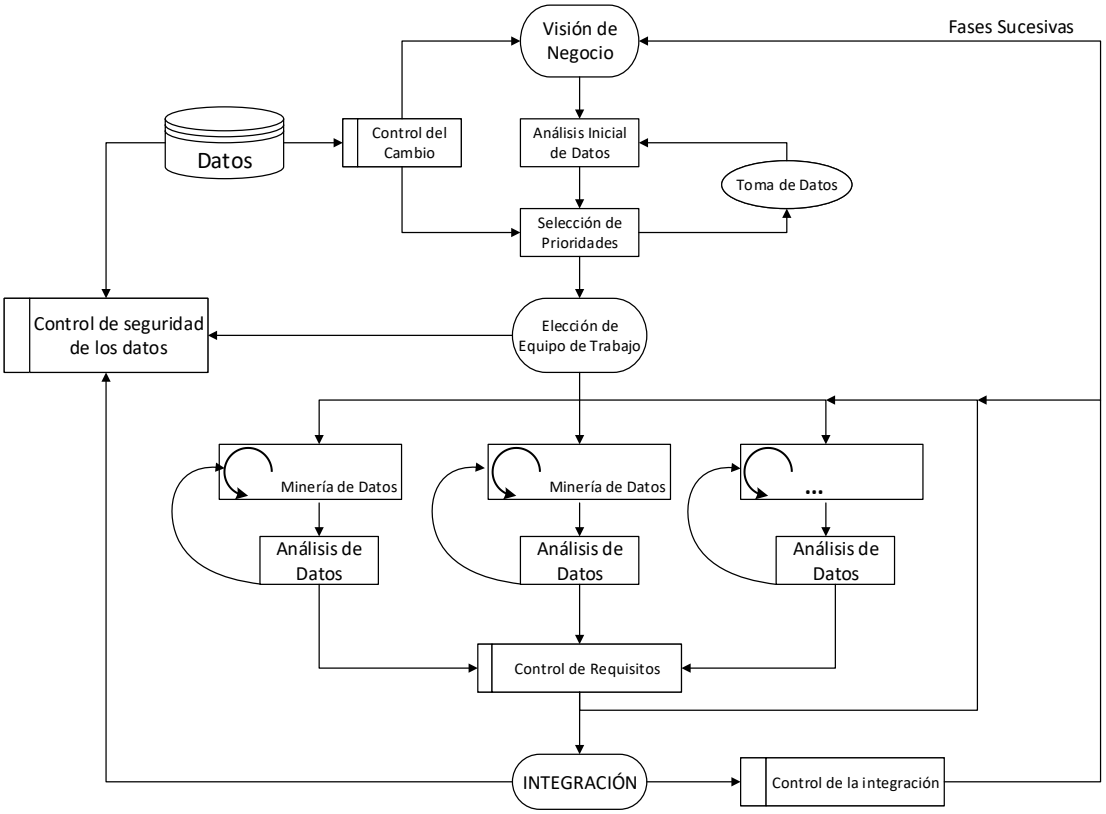

Figura 23. Esquema propuesto para minería de datos en auditorías energéticas

6.2 DESCRIPCIÓN DE FASES Y PROCESOS DE CONTROL

En este apartado se realizará una descripción de las fases que conforman la metodología propuesta, así como de los principales involucrados que poseen influencia en el desarrollo de cada una de estas fases.

Se detallará de igual modo que metodologías de uso tradicional aportan elementos relevantes a adoptar por la metodología propuesta durante el desarrollo de cada una de las fases a detallar.

6.2.1 VISIÓN DE NEGOCIO

INVOLUCRADOS:

- **Dirección Empresa contratante**
- **Dirección empresa contratada**
- **Director de proyecto**
- **Miembros relevantes de ambas empresas**

DESCRIPCIÓN:

Esta fase se compone de los siguientes pasos:

1. **Toma de contacto:** Se realiza una reunión entre ambas empresas en la que la empresa contratante expone su modelo de negocio, las tecnologías que implica su proceso, los recursos necesarios para el mismo y su situación respecto de la competencia.

