Predictor de ataque cardiaco mediante aprendizaje no supervisado

Cristobal Medina-Meza[1] and Cristóbal Estrada-Salinas[2]

[1, 2] *Tecnologico de Monterrey, Escuela de Ingenieria y Ciencias*

Resumen—El aprendizaje computacional abarca enfoques analíticos para la exploración de hipótesis, impulsados por software que implementa estos métodos en diversos escenarios. Las aplicaciones de estas técnicas pueden ser diversas, dependiendo del contexto en el que se evalúe su efectividad. En este escenario particular, la investigación se centra en determinar la viabilidad de utilizar métodos de aprendizaje no supervisado en el ámbito de la medicina y las posibles implicaciones que esto podría tener para su estudio.

Palabras clave—Heart attack, Kmeans, Mean-Shift

I. Introducción a la problemática

a. Aprendizaje no supervisado en salud

El aprendizaje no supervisado es una técnica de inteligencia artificial que implica la formación de modelos sin la guía de un conjunto de datos previamente etiquetado. A diferencia del aprendizaje supervisado, donde se conocen las respuestas correctas, el aprendizaje no supervisado se basa en analizar patrones y características inherentes a los datos sin etiquetar.

En el contexto del aprendizaje no supervisado en salud, el modelo explora los datos de entrenamiento, como datos clínicos, imágenes médicas o resultados de pruebas, para descubrir patrones subyacentes o agrupamientos. En lugar de aprender a realizar predicciones específicas o clasificaciones conocidas, el modelo busca de manera autónoma estructuras significativas dentro de los datos, lo que puede tener aplicaciones en la identificación de relaciones emergentes, descubrimiento de subgrupos o detección de anomalías en entornos médicos.

b. Descripción del dataset y su preprocesamiento

La base de datos utilizada para este proyecto es un conjunto que explora algunas características de personas que son propensas o han tenido un ataque cardiaco. La base de datos fue extraída de kaggle [1] y sus columnas son: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal y target. Todos ellos son datos de tipo numérico, el df cuenta con solo un dato duplicado.
Para la división entre datos de prueba y de entrenamiento se hace uso de la función RandomOverSampler() y train_test_split(), haciendo una división del 30% y 70%. Para hacer los clústeres se utilizan las variables "age","thalach","oldpeak", se hace uso solo de 3 para poder visualizar los resultados.

c. Análisis exploratorio con gráficos

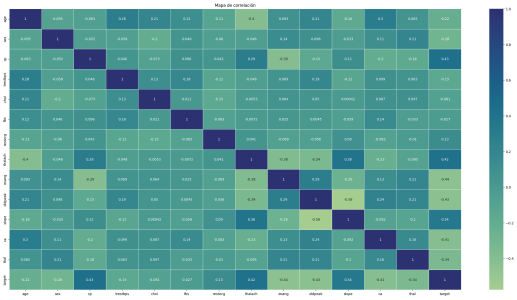

Fig. 1: Mapa de correlación

Analizando el Mapa de correlación nos damos cuenta de que es poco probable que nuestras variables sean linealmente dependientes las unas de las otras.

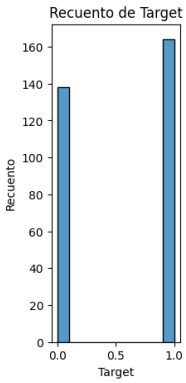

Fig. 2: Recuento de Target

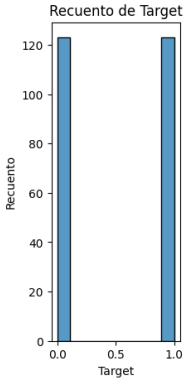

Fig. 3: Recuento de Target en el conjunto de entrenamiento

Analizando los histogramas de la distribución de target, nos damos cuenta de que es similar a una variable aleatoria uniforme, por lo cual, hacemos uso de la función RandomOverSampler() para conseguirla en el conjunto de entrenamiento.

