



UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática



TFG del Grado en Ingeniería
Informática

Semisupervised learning and
instance selection methods



Presentado por Daniel Puente Ramírez
en Universidad de Burgos — 14 de diciembre
de 2021

Tutor: Álgvar Arnáiz González



UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática



D. nombre tutor, profesor del departamento de nombre departamento, área de nombre área.

Expone:

Que el alumno D. Daniel Puente Ramírez, con DNI dni, ha realizado el Trabajo final de Grado en Ingeniería Informática titulado título de TFG.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 14 de diciembre de 2021

Vº. Bº. del Tutor:

D. nombre tutor

Resumen

En este primer apartado se hace una **breve** presentación del tema que se aborda en el proyecto.

Descriptores

Palabras separadas por comas que identifiquen el contenido del proyecto Ej: servidor web, buscador de vuelos, android ...

Abstract

A **brief** presentation of the topic addressed in the project.

Keywords

keywords separated by commas.

Índice general

Índice general	iii
Índice de figuras	iv
Índice de tablas	v
Introducción	1
Objetivos del proyecto	3
Conceptos teóricos	5
3.1. Minería de datos	5
3.2. Función distancia entre instancias	9
3.3. Aprendizaje en <i>machine learning</i>	11
Técnicas y herramientas	15
Aspectos relevantes del desarrollo del proyecto	17
Trabajos relacionados	19
Conclusiones y Líneas de trabajo futuras	21
Bibliografía	23

Índice de figuras

3.1. Enfoque CRISP de la minería de datos.	6
3.2. <i>Machine learning overview</i> [13]	11

Índice de tablas

Introducción

Descripción del contenido del trabajo y del estructura de la memoria y del resto de materiales entregados.

Objetivos del proyecto

Este apartado explica de forma precisa y concisa cuales son los objetivos que se persiguen con la realización del proyecto. Se puede distinguir entre los objetivos marcados por los requisitos del software a construir y los objetivos de carácter técnico que plantea a la hora de llevar a la práctica el proyecto.

Conceptos teóricos

El proyecto tiene una relación directa con la minería de datos y los conceptos que lo rodean.

3.1. Minería de datos

Según IBM [3], podemos definir la minería de datos, o descubrimiento de conocimiento en los datos *knowledge Discovery in Databases*, como el proceso de descubrir patrones y otra información a partir de grandes conjuntos de datos.

Las técnicas de minería de datos principales se pueden dividir en función de sus propósitos principales.

1. Descripción del conjunto de datos objetivo.
2. Predicción de resultados mediante el uso de algoritmos de aprendizaje automático.

Proceso de minería de datos

El proceso seguido en la minería de datos es muy directo. Comienza con la recogida de los datos que van a ser tratados, y finaliza con la visualización de la información extraída de éstos. Los científicos de datos describen los datos a través de sus observaciones de patrones, asociaciones y correlaciones. A su vez se pueden clasificar y agrupar los datos utilizando métodos de clasificación y regresión.

Uno de los marcos de referencia más importantes en el proceso de minado de datos es CRISP-DM, *Cross Industry Standard Process for Data Mining*. Desarrollado por un consorcio de empresas involucradas en la minería de datos. [2]