2. **Determinación de objetivos:** Ambas empresas hacen un análisis previo y determinan los objetivos iniciales del proyecto.
3. **Reunión con interesados:** La empresa contratada realiza una reunión con los jefes de los departamentos relevantes en el consumo de la empresa (mantenimiento, operaciones, logística, adquisiciones…)

Al final de esta fase la empresa contratante cede los datos relevantes para el proyecto.

METODOLOGÍAS QUE APORTAN ELEMENTOS RELEVANTES

Fase de entendimiento de negocio (*Bussiness Understanding)* de CRISP-DM. En concreto las siguientes etapas.

- Background
- Asses Situation
- Data Mining Goals

6.2.2 ANÁLISIS INICIAL DE DATOS

INVOLUCRADOS:

- **Director de proyecto**
- **Áreas funcionales relevantes empresa contratada**

DESCRIPCIÓN:

En esta fase se hace un análisis previo de los datos por parte de las áreas funcionales relevantes de la empresa contratada. Este análisis, preliminar y muy superficial servirá para entender lo aprovechable de estos datos para dar solución a los objetivos del proyecto, los conocimientos y habilidades específicas necesarios para su análisis y aquellos datos adicionales que pudiesen ser necesarios.

METODOLOGÍAS QUE APORTAN ELEMENTOS RELEVANTES

Fase de **Entendimiento de los datos a analizar (Data Understanding)** de CRISP-DM.

6.2.3 TOMA DE DATOS

INVOLUCRADOS:

- **Áreas funcionales relevantes empresa contratada**
- **Áreas funcionales relevantes empresa contratante**

DESCRIPCIÓN:

En esta fase se realiza una toma de datos adicionales necesarios para extraer información que permita la consecución del proyecto. Esta fase puede tener una duración variable dependiendo de si es necesario implantar algún mecanismo específico para la obtención de los datos o del tiempo que pueda ser necesario realizar la obtención de los mismos.

Este proceso puede situarse en paralelo al proyecto utilizándose los datos extraídos durante este en futuras fases del mismo.

METODOLOGÍAS QUE APORTAN ELEMENTOS RELEVANTES

Esta fase comprende un proyecto a parte que puede transcurrir en paralelo al que ya se está ejecutando, su alcance dependerá de la amplitud que abarque el proyecto, así como aquellos recursos y cambios que implique en la empresa.

Para ello se utilizará una versión escalada de PMBOK, adoptando solo las partes que sean útiles para el proyecto de toma de datos.

En general se compondrá al menos de: Esta fase comprende un proyecto a parte que puede transcurrir en paralelo al que ya se está ejecutando, su alcance dependerá de la amplitud que abarque el proyecto, así como aquellos recursos y cambios que implique en la empresa.

Para ello se utilizará una versión escalada de PMBOK, adoptando solo las partes que sean útiles para el proyecto de toma de datos.

En general se compondrá al menos de:

Iniciación:

- Identificar a los Interesados

Planificación

- Recopilar Requisitos
- Definir el Alcance
- Crear la EDT
- Definir las Actividades
- Estimar los Recursos de las Actividades
- Estimar la Duración de las Actividades
- Estimar los Costos

Ejecución

- Gestionar la Ejecución del Proyecto
- Grupo del Proceso de Seguimiento y Control
- Monitorear y Controlar el Trabajo del Proyecto

Cierre

6.2.4 SELECCIÓN DE PRIORIDADES

INVOLUCRADOS:

- **Dirección Empresa contratante**
- **Dirección empresa contratada**
- **Director de proyecto**

- **Miembros relevantes de ambas empresas**

DESCRIPCIÓN:

A partir de las conclusiones extraídas en "Análisis Inicial de Datos" se deciden las prioridades del proyecto. Mediante matrices y usando el concepto de "Mínimo Producto Viable" se distribuye el proyecto en fases o tareas.

Las **fases** serán de aplicación secuencial y corresponderán a futuros ciclos en la vida del proyecto con la idea de mejora del mismo, corresponden por así decirlo a nuevos proyectos que den solución a los objetivos no abordados en el proyecto original o perfeccionen las soluciones dados a estos objetivos en el primer proyecto.

Las **tareas** son fases para las cuáles su aplicación es posible en paralelo dentro del mismo ciclo de vida del proyecto. Cada una de estas tareas será llevada a cabo por un equipo distinto.

