Agrupación de Vinos

Daniel J. Barandica F, *dabarandica@javeriana.edu.co*
Bogotá D.C, Pontificia Universidad Javeriana

***Resumen*—El presente documento tiene como propósito explicar el algoritmo implementado para llevar a cabo la agrupación de un conjunto de vinos, de los cuales solo se tiene conocimiento sobre sus características. Además, se muestran los resultados obtenidos y el análisis realizado a estos.**

I. BUSINESS UNDERSTANDING

El método de agrupación de K-mean permite organizar diferentes grupos según la relación encontrada en las características que le sean ingresadas al algoritmo. No se tienen salidas etiquetadas.

Sabiendo lo anterior, se implementó el algoritmo K-means para la agrupación de diferentes clases de vinos, a partir de 13 características tomadas de cada vino.

El poder crear grupos a partir de las características de los vinos, permite organizas la información de tal manera que se entienda mucho mejor y a la hora de analizar, será más fácil encontrar factores en común que relacionen los datos de cada agrupación obtenida.

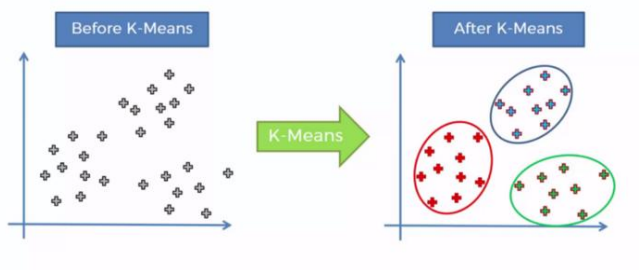

Figura 1. Resultado de realizar K-means. [1]

II. DATA UNDERSTANDING

La base de datos usada para este programa se encuentra en: https://www.kaggle.com/harrywang/wine-dataset-for-clustering.

La base de datos cuenta con 13 características de vinos, que son el resultado de análisis químicos realizados a los mismos. Las características son:

- Alcohol: Variable que mide la cantidad de alcohol que existen en el vino. Comúnmente esta característica también indica que tanta calidez tendrá el vino en la boca
- Acido málico: Variable que mide la cantidad de ácido málico en el vino, el cual es un acido de origen natural, presente en alimentos vegetales [3]
- Fresno: Variable que mide la cantidad de fresno en el vino
- Alcalinidad del fresno: Variable que mide la capacidad del fresno presente en el vino, de neutralizar los ácidos
- Magnesio: Variable que mide la cantidad de magnesio que tiene el vino
- Fenoles totales: Variable que mide los compuestos orgánicos aromáticos que tiene presente el vino [4]
- Flavonoides: Variable que mide los diversos fitonutrientes que se encuentran en el vino
- Fenoles no flavonoides: Variable que mide los compuestos orgánicos aromáticos que no contienen fitonutrientes en el vino
- Proantocianidinas: Variable que mide la cantidad de Proantocianidinas presentes en el vino, debido a la piel de las uvas
- Intensidad de color: Variable que mide el grado de opacidad del vino [5]
- Matiz: Variable que mide el grado de vejez del vino
- OD280/OD315 de vinos diluidos: Variable que indica la cantidad concentración de proteína presente en el vino [6]
- Prolina: Variable que mide la cantidad de prolina (aminoácido) en el vino.

III. DATA PREPARATION

La base de datos contaba con datos únicamente de tipo numérico, por lo tanto, no fue necesario modificar información de estos. Sin embargo, para poder trabajar los datos de manera más acertada, se debió estandarizar los datos a través de la

[*] Inteligencia Artificial 2021.

función standartscaler. Obteniendo datos que tendrían un mínimo de -1 y un máximo de 1.

En la imagen debajo se observa la formula utilizada por la función standartscaler, en donde se resta la media de los datos y se divide sobre la desviación estándar.

$$z = \frac{x - \mu}{\sigma}$$

$$\mu = \text{Mean}$$
$$\sigma = \text{Standard Deviation}$$

Figura 2. Fórmula para estandarización. [1]

IV. MODELLING

De manera resumida, se mostrará los pasos seguidos para la realización del algoritmo presentado en el documento:

1. Lectura de dataset: A través de la librería de pandas se convierte un archivo CSV (comúnmente los dataset vienen en archivos tipo CSV) en un dataframe

2. Estandarización de características: Se estandarizan los datos para que al momento de realizar el algoritmo de K-means no se presenten errores debido a datos con diferentes magnitudes en comparación a los demás.

