Artículo sobre tecnicas de Agrupamiento en habitos de las personas

Leonardo Garcia Muñoz

9 de noviembre de 2025

1. Introducción

Los habitos 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 mas relevantes en un monitoreo medico que incluso la predisposicion genetica de las personas. Existiendo distintas maneras de clasificar los distintos habitos con los que cuentan las personas.

El siguiente artículo tiene como principal proposito el aplicar tecnicas de **clustering** o agrupamiento para explorar distintas agrupaciones de los habitos de las personas y las posibles relaciones con los riesgos que tienen las personas de sufrir alguna enfermedad de gravedad. Por lo cual se estaran 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 númericas:

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

Pero tambien presenta las siguientes variables categoricas:

- 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 habitos, condiciones de vida e historias diferentes. Pero esta informacion 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** basandose en la distancia al punto promedio o **centroide** que representa el centro de un grupo, y a cada dato se le asigna el **centroide** mas cercano. Para despues 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 minimizacion del error cuadrado 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

$$\text{mín} \sum_{i \in K} SS_k.$$

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

3.2. Algoritmo *Optics*

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

Formalmente se define de la siguiente manera

$$reachability\_distance(p, o) = max(core\_distance(o), dist(p, o))$$

Donde $core\_distance(o)$ es el parametro de densidad minimo.

3.3. Metrica *Inercia*

Esta métrica que se puede utilizar en *K-Means* mide que 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

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

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

3.4. Metrica *Calinski-Harabasz*

El índice *Calinski-Harabasz* es una métrica para evaluar clusters o grupos que compara la separación entre grupos con que tan cerca estan los puntos entre si, 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 numero 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:

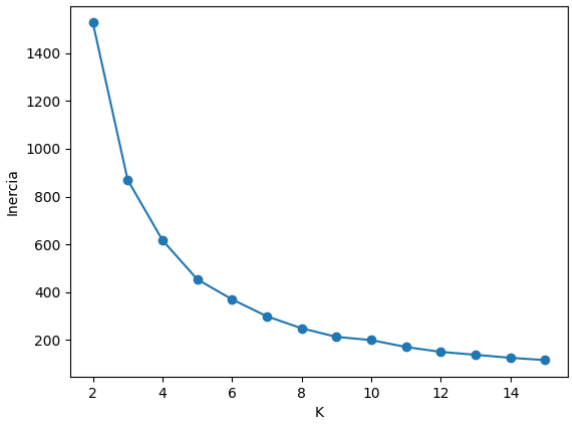

Figura 1: Gráfico generado en Python mediante la metrica *Inercia*.

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

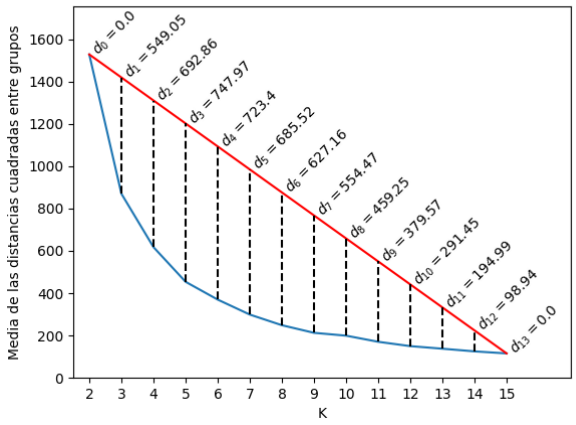

Figura 2: Gráfico generado en Python para aplicar el Método del codo.

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

Por lo tanto se utiliza $k = 5$ para *K-Means*. Obteniendo asi la siguiente tabla con los centroides o valores promedio de las variables de cada grupo.

booktabs

	Edad	Genero	IMC	Pasos Diarios	Horas de sueño	Agua consumida	Calorias
0	0.511751	2.442491e-15	0.513989	0.505706	0.520111	0.500827	0.525152
1	0.508956	1.000000e+00	0.481818	0.465464	0.516241	0.505310	0.504535
2	0.514951	6.550316e-15	0.491878	0.506076	0.478971	0.513320	0.513688
3	0.529392	1.000000e+00	0.490052	0.490275	0.488648	0.499289	0.502031
4	0.491502	5.366972e-01	0.492900	0.486518	0.490531	0.511417	0.526274

	Fumador	Alcohol	Descanso	Presion sistolica	Presion diastolica	colesterol	Historial Familiar
0	0.216828	1.000000e+00	0.494221	0.504200	0.506994	0.512760	3.106796e-01
1	0.221053	1.942890e-15	0.505134	0.499965	0.491383	0.502974	2.220446e-16
2	0.202899	1.942890e-15	0.512063	0.492707	0.477629	0.490767	3.885781e-16
3	0.208754	1.000000e+00	0.537827	0.498506	0.473720	0.517773	3.164983e-01
4	0.192661	1.831868e-15	0.498362	0.516158	0.479708	0.502386	1.000000e+00

Cuadro 1: Centroides de las variables.

Ahora, utilizando la metrica *Calinski-Harabasz*

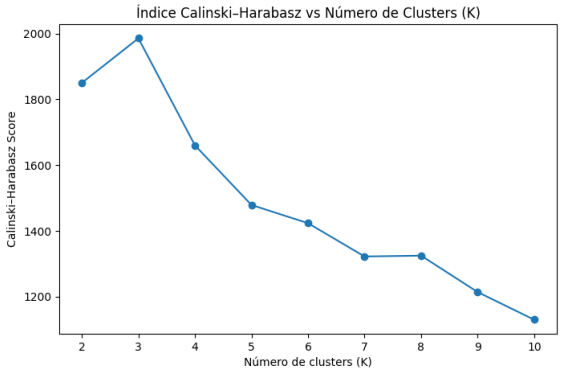

Figura 3: Gráfico generado en Python mediante la metrica *Calinski-Harabasz*.

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

Aplicando al algoritmo *OPTICS* el cual obtiene los clusters necesarios

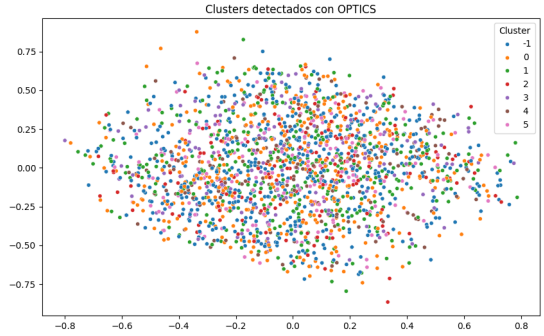

Figura 4: Gráfico generado en Python mediante *OPTICS*.

5. Conclusiones y discusiones

Podemos concluir principalmente 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 presion sistólica mayor, aunque un IMC menor. Lo cual podría hacer sentido debido a que la presion sistólica es la principal en verse afectada al tener una mala noche de sueño.

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

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

Referencias

url

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

Mayo Clinic. (2025, marzo 07). ¿La falta de sueño puede provocar presión arterial alta? https://www.mayoclinic.org/es/diseases-conditions/high-blood-pressure/expert-answers/sleep-deprivation/faq-20057959