La división del proyecto en fases y tareas se llevará a cabo mediante apoyo de elementos de decisión como pueden ser matrices, en base distintos criterios, como pueden ser:

- Ahorro económico.
- Inversión necesaria.
- Tiempo necesario para el análisis de datos.
- Tiempo de implantación de la solución.
- Necesidad de parada de actividad productiva de la empresa.
- …

Se realizará una discusión de las soluciones entre ambas empresas elaborando el **plan del proyecto** de común acuerdo.

METODOLOGÍAS QUE APORTAN ELEMENTOS RELEVANTES

Se utilizarán matrices de decisión como para ponderar la viabilidad de las distintas opciones.
Posteriormente se elaborará un plan de proyecto según el criterio del **"Mínimo Producto Viable"**.
Se producirá un plan de proyecto para lo cual podemos apoyarnos en el paso "**Producir un plan de proyecto (*Produce Project Plan*)"** de la fase Bussiness Understanding de CRISP-DM.

Ejemplo de matriz de prioridad de fases	Nivel de prioridad						
	Multiplicador		1	2	3	4	5
Interés del contratante	x1		Muy bajo	Bajo	Moderado	Alto	Muy alto
Nivel de impacto en objetivo de proyecto	x1,5		Muy bajo	Bajo	Moderado	Alto	Muy alto
Coste de implantación	x1		Muy bajo	Bajo	Moderado	Alto	Muy alto
Tiempo de implantación	x1		Muy bajo	Bajo	Moderado	Alto	Muy alto
Disponibilidad de recursos	x1		Muy mala	Mala	Moderada	Buena	Muy buena
Afección a sistema	x1		Muy baja	Baja	Moderada	Alta	Muy alta
Paradas programadas	Si afeccion a sistema ≥ 4 ptos	x1,5	> 18meses	12 - 18 meses	6 - 12 meses	3 - 6 meses	1-3 meses
	Si afeccion a sistema = 3 ptos	x1					
	Si afeccion a sistema ≤ 2 ptos	x0,5					

Figura 24. Ejemplo de matriz de selección de prioridades (los multiplicadores y elementos de evaluación pueden variar según el tipo de proyecto)

6.2.5 ELECCIÓN DEL EQUIPO DE TRABAJO

INVOLUCRADOS:

- **Director de proyecto**

DESCRIPCIÓN

Los equipos de trabajo por el director de proyecto en base a los conocimientos necesarios para la consecución de cada tarea. Se priorizará la formación de equipos pequeños para facilitar la comunicación entre sus miembros.

Cada uno de los equipos formados (equipos de proyecto) dispondrá de un miembro "coordinador" de la tarea, encargado de realizar la comunicación con el director del proyecto.

METODOLOGÍAS QUE APORTAN ELEMENTOS RELEVANTES

Paso de ***"formación del equipo multidisciplinar"*** de la metodología Ramco Cements (Fase Trabajo estratégico) o "Adquirir el Equipo de Proyecto" del PMBOK (Proceso de Ejecución).

6.2.6 MINERÍA DE DATOS Y ANÁLISIS DE DATOS

INVOLUCRADOS:

- **Equipos de proyecto**

DESCRIPCIÓN

Estas dos fases son realizadas por cada uno de los respectivos equipos de proyecto siguiendo el método clásico de análisis de datos. Se puede entender cada uno de estos proyectos como un pequeño proyecto de minería de datos, formado por los siguientes pasos:

Fase de **minería de datos**:

0. **Análisis previo**
1. **Selección de datos**
2. **Pre-procesado**
3. **Transformación**
4. **Minería de Datos**

Fase de **análisis de datos:**

5. **Interpretación/Evaluación**
6. **Propuesta de soluciones**

METODOLOGÍAS QUE APORTAN ELEMENTOS RELEVANTES

Se elaborarán en base a CRISP-DM o KDD, adoptando una estructura ágil como puede ser cualquiera de las descritas en el apartado 3.3.1

6.2.7 INTEGRACIÓN

INVOLUCRADOS:

- **Director de proyecto**
- **Equipos de proyecto de cada tarea**

DESCRIPCIÓN

El equipo de proyecto realiza un plan de integración de las soluciones extraídas en la etapa de análisis de datos el cuál debe de ser validado por el director de proyecto.

