Proyecto Final - SIA

K-Medias

Asignatura:
Sistemas Inteligentes Artificiales
Profesor:
Fernández, Agustín

Escrito por:
Fernández, Florencia
Ventura, Gino

Fecha de entrega: 05/07/2024

UBP UNIVERSIDAD Blas Pascal

CONTENIDO

INTRODUCCIÓN

QUÉ ES K-MEDIAS

Es un método de agrupamiento, que tiene como objetivo organizar los datos de manera que los puntos dentro de cada grupo, llamados clústeres, sean lo más similares posible entre sí. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster. Algunas características principales son:

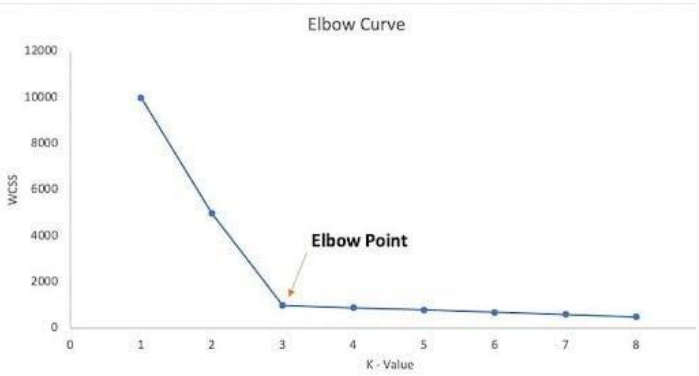
- Permite solo datos numéricos.
- Se suele usar la distancia cuadrática para ajustar la distancia de los centroides durante el proceso iterativo.
- Requiere normalización.
- Es no supervisado, es decir, no requiere etiquetas predefinidas para entrenar el modelo, ya que agrupa los datos basándose en su similitud.
- Depende mucho de las semillas.
- Basado en los grafos de Voronoi, es una manera de dividir un espacio en regiones basadas en la distancia a un conjunto especifico de puntos.

PASOS DEL ALGORITMO

1. ***Inicialización:*** selecciona ***k*** centroides iniciales, que pueden ser puntos de datos aleatorios del dataset o puntos generados de alguna manera específica entre los datos de entrada.
2. ***Asignación de clústeres:*** asigna cada punto de datos al centroide más cercano, formando *k* clústeres.
3. ***Actualización:*** se calcula la nueva posición de cada centroide, desplazando las semillas a los centros de cada grupo.
4. ***Repetición:*** se repiten los pasos de asignación de clústeres y actualización de centroides hasta que los centroides ya no cambien significativamente entre iteraciones consecutivas, o hasta que alcance un número máximo de iteraciones.

MÉTODO DEL CODO

El algoritmo de K-Medias depende de encontrar la cantidad de grupos adecuada para agrupar los datos. El método del codo es un método gráfico para encontrar el valor de ***k*** óptimo. Muestra los valores de suma de cuadrados dentro del grupo (***WCSS, Within Clúster Sum of Squares***) en el eje *y* correspondientes a los diferentes valores de ***k*** *(*en el eje *x)*. El valor óptimo de ***k*** es el punto en el que el gráfico forma un “codo”.



INTRODUCCIÓN AL PROYECTO

El laboratorio de Inteligencia Artificial de la empresa Ultralistic nos ha solicitado realizar nuestra propia implementación del algoritmo K-Medias, ya que posee un dataset sobre la calidad vitivinícola y pretende analizar cuáles son las agrupaciones relevantes y "naturales" del mismo, más allá de las clases presentes en dicho dataset, y para ello desea utilizar el mencionado algoritmo.

