Universidad Autónoma de Querétaro

Facultad de Ingeniería

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

6 de octubre de 2023

Autor: *Ing. Carlos Manuel Sánchez Martínez*

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

Practica 3: Clustering con K-Means

Resumen

En esta práctica, abordaremos el tema de clusterización, el cual, fue aplicado a una base de datos pública que contiene información salarial de desarrolladores de software, la cual nos brinda una visión de los sueldos en distintos roles laborales, y cómo estos sueldos se vinculan con el índice de popularidad de las empresas. Se busca agrupar los distintos atributos propuestos y realizar un análisis exploratorio para reconocer distintos patrones en los datos. Este documento no solo busca ofrecer una visión analítica y objetiva de las empresas de software a través de la minería de datos y el algoritmo k-Means, sino también aspira a ser una herramienta útil y confiable para todos los interesados en invertir, colaborar o simplemente comprender el complejo y fascinante mundo de las empresas de tecnología de software.

Palabras clave: K-Means, Minería de Datos, Lógica Difusa, Clustering, Exploración de Datos.

1. Introducción

El constante crecimiento y evolución en el área de las TIC's ha generado un mercado competitivo de empresas que buscan destacarse por su calidad y excelencia en servicios de software [18]. En esta práctica se explora una metodología para evaluar y clasificar empresas de software, mediante la aplicación de técnicas de minería de datos, con la que se busca encontrar patrones, tendencias o relaciones ocultas que puedan ser útiles para la toma de decisiones y la generación de conocimientos.

El algoritmo k-Means es una herramienta de clasificación no supervisada, ampliamente reconocida por su eficacia en la agrupación de datos a través de características seleccionadas [5]. En nuestro proyecto, los datos de las empresas se han segmentado conforme a algunos atributos clave que influyen en la percepción y desempeño en el mercado. Estos atributos, seleccionados por su relevancia y su impacto sobre el desempeño de las empresas, incluyen, la ubicación geográfica (país), la calificación o rating recibido, los puestos de trabajo ofertados y los rangos salariales asociados, etc. Cada uno de estos factores proporciona una perspectiva valiosa sobre la posición y la estrategia de las empresas en el mercado, influyendo significativamente en su valoración y percepción por parte de las personas interesadas en el área.

2. Marco teórico

La ciencia, la tecnología y la innovación son principales factores del desarrollo económico sustentable [5]. 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 [12]. 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 [11]. 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 [17]. 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 [13].
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 [17]. Tenemos diferentes ventajas al aplicar este tipo de entrenamiento, ejemplo: Clustering: Agrupa datos similares basados en características [6]. 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[9].
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 [1]. 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[17].

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 [5]. Mediante algoritmos como K-Means[10], DBSCAN [14], y jerárquicos [15], 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 [6].

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 [3].

   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-means

El algoritmo K-Means es una técnica popular de clustering utilizada para segmentar un conjunto de datos flotantes o enteros en K grupos o clusters, donde K es predefinido por el usuario. Se seleccionan K puntos aleatorios como centroides iniciales. Luego, cada punto del conjunto de datos se asigna al centroide más cercano, formando K clúster. Los centroides se recalculan como el promedio de los puntos asignados a cada cluster, y el proceso se repite hasta que los centroides no cambien significativamente o se alcance un número máximo de iteraciones [5][10]. Definición Matemática: Para un conjunto de datos $X = \{x_1, x_2, ..., x_n\}$ y centroides $C = \{c_1, c_2, ..., c_k\}$, se define como:

$$J(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - c_i||^2$$

Donde $||x_j - c_i||$ es la distancia euclidiana entre un punto x_j y un centroide c_i, y C_i es el i-ésimo cluster.

Otros modelos: Kmeans++,Kmeans esferico,Kmenas basado en kernel,Kmeans difuso.

2.0.3. Clustering Difuso

1. Fuzzy cluster means (FCM): Es un método de agrupamiento que permite que un punto de datos pertenezca a múltiples clusters con diferentes grados de pertenencia. Esto nos proporciona una solución más flexible y generalizada para problemas de clasificación donde los limites entre los clusters no son claros. A diferencia de K-Means, donde cada punto pertenece a un solo clúster, FCM asigna a cada punto un grado de pertenencia a cada cluster, con valores entre 0 y 1 [16]. Definición: Fuzzy C-Means (FCM) Clustering minimiza la siguiente función objetivo:

$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{c}(u_{ij})^m ||x_i - c_j||^2$$

   Donde:

   - n es el número total de datos.
   - c es el número de clusters.
   - u_{ij} es el grado de pertenencia del i-ésimo dato al j-ésimo clúster.
   - x_i es el i-ésimo dato.
   - c_j es el centroide del j-ésimo clúster.
   - m es un exponente (mayor que 1) que determina el grado de "difusión" de los clusters.

