Análisis de Clusteres en Vinos mediante PCA y K-Means

Delfina Solís, Thiago Corte

26 de diciembre de 2025

Resumen

En el siguiente informe se presenta el estudio y la aplicacion de tecnicas de reduccion y clustering para identificar grupos de vinos para el dataset presentado. Se utilizo el metodo de PCA para reducir la dimensionalidad, para luego aplicar el metodo de K-means sobre las componentes principales. Para poder obtener la cantidad optima de clusters se utilizaron los metodos del codo (inercia) y silhouette score, sumado a una inspeccion visual de la separacion proyectada, por lo que se concluyo que el numero optimo era de dos clusters. La interpretación de los clusters se realizó relacionando los promedios de las componentes principales por cluster con las variables originales más relevantes, identificando diferencias claras en niveles de dióxido de azufre, densidad y contenido alcohólico entre los grupos.

1. Introducción

En este análisis se busca identificar agrupaciones de vinos a partir de algunas de sus propiedades utilizando técnicas de aprendizaje no supervisado. Se trabajó sobre un set de datos públicos de vinos, *Wine Data Set*, obtenido del *UCI Machine Learning Repository*.

El dataset cuenta con características como cantidad de alcohol, sulfatos, ph, acidez, entre otras. Se utilizó **PCA** para reducir la dimensionalidad del conjunto de datos y luego mediante el metodo de **K-means** se realizo la clusterización con el objetivo de encontrar grupos de vinos similares.

2. Preprocesamiento y PCA

El dataset utilizado contiene 6497 filas y 11 columnas con variables numericas, dentro de las cuales esta la cantidad de alcohol, el ph, los sulfatos, etc. Dado que K-Means utiliza la distancia euclidiana como métrica principal, es importante reducir el ruido y la correlación entre los atributos. Luego de que se hayan escalado las variables mediante z-score, se realizo el analisis de componentes principales (PCA) para reducir las dimensiones manteniendo la mayor cantidad de informacion. El análisis de la varianza explicada mostró que con las primeras 9 componentes principales se conserva aproximadamente el 95 % de la varianza total del dataset. Teniendo en cuenta los resultados del PCA se transformaron los datos originales a este nuevo espacio reducido, sobre el cual ahora si, se aplico el algoritmo de K-means.

3. Determinación del número de clusters

Para la determinacion de optima de la cantidad optima de clusters a utilizar en el metodo de K-means se utilizaron los metodos de *inercia* y de *silhouette score*, lo cual se puede ver en los siguientes graficos.

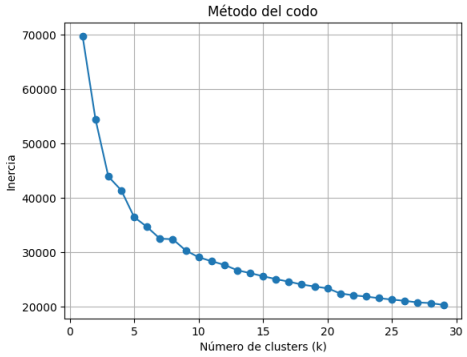

Figura 1: El gráfico muestra la inercia en función del número de clusters. El çodo"sugiere el punto a partir del cual añadir más clusters no mejora significativamente la compactación de los grupos.

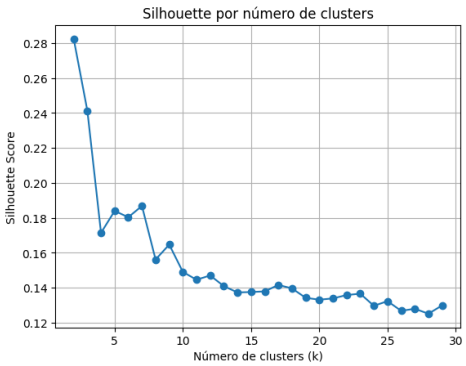

Figura 2: Evaluación del número óptimo de clusters mediante el silhouette score. El número óptimo de clusters se asocia con el valor máximo del coeficiente.

Como ya se menciono, los graficos fueron generados a partir de las componentes principales obtenidas tras la reduccion de dimensionalidad del PCA. Se puede observar en la 1 que para el metodo basado en la inercia el valor optimo de los clusters es de 3, mientras que en el grafico 2 el metodo de silhouette score sugiere que el valor optimo es de 2. Para decidir que clusterizacion utilizar se hizo una inspeccion visual de cada una de las clusterizaciones diferentes, con lo que se decidio que el valor optimo deberia ser el de 2 ya que ofrece una agrupacion mas clara y consistente. En el siguiente grafico 3 se puede observar el resultado de la clusterizacion mas optima obtenida utilizado el metodo de K-means para 2 clusters.

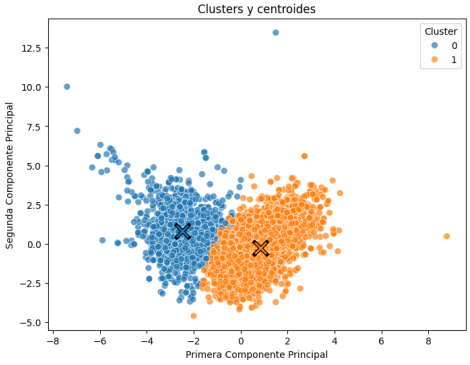

Figura 3: Clusters obtenidos por K-means proyectados sobre las dos primeras componentes principales. Los puntos están coloreados según su cluster y las cruces indican los centroides.

4. Clustering y Resultados

Una vez obtenidos los clusters, se calculó para cada uno el valor promedio de cada componente principal, lo que se puede ver en las cruces que se encuentran en los centroides de cada uno. Luego, se analizaron los coeficientes obtenidos para cada componente princiapal con respecto a las variables originales, para tener una relacion entre ambos. Se aplicaron umbrales tanto para el promedio de la componente (umbral de 0.3) como para la magnitud de los loadings (umbral de 0.4), puediendo asi identificar las variables originales mas relevantes de cada cluster, las cuales caracterizan de mejor manera a cada uno de estos. Se observo que para el cluster 0 se distingue por valores bajos en *free sulfur dioxide, total sulfur dioxide* y *densidad*, y un contenido *alcoholico* alto. Para el cluster 1 se obtienen altos niveles de *free sulfur dioxide, total sulfur dioxide* y *densidad*, al igual que en el cluster 0, pero

por otro lado, se tiene un menor nivel de *alcohol*, y diferencias significantes en *fixed acidity, citric acid* y *pH*.

5. Conclusiones

El uso combinado de PCA y K-means permitió agrupar eficazmente los vinos en dos clusters con características diferenciadas. La reducción de dimensionalidad facilitó el manejo de los datos permitiendo a la clusterización identificar grupos con propiedades claramente contrastantes, especialmente en relación con el dióxido de azufre, densidad y alcohol. Este enfoque demuestra la utilidad de combinar técnicas de reducción de dimensionalidad con clustering para el análisis exploratorio de los datos.

En conclusion, se puede decir que los metodos de PCA y K-means permitieron agrupar de la forma mas efectiva y minimizando la cantidad de dimensiones a los vinos utilizando diferentes caracteristicas de los mismos. La obtencion de los mismos se facilito debido a la reduccion de la dimensionalidad permitiendo representar la mayor cantidad posible de datos con la menor cantidad de variables