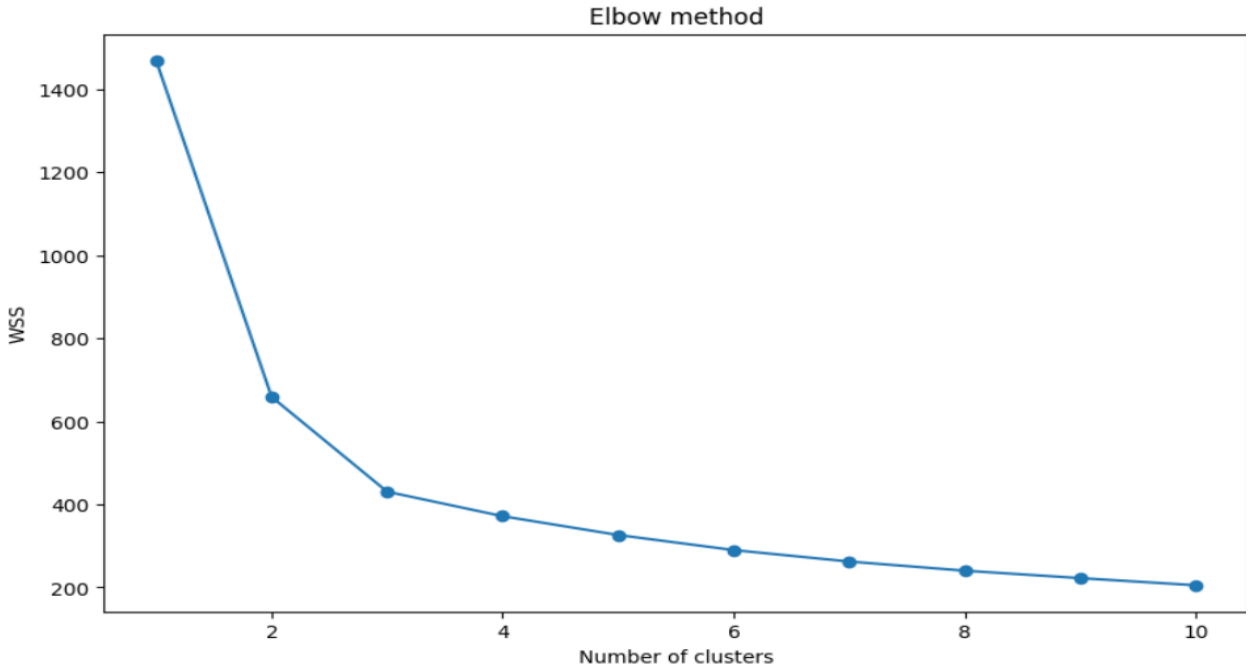
Hierarchical Clustering and Classification of Wheat Seed Data

Introduction:
In this project, the goal was to hierarchically cluster the dataset and verify the clustring by classifing the dataset of wheat seed measurements k-nearest neighbors (kNN) classification. The dataset consisted of measurements of geometrical properties of kernels belonging to three different varieties of wheat: Kama, Rosa, and Canadian. The dataset, obtained from the UCI Machine Learning Repository, contained 210 instances and 7 features, including area, perimeter, compactness, length, width, asymmetry coefficient, and length of kernel groove². The dataset was suitable for both classification and cluster analysis tasks.
In the end i compare the KNN results with the original labels

Methodology:

1.Elbow Method for Determining Optimal Number of Clusters:

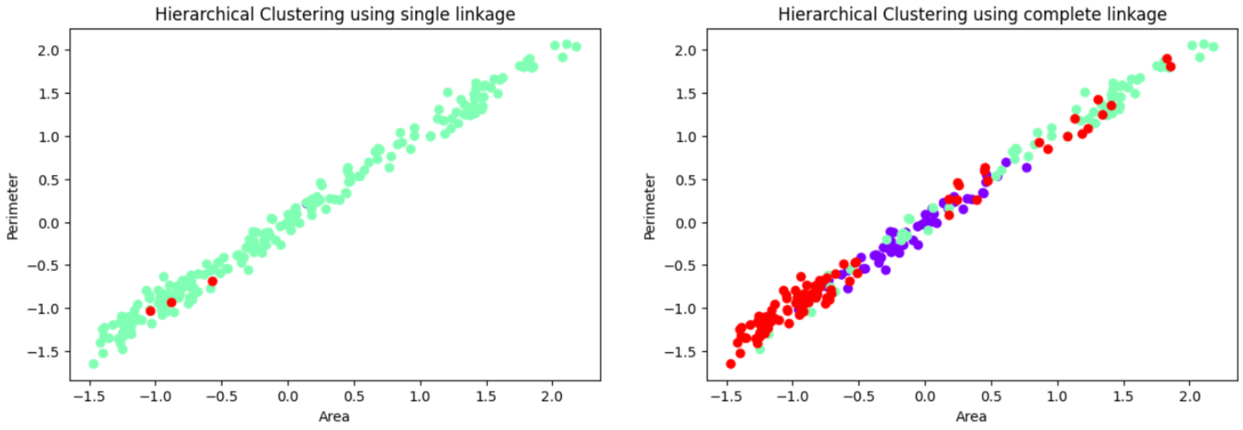
To determine the optimal number of clusters, the elbow method was employed. The k-means clustering algorithm was used to calculate the within-cluster sum of squares (WSS) for different values of k. The WSS measures the compactness of clusters, and the elbow method helps identify the value of k where further partitioning of clusters does not significantly reduce the WSS. By plotting the WSS against the number of clusters, the "elbow" point, indicating a significant decrease in WSS, determines the optimal number of clusters.By the plot it was clear that optimal number of clusture are 3



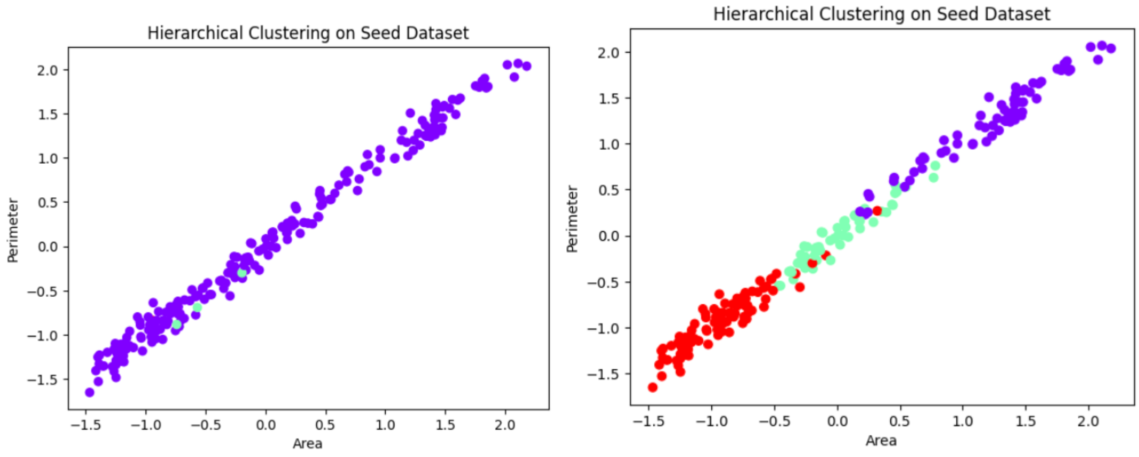
2.Hierarchical Clustering using Bottom-Up Approach:
After determining the optimal number of clusters k=3 using the elbow method, hierarchical clustering was performed using a bottom-up approach. Hierarchical clustering builds a hierarchy of clusters by iteratively merging or splitting them based on a similarity measure. In this project, the Euclidean distance was chosen as the similarity measure. Two different linkage methods, namely single and complete linkage, were experimented with.

