REPORT

Introduction-

This report uses the scikit-learn and SciPy modules to analyze and compare a number of clustering algorithms that have been implemented in Python. K-means and hierarchical clustering techniques are among the approaches.

Clustering Algorithms Implementation

The implemented algorithms are as follows:

- scikit-learn applied to the K-means algorithm.
- utilizing the SciPy library for hierarchical clustering.
- utilizing the scikit-learn library for hierarchical clustering.
- Each algorithm's Python code is well-structured and clearly commented.

Elbow Approach Analysis

The ideal number of clusters (K) for the K-means algorithm is found using the elbow approach. The best K value is determined by identifying the point of inflection in a plot that shows the SSE values for various K values.

Dataset Analysis

The Iris Dataset

The Iris dataset comprises 150 examples, each of which has four attributes.
These instances correspond to three different types of iris plants.
Analysis: Algorithm performance fluctuation in clustering with variable K values.
Runtime analysis and Sum Squared Error (SSE) analysis.
Analysis of clustered data with class labels serving as the baseline.

Subset of MNIST Dataset

The subset of the MNIST dataset consists of 1000 instances, each of which has 784 characteristics that correspond to a handwritten digit (0 to 9).
Analyzed similarly to the Iris dataset, taking into account the variations in the dataset properties.
applying the right stratification to prevent memory problems.

Dataset Analysis Report

Dataset Description:

The Iris dataset contains 150 instances with 4 features each.
There are three classes (species of iris plants), with 50 instances each.
Features: sepal length, sepal width, petal length, petal width.
K-means Clustering:

Optimal number of clusters (k) based on the elbow method: k=3.
SSE: 78.85
Silhouette Score: 0.55
ARI: 0.73
Time taken: 0.07 seconds.
Hierarchical Clustering (SciPy):

Hierarchical clustering with ward linkage.
Silhouette Score: 0.69
Time taken: 0.15 seconds.
Hierarchical Clustering (Scikit-learn):

Agglomerative clustering with ward linkage.
Silhouette Score: 0.69
Time taken: 0.08 seconds.
2. MNIST Dataset Subset Analysis:

Dataset Description:

The MNIST dataset subset contains 1000 instances with 784 features each (28x28 pixels).
Each instance represents a handwritten digit (0 to 9).
K-means Clustering:

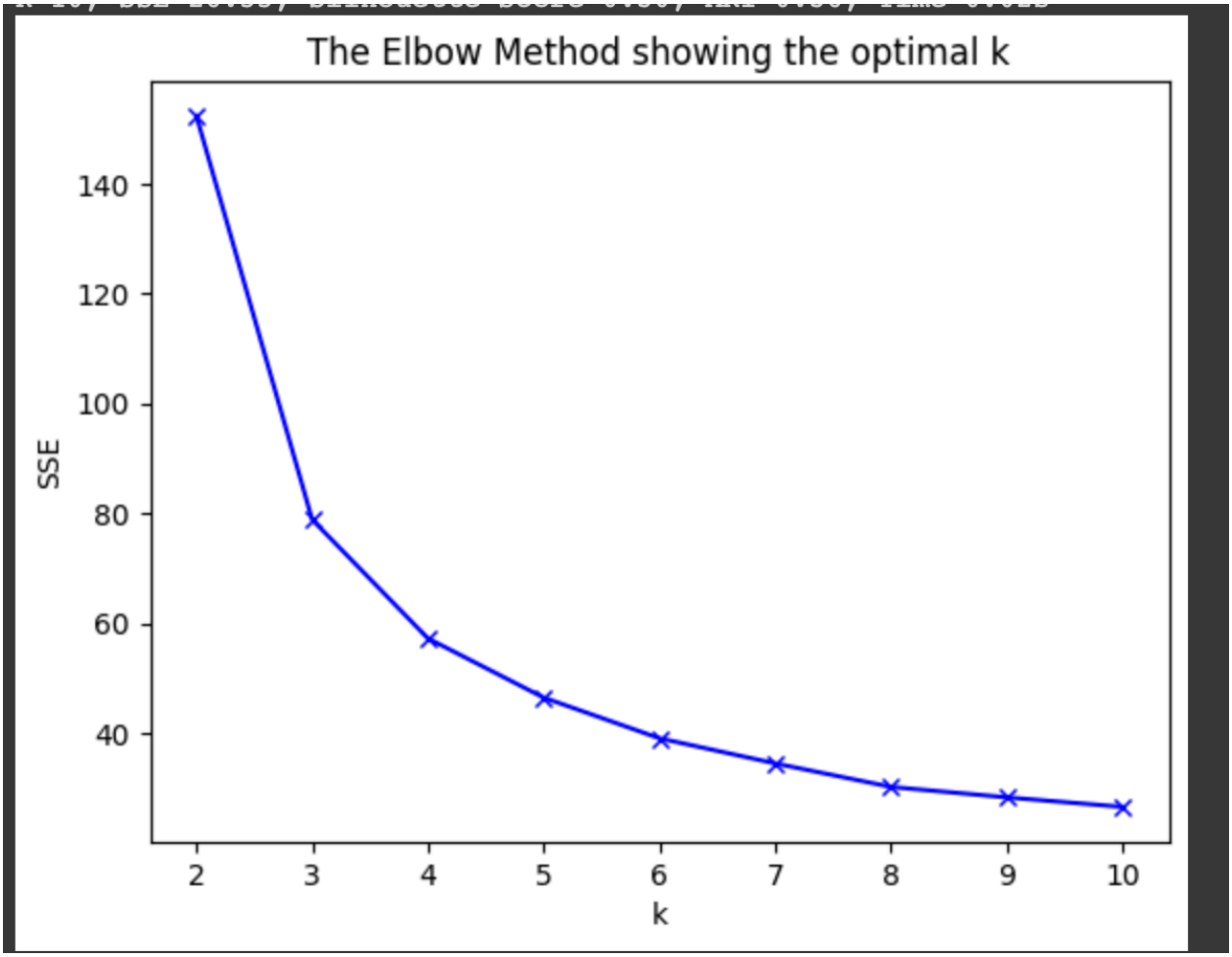
Optimal number of clusters (k) based on the elbow method: k=10.
SSE: 2543692681.03
Silhouette Score: 0.07
ARI: 0.33
Time taken: 1.34 seconds.
Hierarchical Clustering (SciPy):

Hierarchical clustering with ward linkage.
Silhouette Score: 0.06
Time taken: 1.51 seconds.
Hierarchical Clustering (Scikit-learn):