REQUERIMIENTOS

Para la implementación se establecieron los siguientes requerimientos:

- Se puede escribir en el lenguaje de programación preferido por el grupo de trabajo.
- No se admiten implementaciones que posean la utilización de librerías o frameworks (o cualquier cosa parecida) que ya contengan el algoritmo solicitado.
- Debe realizarse una implementación del algoritmo sin "vectorizar" y otra "vectorizada".
- Esta permitido utilizar librerías o frameworks (o cualquier cosa parecida) para soporte de los cálculos matemáticos, como pueden ser las operaciones de matrices, cálculos estadísticos, etc.
- Las implementaciones deben ser flexibles, es decir, que se debe poder determinar la cantidad de clústeres que se desean y además se debe poder manejar cualquier número de ejemplos y atributos del dataset correspondiente.
- Los tipos de datos que deben soportar las implementaciones propias de K-Medias, por supuesto que deben ser normalizados de alguna forma para evitar inconvenientes.
- Se debe realizar una comparativa entre las implementaciones propias y la implementación de alguna librería o framework disponible en el mercado (dicha librería o framework puede ser que se encuentre escrita en otro lenguaje de programación).

Teniendo en cuenta dichos requerimientos se debe lograr:

1. Construir las implementaciones de K-Medias solicitadas cumpliendo con lo establecido en los requerimientos.
2. Realizar una comparativa (puede ser gráfica) que contenga múltiples ejecuciones del dataset para diferentes números de clústeres utilizando tanto las propias implementaciones de K-Medias como las de terceros.
3. Se espera que las implementaciones propias de K-Medias posean una interfaz amigable y permitan "predecir" a qué grupo (o clúster) pertenecería un elemento (o registro) que se encuentre fuera del dataset .

OBJETIVOS GENERALES

1. ***Desarrollar implementación del algoritmo K-Medias:*** implementar el algoritmo K-Medias desde cero, cumpliendo con los requisitos establecidos, sin utilizar bibliotecas o frameworks que ya contengan el algoritmo solicitado.
2. ***Realizar implementaciones No Vectorizadas y Vectorizadas:*** crear dos versiones del algoritmo K-Medias para evaluar las diferencias en términos de eficiencia y rendimiento.

3. ***Soporte para diferentes números de clústeres y atributos:*** diseñar las implementaciones de tal manera que permitan configurar el número de clústeres y establecer los atributos a utilizar.
4. ***Comparación con implementaciones de terceros:*** realizar una comparativa entre la implementación propia y una implementación de terceros.
5. ***Interfaz gráfica:*** desarrollar una interfaz para poder visualizar los atributos del dataset, las estadísticas de cada atributo, controlar qué atributos se van a utilizar para la ejecución del algoritmo, establecer la cantidad de clústeres, visualizar resultados en forma de gráficos, etc.

INFORMACIÓN DEL DATASET

DESCRIPCIÓN DEL DATASET

El dataset utilizado para este proyecto es conocido como "Wine Quality Dataset" y consta de múltiples registros, cada uno representando un vino con diversas características químicas medidas y una clasificación de calidad asignada (no se tendrá en cuenta para el algoritmo). Cada atributo en el dataset es una propiedad específica del vino que puede influir en su calidad final.

DESCRIPCIÓN DE LOS ATRIBUTOS

- ***fixedacid (Ácido Fijo):***
  - **Descripción:** concentración de ácidos no volátiles en el vino, como el ácido tartárico. Los ácidos fijos contribuyen a la acidez total del vino.
- ***volacid (Ácido Volátil):***
  - **Descripción:** concentración de ácidos volátiles, principalmente ácido acético. Altas concentraciones pueden dar al vino un sabor a vinagre.
- ***citricacid (Ácido Cítrico):***
  - **Descripción:** cantidad de ácido cítrico en el vino. Este ácido puede añadir frescura y sabor al vino.
- ***residualsugar (Azúcar Residual):***
  - **Descripción:** cantidad de azúcar que queda después de la fermentación. Niveles altos pueden indicar vinos dulces.
- ***chlorides (Cloruros):***
  - **Descripción:** concentración de cloruros (sal) en el vino. Altas concentraciones pueden afectar negativamente el sabor.
- ***freesulfur (Dióxido de Azufre Libre):***
  - **Descripción:** cantidad de dióxido de azufre libre, que actúa como conservante para prevenir la oxidación y el crecimiento microbiano.
- ***totalsulfur (Dióxido de Azufre Total):***
  - **Descripción:** cantidad total de dióxido de azufre presente. Incluye tanto el dióxido de azufre libre como el combinado.
- ***density (Densidad):***
  - **Descripción:** la densidad del vino, que puede estar relacionada con el contenido de azúcar y alcohol.

