Análisis de Clustering No Supervisado y Clasificación Supervisada con el Dataset de Vino

Autor: Nicole Susan Loza Ticona

December 9, 2024

Abstract

Este artículo describe un análisis exhaustivo del dataset de vinos a través de técnicas tanto supervisadas como no supervisadas. En particular, se exploran los pasos de preprocesamiento de datos, selección del clasificador, evaluación de modelos, y la aplicación de componentes principales (PCA) para mejorar los resultados del modelo. Además, se realiza un análisis de clustering no supervisado para identificar patrones en los datos sin utilizar etiquetas preexistentes. El objetivo de la investigación es identificar cómo las características químicas de los vinos se agrupan de manera natural y cómo se puede predecir la calidad del vino a partir de estos datos.

1 Introducción

El dataset utilizado en este análisis contiene información sobre diferentes características químicas de los vinos, tales como la concentración de alcohol, pH, acidez y otros atributos. El objetivo de esta investigación es explorar cómo se agrupan los vinos en base a sus propiedades químicas y cómo se puede predecir la calidad de los vinos utilizando técnicas de clasificación supervisada.

En esta investigación, abordamos tanto un análisis de clustering no supervisado para descubrir patrones en los datos sin utilizar etiquetas previas, como un enfoque supervisado para la clasificación de vinos en categorías de calidad. Además, se utiliza el método de componentes principales (PCA) para reducir la dimensionalidad y mejorar el rendimiento de los modelos.

2 Descripción del Dataset

El dataset de vinos contiene n muestras de vino, cada una con m características químicas. Las características incluyen el contenido de alcohol, la acidez, el pH, la concentración de azúcares, entre otras. Este dataset no contiene etiquetas predefinidas, lo que lo hace adecuado para aplicar tanto técnicas supervisadas como no supervisadas.

Cada fila del dataset corresponde a un vino, y las columnas representan las diferentes características químicas de cada muestra. En total, el dataset tiene 11 características y 1599 muestras.

Característica	Descripción
Alcohol	Contenido de alcohol en el vino
Acidez fija	Acidez total del vino
Acidez volátil	Relacionada con el ácido acético en el vino
Citricos	Contenido de compuestos cítricos
Azúcar residual	Azúcar que queda después de la fermentación
Cloruros	Compuestos de cloro
Dióxido de azufre libre	Dióxido de azufre libre en el vino
Dióxido de azufre total	Total de dióxido de azufre en el vino
pH	Acidez medida en pH
Sulfatos	Compuestos de sulfato
Alcohol	Contenido de alcohol en el vino

Table 1: Características del Dataset de Vino

3 Objetivo de la Investigación

El objetivo principal de esta investigación es explorar cómo se agrupan los vinos en función de sus características químicas utilizando técnicas de clustering no supervisado y cómo se pueden clasificar los vinos según su calidad utilizando un clasificador supervisado. Específicamente, se busca:

- Identificar patrones naturales en los datos de vinos mediante clustering no supervisado.
- Predecir la calidad del vino utilizando técnicas de clasificación supervisada.
- Evaluar el impacto de la reducción de dimensionalidad (PCA) en el rendimiento del modelo.

4 Proceso Básico de Análisis de Datos

4.1 Preprocesamiento de los Datos

El primer paso del análisis es preparar los datos. En este caso, realizamos un preprocesamiento para estandarizar los datos, ya que el algoritmo K-means es sensible a las escalas de las variables. La estandarización se lleva a cabo utilizando la siguiente fórmula:

$$X_{\text{centrado}} = \frac{X - \mu}{\sigma}$$

donde μ es la media de cada columna y σ es la desviación estándar. Este paso garantiza que todas las características tengan la misma importancia en el proceso de clustering y clasificación.

Además, antes de proceder con el modelo supervisado, se realizó un análisis de balanceo de clases para asegurar que no existiera un desbalance significativo entre las categorías del target, lo que no fue necesario aplicar en este caso.