METODOLOGÍAS QUE APORTAN ELEMENTOS RELEVANTES

Pasos de "Aprobación de Modelos" y de "Determinación de nuevos pasos" de CRISP-DM (Fase de Evaluación).

6.2.8 CONTROL DE CAMBIO

INVOLUCRADOS:

- **Dirección Empresa contratante**
- **Miembros relevantes de empresa contratante**
- **Director de proyecto**
- **Miembros relevantes de empresa contratada**

DESCRIPCIÓN

El control del cambio está pensado para realizar cambios en el proyecto en el caso de que algo cambie en los datos, o procesos analizados, en el área de negocio de la empresa, en aquellos factores externos a la empresa (legislación, precios energía, materias primas), en el organigrama de la empresa (ciclos de paradas de mantenimiento) o en cualquier otro factor que pueda afectar al proyecto.

Se establecerán las vías necesarias de comunicación de la empresa contratante con el director de proyecto, así como los cambios aceptables sin que conlleven aumentos de plazo o coste del proyecto.

METODOLOGÍAS QUE APORTAN ELEMENTOS RELEVANTES

Uso de conceptos ágiles y los siguientes procesos del PMBOK:

- Planificar las Comunicaciones. (Grupo del Proceso de Planificación)
- Realizar el Control Integrado de Cambios. (Grupo del Proceso de Seguimiento y Control)
- Monitorear y Controlar los Riesgos. (Grupo del Proceso de Seguimiento y Control)

6.2.9 CONTROL DE REQUISITOS

INVOLUCRADOS:

- **Director de proyecto**
- **Coordinadores de tareas**

DESCRIPCIÓN

El director del proyecto se encargará de verificar junto a los coordinadores de las tareas que las soluciones hacia las que derive el proceso de minería y análisis de datos no se desvíen en exceso de los establecidos al inicio del proyecto.

METODOLOGÍAS QUE APORTAN ELEMENTOS RELEVANTES

Uso de los siguientes procesos del PMBOK:

- Identificación de Interesados (Proceso de Iniciación de PMBOK)
- Monitorear y Controlar el Trabajo del Proyecto.
- Verificar el Alcance.
- Controlar el Alcance.

6.2.10 CONTROL DE LA INTEGRACIÓN

INVOLUCRADOS:

- **Director de proyecto**
- **Miembros relevantes de empresa contratante**

DESCRIPCIÓN

El director del proyecto se reúne con los miembros relevantes de la empresa contratante para determinar que la integración de las diversas soluciones propuestas en las fases de implantación de cada tarea no choque entre sí y sean compatibles en su aplicación en la empresa.

Juntos determinarán también la posibilidad de planes de "entrenamiento" para los trabajadores de la empresa.

METODOLOGÍAS QUE APORTAN ELEMENTOS RELEVANTES

Uso de los siguientes procesos del PMBOK:

- Gestionar las Expectativas de los Interesados
- Dirigir y Gestionar la Ejecución del Proyecto
- Realizar el Aseguramiento de Calidad

6.2.11 CONTROL DE SEGURIDAD DE LOS DATOS

INVOLUCRADOS:

- **Director de proyecto**
- **Coordinadores de tarea**

DESCRIPCIÓN

El director de proyecto llevará a cabo un control de las personas que tienen acceso a los datos sensibles de la empresa contratante, específicamente en lo que respecta a los equipos de trabajo y aquellas personas que puedan entrar en contacto con estos datos durante la fase de "integración".

METODOLOGÍAS QUE APORTAN ELEMENTOS RELEVANTES

Uso de metodología **"Hermes 5", módulo de seguridad y protección de datos**. Este módulo se compone de los siguientes pasos:

- Determinar los requerimientos de seguridad y protección de los datos necesarios
- Determinar riesgos
- Implementar medidas de seguridad que den cumplimiento a los riesgos
- Crear una metodología de protección de los datos
- Documentar continuamente los resultados.

