¿Quién es mas propenso a tener diabetes? Análisis y agrupamiento mediante K-means.

Antonio Chacón Flores
Lic. Tecnologias para la Información en Ciencias
ENES Morelia, UNAM.
chacon.floresantonio@gmail.com

Figure 1: Podemos observar como se realiza la prueba del azúcar en alguna persona.

ABSTRACT

En este proyecto, que lleva por nombre: ¿Quién es mas propenso a tener diabetes? Análisis y agrupamiento mediante K-means, tenemos dos propósitos, el primero es trabajar con aprendizaje no supervisado, el cual tiene varios métodos, pero el elegido por nosotros fue K-means, del cual hablaremos a continuación.
Además me gustaría mencionar algunas cifras de las grandes cifras que se manejan de diabetes en todo el mundo, pero mas en especifico en México, este reporte fue pensando para trabajar con todos estos datos de mujeres de 21 años o mas, las cuales fueron expuestas a una prueba de diabetes en la cual existen tanto casos positivos que se manejan con un 1, como casos negativos que es un 0.

ACM Reference Format:
Antonio Chacón Flores. 2021. ¿Quién es mas propenso a tener diabetes? Análisis y agrupamiento mediante K-means.. In *Proceedings of Proyecto Minería de Datos.*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Proyecto Minería de Datos.,

https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 KEYBORDS

- Clustering
- K-means

2 INTRODUCCIÓN

De forma simple, en el aprendizaje no supervisado, un algoritmo segrega los datos en un conjunto de datos en el que no están etiquetados en función de algunas características ocultas en los datos. Esta función puede ser útil para descubrir la estructura oculta de los datos y para tareas como la detección de anomalías. Las técnicas de aprendizaje no supervisado se pueden aplicar sin necesidad de tener los datos etiquetados para el entrenamiento.

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.

Algunas de las ventajas de utilizar Aprendizaje no Supervisado:

- El Aprendizaje no Supervisado encuentra todo tipo de patrones desconocidos en los datos.
- Los métodos no supervisados te ayudan a encontrar características que pueden ser útiles para la categorización.

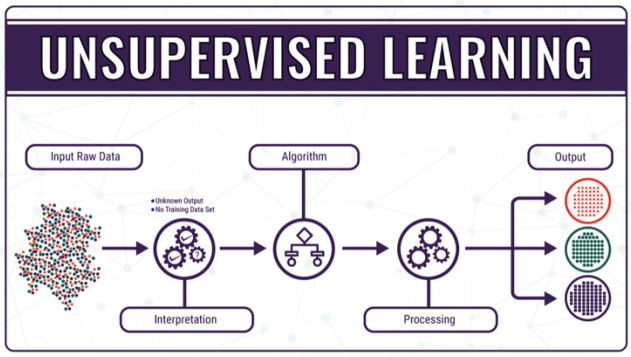

Figure 2: En esta imagen se puede apreciar un poco como es que trabaja el Aprendizaje No Supervisado.

- Es más fácil obtener datos no etiquetados que los datos etiquetados.

3 DATOS Y ANTECEDENTES

Comenzaremos hablando de nuestros datos, estos fueron un dataset de Kaggle **datos** con los cuales no tuvimos tanto problema de procesamiento en cuanto a valores vacíos o nulos ya que no existían. Quisimos comprobar que nuestro dataset no tenia valores nulos, así que nos ayudamos con un mapa de calor(ver fig.3).