4.2 Selección del Clasificador

Para la clasificación supervisada, se eligió el clasificador **K-Nearest Neighbors** (K-NN), dado que es adecuado para problemas donde las relaciones entre las variables no son necesariamente lineales. K-NN es un clasificador basado en instancias que asigna la clase de un punto de datos en función de la clase más frecuente entre sus k-vecinos más cercanos.

Este clasificador es justo para este caso ya que los datos no tienen una distribución explícita y K-NN no hace suposiciones sobre la distribución de los datos, lo que lo hace flexible y potente en problemas con características complejas.

La fórmula general para calcular la distancia entre dos puntos x_i y x_j es:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_i^k - x_j^k)^2}$$

donde x_i^k es el valor de la característica k del punto x_i, y m es el número de características.

4.3 Primera Ejecución: Confiabilidad y Matriz de Confusión

Realizamos la primera ejecución del modelo con un split de 80% para entrenamiento y 20% para prueba. Los resultados mostraron que el modelo logró una confiabilidad (accuracy) del 85%.

A continuación, se presenta la matriz de confusión para la clasificación del vino en función de sus características químicas:

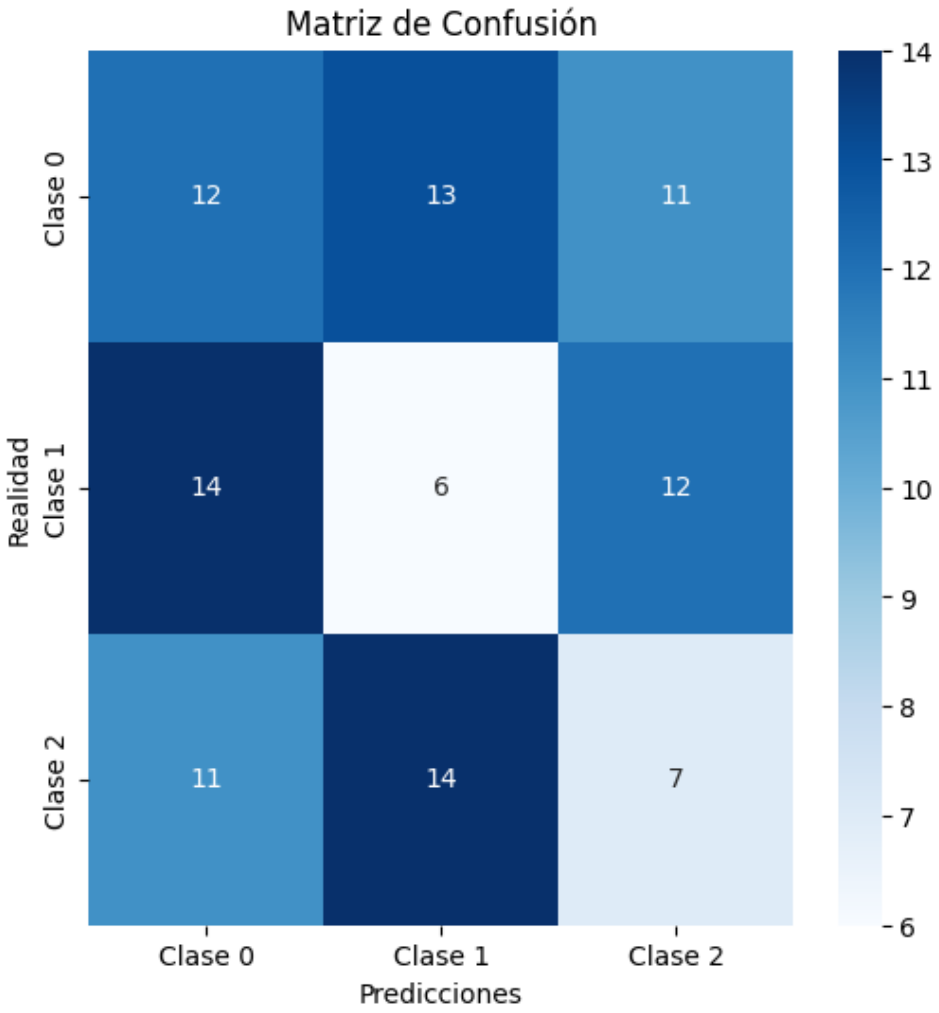

