Universidad Autónoma de Querétaro

Facultad de Ingeniería

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

21 de octubre de 2023

Autor: *Ing. Carlos Manuel Sánchez Martínez*

Machine Learning (curso), Dr. Marco Antonio Aceves Fernández

Examen 1: Clustering con K-Nearest Neighbors

Resumen

En esta práctica, abordaremos el tema de clusterización, el cual, fue aplicado a una base de datos pública que contiene información del Titanic. A través del análisis de la base de datos, se ha demostrado la importancia de un riguroso preprocesamiento para tratar las anomalías y garantizar la calidad de los datos. El algoritmo k-NN se aplicó con éxito para predecir la supervivencia de los pasajeros basándose en características clave, destacando la relevancia de seleccionar un número óptimo de vecinos, en este caso 4. Esta investigación no solo resalta la eficacia de las técnicas de minería de datos en contextos históricos, sino que también enfatiza el potencial de tales técnicas para proporcionar insights valiosos y enriquecer nuestra comprensión de eventos pasados.

Palabras clave: K-NN, Minería de Datos, Titanic , Clustering, Exploración de Datos.

1. Introducción

El trágico suceso del Titanic ha capturado la atención del mundo durante más de un siglo. Gracias a los avances en el área de las TIC's, contamos con bases de datos sobre los pasajeros que estuvieron a bordo de este icónico barco con mayor facilidad y la disponibilidad de cualquier usuario de internet [22][23]. En esta práctica, exploramos una metodología para evaluar y clasificar a los pasajeros del Titanic, mediante la aplicación de técnicas de minería de datos. Con ello, buscamos descubrir patrones, tendencias o relaciones ocultas que puedan ofrecer una comprensión más profunda sobre el evento y las circunstancias de quienes viajaban en él, lo que puede ser útil para investigaciones históricas y la generación de nuevos conocimientos.

Al abordar la base de datos del Titanic, es esencial reconocer que, al igual que muchas bases de datos históricas, no está exenta de imperfecciones. Debido a la naturaleza del evento y las diversas fuentes de donde se han recolectado los datos, es común encontrar incongruencias, datos faltantes y elementos no normalizados. El preprocesamiento de datos se convierte en una etapa crucial en este proyantes de realizar cualquier tipo de análisis con técnicas de minería de datos.

El algoritmo k-NN (k-Nearest Neighbors) es una técnica de clasificación supervisada que se utiliza para asignar una clase a una nueva instancia basada en las clases de sus k vecinos más cercanos en el conjunto de datos [8]. Es ampliamente valorado por su simplicidad y eficacia en la predicción para calcular la probabilidad de las densidades, usado por US Air Force School of Aviation Medicine, y reportado en 1951 [18]. En nuestro caso, el algoritmo propuesto podría ser utilizado para predecir la supervivencia de un pasajero basándose en ciertas características, como la clase del pasaje, edad, género, entre otros. Dado un nuevo registro, o un pasajero cuyos datos de supervivencia no se conocen, k-NN buscaría en el conjunto de datos los k vecinos más cercanos (por ejemplo, otros pasajeros) que tienen características similares y basaría la predicción en el resultado más común entre estos vecinos (por ejemplo, si sobrevivieron o no).

2. Marco teórico

La ciencia, la tecnología y la innovación son principales factores del desarrollo económico sustentable [7]. En las últimas décadas, el campo del análisis de datos multivariados ha experimentado un crecimiento significativo, impulsado por la necesidad de abordar problemas complejos en diversos campos como la estadística, la economía, la Sociología, ingeniería, Medicina, entre otros [PEÑA]. Por ello, Los métodos de clasificación y análisis de observaciones multivariadas se han convertido en herramientas esenciales en el mundo actual.

