Práctica 5. Clustering

Nombre: Gregorio Alberto Alvarez Alvarez

Materia: Análisis estadístico multivariable

1. Introducción

Hasta la actualidad se ha determinado la calidad del vino en base a la percepción sensorial de los usuarios, ya sean patrones ya establecidos o experiencia propia de los usuarios. Entre los factores que se toman en cuenta para la calidad del vino se encuentran el conjunto de sabores, aromas, colores y texturas 1. Debido a que no se tienen patrones bien establecidos para determinar la calidad del vino, esta puede estar sujeta a errores humanos 2 o percepciones subjetivas que no representen los gustos de los consumidores 3. Sin embargo, se ha puesto en evidencia que la calidad del vino se puede determinar en base a su composición química lo cual estandariza la calidad del vino y permite que se pueda generalizar.

2. Objetivo:

El objetivo de esta práctica es aplicar los métodos de clustering para agrupar los datos de vino en base a su composición química. Analizar las características de cada clúster y encontrar sus diferencias. Finalmente, determinar si existe una relación entre los grupos encontrados y la calidad del vino.

2. Descrición de la base de datos

La base de datos fue obtenida de 4. Esta contiene el puntaje de calidad de los vinos con un rango de 0 a 10, que sirve para describir desde un vino muy malo (0) hasta un vino excelente (10). También contiene 11 atributos psico químicos de los mismos. La base de datos contiene 1599 registros y 12 columnas. Las columnas son:

- fixed acidity: Acidez fija del vino.
- volatile acidity: Acidez volátil del vino.
- citric acid: ácido cítrico presente en el vino.
- residual sugar: cantidad de azúcar restante después de la fermentación.
- chlorides: cantidad de cloruros.
- free sulfur dioxide: cantidad de dióxido de azufre libre.
- total sulfur dioxide: cantidad de dióxido de azufre libre y combinado.
- density: densidad del vino.
- pH: pH del vino.
- sulphates: cantidad de SO2.
- alcohol: porcentaje de alcohol.
- quality: calidad del vino, rango de 0 a 10.

3. Desarrollo de los datos

3.1. Preprocesamiento de los datos

Con el fin de visualizar gráficamente la separación entrando por el proceso de cauterización, se aplicó una reducción de variables mediante el análisis de la varianza. Se obtuvieron 3 variables poco correlacionadas, las cuales se utilizaron para el análisis de cauterización:

- pH
- free sulfur dioxide
- residual sugar

3.2. Clustering

Se aplicaron los siguientes métodos de clustering:

- Aglomerativo
- K-means

3.2.1. Elección del número de clústeres

Se normalizaron los datos y se obtuvo un endógama por medio del método de Ward con una métrica euclidiana (Figura 1). Mediante una prueba visual se decidió utilizar 2 clústeres como primera aproximación.

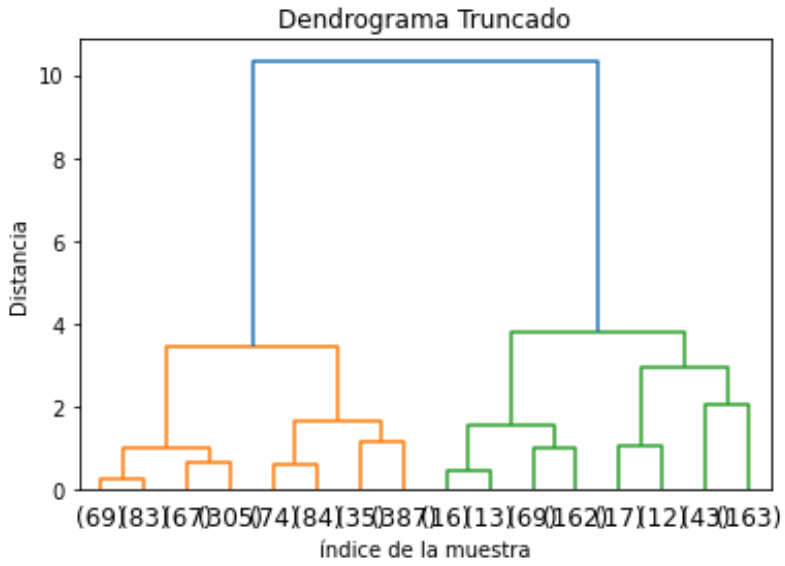

Figura 1. Endógama obtenido mediante el método de Ward.Truncado en una profundidad de 3.

Asi mismo, mediante un gráfico de codo, obtenido con el método de k-means++, se obtuvo que un numero adecuado de clúster es 2 (Figura 2).

