Informe de aplicación de k-means

Introducción

En este informe, se presenta un análisis de agrupamiento de un conjunto de datos que contiene información sobre tres tipos de flores Iris (Iris-setosa, Iris-versicolour e Iris-virginica) utilizando el algoritmo K-Means. El objetivo principal es agrupar las flores en clústeres basados en sus características y visualizar los resultados. El análisis incluye la determinación del número óptimo de clústeres, el entrenamiento del modelo K-Means y la visualización de los resultados.

El lenguaje de programación que fue utilizado es Python

El dataset lo sacamos de la página de kaggle. Este contiene 50 muestras de cada una de las tres especies de flor “Iris”, estas son:

Iris Versicolor

Iris Setosa

Iris Virginica

Para cada una de estas especies se tienen cuatros variables, la longitud y el ancho del petalo y el sépalo: (El sépalo es lo que envuelve a las otras piezas florales en las primeras etapas de desarrollo de la flor).

Sepal\_length

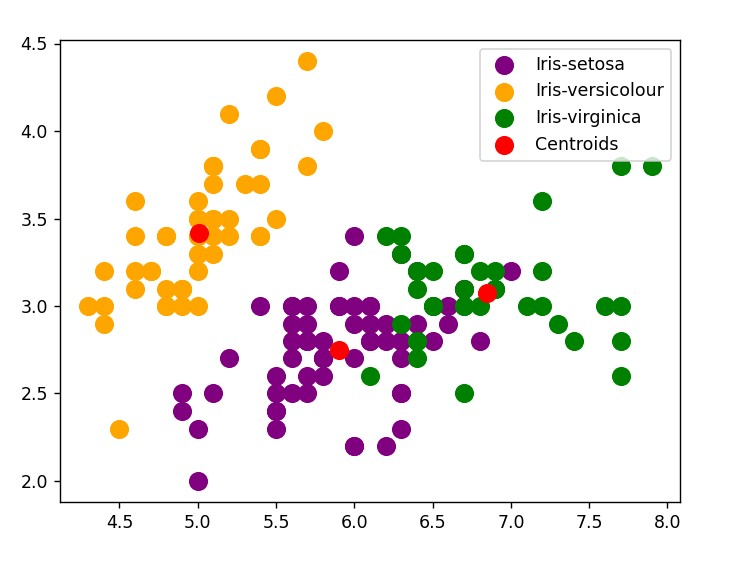
Sepal\_width

Petal\_length

Petal\_width

Ahora, el clasisficador k-means en base a estas características determinara de que clase de flor se trata.

Después de realizar la aplicación del K-means tenemos los siguientes resultados:



Visualización de los resultados:

Se utilizan gráficos de dispersión para visualizar los clústeres encontrados.

Se muestran tres tipos de flores del conjunto de datos Iris: Iris-setosa (en púrpura), Iris-versicolour (en naranja) e Iris-virginica (en verde).

Los centroides de los clústeres se muestran en rojo.

Codigo:

Importación de librerías:

pandas se usa para cargar y manipular conjuntos de datos en formato tabular.

matplotlib.pyplot se utiliza para crear gráficos y visualizar los resultados.

KMeans se importa del módulo sklearn.cluster y se usa para realizar el agrupamiento K-Means.

Carga de datos:

Se carga un conjunto de datos llamado "IRIS.csv" en un DataFrame de pandas llamado "iris".

Se seleccionan las columnas 0 a 3 (las características) y se almacenan en la variable "x".

Determinación del número óptimo de clústeres:

Se crea una lista vacía llamada "wcss" (Within-Cluster-Sum-of-Squares).

Se ejecuta un bucle que va desde 1 hasta 10, donde se ajusta el modelo K-Means para diferentes números de clústeres y se calcula la suma de los cuadrados de las distancias entre puntos y los centroides (wcss) para cada número de clústeres.

El objetivo es encontrar el "codo" en el gráfico de wcss, que indica el número óptimo de clústeres. El codo es el punto donde la disminución en wcss se estabiliza.

Entrenamiento del modelo K-Means:

Se crea una instancia de K-Means con el número óptimo de clústeres encontrado en el paso anterior (en este caso, 3).

El modelo se ajusta a los datos "x" utilizando el método fit\_predict, que asigna cada punto de datos a uno de los clústeres.

Conclusiones

El análisis de agrupamiento utilizando K-Means ha sido exitoso en la agrupación de flores Iris en tres clústeres. La visualización de los clústeres y sus centroides proporciona una representación efectiva de la estructura de los datos y permite una fácil identificación de los diferentes grupos de flores.