3. Comprobación de PCA: Con el fin de verificar si al reducir las características del dataset se obtenían mejores resultados en el coeficiente de silueta, se fue probando con diferentes números de componentes hasta obtener un 90 porciento de varianza. Luego, se compararon los resultados obtenidos con PCA y sin PCA

4. K-means: Por último, se implemento el algoritmo de K-mean con la librería Sklearn, y se comprobó por medio del coeficiente de silueta, cual sería el mejor K y Random State para el dataset estudiado.

V. EVALUATION

Para verificar que los resultados obtenidos al realizar el algoritmo de K-means, se opto por hacer uso del coeficiente de silueta, el cual nos indica que tan buena es la predicción de que los datos pertenezcan a los grupos, en donde fueron agrupados.

En Fig. 3 se observa los valores de coeficiente de silueta obtenidos de la siguiente forma:

- En la función de K-means de la librería Sklearn, se varía el parámetro Random State desde 0 hasta 20, y, para cada Random State se variaría el parámetro K desde 2 hasta 20.

Teniendo claro lo mencionado, se puede deducir de la gráfica que el K con valor igual a 3, es el mejor para el dataset estudiado.

Por otro lado, en Fig. 4 se observa los valores de coeficiente de silueta con PCA, los cuales mejoraron en un 13% aproximadamente, en comparación con los coeficientes de silueta sin PCA. Esto se vuelve a confirmar al observar las figuras 5 y 6, en donde vemos de manera mas evidente, como los datos con PCA son agrupados de manera más acertada.

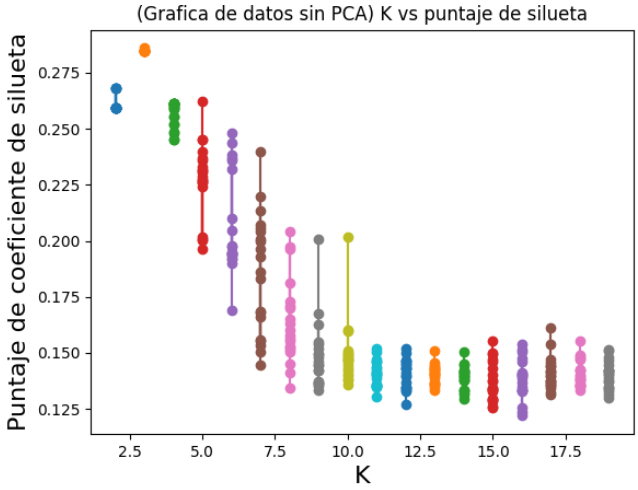

Figura 3. Coeficientes de silueta obtenidos para datos sin PCA.

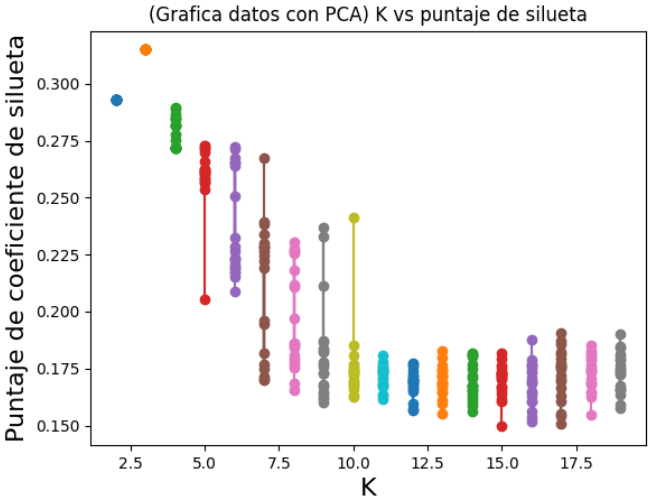

Figura 4. Coeficientes de silueta obtenidos para datos con PCA.

Las graficas mostradas en Fig. 5 y Fig. 6, fueron realizadas con solo dos características, de las 13. Con el propósito de mostrar así sea con dos características, como mejoraba la agrupación entre las características con PCA y sin PCA.

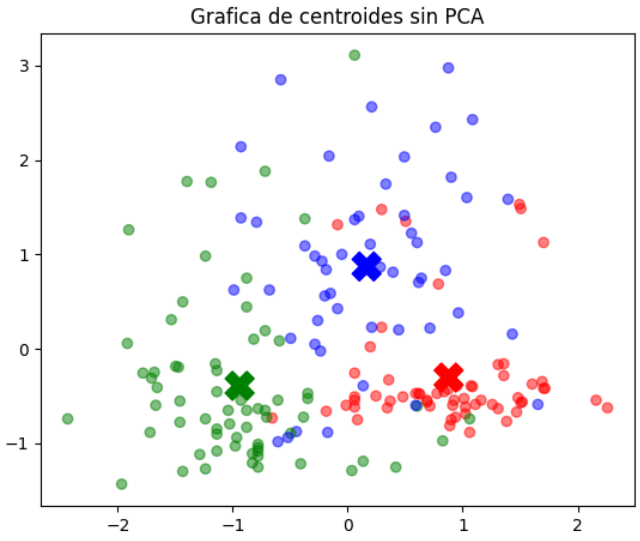

Figura 5. Datos agrupados sin PCA.

