

Análisis de hábitos saludables de las personas mediante Algoritmos de Aprendizaje Supervisado y No Supervisado

Leonardo Garcia Muñoz

10 de noviembre de 2025

1. Introducción

Los hábitos con los que cuentan las personas en los tiempos actuales son consideradas una de las principales razones de las enfermedades que actualmente presentan una alta incidencia, siendo estos factores mucho más relevantes en un monitoreo médico que incluso la predisposición genética de las personas. Existiendo distintas maneras de clasificar los distintos hábitos con los que cuentan las personas.

El siguiente artículo tiene como principal propósito el aplicar técnicas de **clustering** o agrupamiento para explorar distintas agrupaciones de los hábitos de las personas y las posibles relaciones con los riesgos que tienen las personas de sufrir alguna enfermedad de gravedad. Por lo cual se estarán utilizando principalmente los algoritmos **Optics** y **K-Means** para realizar las agrupaciones.

2. Descripción de los datos

El conjunto de datos se compone principalmente por variables numéricas:

- age,bmi,daily steps, sleep hours, water intake, calories consumed, resting hr, systolic bp, diastolic bp, cholesterol

Pero también presenta las siguientes variables categóricas:

- gender, smoker, alcohol, family history

Todas estas variables fueron registradas por persona, por lo cual cada registro de estas variables es una persona con distintos hábitos, condiciones de vida e historias diferentes. Pero esta información para poder ser utilizada en los algoritmos a aplicar, se estandarizaron para evitar problemas de escalabilidad entre variables y presentarlas todas en las mismas dimensiones.

3. Metodología

3.1. Algoritmo K-Means

Este algoritmo de **clustering** o agrupamiento agrupa los datos en k grupos o **clusters** basándose en la distancia al punto promedio o **centroide** que representa el centro de un grupo, y a cada dato se le asigna el **centroide** más cercano. Para después recalcular los **centroide** de manera iterativa. Buscando que los datos que componen cada grupo sean similares.

Formalmente, la distancia más cercana respecto a los centroides se define como la minimización del error cuadrático para cada grupo.

$$SS_k = \sum_{i \in K} (x_i - \hat{x}_k)^2 + (y_i - \hat{y}_k)^2.$$

Teniendo como función objetivo

$$\min \sum_{i \in K} SS_k.$$

Donde \hat{x}_k es un centroide compuesto por (\hat{x}, \hat{y})

3.2. Algoritmo Optics

El algoritmo *Ordering Points To Identify the Clustering Structure* o **OPTICS** está en la identificación de regiones con una alta densidad de datos. Cada dato o punto tiene una densidad que se calcula como el valor máximo entre la distancia al dato vecino más cercano y un parámetro de densidad mínimo. El algoritmo recorre los datos, expandiendo los clusters desde los puntos más densos y ordenándolos de manera que refleja la estructura de densidades.

Formalmente se define de la siguiente manera

$$\text{reachability_distance}(p, o) = \max(\text{core_distance}(o), \text{dist}(p, o))$$

Donde **core_distance**(o) es el parámetro de densidad mínimo.

3.3. Métrica Inercia

Esta métrica que se puede utilizar en **K-Means** mide qué tan compactos están los clusters, calculando la suma de las distancias al cuadrado de cada punto a su centroide. Para elegir el número óptimo de clusters, se prueba K-Means con varios k y se grafica la inercia contra k . Mediante el método del codo se busca el punto de quiebre, en donde la disminución de la inercia se vuelve menos pronunciada en escala de valores.

Formalmente para k clusters o grupos

$$\text{Inercia} = \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$

Donde C_i son los puntos que componen el grupo o cluster i y μ_i su centroide.

3.4. Métrica Calinski-Harabasz

El índice **Calinski-Harabasz** es una métrica para evaluar clusters o grupos que compara la separación entre grupos con que tan cerca están los puntos entre sí, dentro de cada grupo. Se usa para elegir el número óptimo de clusters k . Calculando CH para varios k , el valor más alto indica la mejor separación y cohesión.