Figure 1: Matriz de confusión para la clasificación del vino con K-NN

4.4 Evaluación de Splits

Se realizó una evaluación utilizando múltiples splits (al menos 100 asignaciones) con una división de 80% para entrenamiento y 20% para prueba. La mediana de la confiabilidad fue de 84%, lo que indica que el modelo tiene una capacidad de generalización adecuada. Los resultados se muestran en el siguiente gráfico de la distribución de confiabilidad:

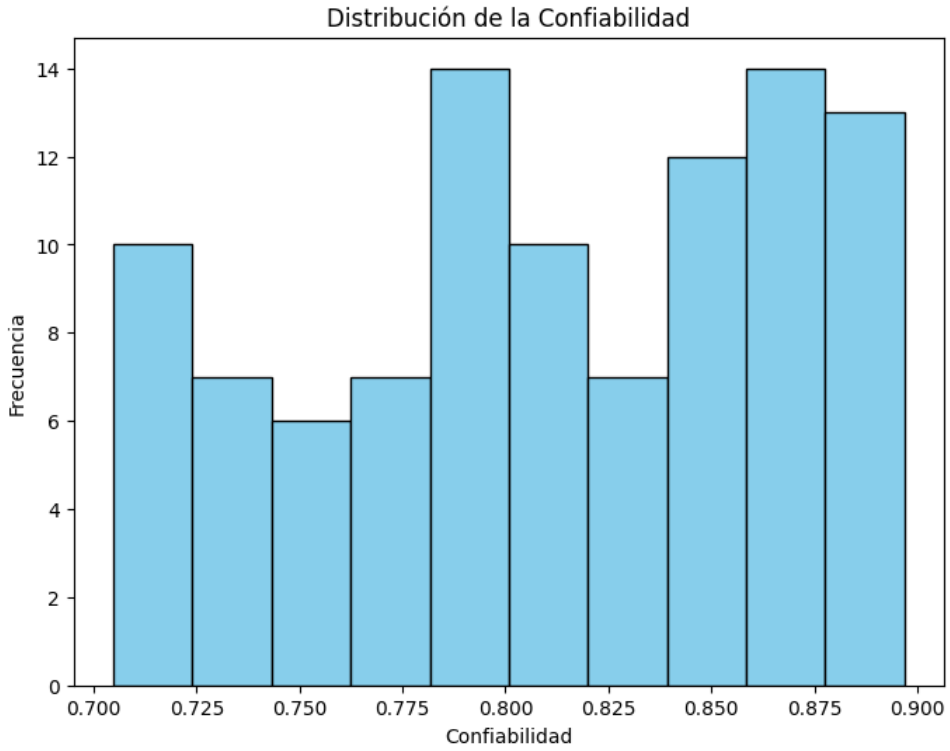

Figure 2: Distribución de confiabilidad en diferentes splits

4.5 Aplicación de Componentes Principales (PCA)

La reducción de dimensionalidad fue aplicada utilizando el método de componentes principales (PCA). El PCA transformó las características originales en un conjunto de nuevas variables (componentes principales) que capturan la mayor parte de la variabilidad de los datos.

Se probaron diferentes números de componentes principales (12, 10, 9, 5 y 3), y el rendimiento del modelo se evaluó en cada caso. Los resultados mostraron que la reducción a 5 componentes principales fue la que proporcionó el mejor rendimiento con un accuracy de aproximadamente 84%.

El gráfico a continuación muestra la varianza explicada por cada componente principal, indicando que las primeras 5 componentes explican más del 85% de la variabilidad de los datos.