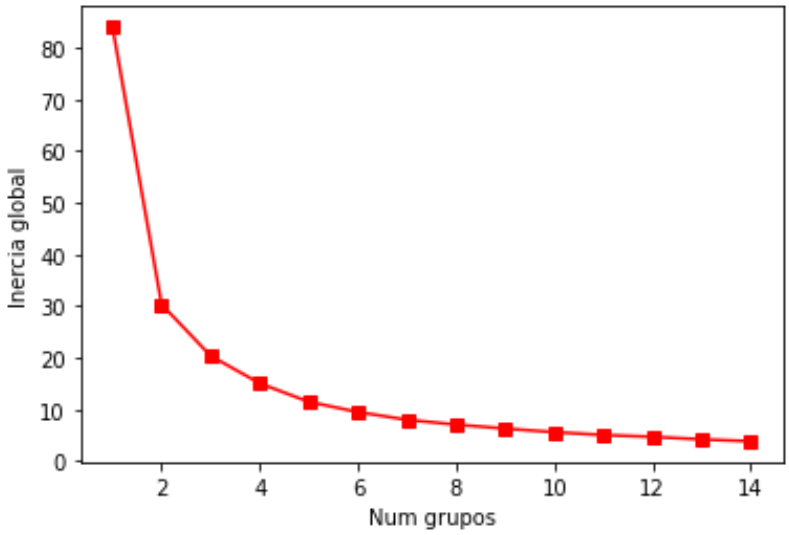

Figura 2. Método del codo para determinar el número de clústeres.

3.2.2. Clustering por aglomeración

Se aplico el método de aglomeración con un método de Ward y una métrica euclidiana. Se obtuvo la siguiente representación gráfica (Figura 3).

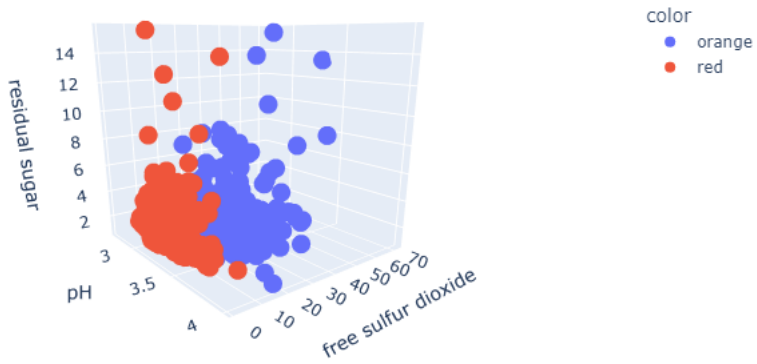

Figura 3. Representación grafica del clustering por aglomeración.

Como se ve en la figura, los clústeres se encuentran bien separados, sin embargo, se observa que hay un clúster que contiene la mayoría de los datos. También se observa que la separación entre los clústeres es muy importante en la variable `free sulfur dioxide`, siendo el resto de las variables menos significativas para la separación de los clústeres. Esto es un comportamiento esperado, ya que se sabe que la metodología Ward toma en cuenta la dirección de la varianza para clisterizar y se sabe que la variable `free sulfur dioxide` representa una parte importante de la varianza de los datos.

3.2.3. Clustering por K-means

Se aplico el metodo de k-means con un numero de clústeres igual a 2. Se obtuvo la siguiente representación grafica (Figura 4).

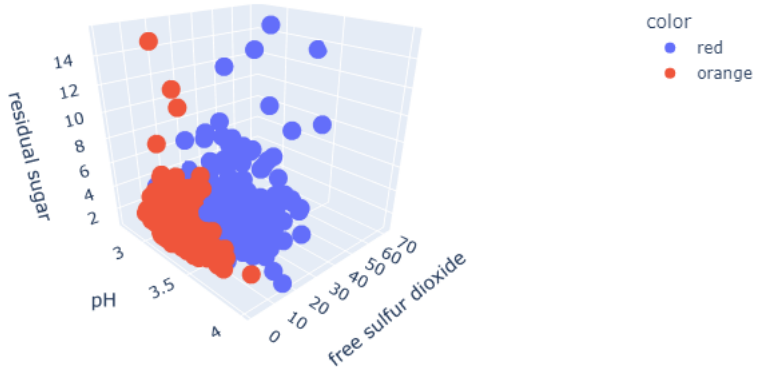

Figura 4. representación grafica del clustering por k-means.

De la Figura 4, se puede observar a primera vista que la separación de los clústeres con ambas metodologías es muy parecida.

