Análisis de Clustering No Supervisado con K-means para el Dataset de Vino

Autor: Nicole Susan Loza Ticona

December 9, 2024

Abstract

El presente artículo presenta un análisis de clustering no supervisado utilizando el algoritmo K-means sobre un dataset relacionado con las características de vinos. El objetivo es explorar patrones inherentes en los datos sin recurrir a una variable objetivo etiquetada. Se utiliza el algoritmo K-means para agrupar los datos en clusters y se analiza el rendimiento del modelo con la ayuda del método del codo para la elección del número óptimo de clusters. A través de este análisis, se exploran las relaciones entre las variables del dataset y cómo se pueden identificar subgrupos dentro de los mismos.

1 Introducción

En el ámbito del análisis de datos, uno de los objetivos comunes es encontrar patrones y estructuras dentro de los datos sin la necesidad de utilizar etiquetas predefinidas. Este tipo de problemas corresponde al aprendizaje no supervisado, una de las ramas fundamentales del aprendizaje automático. Un ejemplo clásico de aprendizaje no supervisado es el *clustering*, que busca agrupar datos similares en subconjuntos o clusters.

El algoritmo de *K-means clustering* es uno de los métodos más populares para realizar agrupamientos. K-means tiene como objetivo dividir un conjunto de datos en k clusters en los cuales los datos dentro de cada cluster son lo más similares posible, y los datos de diferentes clusters son lo más distintos posible.

Este artículo tiene como objetivo aplicar el algoritmo K-means sobre un dataset relacionado con las características químicas de vinos para explorar cómo se agrupan en base a sus características, sin hacer uso de etiquetas. Además, se utilizará el método del codo para determinar el número óptimo de clusters.

2 Metodología

2.1 Descripción del Dataset

El dataset utilizado en este análisis contiene diversas características de vinos, incluyendo atributos químicos como el pH, la acidez, el alcohol, entre otros. Este conjunto de datos no tiene etiquetas (no es supervisado), por lo que la tarea consiste en identificar grupos de vinos similares basados únicamente en sus características. El número de características es

relativamente alto, lo que sugiere la posibilidad de realizar una reducción de dimensionalidad, pero para este análisis nos enfocaremos en el uso directo del algoritmo K-means.

El dataset contiene n muestras y m características, donde n es el número de vinos y m es el número de atributos medidos. Cada registro en el dataset corresponde a un vino con sus valores en cada una de las características, como se muestra a continuación:

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \dots & x_1^m \\ x_2^1 & x_2^2 & \dots & x_2^m \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^m \end{bmatrix}$$

2.2 Preprocesamiento de los Datos

El primer paso en la aplicación de K-means es la estandarización de los datos. K-means es sensible a la escala de las variables, por lo que es necesario estandarizar las características para que todas tengan la misma importancia. La estandarización se realiza restando la media de cada característica y dividiendo por su desviación estándar:

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

donde μ es la media de cada columna y σ es la desviación estándar de cada columna.

2.3 Algoritmo K-means

El algoritmo K-means sigue un proceso iterativo en el cual se buscan los k centroides de los clusters que minimizan la suma de las distancias cuadradas entre los puntos de datos y los centroides correspondientes. El proceso general es el siguiente:

1. Inicializar k centroides de manera aleatoria.
2. Asignar cada punto de datos al centroide más cercano.
3. Recalcular los centroides de los clusters basándose en los puntos asignados.
4. Repetir los pasos 2 y 3 hasta que los centroides ya no cambien significativamente.

El resultado final es un conjunto de k clusters, cada uno de los cuales contiene los puntos de datos que son más cercanos a su respectivo centroide.

2.4 Determinación del Número Óptimo de Clusters

Para determinar el número óptimo de clusters k, se utiliza el *método del codo*, que consiste en graficar la suma de los errores cuadráticos (SSE) en función de k y observar el punto donde la tasa de disminución de SSE comienza a ser más lenta. Este punto indica el valor óptimo de k.

$$SSE(k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \|x_i - c_j\|^2$$

donde x_i es un punto de datos y c_j es el centroide del cluster al que pertenece.