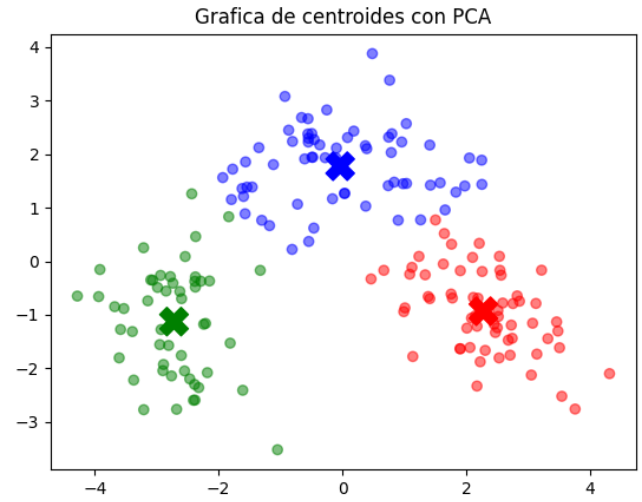

Figura 6. Datos agrupados con PCA

En la tabla 1 se observan los coeficiente de siluetas obtenidos con el valor de K que presento un mayor coeficiente (K=3), y, como este varia cuando se cambia el parametro Random State. Sin embargo, se observa que al modificar este ultimo parametro no influye en el valor del coeficente de silueta, exceptuando el Random State igual a 8, el cual presentó una mejoria en el coeficiente.

Tabla 1. Coeficiente de silueta sin PCA.

K	Silueta_Score	Random state
3	0.284859	0
3	0.284859	1
3	0.284859	2
3	0.284859	3
3	0.284859	4
3	0.284859	5
3	0.284859	6
3	0.284859	7
3	0.285941	8
3	0.284859	9

En la tabla 2 se observa que el parametro Random State no realiza ningun cambio en el valor del coeficente de silueta, por lo tanto, este no tiene mucha repercusion para el dataset de los vinos. Sin embargo, siempre es necesario probar con diferentes semillas (Random state) con el fin de verificar si mejora la agrupación o no.

Tabla 2. Coeficiente de silueta con PCA.

K	Silueta_Score	Random state
3	0.314969	0
3	0.314969	1
3	0.314969	2
3	0.314969	3
3	0.314969	4
3	0.314969	5
3	0.314969	6
3	0.314969	7
3	0.314969	8
3	0.314969	9

VI. Conclusiones

A partir de los resultados obtenidos, se logro llegar a las siguientes conclusiones:

- El parámetro random state es importante, ya que dependiendo de donde se inicialice el algoritmo K-means, así mismo podrá variar la predicción de agrupaciones que se realicen. Sin embargo, esto variará dependiendo del tipo de dataset que se este trabajando, en este caso se observo que no importó mucho el parámetro random state. Esto puede ser debido a que se tenían demasiados datos y sin importar en donde se inicializará el algoritmo, este agruparía de igual manera los datos.

- El algoritmo de agrupación K-means tendrá un mejor coeficiente de silueta al hacer uso de la reducción de dimensionalidad de las características (PCA).

Referencias

[1] S. Srivastava, "Feature Scaling in Scikit-learn", *Data Science Journey*, 2019. [Online]. Available: https://datasciencewithshobhit.blogspot.com/2019/02/feature-scaling-in-scikit-learn.html.

[2] E. Carreras, "Las 5 características básicas de un vino", *Maset*, 2019. [Online]. Available:

https://www.maset.com/es/blog/las-5-caracteristicas-basicas-de-un-vino.

[3] "¿Qué es el Ácido Málico? Beneficios y propiedades", *Blog.nutritienda.com*, 2010. [Online]. Available: https://blog.nutritienda.com/acido-malico/#:~:text=El%20%C3%A1cido%20m%C3%A1lico%20(C4H6O5)%2C,derivados%2C%20como%20en%20el%20vino.

[4]"Fenoles", *Prtr-es.es*. [Online]. Available: https://prtr-es.es/fenoles,15658,11,2007.html.

[5]G. Gómez, "El singular vino de René Ormazabal Moura", *catadelvino*, 2021. [Online]. Available: https://www.catadelvino.com/blog-cata-vino/singular-vino-rene-ormazabal-moura.

[6]B. Xueting and L. Hanning, *Identification of red wine categories based on physicochemical properties*. 2019.