- ***pH:***
  - **Descripción:** medida de la acidez/alcalinidad del vino. Los valores típicos oscilan entre 3 y 4.
- ***sulphates (Sulfatos):***
  - **Descripción:** concentración de sulfatos en el vino, que puede contribuir a la frescura y también actuar como conservante.
- ***alcohol:***
  - **Descripción:** contenido de alcohol en el vino (% vol).
- ***class (Calidad del Vino):***
  - **Descripción:** clasificación de la calidad del vino en una escala discreta de 0 a 10, donde 0 es la calidad más baja y 10 la más alta.

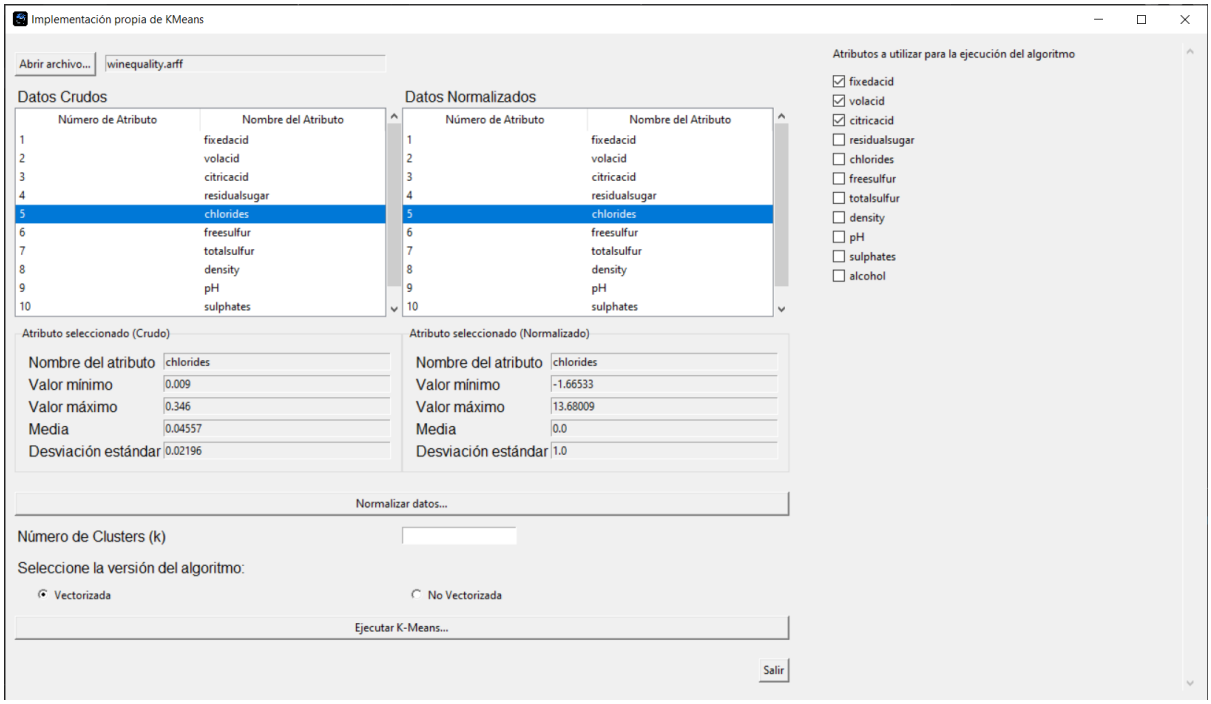
DESARROLLO DEL PROYECTO

1. CARGA DE DATOS

La aplicación permite al usuario cargar archivos (en formato ***.ARFF***, utilizado para conjuntos de datos en aprendizaje automático).