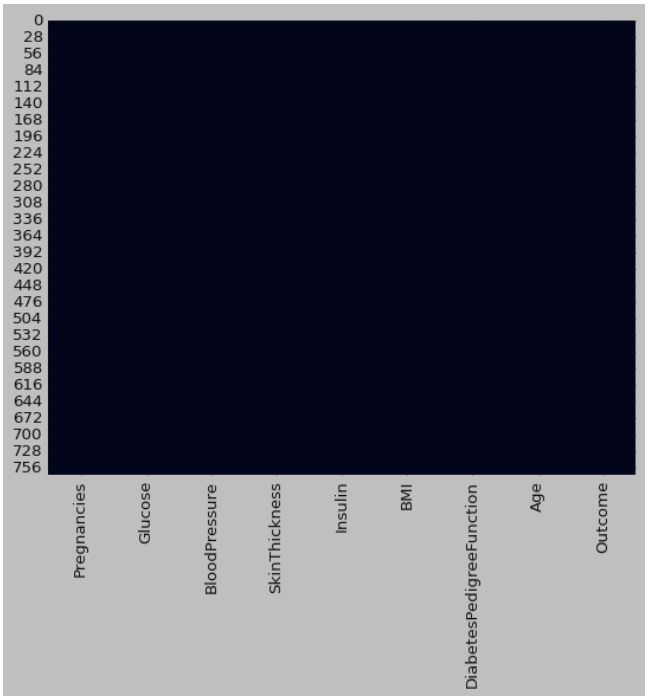

Figure 3: En esta imagen se puede apreciar un mapa de calor el cual nos ayuda a verificar que no existen datos nulos en el dataset.

Pero tuvimos que decidir cuales eran las columnas adecuadas para trabajar con ellas y no generar "basura" en nuestros datos. Para ello ocupamos de una matriz de correlación, la cual nos ayudo a decidir cuales eran las columnas que mas nos podrían servir. Después de ello pudimos quedarnos con las columnas que mas nos servían, a pesar de que la matriz nos servía, nos pudimos dar cuenta que por obvias razones las columnas que nos dios la matriz de correlación eran las que por lógica elegiría una persona que conoce un poco de la enfermedad(ver fig.4).

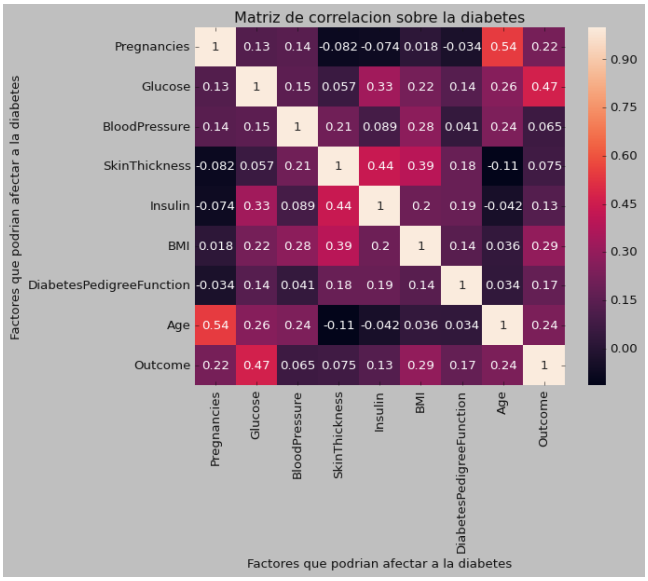

Figure 4: En esta imagen se puede apreciar la matriz de correlación con todas las columnas.

Después de haber observado la matriz nos pudimos dar cuenta que la mejor correlación era en los factores que se interceptaban los factores y los después mencionados eran los más frecuentes, por lo tanto trabajaremos con las siguientes columnas: Pregnancies, Glucose, Age,Insulin, junto con Outcome. Y si lo pensamos por obvias razones son los factores que más son frecuentes en personas que tienen diabetes. Así que ahora hicimos lo mismo pero con las columnas que si trabajaremos(ver fig.5).

4 EXPERIMENTOS

Para comenzar, realizamos el método del codo el cual es unos de los métodos para elegir un k ideal, aunque no es perfectamente exacto pero busca ser lo mas preciso, así que para eso realizamos un pequeño código para crear una gráfica donde nos pueda ayudar a tomar la decisión de que k elegir.

Con el método del codo antes realizado pudimos darnos cuenta que un valor ideal de k, es k = 6. Bien sabemos que no existe un valor exacto de k, pero mas adelante en la interpretación de nuestros clusters podremos darnos cuenta si elegimos un buen k(ver fig.6).