Formalmente para k clusters o grupos.

$$CH = \frac{B_k/(k-1)}{W_k/(n-k)}$$

Donde B_k es la varianza entre clusters, W_k la varianza dentro de los clusters, n el número de puntos y k el número de clusters o grupos. Un CH alto indica grupos altos y bien separados de otros.

4. Resultados

Primero realizando la tecnica de *k-Means* y utilizando la metrica *Inercia* obtenemos los siguientes resultados:

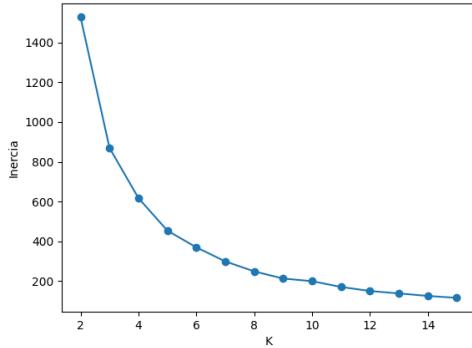


Figura 1. Gráfico generado en Python mediante la metrica *Inercia*.

Para obtener el numero de k grupos a utilizar primero debemos aplicar el Método del codo.

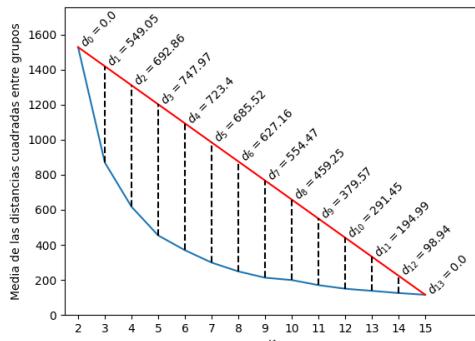


Figura 2. Gráfico generado en Python para aplicar el Método del codo.

Se aprecia como el punto de quiebre ocurre en $k = 5$ debido a que apartir de ahí, la disminucion ya no es tan pronunciada.

Por lo tanto se utiliza $k = 5$ para *K-Means*.

Ahora, utilizando la metrica *Calinski-Harabasz*

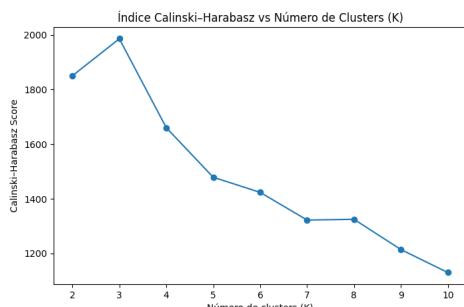


Figura 3. Gráfico generado en Python mediante la metrica *Calinski-Harabasz*.

Obtenemos que el CH mas alto pertenece a $k = 3$, por lo que bajo esta metrica tendrian que utilizarse 3 clusters.

Pero tambien aplicaremos al algoritmo *OPTICS* el cual obtiene los clusters necesarios

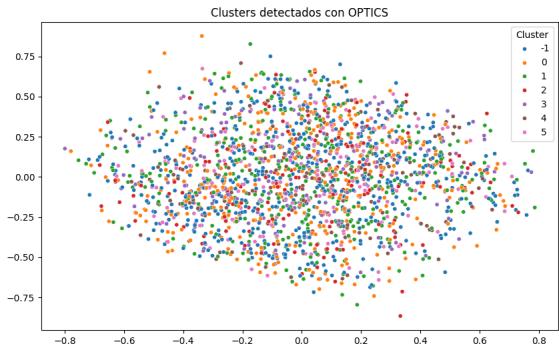


Figura 4. Gráfico generado en Python mediante *OPTICS*.

Calibrando un modelo de regresion lineal para realizar predicciones sobre la variable de *Presión Arterial diastólica* obtenemos los siguientes resultados reflejados en el grafico.