Al cargar el archivo, se muestran los atributos en una tabla, donde el usuario puede seleccionar cualquiera de los atributos para visualizar el valor mínimo, máximo, la media y la desviación estándar de este. La lectura y carga de estos archivos se realiza utilizando la biblioteca Pandas, que permite convertir los datos en un DataFrame para luego poder manipularlos.

Además, se cargan todos los atributos para que posteriormente el usuario pueda seleccionar aquellos que se van a utilizar para la ejecución del algoritmo.



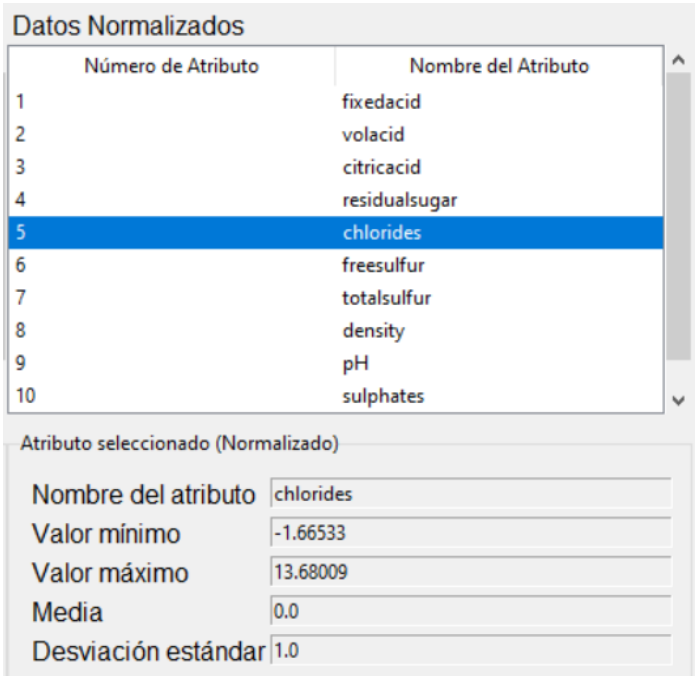
2. NORMALIZACIÓN DE LOS DATOS

Como vimos anteriormente, una de las claves del algoritmo K-Medias, es que los datos estén normalizados, es decir, que tengan una escala similar. Para ello, utilizamos la clase

"StandardScaler" de la librería ***"scikit-learn"*** para normalizar los datos. Específicamente, realiza la normalización ***"Z-score"*** o también conocida como ***"Escalado estándar"***,

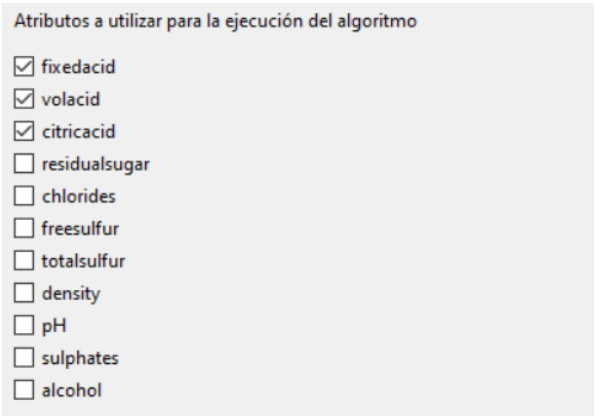
Este tipo de normalización transforma los datos de tal manera que cada característica tenga una media de 0 y una desviación estándar de 1. Esto se realiza mediante la fórmula:

$$X_{normalizado} = \frac{X - X_{media}}{X_{desvEstd}}$$



3. SELECCIÓN DE LOS ATRIBUTOS

El usuario puede seleccionar los atributos que desea utilizar para la ejecución del algoritmo K-Medias. Esto se implementa mediante una lista de todos los atributos que contiene el dataset donde el usuario tiene que tildar aquellos que quiera o necesite utilizar.