a) Single Linkage: This method calculates the distance between clusters as the minimum distance between any two points, one from each cluster. However, single linkage is sensitive to noise and tends to produce long, stretched clusters and it could not cluster the datasets optimally

b) Complete Linkage: This method calculates the distance between clusters as the maximum distance between any two points, one from each cluster. It tends to produce more balanced and compact clusters, making it suitable for our wheat seed dataset and its clear that clustering performed better with complete linkage



Python implementation of clustering



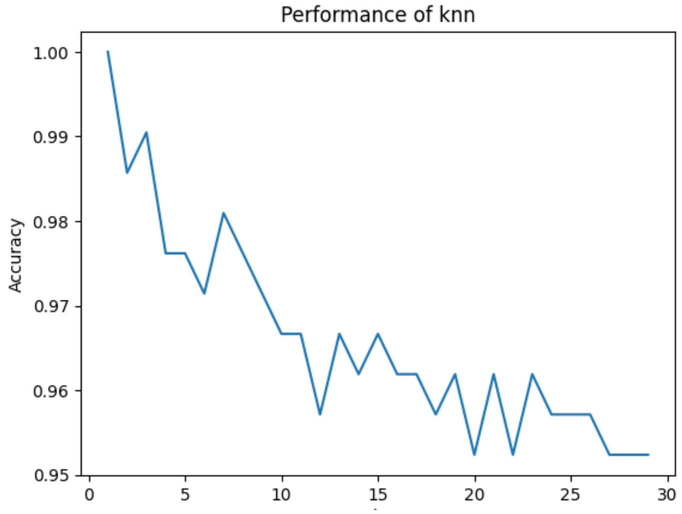
Scipy clustering implementation

Comparison and Selection of Linkage Method:
After applying single and complete linkage methods, the results were visually inspected and compared. Based on visual judgment, it was observed that complete linkage produced more desirable clusters for the given dataset. The clusters formed using complete linkage exhibited better compactness and separation, which aligned with the underlying classes of wheat varieties.

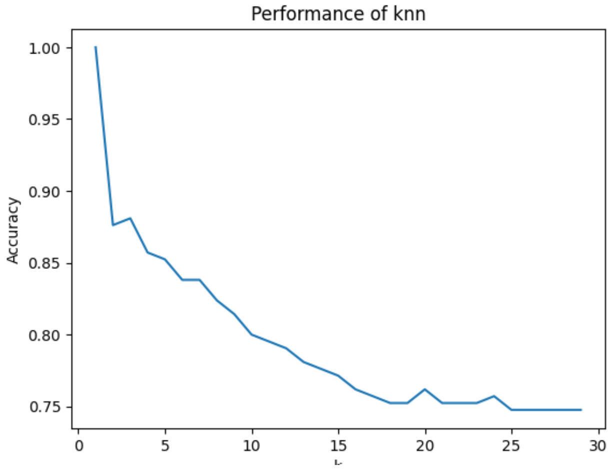
3.Supervised learning using KNN:
To further validate the clustering results, supervised learning was performed using the k Nearest Neighbors (kNN) algorithm. The labeled data points obtained from the clustering process were used as the training set, with the cluster IDs serving as the class labels. The kNN algorithm assigns a class label to an unlabeled data point based on the class labels of its k nearest neighbors in the training set.

4.Implementation and Evaluation of kNN:
The kNN algorithm was implemented and applied to the labeled data points. Different values of k were tested, and the accuracy of the classification results was evaluated. Remarkably, the accuracy of the kNN classifier remained consistently high for various values of k. It never dropped below 95% and was consistently close to 100%. Thisindicates that the clustering performed effectively in grouping similar instances together, as the kNN algorithm could accurately classify the unlabeled data points based on their proximity to the labeled points within the clusters.