7 CONCLUSIONES Y LÍNEAS DE FUTURO

A partir del análisis del "Estado del Arte" de la gestión de proyectos de minería y análisis de datos se ha llegado a las siguientes conclusiones:

1. Las metodologías tradicionales de análisis de datos son perfectamente válidas para sectores tradicionalmente vinculados con el área de "Tecnologías de la Información". Estas metodologías siguen un proceso predefinido y son perfectamente válidas para proyectos y sectores tradicionalmente ligados con la minería de datos como pueden ser el de los seguros, el financiero, el matemático.
2. La aplicación de los principios ágiles beneficia ampliamente a los proyectos de minería de datos cuando los equipos que los desarrollan son de tamaño comedido. Esto es debido a las similitudes entre las metodologías ágiles y las metodologías tradicionales de minería de datos.
3. Cuanto mayor es el tamaño del proyecto y más específico en conocimientos necesarios para ser abordado las metodologías tradicionales de minería de datos empiezan a presentar problemas. La ausencia de procesos específicos que permitan abordar equipos multidisciplinares y organizar grandes grupos de trabajo se considera como el autor su gran "talón de Aquiles".
4. Hoy en día la información es cada vez más importante para las empresas y es por ello que se considera que las metodologías de gestión de proyectos en los cuáles se trabaje con "datos sensibles" para las empresas han de poseer procesos que garanticen el control y seguridad de la información tratada.

5. La integración del proyecto en la organización es uno de los pasos más importantes que presenta un proyecto de estas características y por tanto se cree que todo proyecto de minería y análisis de datos debe de presentar procesos que garanticen la integración de los resultados o medidas propuestas en la organización, prestando especial importancia a la armonización de distintas fases o etapas del proyecto en caso de existir estas.

Se considera que con la metodología propuesta se da solución a estas inquietudes. Esta presenta una simple aproximación inicial y simplificada. Se considera pudiera ser interesante como línea de futuro la redacción de una metodología completa que diese una solución más detallada a dichas inquietudes, aportando las herramientas necesarias para el uso de aquellos "Directores de Proyecto" que pudieran tener las mismas inquietudes que el autor de este Trabajo de Fin de Master.

8 REFERENCIAS

[1] R. Finos, «Statista,» 2017. [En línea]. Available: https://www.statista.com/statistics/254266/global-big-data-market-forecast/.

[2] L. Columbus, «Forbes - Where Big Data Jobs Will Be In 2015,» 2014. [En línea]. Available: https://www.forbes.com/sites/louiscolumbus/2014/12/29/where-big-data-jobs-will-be-in-2015/#33e85b50493c.

[3] Evans Data Corporation, Big Data and Advanced Analytics Survey 2015, Volume I, 2015.

[4] B. Marr, Big Data: 20 Mind-Boggling Facts Everyone Must Read, Forbes, 2015.

[5] «The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things,» IDC - Analyce the Future, 2014. [En línea]. Available: https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

[6] Cisco Systems, «The Zettabyte Era — Trends and Analysis,» 2016.

[7] C. Hagen, H. Evans, J. Miller, M. Ciobo, D. Wall y A. Yadav, «Big Data and the Creative Destruction of Today's Business Models,» *ATKearney,* 2013.

[8] G. Piatetsky-Shapiro, «Knowledge Discovery in Databases: 10 years after,» *KDnuggets,* 2000.

[9] R. J. Brachman y T. Anand, «The process of knowledge discovery in databases,» *Advances in knowledge discovery and data mining,* 1994.

[10] U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, «From Data Mining to Knowledge Discovery in Databases,» *American Assosiation for Artificial Science,* 1996.

[11] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, S. Colin y R. Wirth, «CRISP-DM 1.0: Step-by-step data mining guide,» 2000.

[12] N. Leaper, «A visual guide to CRISP-DM methodology,» 2009. [En línea]. Available: https://exde.files.wordpress.com/2009/03/crisp_visualguide.pdf.

[13] SAS Institute Inc., Data Mining Using SAS Enterprise Miner. A Case Study Approach. Second Edition, 2013.

[14] J. S. Saltz, «The Need for New Processes, Methodologies and Tools to Support Big Data Teams,» de *International Conference on Big Data (Big Data), IEEE*, Santa Clara, CA, USA , 2015.

[15] J. Kaskade, «CIOs & Big Data: What IT Teams Want Their CIOs to Know,» CSC - Infochimps, 2013. [En línea]. Available: blog.infochimps.com/2013/01/24/cios-big-data/.

[16] J. Pérez Segovia, «Definition and instantiation of an integrated data mining process,» de *Jornadas de Seguimiento de Proyectos*, Universidad Politécnica de Madrid, 2007.