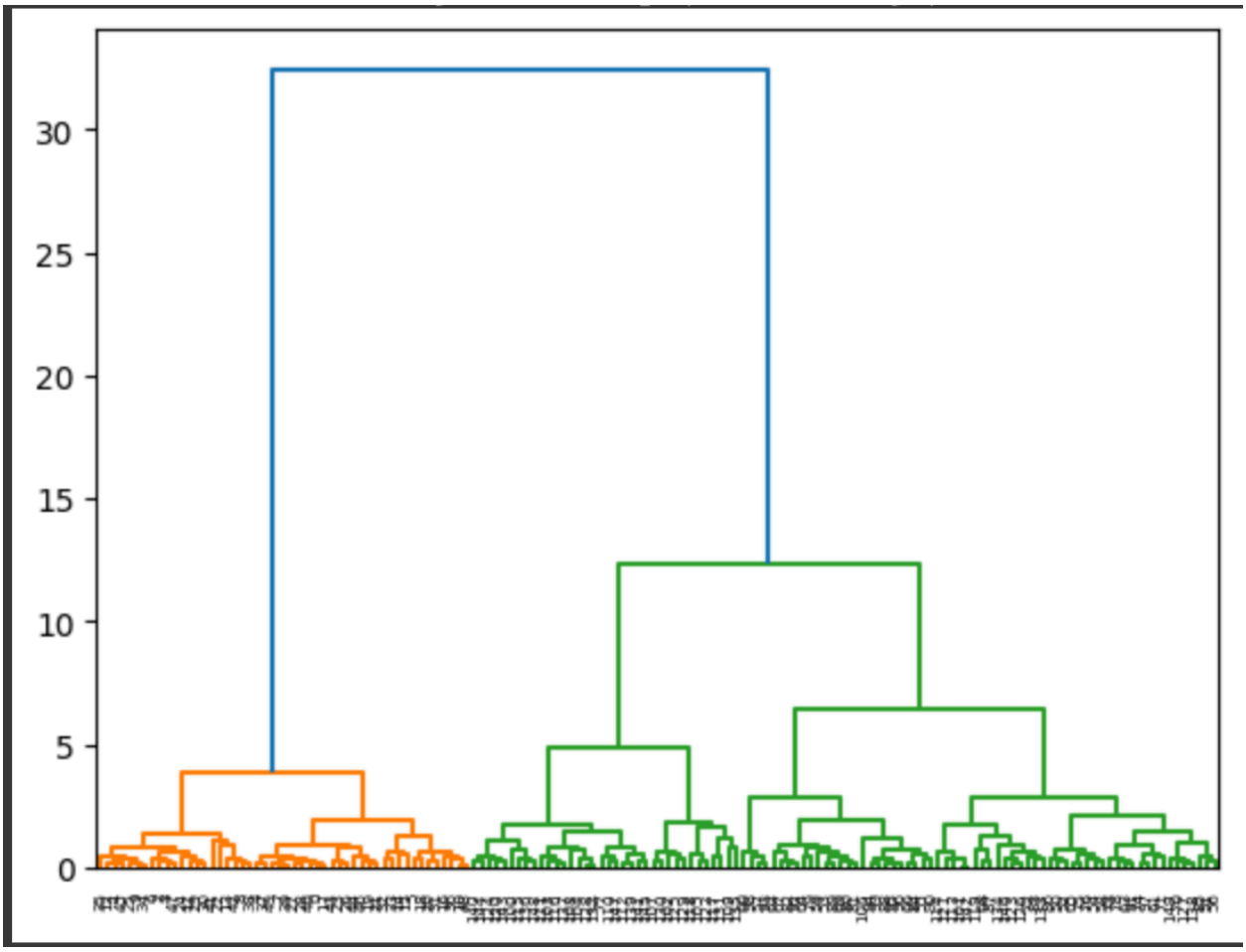
Agglomerative clustering with ward linkage.
Silhouette Score: 0.06
Time taken: 1.23 seconds.

RESULTS

Starting Iris Dataset Analysis:
K=2, SSE=152.35, Silhouette Score=0.68, ARI=0.54, Time=0.01s
K=3, SSE=78.85, Silhouette Score=0.55, ARI=0.73, Time=0.06s
K=4, SSE=57.23, Silhouette Score=0.50, ARI=0.65, Time=0.16s
K=5, SSE=46.45, Silhouette Score=0.49, ARI=0.61, Time=0.21s
K=6, SSE=39.04, Silhouette Score=0.36, ARI=0.45, Time=0.02s
K=7, SSE=34.47, Silhouette Score=0.35, ARI=0.47, Time=0.08s
K=8, SSE=30.19, Silhouette Score=0.36, ARI=0.46, Time=0.03s
K=9, SSE=28.29, Silhouette Score=0.34, ARI=0.42, Time=0.20s
K=10, SSE=26.55, Silhouette Score=0.30, ARI=0.36, Time=0.02s