El Aprendizaje de Máquinas (AM) es un subcampo de la inteligencia artificial (IA) dedicado al desarrollo de algoritmos y modelos que permiten a los sistemas mejorar su rendimiento en una tarea específica mediante la experiencia [15]. El Aprendizaje de Máquina puede categorizarse principalmente según el tipo de aprendizaje o entrenamiento que implementan los algoritmos:

1. Aprendizaje Supervisado: El Aprendizaje Supervisado es un enfoque donde los modelos se entrenan utilizando un conjunto de datos etiquetado, es decir, cada muestra de entrenamiento está asociada con una etiqueta o resultado [26]. Este método es ampliamente aplicado en tareas como clasificación y regresión. Lo podemos definir como un proceso en el que un modelo es entrenado mediante un conjunto de datos de entrada donde las respuestas correctas son conocidas. El modelo hace predicciones o clasificaciones basadas en la entrada y es corregido cuando sus predicciones son incorrectas [19].
2. Aprendizaje No Supervisado: Este tipo de enfoque involucra modelos que trabajan con conjuntos de datos sin etiquetas previas. El objetivo principal es explorar la estructura subyacente, así mismo busca identificar patrones, agrupaciones o anomalías en el conjunto de datos [26]. Tenemos diferentes ventajas al aplicar este tipo de entrenamiento, ejemplo: Clustering: Agrupa datos similares basados en características [11]. Y otro puede ser la reducción de dimensionalidad: Reduce el número de variables en un conjunto de datos mientras retiene la mayoría de la información[13].

3. Aprendizaje Semi-Supervisado: El Aprendizaje Semi-Supervisado (ASS) se sitúa entre el aprendizaje supervisado y no supervisado, aprovechando tanto datos etiquetados como no etiquetados para construir modelos más efectivos. Este enfoque es esencial cuando se cuenta con una cantidad limitada de datos etiquetados y una gran cantidad de datos no etiquetados, lo cual es un escenario común en la realidad [3]. Al hacerlo, el ASS puede mejorar significativamente la eficiencia del aprendizaje y la calidad del modelo resultante, mientras minimiza la necesidad de datos etiquetados extensos y costosos[26].

2.0.1. Clustering

El Clustering es un algoritmo que pertenece a las técnicas aprendizaje no supervisado, la cual agrupa datos con base en similitudes, facilitando la identificación de patrones y la toma de decisiones [7]. Mediante algoritmos como K-Means[14], DBSCAN [20], y jerárquicos [24], los datos se organizan en grupos llamados clusters que reflejan relaciones intrínsecas, permitiendo la identificación de estructuras subyacentes. El clustering se aplica extensamente en segmentación de mercado, análisis de redes sociales y detección de anomalías [11].

1. Centroide: El centroide es un concepto fundamental en la geometría y el análisis de datos. Se refiere al punto central o núcleo de un clúster, representando una síntesis o promedio de todos los puntos que pertenecen a ese clúster [5].

   Definición Matemática: En términos matemáticos, el centroide de un conjunto de puntos en un espacio N-dimensional se calcula como el promedio aritmético de las coordenadas de todos los puntos en ese conjunto. Para un conjunto de puntos $X = \{x_1, x_2, ..., x_n\}$, el centroide C se calcula como:

$$C = \frac{1}{n} \sum_{i=1}^{n} x_i$$

2.0.2. k-Nearest Neighbors (K-NN)

El algoritmo K-NN es una técnica supervisada de clasificación ampliamente utilizada para categorizar una nueva instancia basada en las clases de sus k vecinos más cercanos presentes en el conjunto de entrenamiento. El número k es definido por el usuario. La decisión sobre la clase de la nueva instancia se toma por mayoría, es decir, la clase que más prevalezca entre los k vecinos más cercanos es la que se asignará a la nueva instancia. La métrica de distancia, comúnmente la distancia euclidiana, se utiliza para identificar los vecinos más cercanos, aunque otras métricas pueden ser adoptadas [8].

Definición Matemática: Dado un conjunto de entrenamiento $T = \{t_1, t_2, ..., t_n\}$ con etiquetas $L = \{l_1, l_2, ..., l_n\}$ y una nueva instancia x, la tarea es encontrar los k ejemplos en T más cercanos a x.