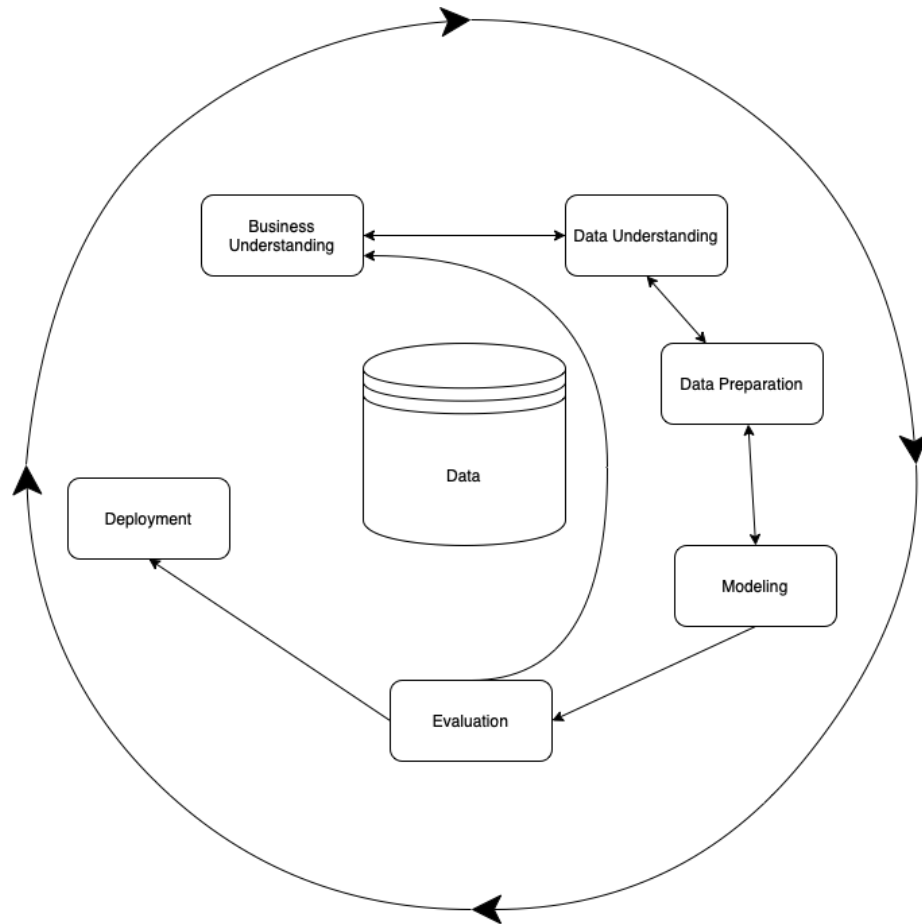


Figura 3.1: Enfoque CRISP de la minería de datos.

En [8] se divide el proceso de la minería de datos en 5 etapas o pasos principales: establecimiento de los objetivos y comprensión del problema, recopilación y preparación de los datos, desarrollo del modelo, aplicación del modelo y la evaluación de los resultados y despliegue en producción.

1. **Establecer los objetivos y comprensión del problema.** La primera etapa puede resultar la más complicada del proceso. Todas las partes interesadas deben de estar presentes y de acuerdo en la definición del problema que se va tratar, esto incluye tanto a los científicos

de datos como las terceras partes involucradas o interesadas. Este procedimiento ayuda a la formulación de las preguntas de los datos y los parámetros a utilizar en el proyecto. Si se trata de un proyecto empresarial, se debe hacer un estudio o investigación adicional para comprender el contexto de la empresa.

2. **Preparación de los datos.** Con el alcance del problema definido ya se puede comenzar a identificar qué conjunto de datos será el más efectivo o representativo con el fin de comenzar a dar respuesta a las preguntas formuladas en el proceso anterior.

Una vez se tienen todos los datos recogidos comienza el proceso de pre-procesado de los mismos. Este proceso se basa en la limpieza de los datos con el fin de eliminar cualquier posible ruido, entendiéndose por ruido los datos duplicados, los valores perdidos y aquellos atípicos; aquellos que puedan causar problemas a la resolución del problema o generen incertidumbre. En determinados conjuntos de datos se puede hacer una reducción de dimensiones. Consiste en la reducción del número de dimensiones que poseen las instancias recogidas, con el fin de eliminar aquellas que no sean realmente representativas o significativas, este proceso reduce la complejidad de los cálculos posteriores. Por contrapartida hay que conocer cuáles serán los predictores con mayor relevancia en el problema para garantizar una precisión "óptima" del modelo.