2. Fuzzy cluster substractive: Es una técnica de agrupación basada en un enfoque de densidad que sirve como método rápido para estimar el número y los centros de clusters en un conjunto de datos. Este método es particularmente útil para identificar agrupaciones en un espacio de características multidimensional [8]. La densidad en una ubicación particular es proporcional a la suma de las distancias a todos los otros puntos en el conjunto de datos. Se selecciona como centro del clúster el punto con la densidad más alta, luego se reduce la densidad en la región cercana a este punto y se repite el proceso para encontrar más centros [2][8]. Subtractive Clustering calcula la densidad de puntos $D(x)$ en la ubicación x mediante la fórmula:

$$D(x) = \sum_{i=1}^{n} e^{-\frac{||x - x_i||^2}{r_a^2}}$$

   Donde:

   - n es el número total de puntos de datos.
   - x_i representa cada punto de datos.
   - r_a es un radio de influencia definido por el usuario que determina la escala de distancia sobre la que un punto puede influir en la densidad de otros puntos.

2.0.4. Within-Cluster Sum of Squares (WCSS)

El WCSS sirve como criterio de optimización para seleccionar el número de clusters. Un valor bajo de WCSS indica que los puntos dentro de los clusters están cercanos entre sí, lo cual es un indicativo de buena segmentación. Sin embargo, es esencial considerar que minimizar excesivamente el WCSS puede llevar a un sobreajuste, resultando en un número excesivo de clusters [4]. El WCSS se calcula sumando las distancias cuadradas entre cada punto y el centroide del cluster al cual pertenece, para todos los puntos y todos los clusters. Definición:

$$WCSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} ||x_{ij} - c_i||^2$$

Donde:

- k es el número de clusters.
- n_i es el número de puntos en el i-ésimo cluster.
- x_{ij} es él j-ésimo punto en el i-ésimo cluster.
- c_i es el centroide del i-ésimo cluster.
- $||x_{ij} - c_i||^2$ es la distancia cuadrada entre el j-ésimo punto y el centroide del i-ésimo cluster.

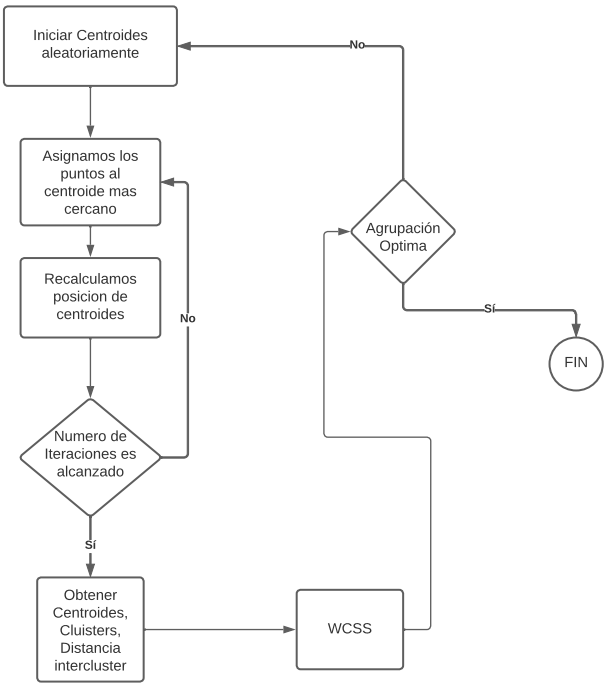
3. Materiales y métodos

1. Base de Datos: En esta práctica, se utilizó una base de datos pública, la cual contiene información de más de, 22700 profesionales de software con diferentes características, como sus salarios (rupias), puestos de trabajo, nombre de la empresa, calificación de la empresa, número de veces que se informan los salarios y ubicación de la empresa.
   - **Enlace**: https://www.kaggle.com/datasets/whenamancodes/software-professional-salary-dataset.
   - **Preprocesamiento**: Anteriormente, fue modificada la base de datos en la práctica 2 (Imputación de datos), con la cual fue necesario codificar los valores categóricos debido a las condiciones de uso previamente mencionadas del algoritmo de k-means, así mismo fue necesario el balanceo entre clases y fue utilizado el algoritmo SMOTE.