La distancia entre dos puntos t_i y x está dada por:

$$D(t_i, x) = \sqrt{\sum_{j=1}^{m} (t_{i,j} - x_j)^2}$$

Donde m es el número de características o dimensiones y $t_{i,j}$ y x_j son las coordenadas de t_i y x en la j-ésima dimensión, respectivamente. El conjunto de los k vecinos más cercanos es determinado por las k menores distancias $D(t_i, x)$, y la etiqueta de x se decide por votación mayoritaria entre las etiquetas de estos k vecinos.

2.0.3. Otros modelos:

Bosque aleatorio, Árbol de decisión, Perceptrón multicapa.

Dentro de la gran cantidad de modelos predictivos que se han utilizado para analizar bases de datos, como la del Titanic, se encuentran algunos que usan redes neuronales. Estos modelos, reconocidos por su capacidad para capturar relaciones complejas y no lineales, han sido explorados en diversas investigaciones [2][21][10]. Sin embargo, es esencial señalar que algunos resultados derivados de estos modelos han mostrado signos de posible sobreentrenamiento, lo cual puede limitar su generalización en datos no vistos. Esta observación no pretende desacreditar, sino subrayar la importancia de la validación y evaluación rigurosa del modelo. Frente a este panorama, hemos decidido explorar el algoritmo k-NN (k-Nearest Neighbors) en nuestro estudio, el cual abordará una exploración siguiendo otro tipo de metodología, con eso compararemos los resultados con los obtenidos en otras investigaciones en donde se aborda una rigurosa aplicación de diferentes modelos y como la aplicación de diferentes metodologías de preprocesamiento afectan a los resultados de los entrenamientos [10].

2.0.4. Within-Cluster Sum of Squares (WCSS)

En el algoritmo K-NN, el parámetro k determina el número de vecinos más cercanos a considerar al clasificar una nueva instancia. El método del codo se utiliza para encontrar un valor óptimo de k, minimizando el error de clasificación [1].

Para aplicar el método del codo en K-NN, se traza el error de clasificación (o cualquier otra métrica de error deseada) en función de diferentes valores dé k. A medida que k aumenta, el error tiende a disminuir, pero llega un punto en el que el beneficio marginal de aumentar k es mínimo. Este punto se reconoce como el codo y generalmente se considera un buen candidato para el valor óptimo dé k.

Definición Matemática:

$$E(k) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i(k))$$

Donde:

- n es el número total de instancias en el conjunto de datos.
- y_i es la etiqueta real de la i-ésima instancia.
- $\hat{y}_i(k)$ es la etiqueta predicha de la i-ésima instancia cuando se utiliza k vecinos.
- I es la función indicadora, que toma el valor de 1 si y_i no es igual a $\hat{y}_i(k)$ y 0 en caso contrario.

Las ventajas de usar el método del codo en K-NN incluyen:

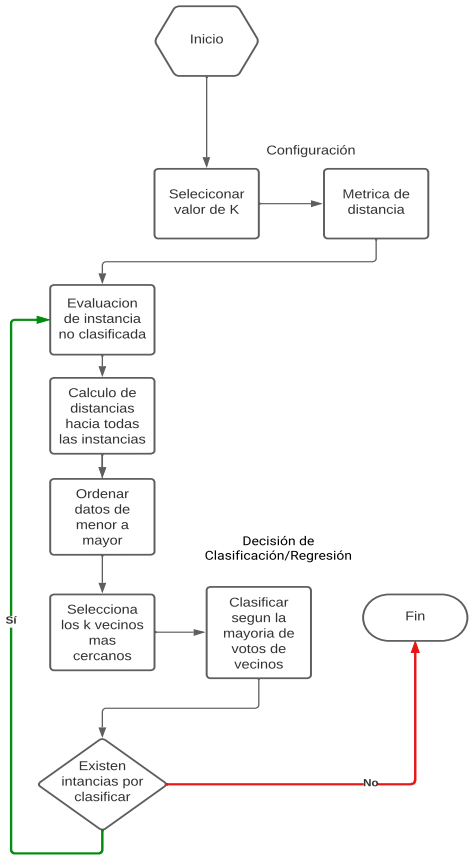
- Proporciona una técnica sistemática para seleccionar k en lugar de depender de elecciones arbitrarias.
- Ayuda a equilibrar entre sensibilidad al ruido y suavizado excesivo.
- Es fácil de visualizar e interpretar.

