INFORME DESARROLLO CASO DE ESTUDIO MÓDULO 3

APRENDIZAJE NO SUPERVISADO

ANÁLITICA PARA LA TOMA DE DECISIONES

OSIRIS CONTRERAS TRILLOS

DAVID STEEVEN TAMAYO TORO

JUAN JOSÉ MOLINA OCAMPO

MARITZA ZAPATA GONZÁLEZ

DOCENTE

MANUELA LONDOÑO OCAMPO

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

2024 - S1

INTRODUCCIÓN

La clusterización, también conocida como segmentación, es una técnica fundamental en el análisis de datos. Al explorar una base de datos que contiene características de vinos rojos y blancos, obtenidos a través de análisis químicos, proporcionan información valiosa sobre vinos, que nos permitirán aplicar las técnicas de análisis de datos del aprendizaje no supervisado. Para el desarrollo de la actividad se realizó un estudio de caso para la clasificación de vinos.

Contexto: Una empresa vinícola te contrata con el objetivo de clasificar diferentes vinos en grupos o "clusters" según sus características químicas y sensoriales. Esto les ayudará a comprender mejor las relaciones entre los vinos y a tomar decisiones informadas sobre la producción y catalogarlos para su comercialización.

Las columnas en el dataset representan diferentes características fisicoquímicas de los vinos. Aquí está una breve descripción:

1. Acidez Fija: La cantidad de ácidos no volátiles en el vino.
2. Acidez Volátil: La cantidad de ácidos volátiles en el vino.
3. Ácido Cítrico: La concentración de ácido cítrico en el vino.
4. Azúcar Residual: La cantidad de azúcar que queda después de la fermentación.
5. Cloruros: La concentración de cloruros en el vino.
6. Dióxido de Azufre Libre: La cantidad de SO_2 libre en el vino.
7. Dióxido de Azufre Total: La cantidad total de SO_2 en el vino.
8. Densidad: La densidad del vino.
9. pH: El nivel de acidez o alcalinidad del vino.
10. Sulfatos: La concentración de sulfatos en el vino, Puede afectar la estabilidad y el sabor.
11. Alcohol: El contenido de alcohol en el vino.
12. Calidad: Una puntuación subjetiva de calidad (por ejemplo, de 1 a 10).
13. Color: El color del vino (blanco o tinto).

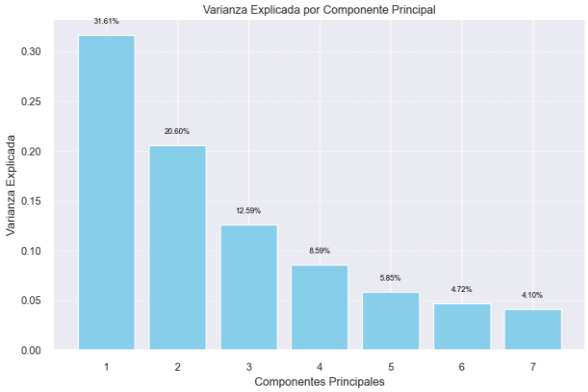
Objetivo Utilizar técnicas de clusterización para agrupar los vinos en categorías similares. Explorando algoritmos como k-means, DBSCAN y clustering jerárquico y aplicando técnicas para la reducción de dimensionalidad.

Análisis exploratorio: Eliminación de duplicados, graficas de correlación, histogramas, boxplots, análisis de la variable categórica color, contenido de alcohol por color.

Procesamiento de los datos: Se transforma la variable categórica color con la proporción respectiva de cada color dentro del conjunto de datos original y luego se procede a escalar las variables. Posteriormente se tratan lo datos atípicos utilizando dos métodos: rango interquantilico y winsorizado.

Reducción de la dimensionalidad por PCA: Se realizó el análisis de componentes principales para el dataset escalado y contratamiento de atípicos por los dos métodos, los resultados son los siguientes:

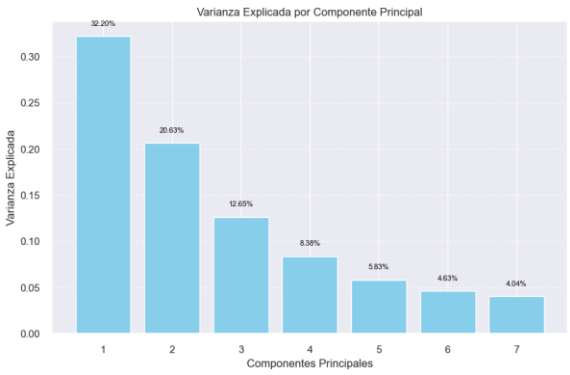
Dataset original quantiles



	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
fixed_acidity	0.26	0.19	0.49	0.13	0.12	0.10	-0.06
volatile_acidity	0.36	0.01	-0.25	0.10	0.34	0.38	0.09
citric_acid	-0.11	0.17	0.59	-0.09	-0.33	-0.20	-0.12
residual_sugar	-0.19	0.41	-0.07	-0.19	0.55	-0.30	0.24
chlorides	0.38	0.19	-0.00	-0.04	-0.13	0.20	-0.42
free_sulfur_dioxide	-0.31	0.24	-0.12	-0.30	-0.07	0.54	-0.26
total_sulfur_dioxide	-0.36	0.28	-0.12	-0.13	-0.13	0.30	0.05
density	0.23	0.50	-0.02	-0.19	0.17	-0.24	-0.01
pH	0.15	-0.19	-0.43	-0.45	-0.27	-0.40	-0.19
sulphates	0.27	0.01	0.16	-0.53	-0.28	0.21	0.65
alcohol	-0.07	-0.49	0.19	-0.10	0.27	0.15	0.18
quality	-0.12	-0.28	0.27	-0.52	0.42	0.01	-0.39
color	-0.46	0.01	-0.02	0.11	-0.08	-0.07	0.15

Para este dataset hay 7 componentes principales, el primero de ellos explica el 31.61% de la varianza, y disminuye hasta llegar a el último, que explica el 4.10%. El componente 1 tiene una relación positiva fuerte con la acidez volátil y los cloruros, con valores de 0.36 y 0.38 respectivamente. Y una relación inversa con las variables color y total dióxido de azufre valorada respectivamente en -0.46 y -0.36.

Dataset original winsorizado



	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
fixed_acidity	0.26	0.16	0.50	-0.12	0.12	0.08	-0.12
volatile_acidity	0.36	0.02	-0.25	-0.10	0.34	0.37	-0.01
citric_acid	-0.12	0.14	0.60	0.09	-0.32	-0.22	-0.17
residual_sugar	-0.19	0.40	-0.05	0.21	0.55	-0.27	0.23
chlorides	0.38	0.19	0.01	0.03	-0.12	0.18	-0.40
free_sulfur_dioxide	-0.31	0.24	-0.10	0.30	-0.06	0.51	-0.35
total_sulfur_dioxide	-0.37	0.27	-0.10	0.13	-0.13	0.30	-0.02
density	0.23	0.50	0.00	0.20	0.17	-0.24	0.00
pH	0.16	-0.18	-0.42	0.46	-0.27	-0.43	-0.24
sulphates	0.27	-0.00	0.16	0.54	-0.27	0.29	0.65
alcohol	-0.07	-0.49	0.18	0.09	0.27	0.16	0.05
quality	-0.11	-0.31	0.25	0.50	0.41	-0.02	-0.31
color	-0.46	0.01	-0.02	-0.10	-0.07	-0.06	0.17

Para este dataset también hay 7 componentes principales, el primero de ellos explica el 32.2% de la varianza, y disminuye hasta llegar a el último, que explica el 4.04%. Las variables más influyentes dentro del componente 1 son la acidez volátil, el color, cloruros y el total de dióxido de azufre, con valores respectivos de 0.36, -0.46 y 0.38 y -0.37.

Algoritmos de pruebas:

K-Means: Se aplicaron los algoritmos de K-means, considerando diferentes datasets, con el fin de identificar el adecuado para construir el resto de los modelos. Los datasets considerados fueron:

1. Dataset original quantiles: se realiza un tratamiento de datos con rango intercuartílico, esta técnica consiste en reemplazar los atípicos por el limite superior o inferior.
2. Dataset original winsorizado: se realiza un tratamiento de datos atípicos realizando una transformación, llevando muchos atípicos a datos más cercanos al conjunto.
3. Dataset reducido quantiles: se reduce la dimensionalidad del dataset original quantiles con la técnica PCA.
4. Dataset reducido winsorizado: se reduce la dimensionalidad del dataset original winsorizado con la técnica PCA.

Los resultados de estos cuatro datasets son los siguientes:

Métrica / Dataset	Original quantiles	Original winsorizado	Reducido quiantiles	Reducido winsorizado
Inercia	41,050.00	40,446.86	32,883.27	32,476.17
Coeficiente de silueta	0.2407	0.2429	0.2760	0.2785
Coeficiente de Calinski	1,820.00	1,887.26	2,265.38	2,343.70

Luego de comparar el rendimiento del modelo K-means para cada uno de los dataset anteriores, se encontró que el **Dataset reducido winsorizado,** arroja mejores resultados.

Selección de variable: se comparan los resultados del modelo K-means para diferentes modificaciones en las variables del Dataset original winsorizado. El conjunto con mejor desempeño es el **Dataset original winsorizado sin la variable calidad**. De aquí en adelante el conjunto de datos utilizado no considerará la variable calidad y recibirá por nombre **nuevo dataset** (winsorizado sin la variable calidad).