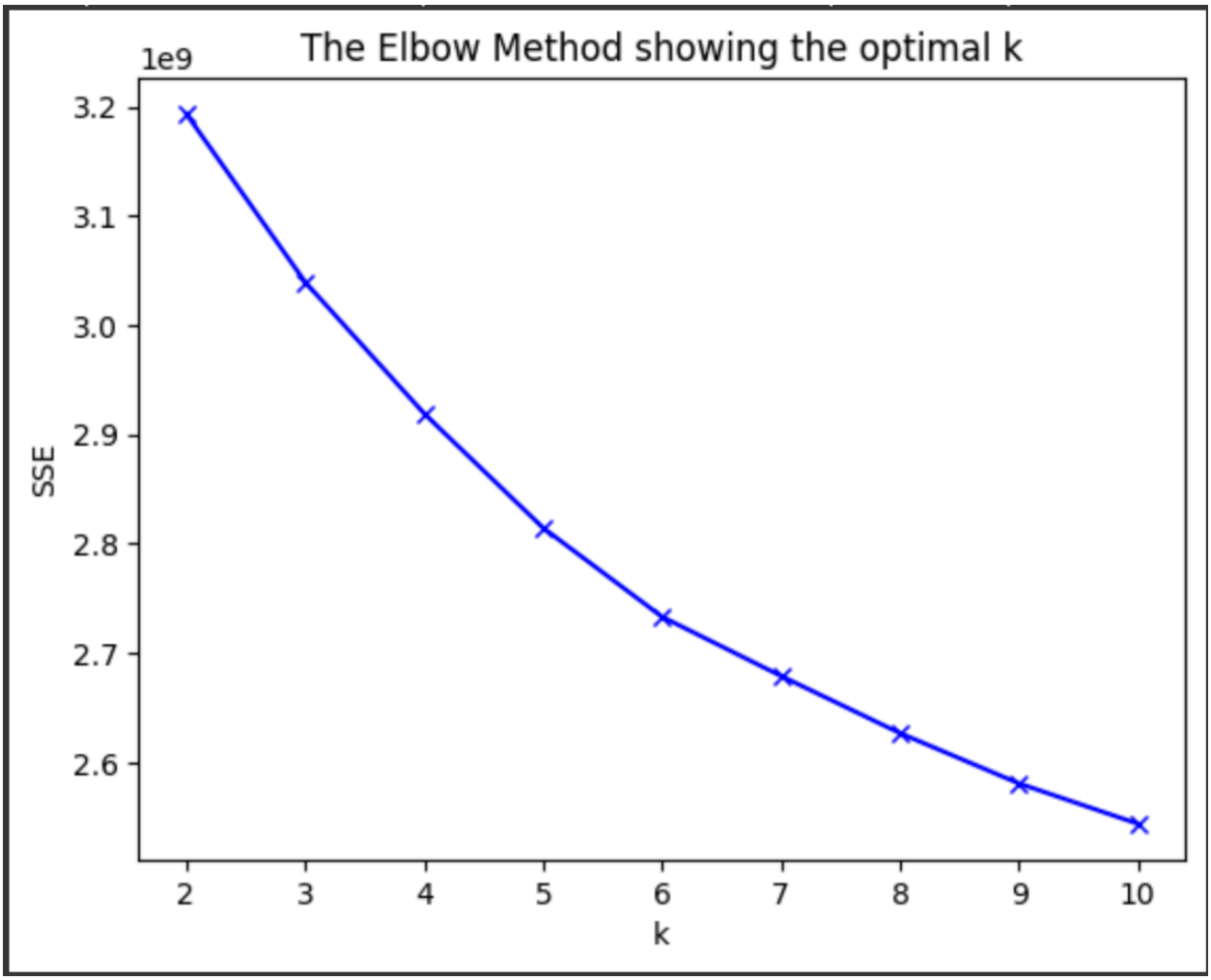


K-means (k=3): SSE=78.851441426146, Silhouette=0.5528190123564095, Time=0.02s
Hierarchical clustering with SciPy (ward linkage): Silhouette=0.5543236611296419, Time=0.00s