3. **Desarrollo del modelo.** Según [8] el modelo es la representación abstracta de los datos y sus relaciones en un conjunto de datos concreto. Actualmente existen cientos de algoritmos que se pueden utilizar, habitualmente proceden de campos como la ciencia de datos, *machine learning*, o la estadística. Se debe tener el conocimiento suficiente para entender como funciona el algoritmo para poder configurar correctamente los parámetros que este va a utilizar en base a los datos y el problema de negocio que estamos resolviendo.

Los modelos en función de como resuelvan el problema que se les presenta se pueden clasificar en:

- a) Regresión.
- b) Análisis de asociación.
- c) *Clustering*.
- d) Detección de anomalías.

El modelo debe ser creado con especial cuidado para evitar el *overfitting*, i.e. el modelo memoriza el conjunto de entrenamiento y no tendrá un rendimiento correcto una vez desplegado en producción. Se desea que el modelo sea lo más generalizado posible de cara a *aprender* de los datos del conjunto de entrenamiento.

4. **Aplicación del modelo.** El momento de la aplicación del modelo es cuando de verdad se comprueba si realmente el modelo está listo para pasar al siguiente punto, en otras palabras, si es apto para ser desplegado en producción. Para ello se tienen en cuenta métricas como la calidad del modelo ante el problema, su tiempo de respuesta, etc.
5. **Evaluación de los resultados y despliegue en producción.** Una vez que el modelo se encuentra listo es desplegado en producción. Es habitual que los parámetros con los que el modelo fue entrenado con el paso del tiempo dejen de ser los más interesantes, pudiendo ser comprobado el error proporcionado por el modelo con los datos de prueba. Cuando ese error sea excesivo o fuera de un margen dado se deberá de volver a entrenar el modelo, comprobar, y desplegar. De esta forma se puede comprobar como el ciclo de vida del modelo es circular.

El proceso aplicado en la minería de datos proporciona un marco de trabajo mediante el cual se permite extraer información aparentemente no trivial de grandes conjuntos de datos. Es un campo de aprendizaje constante, tanto el aplicar los conocimientos del analista para reducir las dimensiones del conjunto de datos, como una vez que se ha entrenado el modelo y puesto en producción aprender los puntos fuertes de este y el porqué de éstos.[2]

Técnicas utilizadas en la minería de datos

Como se ha comentado anteriormente, uno de los mayores problemas de cara al minado de datos es la dimensionalidad que en muchas ocasiones tienen estos. Es por ello que se aplican técnicas o algoritmos que faciliten la extracción de la información útil. Algunos de ellos son:

1. **Reglas de asociación.** Dado un conjunto de datos concreto, consiste en la aplicación de reglas para encontrar relaciones entre las variables.
2. **Redes neuronales.** Principalmente utilizadas en *deep learning*, simulan la interconectividad propia del cerebro humano utilizando capas

de nodos. Cada nodo está compuesto por x_n entradas, w_n pesos y un sesgo o umbral, el cual al ser superado activa la neurona, pasando los datos del nodo a la siguiente neurona. Habitualmente con una única iteración sobre la red neuronal, esta es capaz de obtener una solución medianamente buena de un conjunto de datos de tamaño considerable.

3. **Árboles de decisión.** Mediante el uso de métodos de clasificación y regresión se clasifican o predicen potenciales resultados en función de un conjunto de decisiones. Utiliza una visualización en forma de árbol para representar los posibles resultados de estas decisiones.
4. **k-vecinos más cercanos. KNN (*k-nearest neighbors*)** Método no paramétrico de clasificación, sencillo pero eficaz. [5] Para clasificar un conjunto de datos T , se recuperan sus k vecinos más cercanos, que forman una vecindad de T . Se suele utilizar la votación por mayoría entre los registros de datos de la vecindad para decidir la clasificación de T con o sin consideración de la ponderación basada en la distancia. [4]

3.2. Función distancia entre instancias

La función distancia proporciona la proximidad entre dos instancias en función de todos sus parámetros. Si la distancia que separa dos instancias es cero, ambas instancias son idénticas. Se tiende a trabajar con conjuntos de datos normalizados, i.e. todos los datos son ajustados a una escala común, independientemente de la escala en la que hayan sido medidos.