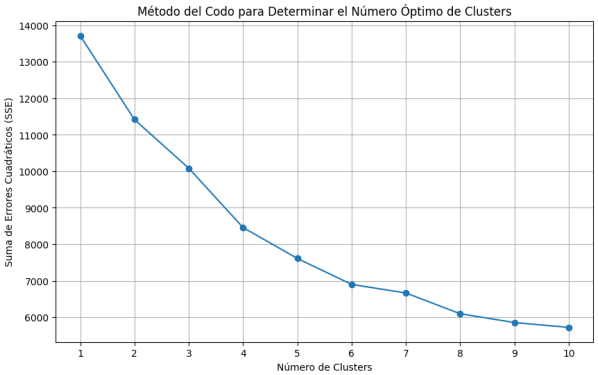

Figure 1: Método del codo para determinar el número óptimo de clusters

3 Resultados

3.1 Aplicación del Algoritmo K-means

Después de aplicar el algoritmo K-means con un valor de $k = 3$, se obtienen tres clusters. Los centroides de estos clusters se encuentran en las siguientes posiciones en el espacio de características:

$$\text{Centroides} = \begin{bmatrix} c_1^1 & c_1^2 & \dots & c_1^m \\ c_2^1 & c_2^2 & \dots & c_2^m \\ c_3^1 & c_3^2 & \dots & c_3^m \end{bmatrix}$$

Esto indica que los datos se agrupan en tres categorías, y cada una de ellas tiene sus propias características promedio.

3.2 Visualización de los Clusters

La visualización de los clusters se realiza utilizando un gráfico de dispersión, donde se representan los puntos de datos de acuerdo con sus características, y los puntos se colorean según el cluster al que pertenecen. La siguiente figura muestra la distribución de los datos a través de los tres clusters identificados:

En este gráfico se puede observar cómo los puntos de datos se agrupan en torno a los centroides de cada cluster.

3.3 Evaluación del Método del Codo

El método del codo indica que el número óptimo de clusters es $k = 3$, ya que después de ese valor, la disminución de SSE se estabiliza y se vuelve más lenta. La siguiente figura muestra la evolución de SSE con respecto a k:

4 Discusión

El análisis realizado con K-means ha sido exitoso en la identificación de tres grupos principales de vinos basados en sus características químicas. La aplicación del método del codo ha sido crucial para determinar que $k = 3$ es el número óptimo de clusters,

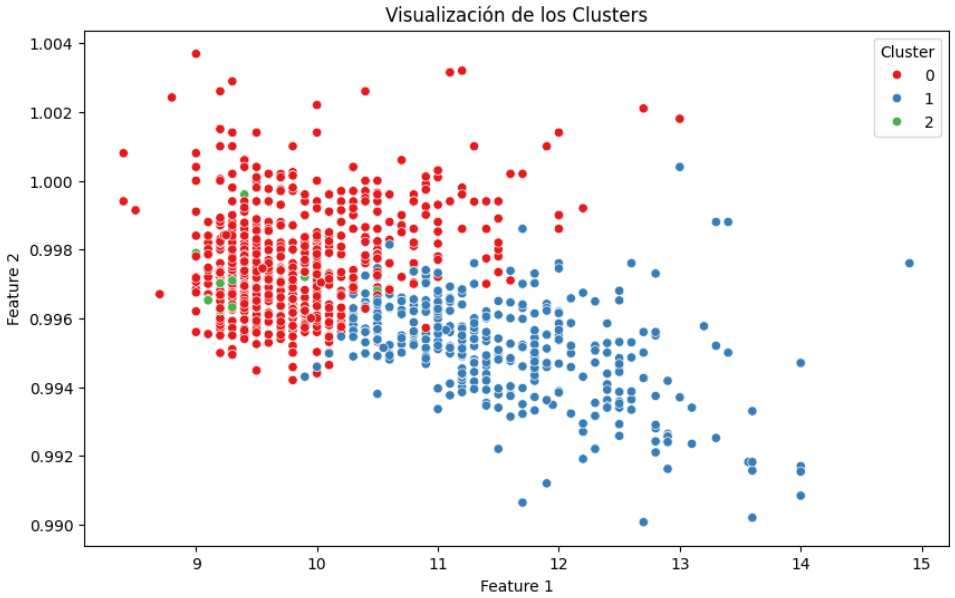

Figure 2: Visualización de los clusters generados por el algoritmo K-means

lo que sugiere que los vinos pueden clasificarse en tres categorías distintas según sus características.

Sin embargo, es importante señalar que la elección del número de clusters depende de la naturaleza del dataset y de los resultados obtenidos mediante métodos de validación adicionales. En el caso de este dataset, el K-means ha demostrado ser una herramienta efectiva para descubrir patrones ocultos en los datos sin la necesidad de etiquetas preexistentes.

5 Conclusión

El aprendizaje no supervisado es una herramienta poderosa para explorar y analizar datos sin la necesidad de etiquetas. El uso de K-means para realizar clustering en el dataset de vinos ha permitido identificar grupos con características similares. El método del codo ha sido fundamental para determinar el número adecuado de clusters, lo que facilita la interpretación de los resultados. Este tipo de análisis puede ser útil en diversas áreas, como la segmentación de clientes, la biología, y la exploración de datos en general.