Esta carpeta contiene una implementación del algoritmo de K-Means clustering mediante el uso de librerías especializadas como lo es sklearn. Asímismo, se usa el conjunto de datos Wine Dataset obtenido de Kaggle, el cuál muestra los datos de análisis químicos para tres tipos diferentes de vinos. El objetivo del proyecto es generar la clusterización de los datos y posteriormente la predicción de un nuevo valor.

Las librerías importadas de sklearn son las siguientes:

- KMeans para poder realizar la separación de clusters con este método.
- PCA: para poder reducir la dimensión del dataset y realizar el KMeans de manera más sencilla.
- silhouette_samples y silhouette_score: necesarias para obtener tanto los valores como la gráfica de silueta y poder analizar lo bien que se conforman los clusters, así como el número de clusters óptimo para una buena separación de los datos.
- KNN: por último, necesario para la predicción de un nuevo valor mediante el algoritmo KNN.

Para empezar con el código, lo primero que se hizo, después de leer los datos, es reducir la dimensionalidad de los datos y pasar de 13 columnas a tan solo 2 columnas usando la función de PCA. La intención de llevarlo a dos dimensiones es para facilitar la visualización de los datos y la representación visual de los clusters.

Posteriormente, se realizan los clusters con la función KMeans, esto se hace con el número de clusters que el usuario indique, se utiliza la función fit para entrenar el modelo y la función labels_ para obtener las etiquetas de los clusters y estos ponerlos en una nueva columna. En la siguiente función se grafican los datos con los colores representativos de cada uno de los clusters.

En la siguiente función se calcula el valor de la silueta de 3 a 10 clusters para determinar el número óptimo de clusters, esto se hace con un ciclo for en el que se hace el kmeans con el número de clusters dependiendo de la iteración y calculando al final su valor de silueta. Una vez finalizado el ciclo, se presentan los resultados en una gráfica. Con el número de clusters elegido por el usuario, se calcula el mismo valor de silueta pero en este caso para cada uno de los clusters, y este resultado se va a graficar con la intención de evaluar la calidad de cada uno de los clusters en específico.

Como último paso en el proceso está la función para predecir un nuevo punto, en este caso ya se tiene un nuevo array de datos predefinido, o bien, el usuario puede ingresarlos a mano (no recomendado por la cantidad de variables). Una vez con este array de datos, se aplica en ellos la imsma transformación del PCA y ahora, con el método KNN y considerando 3 vecinos, se predice el cluster del nuevo punto. Finalmente se grafica de nuevo todos los puntos agregando el nuevo punto predecido.

Todo esto integrado y utilizado desde el main, donde las funciones se van llamando a la par que se pregunta al usuario valores de las variables antes mencionadas.