Los resultados de estos datasets son:

Métrica / Dataset	Original winsorizado sin calidad	Original winsorizado sin color	Original winsorizado s/calidad y s/color
Inercia	35,637.43	40,111.76	31,538.52
Coeficiente de silueta	0.2619	0.2276	0.2468
Coeficiente de Calinski	2,103.87	1,572.64	1,515.96

Los resultados podemos evidenciar que el dataset sin las variables calidad y color, es mejor respecto a la inercia, pero el dataset sin la variable calidad se destaca por mejores coeficientes de silueta y de Calinski.

Se realizó un K-means considerando el nuevo dataset con reducción de la dimensionalidad para comparar los resultados con los modelos anteriores, y se verificó que las métricas mejoraron. I**nercia 27.316,75, coeficiente de silueta 0.3076 y calinski 2730.69.**

Otros algoritmos de clusterización:

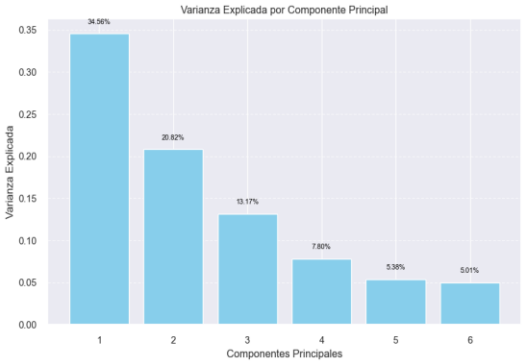
Clúster jerárquico: Se aplicaron los algoritmos de hierarchical clustering, considerando dos datasets, con el fin de explorar otras opciones para la elección del mejor modelo, los resultados son los siguientes:

Métrica / Dataset	Nuevo dataset sin reducir	Nuevo dataset reducido PCA
Coeficiente de silueta	0.3375	0.3788
Coeficiente de Calinski	2,352.51	2,852.28

De acuerdo con los resultados en la tabla anterior, el modelo con mejores métricas es el dataset reducido PCA, el cual consta de 6 componentes principales, y ha capturado eficazmente la estructura subyacente de los datos, permitiendo una mejor agrupación, inclusive si se compara con todos los modelos de K-means realizados anteriormente.

Las componentes principales y sus pesos por variable se muestran a continuación:

El primero de ellos explica el 34.56% de la varianza, y disminuye hasta llegar a el último, que explica el 5.01%. Las variables más influyentes dentro del componente 1 son el color, cloruros y el total de dióxido de azufre, con valores respectivos de -0.46, 0.38 y -0.38.



	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
fixed_acidity	0.26	0.23	0.48	-0.11	-0.08	0.07
volatile_acidity	0.36	0.02	-0.25	-0.27	-0.30	0.36
citric_acid	-0.12	0.19	0.59	0.31	0.13	-0.21
residual_sugar	-0.21	0.42	-0.15	-0.06	-0.62	-0.31
chlorides	0.38	0.22	-0.04	-0.01	0.31	0.20
free_sulfur_dioxide	-0.32	0.23	-0.17	0.26	0.02	0.52
total_sulfur_dioxide	-0.38	0.25	-0.14	0.14	0.07	0.30
density	0.21	0.54	-0.11	0.03	-0.12	-0.24
pH	0.17	-0.19	-0.46	0.50	0.10	-0.42
sulphates	0.28	0.06	0.09	0.68	-0.22	0.26
alcohol	-0.04	-0.48	0.23	0.13	-0.57	0.12
color	-0.46	-0.03	0.01	-0.05	0.06	-0.06

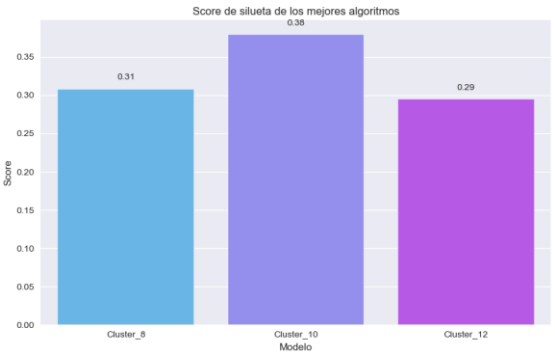
DBSCAN: Se aplicaron los algoritmos de DBSCAN, se utilizan los mismos dos datasets del anterior y los resultados son los siguientes:

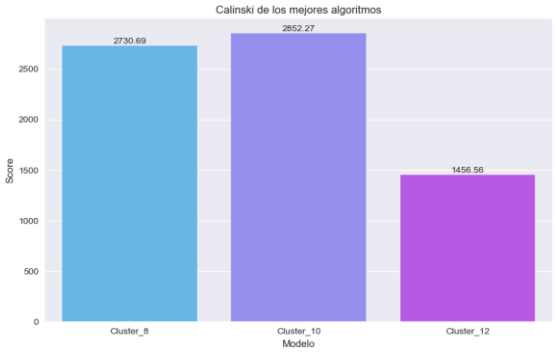
Métrica / Dataset	Nuevo dataset sin reducir	Nuevo dataset reducido PCA
Coeficiente de silueta	0.2786	0.2946
Coeficiente de Calinski	1,194.09	1,456.56
Outliers	50	107

La mayoría de las muestras los dos datasets han sido agrupadas en dos clústeres principales (Clúster 1 y Clúster 0) y una pequeña cantidad de atípicos. En ambos datasets las métricas sugieren una separación y cohesión razonable, siendo mejores las del dataset reducido con coeficiente de silueta de 0.2946 y coeficiente Calinski igual a 1.456,56.

Modelo de mezclas gaussianas: Se realizaron modelos de mezclas gaussianas para los datasets sin reducir y reducidos con PCA. Obteniendo resultados muy desfavorables, en comparación con los modelos anteriores. Posiblemente, se debe a que algunas variables no siguen una distribución normal.

Comparación de los mejores modelos en cada algoritmo



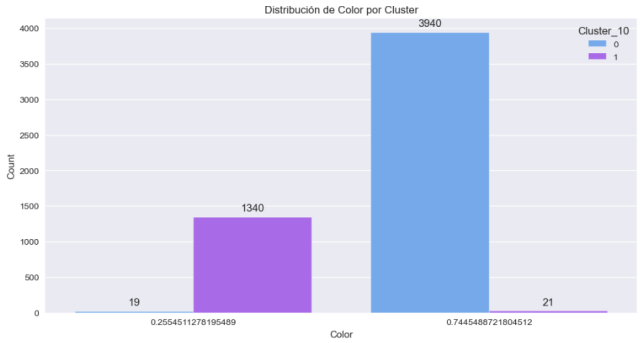


Las dos gráficas anteriores muestran los mejores resultados para las métricas del coeficiente de silueta y el coeficiente de Calinski por cada algoritmo aplicado (K-means, clúster jerárquico, DBSCAN), donde el dataset tratado en todos los casos fue el reducido con PCA. De la gráfica se verifica que el algoritmo que mejor agrupó los datos fue el hierarchical clustering.

Análisis del mejor algoritmo

El clúster generado al aplicar el algoritmo jerárquico reducido PCA a los datos, genera dos clústeres (0 y 1), con 3.959 y 1.361 observaciones respectivamente.

Según la variable color, los clústeres formados tienen el comportamiento que se muestra en la siguiente gráfica:



Podemos observar que los clústeres formados se ven influenciados fuertemente según el color del vino, en el clúster “0” la mayoría de los vinos son de color blanco, 3.940 y sólo 19 vinos son rojos, mientras que en el clúster “1”, hay 1.340 vinos rojos y 21 vinos blancos.

Al analizar el resto de las variables según los clústeres formados, tenemos lo siguiente:

Clúster 0 Vinos blancos	**Clúster 1 Vinos tintos**
3.959 observaciones	1.361 observaciones
3.940 vinos blancos	1.340 vinos rojos
< densidad	> densidad
> azúcar residual	< azúcar residual
< cloruros, sulfatos, pH	> cloruros, sulfatos, pH
< ácidez (fija, volatil y cítrica)	> ácidez (fija, volatil y cítrica)
frec. altas entre 9 y 10 ° alcohol	frec. altas entre 10 y 11 ° alcohol

Conclusión

El análisis revela que la mayoría de los vinos blancos se agrupan en un clúster distinto a los vinos tintos. Esto sugiere que el color es una característica significativa en la diferenciación de los vinos, adicionalmente, los clústeres formados muestran diferencias claras en varias características químicas y sensoriales entre los vinos blancos y tintos. Por ejemplo, los vinos blancos tienden a tener menor densidad, mayor contenido de azúcar residual y menor acidez en comparación con los vinos tintos, que son más densos, tienen mayor acidez y menor contenido de azúcar residual, sin embargo, el contenido de alcohol no se ve afectado por el color, y sigue una distribución similar de los datos en los dos clústeres formados.

Recomendación

Dado el impacto significativo del color en la clusterización de vinos, se recomienda tener en cuenta esta característica al clasificar los vinos en grupos o clústeres. Esto puede ayudar a obtener agrupaciones más coherentes y significativas desde el punto de vista químico y sensorial.