3. Materiales y métodos

1. Base de Datos: En esta práctica, se utilizó una base de datos proporcionada por el docente, la cual contiene información de, 1309 pasajeros con diferentes características, como su identificador, si sobrevivieron o no, clase del boleto, Nombre del pasajero, Género del pasajero, Edad del pasajero, Número de hermanos/cónyuges a bordo, Número de padres/hijos a bordo, Tarifa pagada por el pasajero, Número de cabinas, Puerto de embarque y Número del boleto. No obstante, en internet se pueden encontrar otras bases de datos públicas disponibles, como Kaggle, pero disponen de menor cantidad de instancias.
   - **Enlace**: https://www.kaggle.com/code/artyomkolas/titanic-competition/notebook?scriptVersionId=48346776.

2. Entorno de Ejecución: Con la ayuda de herramientas de desarrollo como Python, utilizamos una de sus muchas librerías, Jupyter Notebooks, el cual nos ayudó en el desarrollar, código interactivo, el cual es ejecutado en aplicaciones web y nos proporciona un kit de herramientas para una visualización intuitiva a los datos descriptivos de nuestra base de datos. Para la lectura de nuestra base de datos (.csv) utilizamos Pandas, la cual es una librería de programación en Python que proporciona herramientas para el análisis y la manipulación de datos, para el manejo de arreglos utilizamos NumPy, la cual nos proporciona una manera optimizada del uso de matrices de nxn. Y por último se utilizaron 2 librerías como medio para representar los datos de manera gráfica, seaborn y matplotlib, las cuales están enfocadas al 100 % en la visualización de datos estadísticos y nos han ayudado en crear gráficos atractivos.
Componentes de Entorno Físico:
   - Procesador: 11th Gen Intel(R) Core(TM) i5-11600K @ 3.90GHz, 3912 Mhz, 6 procesadores principales, 12 procesadores lógicos.
   - Tipo de sistema: x64-based PC.
   - Sistema Operativo: Microsoft Windows 11 Home Single Language.
   - Memoria Física (RAM): 48.0 GB.
   - GPU: NVIDIA GeForce RTX 3060.
   - Memoria GPU Dedicada: 12 GB.
3. Métodos Empleados:
   a) Preprocesamiento:
      - Normalizar Datos:
         - Sex: Cambiar a 0 y 1 (bool).
         - Age: Tiene outlayers, pocos con 80 años y muchos con 0 Años. Pasamos los datos a valores 0,1 con min-max [9].
         - Parch: Tiene datos mayores a la cantidad máxima de personas a bordo [16] y no todos son flotantes, no hay relación con alguna otra columna por lo que se descarta el intercambio de columnas.
         - Ticket: Algunos datos son flotantes y otros string, se pasarán las letras a una codificación numérica, posteriormente se juntará con la otra parte de números y se codifica a valores flotantes.
         - Fare: Contiene datos que no son flotantes y datos intercambiados con columna Cabin.
         - Cabin: Muy pocos datos de Queenstown (Puede que desde que se abordó así fuera o por errores al crear la base de datos y no se tomaron en cuenta a esas personas) y datos intercambiados con embarked.
         - Embarked: Datos encontrados en la columna Cabin.
      - Imputación Datos:
         - Embarked: Se imputan los datos nulos con la moda.
         - Fare: Se imputan los datos nulos con la media.
         - Cabin: Se imputan los datos no numéricos con vecino cercano.
         - Survived: Desbalance de clases, se tienen extra, 300 datos de personas que no sobrevivieron.

   b) Algoritmo de KNN: En la siguiente imagen se muestra el diagrama de flujo describiendo los pasos a seguir para crear nuestra propia librería aplicando el algoritmo de k-Nearest Neighbors.