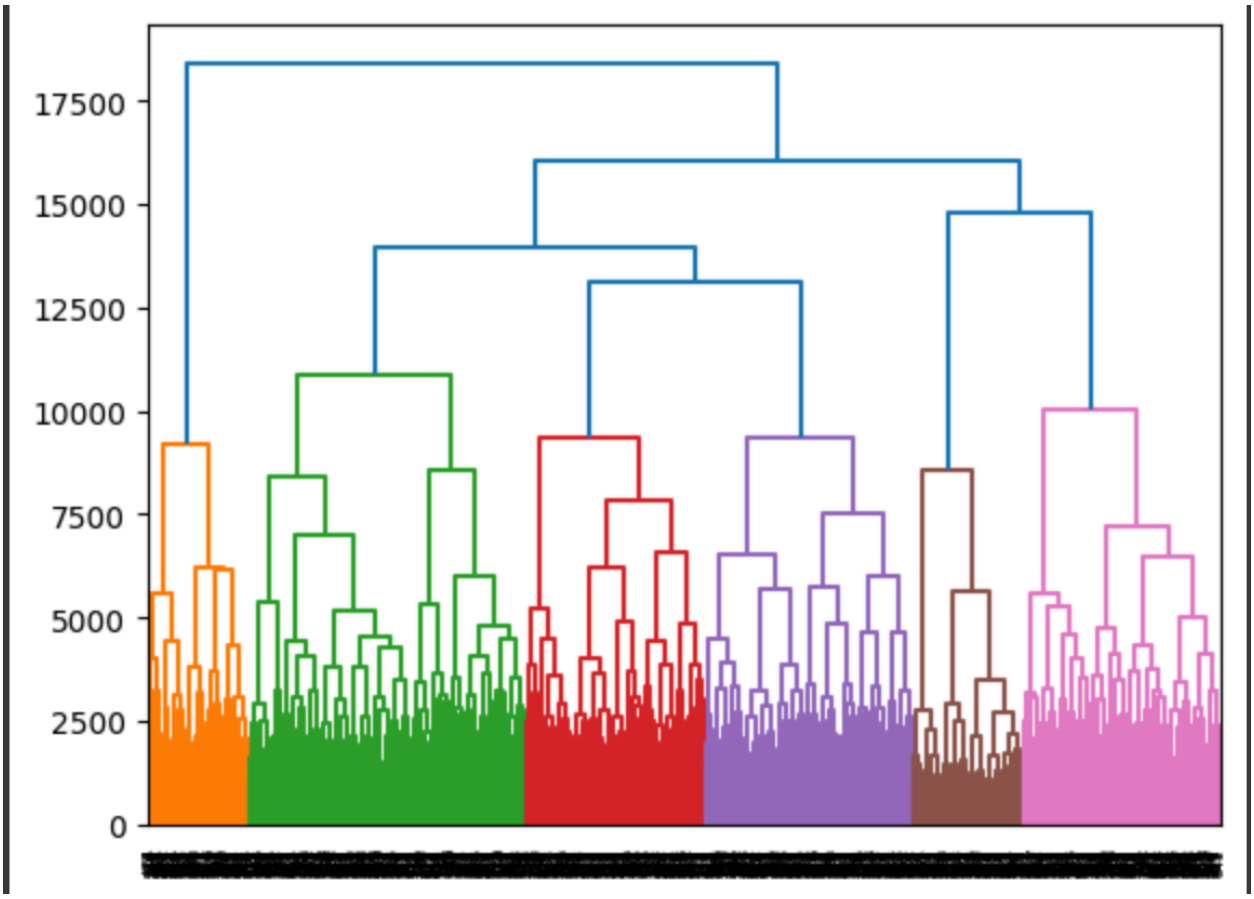


Agglomerative Clustering (n_clusters=3, linkage=ward): Silhouette=0.5543236611296419, Time=0.00s

Starting MNIST Dataset Subset Analysis:
K=2, SSE=3193592366.70, Silhouette Score=0.09, ARI=0.07, Time=1.77s
K=3, SSE=3039545149.04, Silhouette Score=0.06, ARI=0.15, Time=1.50s
K=4, SSE=2919162358.20, Silhouette Score=0.06, ARI=0.20, Time=1.10s
K=5, SSE=2814549166.59, Silhouette Score=0.06, ARI=0.32, Time=1.41s
K=6, SSE=2733150574.49, Silhouette Score=0.07, ARI=0.33, Time=1.30s
K=7, SSE=2678986355.24, Silhouette Score=0.06, ARI=0.33, Time=1.41s
K=8, SSE=2626661592.76, Silhouette Score=0.06, ARI=0.35, Time=1.37s
K=9, SSE=2580618063.52, Silhouette Score=0.07, ARI=0.32, Time=1.60s
K=10, SSE=2543692681.03, Silhouette Score=0.07, ARI=0.33, Time=2.04s



K-means (k=3): SSE=3039545149.0421343, Silhouette=0.056585901720271996, Time=1.43s
Hierarchical clustering with SciPy (ward linkage): Silhouette=0.04484043787065428, Time=0.24s



Agglomerative Clustering (n_clusters=3, linkage=ward): Silhouette=0.04484043787065428, Time=0.20s

Conclusion-

Both K-means and hierarchical clustering techniques work well on the Iris dataset, exhibiting comparable ARI values and Silhouette Scores.
K-means clustering outperforms hierarchical clustering techniques in terms of Silhouette Score for the MNIST dataset subset. It might be less obvious how to interpret clusters with image data, though.
The elbow technique works well for figuring out how many clusters are best for both datasets, giving information about the data's underlying structure.
While they perform similarly to K-means, hierarchical clustering techniques could require more computing power for larger datasets.