Existen multitud de métricas para calcular la distancia, pudiendo variar la distancia en función de cuál se aplique. Se van a analizar las más habituales.

- **Distancia de Minkowski.** La distancia de Minkowski es una métrica en el espacio vectorial normalizado. Es una métrica que se puede modificar con facilidad para calcular la distancia entre dos instancias de diferentes maneras.
 1. $p = 1$, cálculo de la distancia de Manhattan.
 2. $p = 2$, cálculo de la distancia Euclídea.
 3. $p = \infty$, cálculo de la distancia de Chebyshev.

Su fórmula es la siguiente:

$$\mathbb{D}(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

- **Distancia de Manhattan** o distancia del taxista. Es una métrica en un espacio vectorial normalizado, calculándose como la suma de los n segmentos verticales u horizontales que unen dos puntos.

Su fórmula es la siguiente:

$$\mathbb{D}(x, y) = \sum_{i=1}^d |x_i - y_i|$$

- **Distancia euclidiana** o norma L2 o distancia L2. Es la distancia en línea recta entre dos puntos de datos en el espacio euclidiano.

Su fórmula normalizada es la siguiente:

$$\mathbb{D}(x, y) = \sqrt{\sum_{i=1}^d \frac{(x_i - y_i)^2}{\sigma_i^2}}$$

- **Distancia de Chebyshev** o distancia del tablero de ajedrez. La distancia entre dos puntos es la mayor de sus diferencias a lo largo de cualquiera de sus dimensiones coordenadas.

Su fórmula es la siguiente:

$$\mathbb{D}(x, y) = \max_i (|x_i - y_i|)$$

- **Distancia del Coseno**. Mide la similitud o distancia entre dos vectores calculando el coseno del ángulo que forman.

Su fórmula es la siguiente:

$$\mathbb{D}(x, y) = \frac{\sum_{i=1}^d x_i y_i}{\sqrt{\sum_{i=1}^d x_i^2} \sqrt{\sum_{i=1}^d y_i^2}}$$

- **Distribución de Pearson** o χ^2 .

3.3. Aprendizaje en *machine learning*

En [12] se define *machine learning* como una rama dentro del campo de la Inteligencia Artificial que proporciona a los sistemas la capacidad de aprender y mejorar de manera automática, a partir de la experiencia. Estos sistemas transforman los datos en información, y con esta información pueden tomar decisiones. Este tipo de modelos se crean a base del uso masivo de datos. Cuando se dispone de los datos suficientes para entrenar un modelo comienza el proceso de aprendizaje. El objetivo de este aprendizaje es descubrir patrones ocultos en los datos. En muchas ocasiones el resultado del aprendizaje, el modelo, es una función que dadas unos datos de entrada clasifica o predice correctamente una salida. Como se puede ver en la Figura 3.2 el aprendizaje automático, *machine learning*, posee diferentes aproximaciones, cada una de ellas con una aproximación diferente en cuanto al uso de instancias etiquetadas.