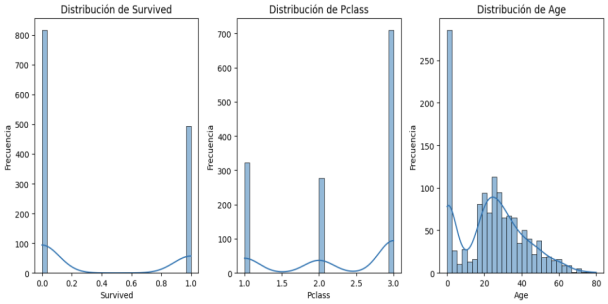
   c) Min-Max: Este tipo de normalización es una técnica utilizada para reescalar las características de un conjunto de datos al rango [0,1]. Esto se hace para asegurar que una característica no influya de manera desproporcionada en el rendimiento de un algoritmo de aprendizaje automático debido a su escala [9]. Definición Matemática:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Una vez expuesta la metodología a seguir para la creación de nuestra librería que aplicaremos a la base de datos del Titanic, pasaremos a mostrar los resultados obtenidos.

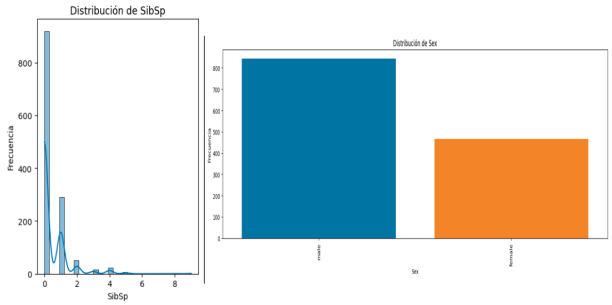
4. Resultados

Con el uso de las herramientas de software y metodologías para el preprocesamiento, se obtuvieron los siguientes resultados después de realizar una estadística descriptiva a la base de datos propuesta, a la cual, posteriormente, fue aplicada una estrategia que solucionó las anomalías de normalización y estandarización antes mencionadas. Consistió, para cada una de ellas, planificar alguna estrategia que nos ayudará a disminuir el impacto que pudieran llegar a ocasionar a nuestro entrenamiento.

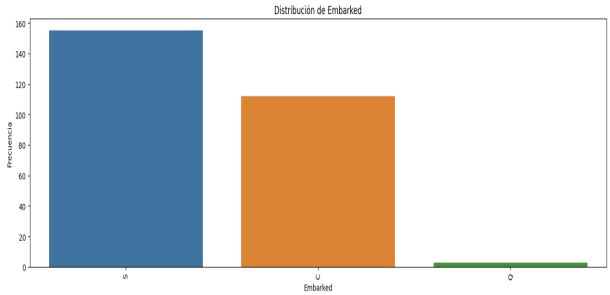


   - Gráfico donde se muestran las distribuciones de

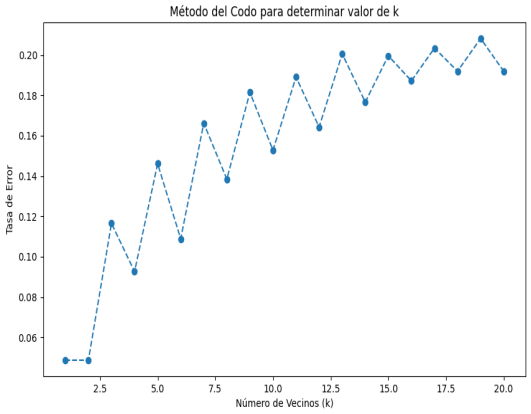
las personas que sobrevivieron, la clase del boleto, y la edad.



- Distribución de hermanos/esposas a bordo y gráfico de barras del sexo de las personas.

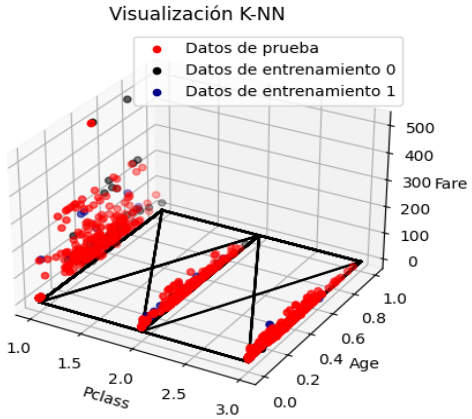


- Distribución del lugar donde abordaron.



- Método del codo.