4. IMPLEMENTACIÓN DEL ALGORITMO K-MEDIAS

VERSIÓN VECTORIZADA

En esta versión, se utilizan operaciones vectorizadas para mejorar la eficiencia del algoritmo.

La inicialización de los centroides se realiza seleccionando puntos aleatorios del conjunto de datos. Luego, se asignan los puntos a los clústeres calculando la distancia entre cada punto y los centroides, utilizando funciones vectorizadas de Numpy para optimizar el cálculo. Finalmente, los centroides se actualizan calculando la media de los puntos asignados a cada clúster.

VERSIÓN NO VECTORIZADA

Esta versión utiliza bucles para realizar los cálculos, donde el conjunto de datos se transforma en una matriz, cada fila representa un "vino" y cada columna una característica de dicho vino.

Como en la versión vectorizada, los centroides se inicializan seleccionando puntos aleatorios, pero, la asignación de puntos a clústeres y la actualización de centroides se realizan de manera secuencial utilizando bucles for.

Esto puede llevar a diferencias en cómo se manejan las actualizaciones de los centroides entre iteraciones, especialmente en términos de cómo se acumulan y promedian los puntos dentro de cada clúster.

Estas diferencias se ven al momento de ejecutar el algoritmo, ya que, en distintas ejecuciones puede generar resultados diferentes.

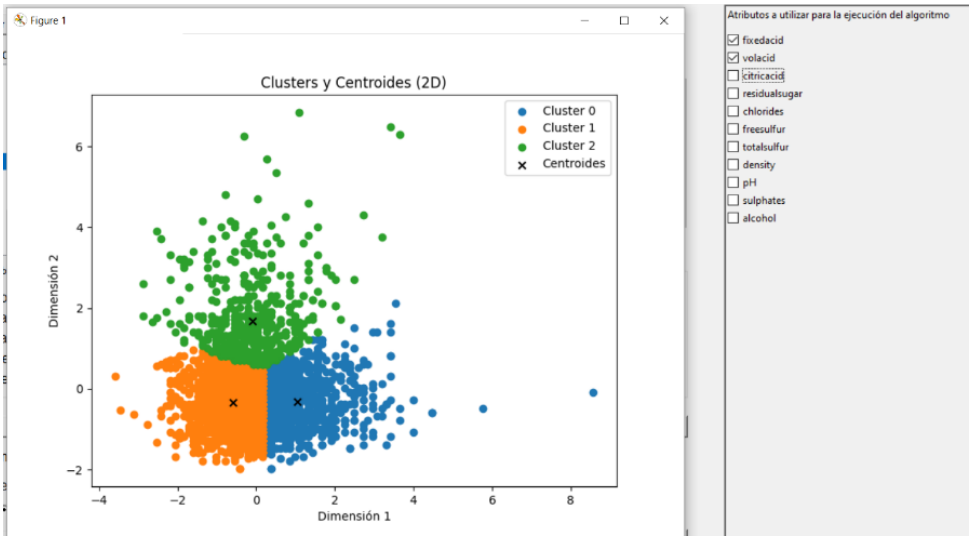
5. VISUALIZACIÓN DE RESULTADOS

Dependiendo de la cantidad de atributos seleccionados, se utilizan diferentes técnicas de visualización. Cuando se seleccionan dos atributos, se utilizan gráficos de dispersión en 2D.