[17] R. Way, Model deployment: the moment of truth, CORIOS REDPAPER, 2013.

[18] O. Niakšu, «Development and application of Data Mining methods in medical diagnostics and healthcare management,» Vilnius University, 2015.

[19] S. W. Lee y L. Kerschberg, «A methodology and life cycle model for data mining andknowledge discovery in precision agriculture,» de *IEEE International Conference on Systems* , 1998.

[20] M. Hofmann y B. Tierney, «Development phases of a generic data mining life cycle,» de *Proceedings of the International Conference on Software Engineering Theory and Practice* , 2007.

[21] M. Beedle, A. v. Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Highsmith, A. Hunt, R. Jeffries, J. Kern, B. Marick, R. C. Martin, K. Schwaber, J. Sutherland y D. Thomas, «Agile Manifesto,» 2001. [En línea]. Available: http://agilemanifesto.org/.

[22] Smartsheet, Agile Project Management 101.

[23] M. Alnoukari, Z. Alzoabi y A. E. Sheikh, «Applying ASD-DM Methodology on Business Intelligence Solutions: A Case Study on Building Customer Care Data Mart,» *Arab Academy for Banking and Financial Sciences,* 2008.

[24] G. Mariscal, O. Marbán y J. Segovia, «Un enfoque Ágil para el Desarrollo de Proyectos de Data Mining,» *Researchgate,* 2013.

[25] Marckx, Gino (Head of Agile Competency Center, EPAM Canada), «Big Data and Agile: The

Perfect Marriage (Webinar),» 2014.

[26] M. Łopuszyński, «Agile Approach to Data Mining Projects,» de *Warsaw Data Science Meetup*, 2016.

[27] G. Santana do Nascimento y A. d. A. Oliveira, «AgileKDD. An Agile Knowledge Discovery in Databases Process Model,» de *IMMM 2012 : The Second International Conference on Advances in Information Mining and Management*, 2012.

[28] P. Guo, «Data Science Workflow: Overview and Challenges,» 2013. [En línea]. Available: ACM.org.

[29] IBM, «Have you seen ASUM-DM?,» 2015. [En línea]. Available: https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/.

[30] IBM Analytics, «Analytics Solutions Unified Method. Implementations with Agile principles,» 2015.

[31] IBM, «IBM Analytics Solutions Unified Method (ASUM),» 2015. [En línea]. Available: https://www-01.ibm.com/marketing/iwm/iwm/web/pick.do?source=swerpba-basimext&lang=en_US.

[32] D. Dutta y I. Bose, «Managing a Big Data project: The case of Ramco Cements Limited,» *International Journal of Production Economics ,* nº 165, pp. 293-306, 2015.

[33] J. Solarte, «A Proposed Data Mining Methodology and its application to Industrial Engineering,» *Master Thesis. University of Tennessee - Knoxville,* 2002.

[34] J. Sim, «Critical Success Factors in Data Mining Projects,» *Tesis. Univesity of North Texas,* 2003.

[35] J. Ranjan y V. Bhatnagar, «Critical Success Factors For Implementing CRM Using Data Mining,» *Journal of Knowledge Management Practice,* vol. 9, nº 3, 2008 .

[36] A. Azevedo y M. F. Santos, «KDD, SEMMA AND CRISP-DM: a pararell overview,» *Researchgate,* 2008.

[37] U. Shafique y H. Qaiser, «A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA),» *International Journal of Innovation and Scientific Research,* vol. 12, nº 1, pp. 217-222, 2014.

[38] S. Ghosh, D. Forrest, T. DiNetta, B. Wolfe y D. C. Lambert, «Enhance PMBOK® by Comparing it with P2M, ICB, PRINCE2, APM and Scrum Project Management Standards,» *PM World Journal,* vol. IV, nº IX, 2015.

[39] Guía de los Fundamentos Para la Dirección de Proyectos (GUÍA DEL PMBOK®) Cuarta

Edición, PMI, 2008.

[40] «Project Management Institute, Inc,» [En línea]. Available: www.pmi.org.

[41] Switzerland's Federal Administration, «HERMES is a project management method for IT, services, products and business organisations.,» 2017. [En línea]. Available: http://www.hermes.admin.ch.