Después de aplicar el método del codo se realizó el entrenamiento con el número k vecinos óptimo. Se obtuvieron los siguientes resultados: k= 3 o 4 como óptimo. Se comparará en un gráfico 3D, las diferentes agrupaciones con nuestros datos de entrenamiento y prueba del dataset y sus correspondientes predicciones.



- Aplicando Algoritmo de Knn con 4 vecinos óptimos, esto se obtuvo a partir del método del codo.

Después de comprobar que el algoritmo funcionaba se buscó realizar métricas para comparar el rendimiento del mismo, por lo que se compararon los resultados de nuestros datos de prueba con los reales, junto con k vecinos = 3.

a) k=3
- Accuracy. 0.883476
- Precision. 0.8745
- Recall. 0.8775
- F1. 0.8760

b) k=4
- Accuracy. 0.907354
- Precision. 0.904617
- Recall. 0.8961
- F1. 0.8999

5. Discusión

Con esta base de datos, fue primordial el seguir una rigurosa metodología de preprocesamiento, debido a las anomalías que esta presentaba, algunas de ellas fueron: Cambiar los datos de edad a 0 y 1 para una codificación a valores numéricos, se detectó que en la columna de edad existían outlayers que pudieran afectar, debido a que se tienen muchos valores cercanos a los 0 años y muy pocos valores en la edad de 80, estos valores nos dificultarían el cálculo en nuestro algoritmo al usar valores con punto decimal y enteros por lo que se optó por utilizar un escalamiento a valores cercanos al 0 y 1, a la columna SibSp no se le realiza nada, aunque se notó que tiene un desbalance, pero en este caso, si llegáramos a modificar los datos estaríamos alterando su distribución y esto afectaría en los resultados de nuestro entrenamiento, en cuanto a la columna Parch, tiene datos mayores a la cantidad máxima de personas a bordo que fueron 2200 según NatGeo [16], no todos son flotantes, por lo que se llenaron con la mediana y también cabe destacar que no era posible tener mayor 9 hijos y existía un promedio de 4 hijos por familia según datos estadísticos del censo de 1911 de United kindom National Archives, National Archives of Ireland, por lo que también se sustituyeron con la mediana; la columna ticket presenta datos en los que el ticket es un flotante, por lo que se tuvieron que eliminar y llenar con alguno que no existiera con base en los otros, se encontró que existe una relación entre los datos string encontrados en la columna de Fare y Cabin, por lo que parece ser que se cometió algún error al llenar los datos y fueron invertidas las columnas para algunos datos, así mismo para Fare, se encontró que tiene datos no numéricos por lo que no corresponde a un precio de algún boleto del Titanic o a alguna cabina, al intercambiar las columnas se tuvo que añadir datos a partir de la mediana. Se encontró que existieron datos nulos en las instancias de las columnas Embarked y que fueron intercambiadas sus columnas con Cabin. Al final se pudo observar que existió un desbalance de personas sobreviviente, no obstante no se buscó mejorar la clase desequilibrada. En cuanto a los resultados obtenidos por el método del codo, se utilizaron dos k, los cuales, a nuestra consideración, eras los que tenían menor error, a partir de esto sé probó el algoritmo con 3 y 4, y obteniendo mejores resultados con los 4 vecinos óptimos, por lo que nos quedamos con ese k para representar en las gráficas.

6. Conclusiones

El análisis de la base de datos del Titanic ha proporcionado una visión valiosa sobre los eventos y las circunstancias de quienes viajaban en el icónico barco. La aplicación de técnicas de minería de datos, específicamente el algoritmo k-NN, ha demostrado ser efectiva

para identificar y predecir la probabilidad de supervivencia de los pasajeros basada en ciertas características clave.

La importancia del Preprocesamiento, la calidad y coherencia de los datos son fundamentales para cualquier proceso de minería de datos. Las discrepancias encontradas en la base de datos subrayan la necesidad de un preprocesamiento cuidadoso. Las anomalías detectadas y las decisiones tomadas, como la codificación y escalamiento de valores, el tratamiento de datos faltantes y la corrección de columnas intercambiadas, son esenciales para garantizar la precisión de los modelos predictivos.