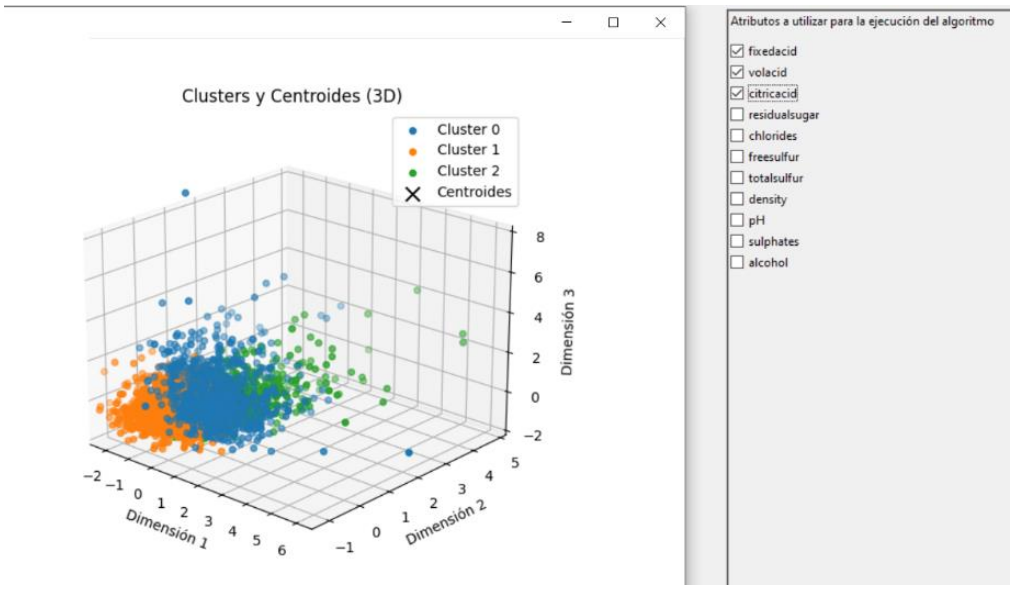
Para tres atributos, se emplean gráficos de dispersión en 3D y en caso de que se hayan seleccionado 4 o más atributos, se aplica el ***Análisis de Componentes Principales (PCA)*** para reducir la dimensionalidad y poder visualizar los resultados en 2D.

El ***Análisis de Componentes Principales (PCA)*** es una técnica estadística de síntesis de la información o reducción de la dimensión. Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible. Los nuevos componentes principales serán una combinación lineal de las variables originales, y además serán independientes entre sí.

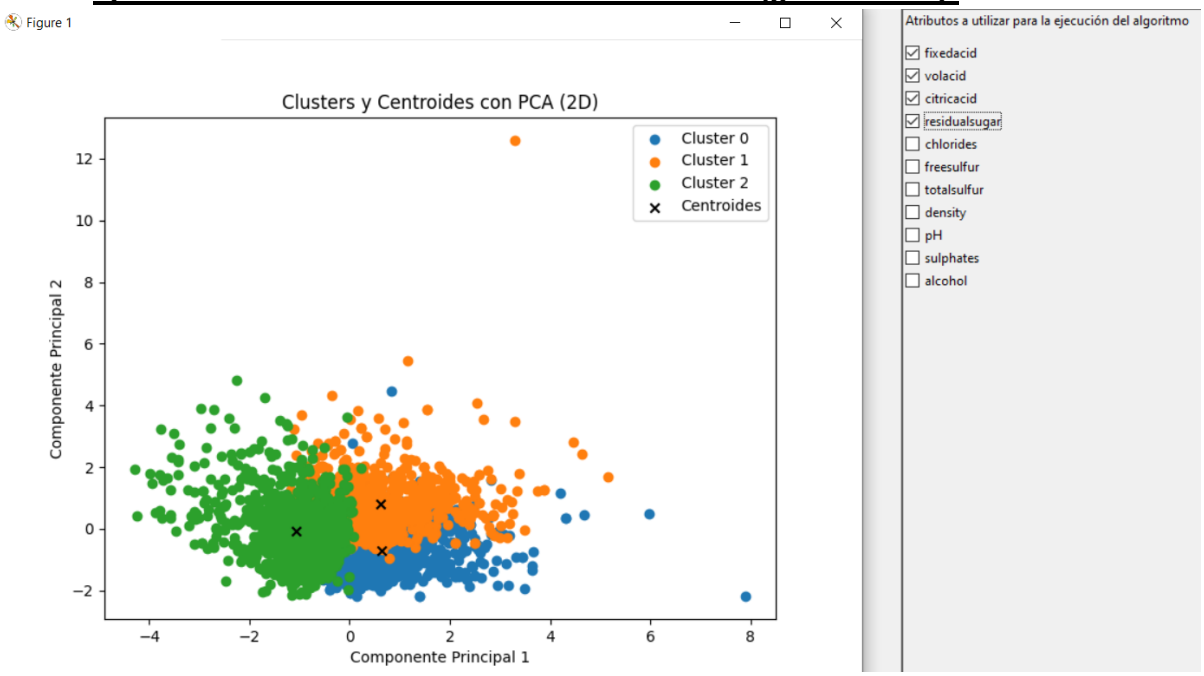
- ***Ejecución con 2 atributos seleccionados (gráfico 2D):***