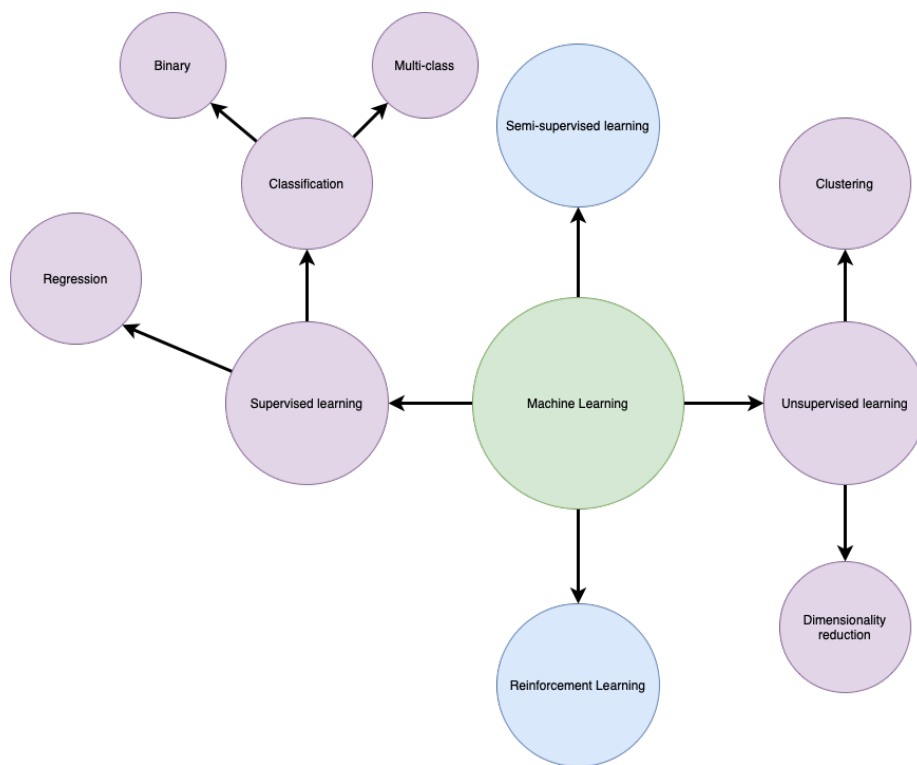


Figura 3.2: *Machine learning overview*[13]

Aprendizaje supervisado

El aprendizaje automático puede ser resumido como aprender de ejemplos. Al programa se le proporcionan dos conjuntos de datos, uno de entrenamiento y otro de validación.[9] El objetivo es simple, debe de “aprender” en función del conjunto de datos etiquetado proporcionado como entrenamiento para posteriormente identificar las correspondientes etiqueta/s de cada instancia del conjunto de validación con la mayor precisión posible.

Dependiendo del tipo de etiqueta, en el aprendizaje supervisado hay dos modelos.[6]

1. **Modelos de clasificación.** Producen como salida una etiqueta discreta, i.e. una etiqueta dentro de un conjunto finito de etiquetas, habitualmente suelen ser o binarias $[0, 1]$, $[sí, no]$... o multi-etiqueta, donde los valores pueden variar $[0...n]$ (no teniendo porque ser exclusivamente numéricas).

Entre los algoritmos de clasificación más frecuentes encontramos:

- Regresión logística.
- *Support Vector Machine, SVM.*
- Redes neuronales.
- Clasificador Naïve Bayes.
- Árbol de decisión.
- Análisis discriminante.
- K vecinos más cercanos, *KNN*.
- Clasificación de ensembles.

2. **Modelos de regresión.** Producen como salida un valor real, numérico. Suelen ser soluciones continuas.

Entre los algoritmos de regresión más frecuentes encontramos:

- Regresión lineal.
- Regresión no lineal.
- Modelo lineal generalizado.
- Árbol de decisión.
- Redes neuronales.
- Regresión con procesos gaussianos.

- Regresión con *support vector machines*.
- Regresión con ensembles.

Aprendizaje no supervisado

En la Sección 3.3 se comenta que, los modelos para que “aprendan” los patrones que se encuentran en los conjuntos de datos, necesitan tener un conjunto de datos etiquetado correctamente para extraer la información de ese conjunto. Pero en los problemas del mundo real no siempre se tienen infinitud de datos disponibles etiquetados correctamente, o simplemente es un proceso muy laborioso y costoso económicamente.