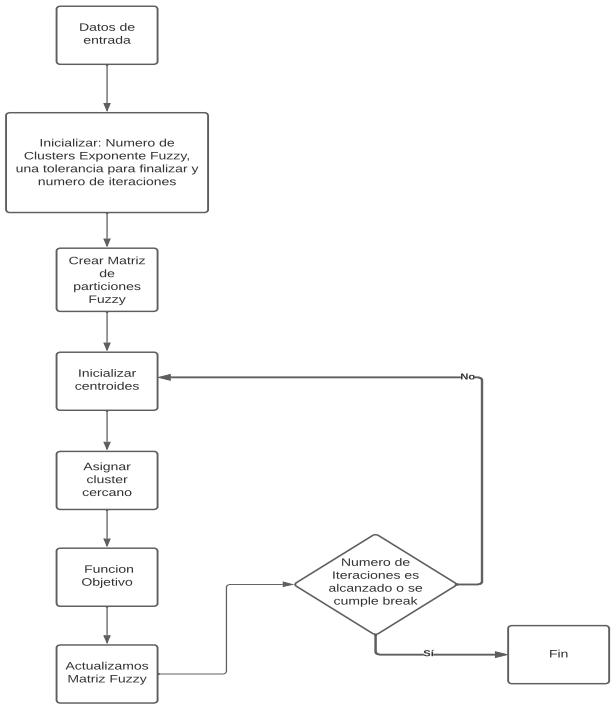
2. Entorno de Ejecución: Con la ayuda de herramientas de desarrollo como Python, utilizamos una de sus muchas librerías, Jupyter Notebooks, el cual nos ayudó en el desarrollar, código interactivo, el cual es ejecutado en aplicaciones web y nos proporciona un kit de herramientas para una visualización 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: En la siguiente imagen se muestra el diagrama de flujo describiendo los pasos a seguir para crear nuestra propia librería aplicando el algoritmo de kmeans.



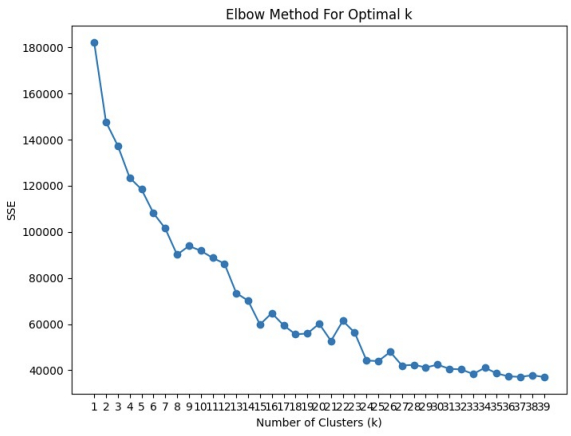
Ahora modificando un poco, pero ahora con lógica difusa, se vería algo así.



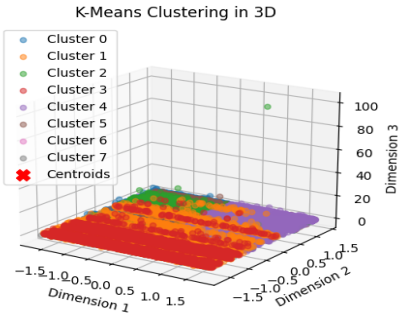
Una vez expuesta la metodología a seguir para la creación de nuestra librería que aplicamos a la base de datos de rating, pasaremos a mostrar los resultados obtenidos.

4. Resultados

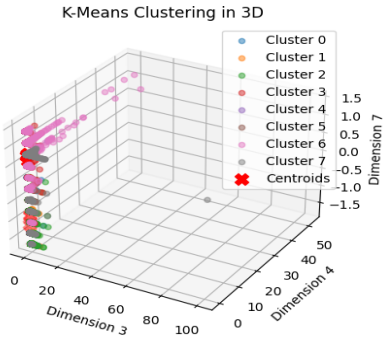
Con el uso de las herramientas de software y metodologías para la agrupación, se obtuvieron los siguientes resultados después de realizar una estadística descriptiva a la base de datos propuesta y posteriormente fue aplicado el preprocesamiento, que consistió en aplicar diferentes técnicas de ciencia de datos como lo fueron: imputación de datos, eliminamos la columna de rating, debido a que queríamos ver si podíamos obtener un resultado en algún atributo ya conocido, normalizar nuestra base de datos, tuvimos que pasar todos los datos a tipo flotante, realizamos un escalado de los datos para disminuir el costo computacional (float¡int), entre otras. Sin conocer un número óptimo de grupos se utilizó wcss (método del codo), para conocer k.