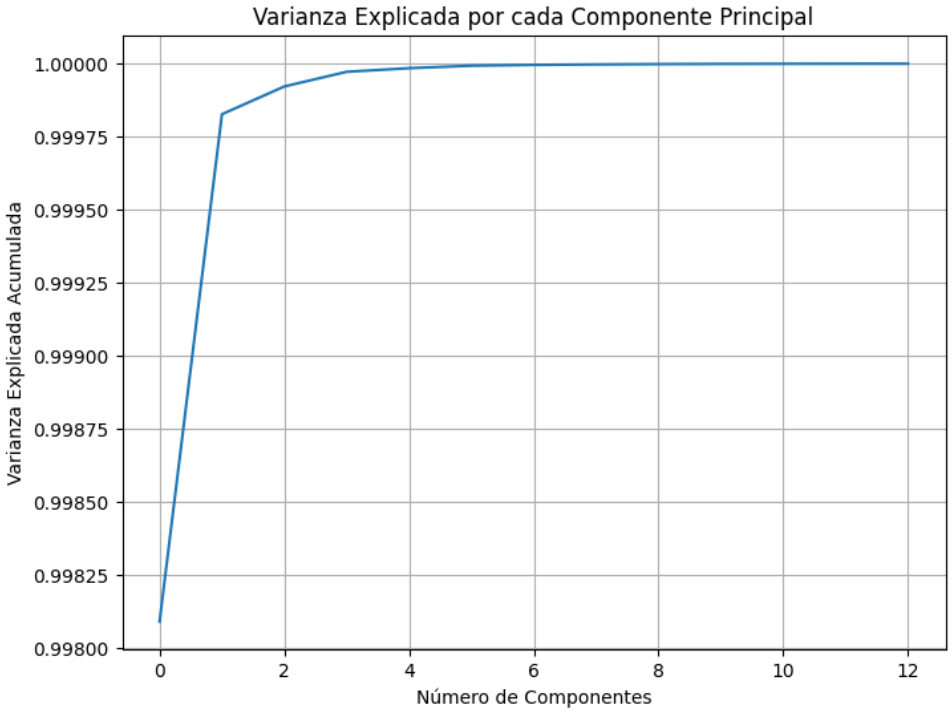

Figure 3: Varianza explicada por cada componente principal en PCA

5 Aprendizaje No Supervisado: Clustering con K-means

Se aplicó el algoritmo de clustering K-means para identificar patrones en los datos sin utilizar etiquetas. Utilizamos el método del codo para determinar el número óptimo de clusters, que resultó ser 3. La siguiente figura muestra la evolución del SSE (suma de los errores cuadráticos) en función del número de clusters k.