3.3. Análisis de los clústeres

3.3.1. Cluster 1

Se obtuvieron los estadísticos de los datos del clúster 1 y se compararon con los datos originales (Tabla 1).

	Cluster_1			Original		
	pH	free sulfur dioxide	residual sugar	pH	free sulfur dioxide	residual sugar
count	495.00	495.00	495.00	1599.00	1599.00	1599.00
mean	3.29	6.01	2.70	3.31	15.87	2.54
std	0.16	2.13	1.60	0.15	10.46	1.41
min	2.86	1.00	1.20	2.74	1.00	0.90
25%	3.19	5.00	1.90	3.21	7.00	1.90
50%	3.29	6.00	2.20	3.31	14.00	2.20
75%	3.38	7.00	2.70	3.40	21.00	2.60
max	3.90	23.00	15.50	4.01	72.00	15.50

Tabla 1. Estadísticos de los datos del clúster 1.

Se puede observar que la variable `free sulfur dioxide` del clustre 1 tienen una media muy por debajo de la media de los datos originales. El pH está un poco por debajo de la media y las azucares residuales un poco por encima de la media. Estos datos representan el límite inferior de dióxido de azufre libre.

Se calculo el porcentaje de registros por calificación de calidad contenidos en el clúster 1 (Tabla 2). Se encontró que este clúster contiene a la mayoría de los vinos con calificación 3 y 8. Con lo que muestra que este clúster contiene a los vinos con calificaciones extremas.

	quality	% Cluster 1
0	3	60

	quality	% Cluster 1
1	4	47.1698
2	5	25.6975
3	6	31.348
4	7	39.6985
5	8	55.5556

Tabla 2. Porcentaje de registros por calificación de calidad contenidos en el clúster 1.

3.3.2. Clúster 2

Se obtuvieron los estadísticos de los datos del clúster 2 y se compararon con los datos originales (Tabla 3).

	Cluster_2			Original		
	pH	free sulfur dioxide	residual sugar	pH	free sulfur dioxide	residual sugar
count	1104.00	1104.00	1104.00	1599.00	1599.00	1599.00
mean	3.32	20.30	2.46	3.31	15.87	2.54
std	0.15	9.65	1.31	0.15	10.46	1.41
min	2.74	8.00	0.90	2.74	1.00	0.90
25%	3.22	13.00	1.90	3.21	7.00	1.90
50%	3.32	17.50	2.20	3.31	14.00	2.20
75%	3.41	26.00	2.60	3.40	21.00	2.60
max	4.01	72.00	15.40	4.01	72.00	15.50

Tabla 3. Estadísticos de los datos del clúster 2.

En la tabla 3 se puede observar que la variable `free sulfur dioxide` del clúster 2 tienen una media muy por encima de la media de los datos originales. El pH está un poco por encima de la media y las azucares residuales un poco por debajo de la media. Estos datos representan el límite superior de dióxido de azufre libre.

Se obtuvo el porcentaje de entradas por calificación de calidad contenidos en el clúster 2 (Tabla 4). Se encontró que este clúster contiene a la mayoría de los vinos con calificación 4, 5, 6 y 7. Con lo que muestra que este clúster contiene a los vinos con calificaciones intermedias.

	quality	% Clúster 2
0	5	74.3025
1	6	68.652
2	7	60.3015
3	4	52.8302
4	8	44.4444
5	3	40

Tabla 4. Porcentaje de registros por calificación de calidad contenidos en el cluster 2.

Conclusiones

Se encontró que ambos métodos de cauterización indican, visualmente, un numero de clústeres igual y la separación obtenida es muy similar.

Se encontró que el clúster 1 contiene a los vinos con calificaciones extremas y el clúster 2 contiene a los vinos con calificaciones intermedias.

Con la información obtenida, no se puede concluir que se sigue una tendencia lineal entre la variable `free sulfur dioxide` y la calidad del vino, ni que esta variable es importante para clasificar los vinos en calidad alta o baja.

Para trabajos futuros, será necesario realizar un análisis más profundo de la métrica que siguen los datos, así como del número de clústeres y de las variables que se utilizan para el clustering, ya que se podrían encontrar subgrupos de vinos en los valores extremos.

Referencias

- Literatura

[1] https://vineroutes.com/wine-rating-system/

[2] https://www.scitepress.org/Papers/2015/55519/55519.pdf

[3] https://www.sciencedirect.com/science/article/abs/pii/S0023643814005817

- Código

[git] https://github.com/Gegori1/analisis-estadistico-multivariado/tree/master/practica_5