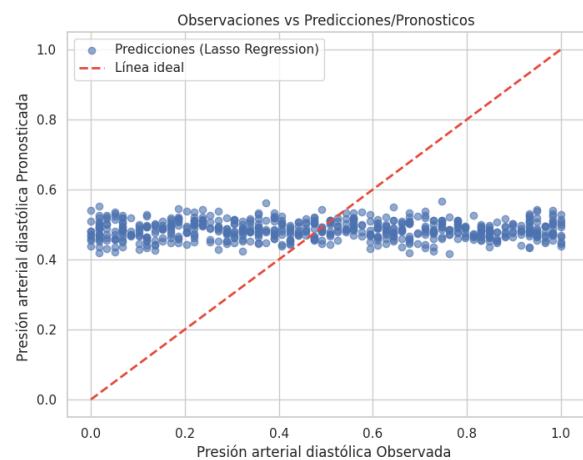


Figura 5. Gráfico generado en Python con base a una Regresion Lineal.

Se observa como los valores pronosticados oscilan entre el 0.4 y el 0.6 mientras que los valores observados se encuentran en un intervalo mucho mas amplio de valores, lo cual nos podria hablar de un posible sobreajuste en el modelo. Ademas de que muy pocos valores realmente se ajustan a la *Línea ideal*

En la siguiente figura se estan comparando distintos algoritmos supervisados para evaluar su desempeño y elegir el mejor modelo a utilizar, adicionalmente del modelo de regresion lineal previamente calibrado.

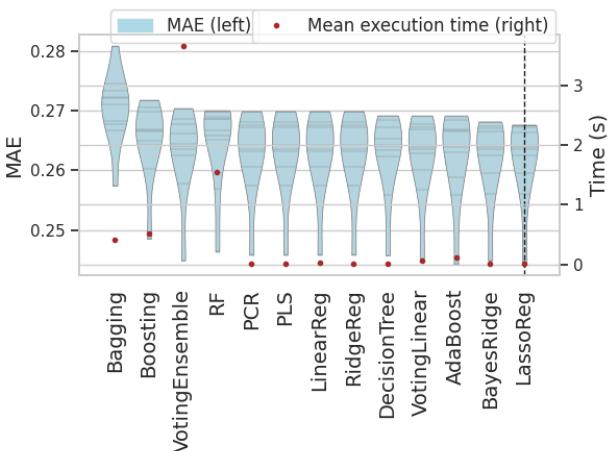


Figura 6. Gráfico de violín generado en Python

El modelo con menor error (MAE) y el tiempo de ejecución más óptimo es el modelo Lasso, por lo que se calibra un modelo Lasso con el cual se pueda predecir la *Presión Arterial diastólica*.

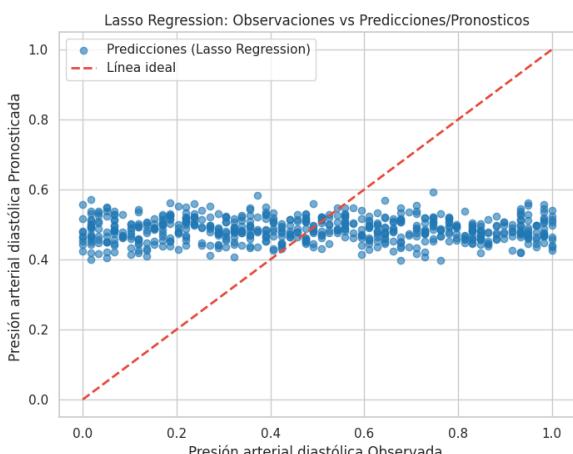


Figura 7. Gráfico generado en Python en base a un modelo Lasso.

En este modelo obtenemos una mayor variación en las predicciones que realiza el modelo, pero sigue existiendo una concentración de los valores pronosticados justo en el rango de 0.38 a 0.6. Además de contar con una menor cantidad de valores que sigan la *Línea ideal*

Con ambos modelos calibrados y utilizados para pronosticar, podemos realizar la comparación de la métrica para medir los errores de los datos.