- ***Ejecución con 3 atributos seleccionados (gráfico 3D):***



- ***Ejecución con 4 o más atributos seleccionados (gráfico PCA):***

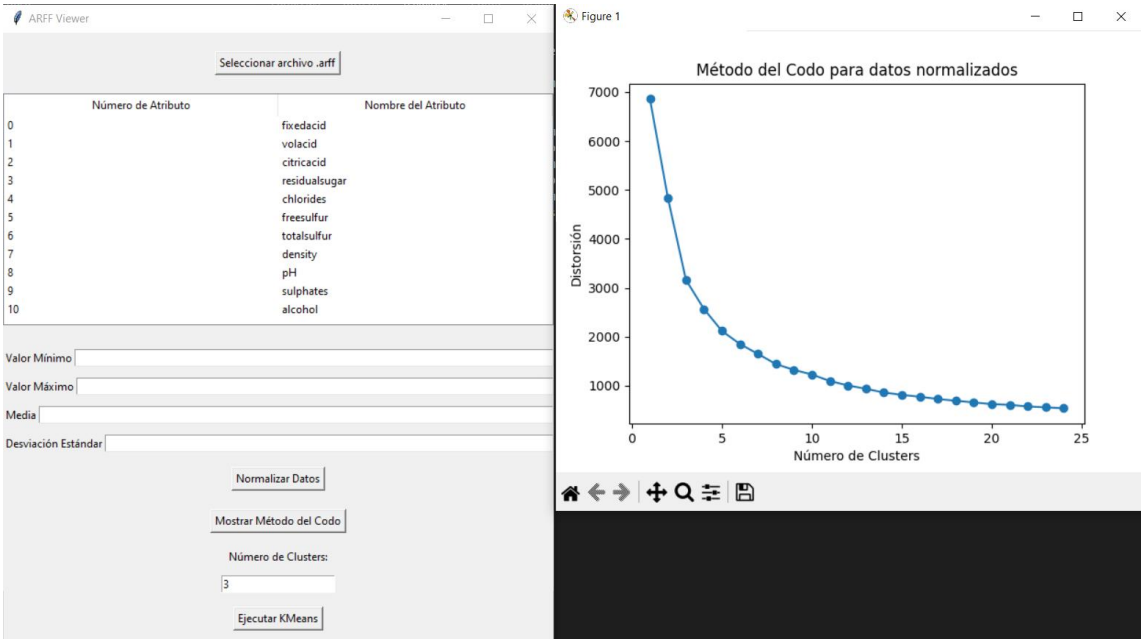


6. VERIFICACIÓN DE RESULTADOS

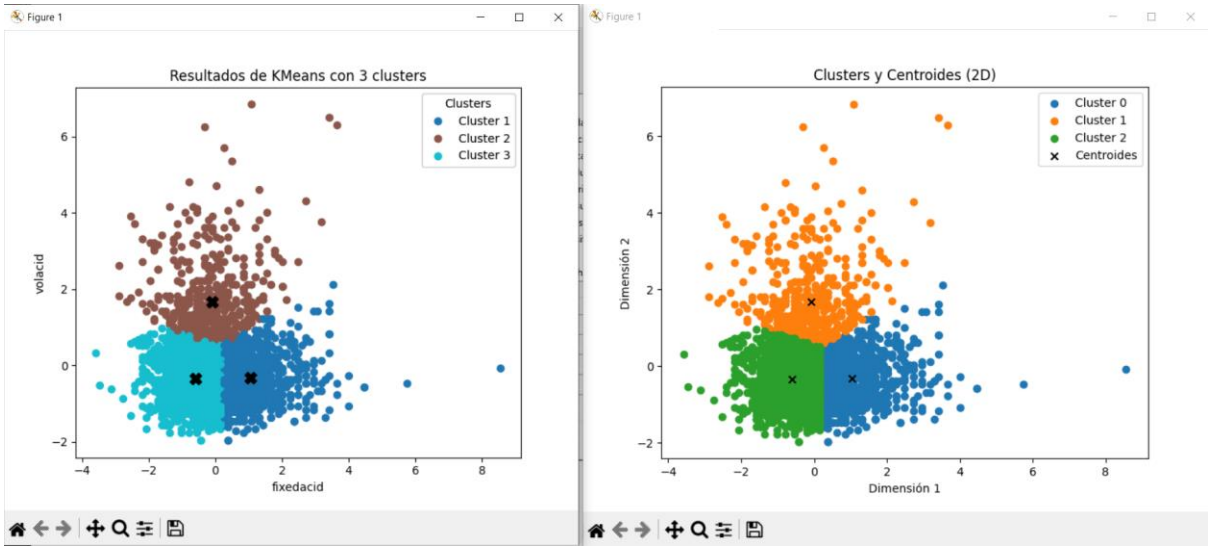
Para asegurar la exactitud de nuestra implementación propia del algoritmo K-Medias, se desarrolló un programa adicional que utiliza la biblioteca ***"sklearn"*** para aplicar K-Medias y poder comparar los resultados. Este programa permite:

- Cargar y normalizar datos de manera similar a nuestra implementación original.
- Permite ejecutar el método del codo para conocer cuál es la cantidad óptima de clústeres para determinado conjunto de atributos.
- Ejecutar el algoritmo K-Medias utilizando ***"sklearn"***.
- Comparar los centroides y las asignaciones de clústeres obtenidas con las de nuestra implementación.

- ***Método del codo para un conjunto de dos atributos:***



- ***Gráficas comparativas entre implementación de "sklearn" (izquierda) y la implementación propia (derecha):***



- ***Comparación de resultados entre implementación de "sklearn" (izquierda) y la implementación propia (derecha):***

```
Error de KMeans: 3162.304493019859
Centroides:
Centroide 1: [ 1.03938117 -0.319843  ]
Centroide 2: [-0.09695987  1.66858742]
Centroide 3: [-0.59413505 -0.33925804]
Puntos asignados a cada cluster:
Cluster 1: 1074 puntos
Cluster 2: 568 puntos
Cluster 3: 1787 puntos
```

```
Error de KMeans: 3162.318391964307
Centroides finales:
Centroide 1: [ 1.03938117 -0.319843  ]
Centroide 2: [-0.08433415  1.6843838 ]
Centroide 3: [-0.59529654 -0.33299508]
Asignaciones de puntos:
Cluster 1: 1074 puntos (31.32%)
Cluster 2: 559 puntos (16.30%)
Cluster 3: 1796 puntos (52.38%)
```