En los datos anteriores podemos traducirlos que son las coordenadas de los 6 clusters.

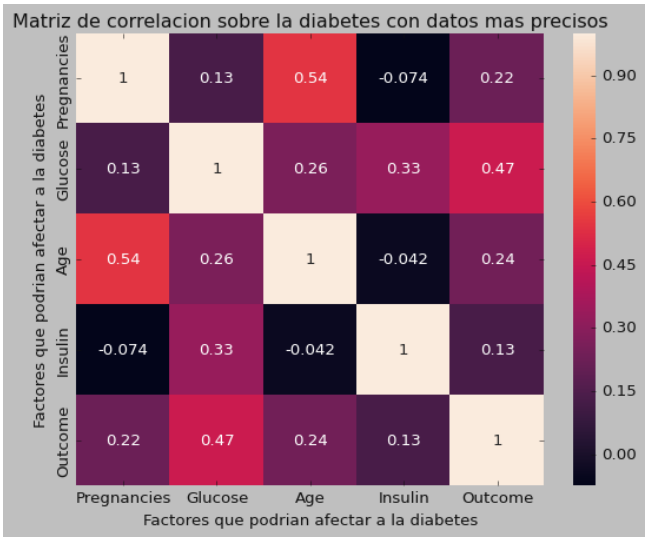

Figure 5: En esta imagen se puede apreciar la matriz de correlación pero solo con las columnas que ocuparemos.

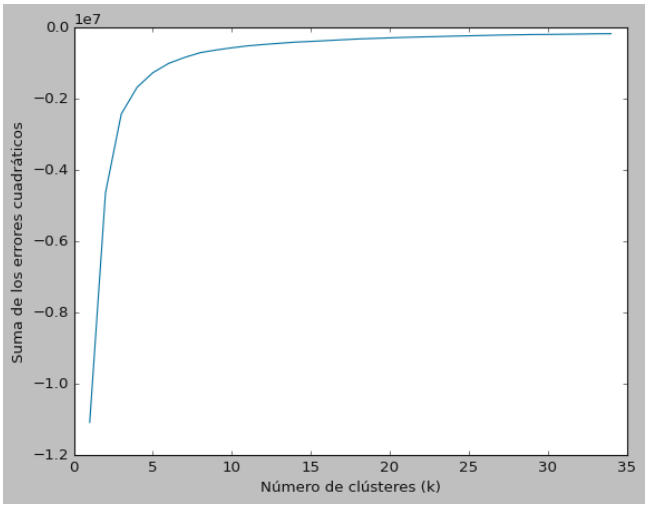

Figure 6: Metodo del codo para elegir un k ideal.

Después de haber elegido nuestro k ideal, corrimos el código de K-means, el cual nos va a hacer que cada uno de nuestros registros les elija el cluster ideal, así que creamos una columna extra en la cual se pondra el cluster para cada registro.

Con esto podemos imprimir nuestros resultados, los cuales son todos los registros que necesitamos con todas las columnas necesarias para obtener el mejor resultado.

A continuación se muestra en que cluster quedo cada fila.

5 RESULTADOS Y DISCUSIÓN

Explicación de cada uno de nuestros clusters. A continuación podemos ver los datos que nos arrojo el algoritmo k-means antes realizado.

En el **primer cluster** pudimos darnos cuenta que la probabilidad de que les diera diabetes es 50% a las persona que en promedio tienen:

- Pregnancies = 4.159091
- Glucose = 150.386364
- Age = 33.863636
- Insulin = 302.772727
- Outcome = 0.500000

En el **segundo cluster** pudimos darnos cuenta que la probabilidad de que les diera diabetes es 50% a las persona que en promedio tienen:

- Pregnancies = 3.800000
- Glucose = 134.264286
- Age = 33.292857
- Insulin = 167.935714
- Outcome = 0.500000

En el **tercer cluster** pudimos darnos cuenta que la probabilidad de que les diera diabetes es 60% a las persona que en promedio tienen:

- Pregnancies = 4.727273
- Glucose = 152.489510
- Age = 39.916084
- Insulin = 0.944056
- Outcome = 0.608392