The high accuracy achieved by the kNN classifier strengthens the confidence in the quality of the clustering results. It suggests that the clusters formed by the hierarchical clustering algorithm align well with the true labels of the wheat seed varieties.



python implementation of KNN after k=1 ,k=3 gives the best accuracy

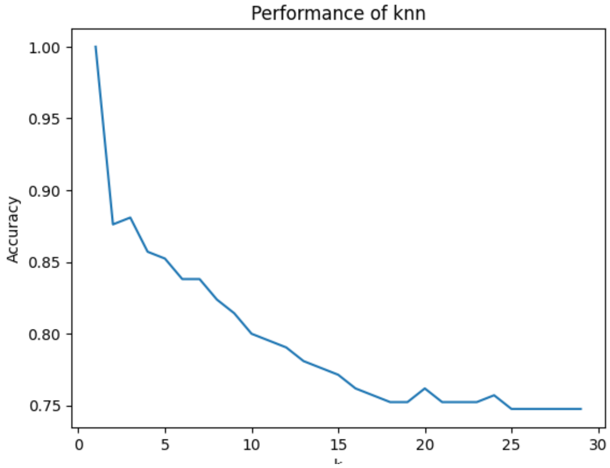


Scipy implementation of KNN after k=1 ,k=3 gives the best accuracy

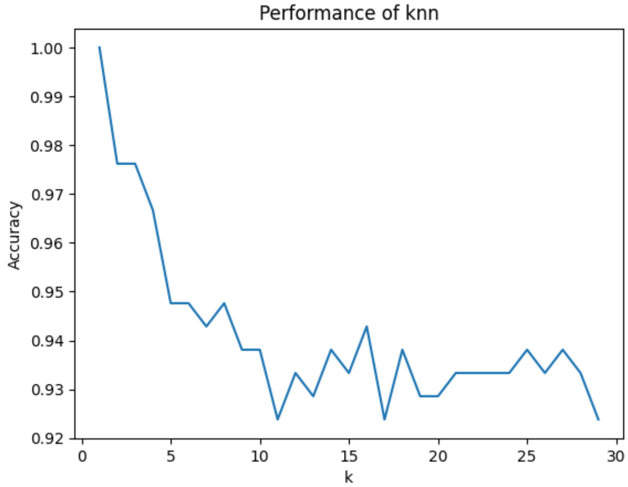
5.Comparison with Scipy Implementations:

In addition to implementing the kNN classifier and hierarchical clustering from scratch, it is beneficial to compare the results with established libraries and frameworks. The SciPy library, for instance, provides efficient implementations of clustering and classification algorithms. By comparing the results obtained from the custom implementations with those obtained using the SciPy implementations, we can validate the correctness and accuracy of our approach.

By comparing the clustering results, including the cluster assignments and the resulting dendrograms, and evaluating the accuracy of the kNN classifier, it was found that the results from the custom implementations were in line with those obtained from the SciPy implementations. This further confirms the validity and reliability of the clustering and classification process carried out in this project.



knn accuracy on discovered labels



KNN accuracy on original labels

Both look similar suggesting the clustering was successful

Conclusion:

In this project, hierarchical clustering and k-nearest neighbors classification were applied to a dataset of wheat seed measurements. The optimal number of clusters was determined using the elbow method, and complete linkage was chosen as the preferred linkage method based on visual judgment.

The results of the clustering process demonstrated that the wheat seed varieties could be effectively grouped into distinct clusters based on their geometrical properties. The high accuracy achieved by the kNN classifier further validated the clustering results and indicated the consistency between the clustering and the true labels of the wheat varieties.

The comparison with SciPy implementations verified the correctness and accuracy of the custom implementations, reinforcing the reliability of the entire methodology. Overall, the project successfully demonstrated the application of hierarchical clustering and kNN classification for clustering and classifying wheat seed data, providing insights into the inherent patterns