Se cree que pueda mejorar los resultados a partir de otro tipo de técnicas, buscaremos abordarlas más adelante.

Referencias

[1] AHUJA, Rishabh; SOLANKI, Arun; NAYYAR, Anand. Movie recommender system using k-means clustering and k-nearest neighbor. En 2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence). IEEE, 2019. p. 263-268.

[2] BARHOOM, Alaa M., et al. Predicting Titanic Survivors using Artificial Neural Network. 2019.
[3] CHAPELLE, Olivier; SCHOLKOPF, Bernhard; ZIEN, Alexander. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Transactions on Neural Networks, 2009, vol. 20, no 3, p. 542-542.
[4] COX, Earl. Fuzzy modeling and genetic algorithms for data mining and exploration. Elsevier, 2005.
[5] DUDA, Richard O., et al. Pattern classification and scene analysis. New York: Wiley, 1973.
[6] EDWARDS, Anthony WF; CAVALLI-SFORZA, Luigi Luca. A method for cluster analysis. Biometrics, 1965, p. 362-375.
[7] FURMAN, Jeffrey L.; PORTER, Michael E.; STERN, Scott. The determinants of national innovative capacity. Research policy, 2002, vol. 31, no 6, p. 899-933.
[8] GUO, Gongde, et al. KNN model-based approach in classification. En On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. Springer Berlin Heidelberg, 2003. p. 986-996.
[9] HAN, Jiawei; PEI, Jian; TONG, Hanghang. Data mining: concepts and techniques. Morgan kaufmann, 2022.

[10] HUR, Tai-Sung; BANG, Suyoung. A Comparative Analysis of the Pre-Processing in the Kaggle Titanic Competition. Journal of The Korea Society of Computer and Information, 2023, vol. 28, no 3, p. 17-24.
[11] HASTIE, Trevor, et al. Overview of supervised learning. The elements of statistical learning: Data mining, inference, and prediction, 2009, p. 9-41.
[12] JANG, Jyh-Shing Roger; SUN, Chuen-Tsai; MIZUTANI, Eiji. Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [Book Review]. IEEE Transactions on automatic control, 1997, vol. 42, no 10, p. 1482-1484.
[13] JOLLIFFE, Ian T. Principal component analysis for special types of data. Springer New York, 2002.
[14] MACQUEEN, James, et al. Some methods for classification and analysis of multivariate observations. En Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967. p. 281-297.
[15] MITCHELL, Tom M. Machine learning. 1997.
[16] NATGEO Vidas Truncadas Titanic https://historia.nationalgeographic.com.es/a/vidas-truncadas-titanic$_1$1387
[17]$PEÑA, Daniel, etal. Análisisdedatosmultivariantes. Cambridge : McGraw - HillEspaña, 2013.$
[18]$PETERSON, LeifE.K - nearestneighbor. Scholarpedia, 2009, vol,4, no2, p,1883.$
[19]$SHALEV-SHWARTZ, Shai; BEN-DAVID, Shai. Understandingmachinelearning : Fromtheorytoalgorithms. CambridgeUniversityPre$
[20]$SCHUBERT, Erich, etal. DBSCANrevisited, revisited : whyandhowyoushould(still)useDBSCAN. ACMTransactionsonDatabaseSystems$
21.
[21]$SHETTY, Jyothi, etal. PredictingtheSurvivalRateofTitanicDisasterUsingMachineLearningApproaches. En20184thInternationalConfere$
5.

[22] Titanic competition, https://www.kaggle.com/code/artyomkolas/titanic-competition/notebook?scriptVersionId=48346776
[23] Titanic$_1$,$https://www.kaggle.com/code/akshayr009/titanic-1/notebook?scriptVersionId=82013284$

[24] WARD JR, Joe H. Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 1963, vol. 58, no 301, p. 236-244.
[25] ZADEH, Lotfi Asker; KLIR, George J.; YUAN, Bo. Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers. World scientific, 1996.
[26] ZHU, Xiaojin; GOLDBERG, Andrew B. Introduction to semi-supervised learning. Springer Nature, 2022.