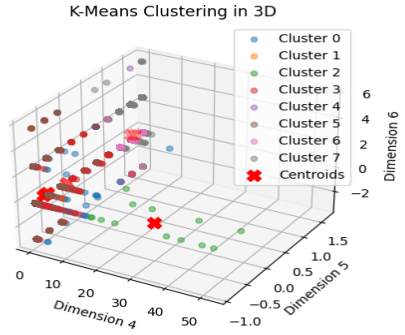
Después de aplicar el método del codo se realizó el entrenamiento con el número k óptimo. Se obtuvieron los siguientes resultados comparando en un gráfico 3D las diferentes columnas del dataset, mostraremos las que seleccionamos como más importantes al momento de evaluar a las empresas de software de nuestro dataset.



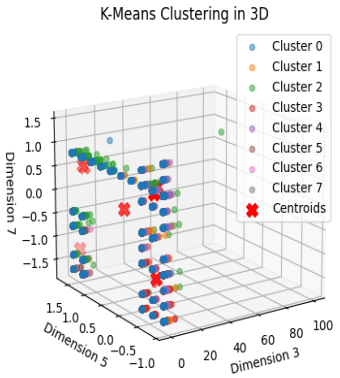
- Dimensión 1: Company Name vs. Dimensión 2: Job Title vs. Dimensión 3: Salary.



- Dimensión 3: Salary vs. Dimensión 4: Salaries Reported vs. Dimensión 7: Job Roles.



- Dimensión 4: Salaries Reported vs. Dimensión 5: Location vs. Dimensión 6: Employment Status.



- Dimensión 7: Job Roles vs. Dimensión 5: Location vs. Dimensión 3: Salary.

Esta matriz tiene dimensiones (C, N), donde C es el número de clusters y N es el número de puntos de datos en tu conjunto de datos

```
[0.02735981 0.0321476  0.03154882 ... 0.06436724 0.07159485 0.07081281]
[0.04247366 0.03629616 0.05145768 ... 0.11730155 0.09498957 0.17541651]
[0.04287654 0.04690358 0.04861366 ... 0.18000282 0.15666473 0.14987316]
...
[0.03525721 0.0510845  0.03845561 ... 0.18830132 0.22459502 0.10431754]
[0.42834112 0.10579152 0.47233988 ... 0.09743869 0.08608139 0.17885714]
[0.0600539  0.06422806 0.06606315 ... 0.0442627  0.05077685 0.05951974]]
```

- Resultados de Fuzzy cluster mean: Esta matriz tiene dimensiones (C, N), donde C es el número de clusters y N es el número de puntos de datos del conjunto de datos.

5. Discusión

En la serie de gráficos tridimensionales presentados, se observa la dinámica de interacción y relación entre diversas variables del conjunto de datos, incluyendo el nombre de la compañía, el título del trabajo, el salario, los salarios reportados, los roles de trabajo, la ubicación y el estado de empleo. Estos gráficos proporcionan una visión visual e intuitiva de cómo estas variables contribuyen a la formación de clusters distintos, en nuestro caso utilizamos el método del codo para obtener el número de clúster óptimo, como resultado se obtuvo que 8 (criterio propio) clusters era la menor cantidad de clusters con la menor distancia intercluster, sin que afecte tanto el rendimiento o se abuse de la cantidad de clusters generados.

El primer gráfico, que despliega la relación entre el nombre de la compañía, el título del trabajo y el salario, ilustra un patrón de agrupamiento significativo en el eje correspondiente al título del trabajo. La expansión sustancial de clusters en esta dimensión sugiere una asociación directa entre el título del trabajo y las demás variables, siendo un factor distintivo y agrupador esencial.

En el segundo gráfico, al analizar la relación entre el salario, los salarios reportados y los roles de trabajo, se observa una tendencia clara y directa: a medida que aumenta la jerarquía o responsabilidad del puesto (reflejada en los roles de trabajo), tanto el salario como la cantidad de salarios reportados experimentan un incremento. Este fenómeno resulta en una separación de clusters más pronunciada, visualmente evidente mediante la diversificación de colores.

En contraste, el tercer gráfico, que compara los salarios reportados, la ubicación y el estado de empleo, no muestra una clusterización tan distintiva y evidente. Esta observación puede indicar que estas variables, aunque relevantes, podrían no interactuar de manera tan significativa como para formar clusters bien definidos y separados en el espacio tridimensional representado.