Para solventar este problema se cuenta con el aprendizaje no supervisado[1], mediante esta técnica no es necesario proporcionar al modelo datos etiquetados. Por definición, el algoritmo encargado de entrenar el modelo “aprenderá” los datos sin conocimiento previo. Para ello el modelo se basará en los datos que tiene disponibles y en la codificación del algoritmo para descubrir los patrones que se encuentren en los datos.

Debido a la forma de trabajar del aprendizaje no supervisado, desde el primer momento en el que el algoritmo tiene los datos comienza a reportar salidas, describiendo la información y categorizando lo que encuentra en los datos.

Principalmente existen dos técnicas de aprendizaje no supervisado.

1. **Clustering.**[11] Proceso por el cual se dividen los datos no clasificado en grupos aparentemente similares. Cuando se identifican datos con algún parecido entre sí, son agrupados. Permite clasificar e identificar atributos únicos de los datos con los que clasificarlos.

Un proceso habitual de agrupamiento es el uso de *K-means*, $K \in \mathbb{R}$, donde se indica en K cuántos *clusters* o grupos se han de hacer con los datos.

Con los datos agrupados el proceso de análisis de éstos puede comenzar. En ocasiones si el número de grupos detectados es muy alto, se pueden encontrar grupos o *clusters* irrelevantes, permitiendo a los científicos de datos eliminar esos datos que los forman, reduciendo la dimensionalidad.

2. **Reducción de la dimensionalidad.** La clasificación en el aprendizaje automático se basa en atributos o características que tienen

los datos, permitiendo su clasificación, valga la redundancia. Cuando los conjuntos de datos poseen múltiples características, más difícil resulta su clasificación. Es por ello que resulta útil identificar aquellos atributos que están fuertemente interrelacionados entre sí para eliminar todos menos un atributo, reduciendo la dimensionalidad.[10]

Aprendizaje semi-supervisado

Semi-Supervised Learning según [15], se define como una forma de entrenamiento de modelos el cual usa tanto datos etiquetados como no etiquetados, i.e. si no sería un aprendizaje supervisado, Sección 3.3, o no supervisado, Sección 3.3.

El uso de aprendizaje semi-supervisado se caracteriza por ser más barato que el supervisado, ya que este último necesita que todo el conjunto de datos que va a utilizar para aprender esté etiquetado, y ese proceso es largo y costoso. Además, obtiene mejores resultados en menor tiempo que el aprendizaje no supervisado. Conseguir datos sin etiquetar es una tarea muy sencilla, mientras que conseguir conjuntos de datos etiquetados es un proceso complejo y actualmente no hay “de todo”.

Para que el aprendizaje sea fructuoso requiere que las instancias se encuentren inter-relacionadas entre sí por alguna de sus características. [7] indica las siguientes suposiciones que se dan en el aprendizaje semi-supervisado.

1. **Continuidad.** Se asume que los objetos cercanos entre sí se encontrarán en el mismo *cluster* o grupo — de etiquetas.
2. **Clustering.** Las instancias son divididas en diferentes grupos discretos, compartiendo todos los elementos de un *cluster* la misma etiqueta.
3. **Manifold** o colectores. Se emplea el uso de distancias y funciones de densidad de forma que las instancias se encuentran en colectores con menos dimensiones que el espacio de entrada.

Dentro de las *best practices* en *semi-supervised learning* se encuentran el uso de diferentes modelos de redes neuronales para el entrenamiento. [14]

Técnicas y herramientas

Esta parte de la memoria tiene como objetivo presentar las técnicas metodológicas y las herramientas de desarrollo que se han utilizado para llevar a cabo el proyecto. Si se han estudiado diferentes alternativas de metodologías, herramientas, bibliotecas se puede hacer un resumen de los aspectos más destacados de cada alternativa, incluyendo comparativas entre las distintas opciones y una justificación de las elecciones realizadas. No se pretende que este apartado se convierta en un capítulo de un libro dedicado a cada una de las alternativas, sino comentar los aspectos más destacados de cada opción, con un repaso somero a los fundamentos esenciales y referencias bibliográficas para que el lector pueda ampliar su conocimiento sobre el tema.