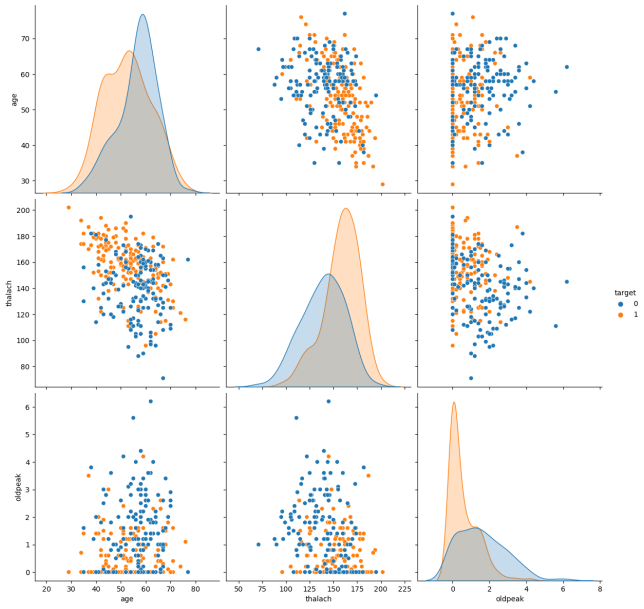

Fig. 4: Combinaciones de las variables seleccionadas

En esta visualización de las combinaciones de las variables seleccionadas, en algunas, se alcanzan a apreciar los clústeres.

II. Metodología

La prueba con la que se realizó el experimento consta de dos modelos, "Kmeans" y "Mean-Shift":

Kmeans: K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática [2].
Para el modelo se hizo uso de 2 clústeres, esta cantidad fue elegida usando el método del codo.

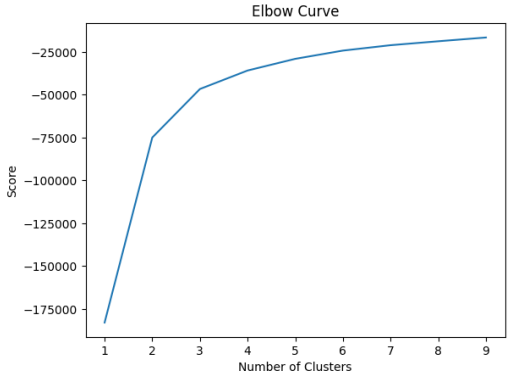

Fig. 5: Gráfica de Codo

Este método creó los siguientes clústeres:

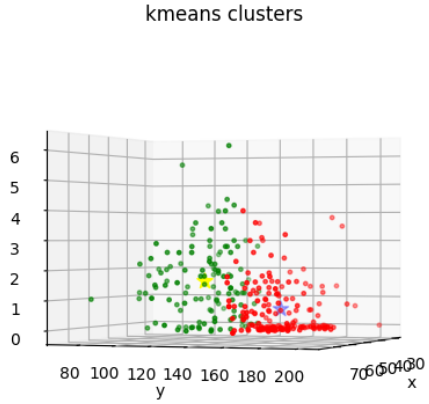

Fig. 6: Clústeres de Kmeans

Mean-Shift: Mean Shift es una técnica no paramétrica para el análisis de un conjunto de puntos d-dimensionales en un espacio de características que obtiene un máximo local de la función de densidad, usualmente se utiliza para clustering [3]. Esta técnica elige por sí misma la cantidad de clústeres y su centro, para este trabajo se hace uso del MeanShift de sklearn.cluster, el cual dio como resultado 2 clústeres.

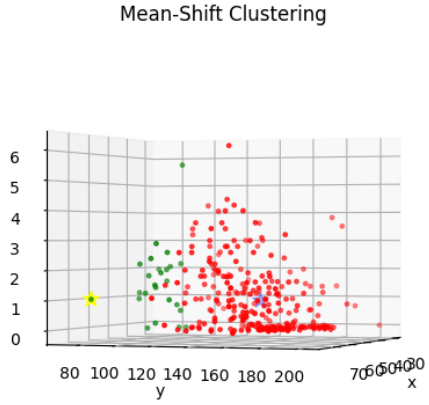

Fig. 7: Clústeres de Mean-Shift

III. Resultados

A pesar de que el dataset es difícil de separar por clústeres, ambos modelos llegan a resultados similares, ambos dividiendo el df en 2 clústeres de características similares, por lo que es probable que los resultados sean correctos a pesar de que no sea tan explícito.

References

[1] BHAT N. Health care: Heart attack possibility. https://colab.research.google.com/drive/1K7-IfpRjEQ2dcNG1AKcVeg$_F$*NUAjLYtE*?*hl* = *esscrollTo* = *QAAnLzwT*0*AXU*.

[2] de Oviedo U. kmeans. https://www.unioviedo.es/compnum/laboratorios$_p$*y*/*kmeans*/*kmeans.html* : : *text* = *K*

[3] Nakama UdBA. Tesisnakama. https://www-2.dc.uba.ar/grupinv/imagenes/archivos/TesisNakama2011: :text=Mean