Modelo	MAE
Regresion Lineal	0.2646
Regresion Lasso	0.2605

Cuadro1. Comparación de los modelos según la métricas de error.

Para ambos modelos se dividió el conjunto de datos de entrenamiento y prueba en una proporción del 70 % y del 30 % respectivamente.

5. Conclusiones y discusiones

5.1. Algoritmos No Supervisados para clustering

Principalmente podemos concluir que el algoritmo *OPTICS* genera 6 clusters distintos, mientras que *Calinski-Harabasz* nos dio 3 y en *K-Means* utilizamos 5. Pero sin importar el algoritmo vemos que los clusters presentan valores muy parecidos entre ellos, aunque se pueden apreciar comportamientos notorios.

El cluster 0 agrupa personas que consumen alcohol con frecuencia, duermen menos y presentan niveles de presión y colesterol moderados, pero no bajos. Mientras que en el cluster 4 se agrupan a personas que no consumen alcohol, con una higiene del sueño similar, pero con una presión sistólica mayor, aunque un IMC menor. Lo cual podría hacer sentido debido a que la presión sistólica es la principal en verse afectada al tener una mala noche de sueño.

Mientras que en el cluster 3 se encuentran las personas con hábitos más saludables, debido a que presentan valores altos en pasos diarios, horas de sueño y agua consumida, teniendo así valores de colesterol bajos.

Concluyendo con que K-Means es un algoritmo muy eficiente bajo la métrica de Inercia para generar grupos en la información, mientras que OPTICS va un paso adelante pues generó un número más alto de grupos, pero no todo lo que brilla es oro. Pues el hecho de tener más agrupaciones no siempre es lo más eficiente y K-Means lo demuestra, con sus agrupaciones da unas buenas descripciones del comportamiento de la información.

5.2. Algoritmos Supervisados para Pronósticos

Concluimos con los resultados obtenidos que en el modelo Lasso las predicciones están más dispersas, tienen un rango un poco más amplio que el que presenta el modelo de Regresión lineal, pero en ambos muy pocos puntos siguen la diagonal o Línea ideal aunque el modelo de Regresión al presentar una menor variabilidad, se podría inferir que más puntos de los pronosticados se encuentran siguiendo esta diagonal, en proporciones generales ambos modelos tienen un número muy reducido de puntos que la siguen.

Mediante la métrica de MAE obtenemos que el modelo Lasso está ligeramente más cerca de los valores reales que el modelo de regresión lineal, aunque es una mejora en el rendimiento muy pequeña, sigue siendo mejor, adicionalmente de que se presenta una mayor variación, lo cual imita mejor el comportamiento real de la presión diastólica.

Por lo que se considera que el modelo Lasso presenta un desempeño mejor que el modelo de Regresión lineal.

Referencias

Fernández, D., Pezzi, J. P., & Caruso, G. (s. f.). *Apnea del sueño e hipertensión arterial. Diagnóstico y terapéutica.* <http://www.saha.org.ar/uploads/pdf/Cap.%2E096.pdf>

Scikit-learn. (s. f.). *Lasso — scikit-learn 1.5.2 documentation.* https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

provocar presión arterial alta? <https://www.mayoclinic.org/es/diseases-conditions/high-blood-pressure/expert-answers/sleep-deprivation/faq-20057959>

Filippi, A., ... (2024). *Smoking cessation decreases arterial blood pressure in hypertensive smokers: A subgroup analysis of the randomized controlled trial GENTSMOKING.* <https://pubmed.ncbi.nlm.nih.gov/38756738/>

Cappuccio, F. P., ... (2023). *Examining Daily Associations Among Sleep, Stress, and Blood Pressure Across Adulthood.* <https://pubmed.ncbi.nlm.nih.gov/36680526/>