El gráfico que compara roles de trabajo, ubicación y salario revela una relación y clusterización bien definidas entre estas variables. Esto sugiere que estas tres variables tienden a agruparse de manera coherente, proporcionando evidencia de una correlación significativa entre ellas y validando la eficacia del algoritmo de clustering en identificar y agrupar datos con características similares.

Por último, una visualización y comparación de los gráficos en 3D de los resultados de clusterización del algoritmo de Fuzzy-C-Means pudimos observar una similitud en los resultados de las gráficas obtenidas con los previos resultados de los algoritmos de K-means, por lo que no decidimos incluirlas en los resultados y solo optamos por incluir la matriz de pertnencia a cada clúster como evidencia del mismo.

6. Conclusiones

A lo largo de la investigación y aplicación de los algoritmos vistos se puede concluir que pueden ser de gran utilidad para la visualización y análisis de grandes cantidades de información, en gráficos que facilitan la interpretación de los patrones subyacentes. Estos patrones, a su vez, pueden ser esenciales para identificar factores críticos que influencian las variables de interés, proporcionando así insights valiosos para futuras investigaciones y decisiones basadas en datos.
Considerando los resultados similares entre ambos algoritmos, podemos concluir que ambos son algoritmos de posicionamiento que dividen un conjunto de datos en grupos o clusters, minimizando una función de costo relacionada con la distancia entre los puntos de datos y los centroides de los clusters, por lo que es obvio el resultado similar entre ellos.
Respecto a los algoritmos empleados en el proyecto podemos definir lo siguiente:

- Ventajas:
  - *a*) Baja Complejidad Computacional:Su complejidad es lineal respecto al número de datos: O(n).
  - *b*) Escalable: Uso con grandes cantidades de datos, derivado de su baja complejidad computacional, k-Means es rápido y eficiente en términos de tiempo de cálculo, especialmente con grandes conjuntos de datos.
  - *c*) Visualización Sencilla: Los clústeres generados son fáciles de interpretar y visualizar. Provee una estructura clara y sencilla que facilita el análisis posterior.
- Desventajas:
  - *a*) Necesidad de especificar el número de clústeres k previamente:
  - *b*) Sensibilidad a la inicialización y la forma de los clústeres:
  - *c*) Dificultad con clústeres de densidades variadas:
  - *d*) Previo Preprocesamiento de la información:

Referencias

[1] 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.
[2] COX, Earl. Fuzzy modeling and genetic algorithms for data mining and exploration. Elsevier, 2005.
[3] DUDA, Richard O., et al. Pattern classification and scene analysis. New York: Wiley, 1973.
[4]EDWARDS, Anthony WF; CAVALLI-SFORZA, Luigi Luca. A method for cluster analysis. Biometrics, 1965, p. 362-375.

[5] FURMAN, Jeffrey L.; PORTER, Michael E.; STERN, Scott. The determinants of national innovative capacity. Research policy, 2002, vol. 31, no 6, p. 899-933.
[6] HASTIE, Trevor, et al. Overview of supervised learning. The elements of statistical learning: Data mining, inference, and prediction, 2009, p. 9-41.
[7] JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. Data clustering: a review. ACM computing surveys (CSUR), 1999, vol. 31, no 3, p. 264-323.
[8] 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.
[9] JOLLIFFE, Ian T. Principal component analysis for special types of data. Springer New York, 2002.
[10] 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.

[11] MITCHELL, Tom M. Machine learning. 1997.
[12] PEÑA, Daniel, et al. Análisis de datos multivariantes. Cambridge: McGraw-Hill España, 2013.
[13] SHALEV-SHWARTZ, Shai; BEN-DAVID, Shai. Understanding machine learning: From theory to algorithms. Cambridge University Press, 2014.
[14] SCHUBERT, Erich, et al. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 2017, vol. 42, no 3, p. 1-21.
[15] 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.
[16] ZADEH, Lotfi Asker; KLIR, George J.; YUAN, Bo. Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers. World scientific, 1996.
[17] ZHU, Xiaojin; GOLDBERG, Andrew B. Introduction to semi-supervised learning. Springer Nature, 2022.
[18] Gartner, Inc. (2021, October 18). Gartner Identifies the Top Strategic Technology Trends for 2022. Gartner Newsroom. https://www.gartner.com/en/newsroom/press-releases/2021-10-18-gartner-identifies-the-top