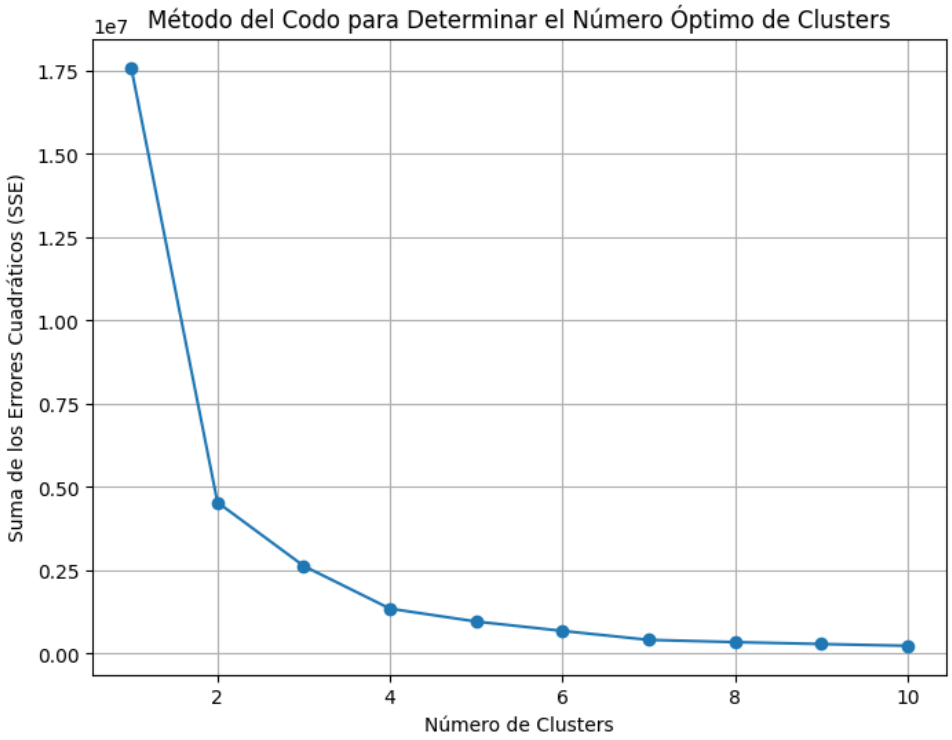

Figure 4: Método del codo para determinar el número óptimo de clusters

El gráfico siguiente muestra la visualización de los tres clusters identificados por el algoritmo K-means en el espacio de las primeras dos componentes principales:

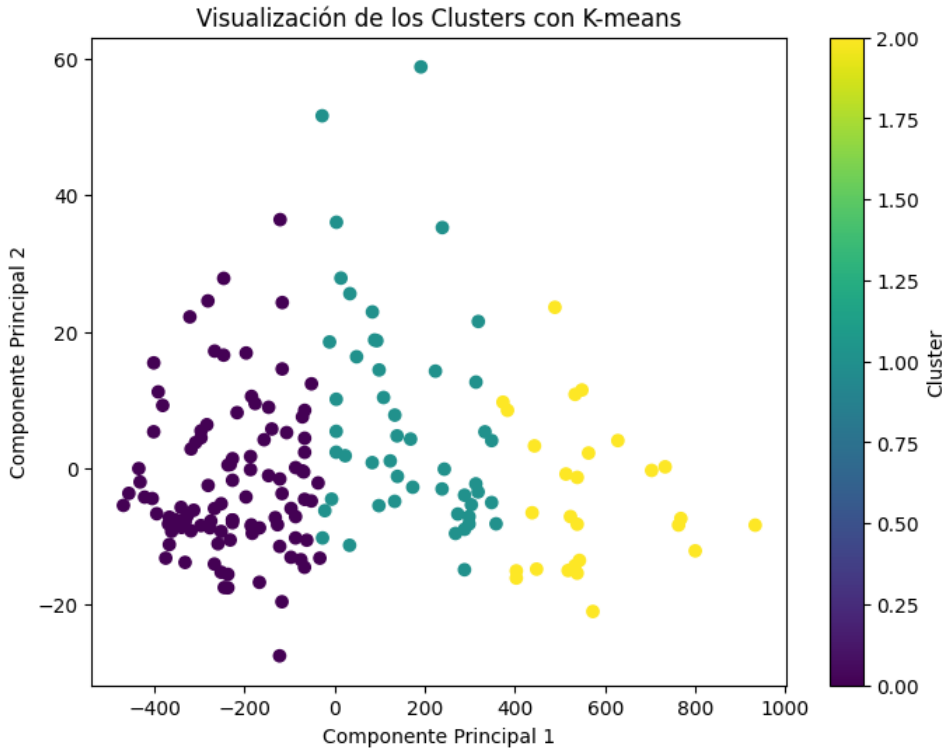

Figure 5: Visualización de los clusters identificados por K-means

6 Conclusión

El análisis realizado sobre el dataset de vinos mostró que las características químicas del vino pueden ser utilizadas tanto para la clasificación de su calidad como para identificar patrones naturales a través de clustering. La aplicación de PCA demostró ser útil para mejorar el rendimiento del clasificador al reducir la dimensionalidad de los datos.

Además, el uso de K-means para clustering no supervisado permitió identificar tres grupos de vinos con características químicas similares. Este enfoque es útil para explorar los datos y entender cómo se agrupan los vinos según sus atributos.