Aspectos relevantes del desarrollo del proyecto

Este apartado pretende recoger los aspectos más interesantes del desarrollo del proyecto, comentados por los autores del mismo. Debe incluir desde la exposición del ciclo de vida utilizado, hasta los detalles de mayor relevancia de las fases de análisis, diseño e implementación. Se busca que no sea una mera operación de copiar y pegar diagramas y extractos del código fuente, sino que realmente se justifiquen los caminos de solución que se han tomado, especialmente aquellos que no sean triviales. Puede ser el lugar más adecuado para documentar los aspectos más interesantes del diseño y de la implementación, con un mayor hincapié en aspectos tales como el tipo de arquitectura elegido, los índices de las tablas de la base de datos, normalización y desnormalización, distribución en ficheros³, reglas de negocio dentro de las bases de datos (EDVHV GH GDWRV DFWLYDV), aspectos de desarrollo relacionados con el WWW... Este apartado, debe convertirse en el resumen de la experiencia práctica del proyecto, y por sí mismo justifica que la memoria se convierta en un documento útil, fuente de referencia para los autores, los tutores y futuros alumnos.

Trabajos relacionados

Este apartado sería parecido a un estado del arte de una tesis o tesina. En un trabajo final grado no parece obligada su presencia, aunque se puede dejar a juicio del tutor el incluir un pequeño resumen comentado de los trabajos y proyectos ya realizados en el campo del proyecto en curso.

Conclusiones y Líneas de trabajo futuras

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas. Además, resulta muy útil realizar un informe crítico indicando cómo se puede mejorar el proyecto, o cómo se puede continuar trabajando en la línea del proyecto realizado.

Bibliografía

- [1] Yoshua Bengio, Aaron C Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, *abs/1206.5538*, 1:2012, 2012.
- [2] Peter Chapman, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas P. Reinartz, Colin Shearer, and Richard Wirth. Crisp-dm 1.0: Step-by-step data mining guide. 2000.
- [3] IBM Cloud Education. What is data mining?, 2021.
- [4] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *OTM Confederated International Conferences. On the Move to Meaningful Internet Systems*, pages 986–996. Springer, 2003.
- [5] David J Hand. Principles of data mining. *Drug safety*, 30(7):621–622, 2007.
- [6] Mathworks Inc. Supervised learning.
- [7] JavaTPoint. Introduction to semi-supervised learning - javatpoint.
- [8] Vijay Kotu and Bala Deshpande. Chapter 2 - data mining process. In Vijay Kotu and Bala Deshpande, editors, *Predictive Analytics and Data Mining*, pages 17–36. Morgan Kaufmann, Boston, 2015.
- [9] Erik G Learned-Miller. Introduction to supervised learning. *I: Department of Computer Science, University of Massachusetts*, 2014.

- [10] Cen Li and Gautam Biswas. Unsupervised learning with mixed numeric and nominal data. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):673–690, 2002.
- [11] Onesmus Mbaabu. Clustering in unsupervised machine learning.
- [12] Jose Antonio Sanchez. ¿cómo aprenden las máquinas? machine learning y sus diferentes tipos, Aug 2020.
- [13] Technovert. Introduction to machine learning, 2020.
- [14] Kiran K Thekumparampil, Chong Wang, Sewoong Oh, and Li-Jia Li. Attention-based graph neural network for semi-supervised learning. *arXiv preprint arXiv:1803.03735*, 2018.
- [15] Xueyuan Zhou and Mikhail Belkin. Semi-supervised learning. In *Academic Press Library in Signal Processing*, volume 1, pages 1239–1269. Elsevier, 2014.