En el **cuarto cluster** pudimos darnos cuenta que la probabilidad de que les diera diabetes es 13% a las persona que en promedio tienen:

- Pregnancies = 2.800000
- Glucose = 104.047059
- Age = 28.247059
- Insulin = 81.682353
- Outcome = 0.135294

En el **quinto cluster** pudimos darnos cuenta que la probabilidad de que les diera diabetes es 21% a las persona que en promedio tienen:

- Pregnancies = 4.071146
- Glucose = 98.632411
- Age = 32.592885
- Insulin = 2.185771
- Outcome = 0.213439

En el **sexto cluster** pudimos darnos cuenta que la probabilidad de que les diera diabetes es 66% a las persona que en promedio tienen:

- Pregnancies = 3.111111
- Glucose = 165.833333
- Age = 34.555556
- Insulin = 548.833333
- Outcome = 0.666667

Cada que volvemos a correr el código podemos darnos cuenta debemos tener conciencia que los datos se modificaran un poco, pero en mi opinión mi algoritmo separo los clusters muy bien(ver fig.7).

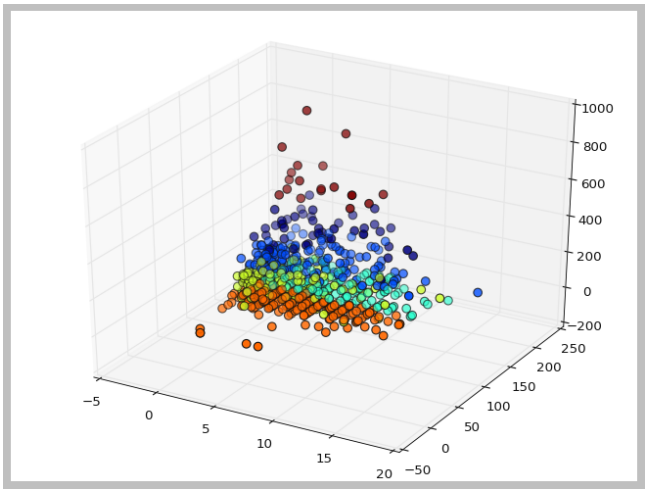

Figure 7: En esta grafica pudimos darnos cuenta de la clasificación de los clusters por colores y se hizo en 3 dimensiones.

6 CONCLUSIONES

Por ultimo, como podemos ver en los pequeños grupos(clusters), los cuales cada uno se refiere a un grupo que tiene características muy similares, de esto se trata lo que hace k-mean, a partir de parámetros que ya tenemos, basándose en características.

En algunos casos es de suma importancia tomar algunas columnas y no todas, ya que en algunos casos los datasets pueden tener columnas que sean "basura" y nos puedan perjudicar si las usamos. En este caso yo utilice la matriz de correlación para solo elegir de las mas importantes y por nuestra gráfica pudimos notar que hicimos una buena elección, tanto de el, como de las columnas seleccionadas.

En conclusión, me gustaría hacer mención que no solo el dejar de producir insulina provoca diabetes, sino el producir en exceso también lo provoca:

Al principio, la resistencia a la insulina hace que el cuerpo produzca insulina adicional para compensar la insulina ineficaz. El exceso de insulina en el torrente sanguíneo puede causar hipoglucemia. Pero la resistencia la insulina tiende a empeorar con el tiempo hasta que finalmente disminuye la capacidad del cuerpo para producir insulina. A medida que los niveles de insulina bajan, los niveles de azúcar suben. Si los niveles no vuelven a la normalidad, la persona puede desarrollar diabetes tipo 2. **Vean este sitio, puede ser de su interes.**

7 REFERENCES

[1] 2020. Diabetes. https://medlineplus.gov/spanish/diabetes.html#:~:text=La%20diabetes%20es%20una%20enfermedad,el%20cuerpo%20no%20produce%20insulina.

[2] 2019. Diabetics prediction using logistic regression. https://www.kaggle.com/kandij/diabetes-dataset