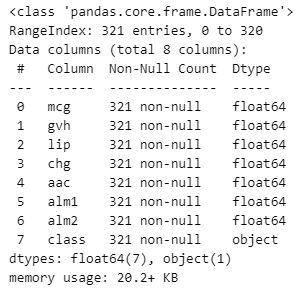
**A**

**Diagram

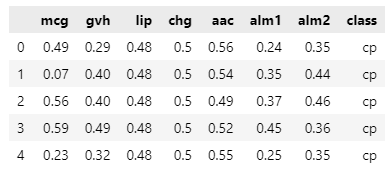
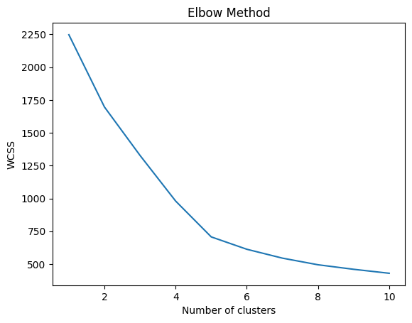
Description automatically generatedTable

Description automatically generated**Description Summary .

Pair plot representation

**Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated**

Header

**Chart, line chart

Description automatically generatedB KNN method**

optimal cluster size: 5

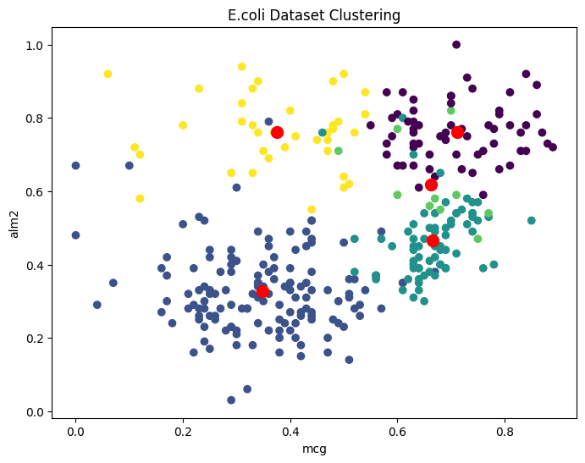
The Elbow method looks for the "elbow" point where the rate of change decreases. The optimal number of clusters in the average silhouette score is the one with the highest average silhouette score. In the Calinski-Harabasz score, we choose the number of clusters that maximizes this ratio as the optimal number. For the DBI score, we choose the number of clusters that minimizes this ratio.

In each plot, the optimal number of clusters is pointed to 5, except for DBI, which is close enough to declare it as 5.

**How well does this method perform in labeling the Ecoli proteins from the training set according to their cellular localization?**

ARI Score: 0.69

An ARI score of 0.69 suggests that the Agglomerative Hierarchical method with complete linkage and Euclidean distance is able to capture some of the underlying structure

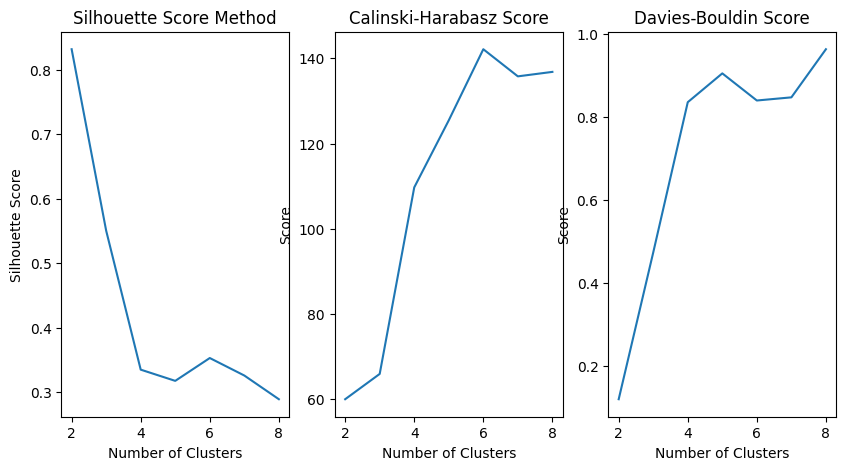
**** Plot of ‘mcg’ and ’alm2’ planes

**C Agglomerative Hierarchical clustering**

Dendrogram for the complete linkage

**Chart, histogram

Description automatically generated**

****

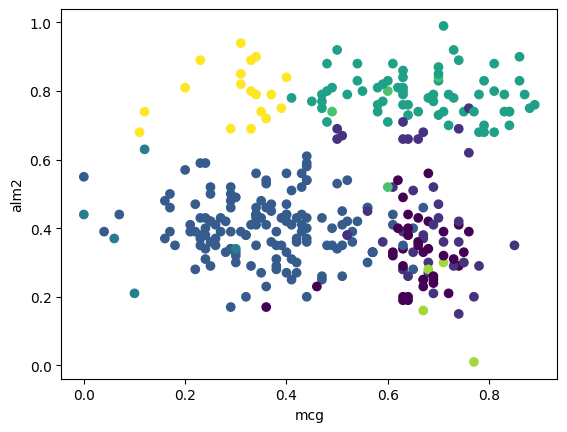
optimal number of clusters: 8

**How well does complete linkage method perform in labeling the Ecoli proteins from the training set according to their cellular localization?**

ARI Score: 0.71

A score of 0.71 for the Adjusted Rand Index indicates a moderately strong agreement between the predicted labels and the true labels.

Plot of ‘mcg’ and ’alm2’ planes

****

**Select one or two other types of linkages beyond complete.**

single linkage ARI: 0. 0419 0.731 for average linkage

The ARI value of 0.0419 for single linkage and 0.731 for average linkage indicate that the Agglomerative Hierarchical clustering with average linkage and Euclidean distance performs much better than single and complete linkage for clustering the E.coli dataset

**D**

**Which clustering method performs best overall on the training Ecoli dataset?**

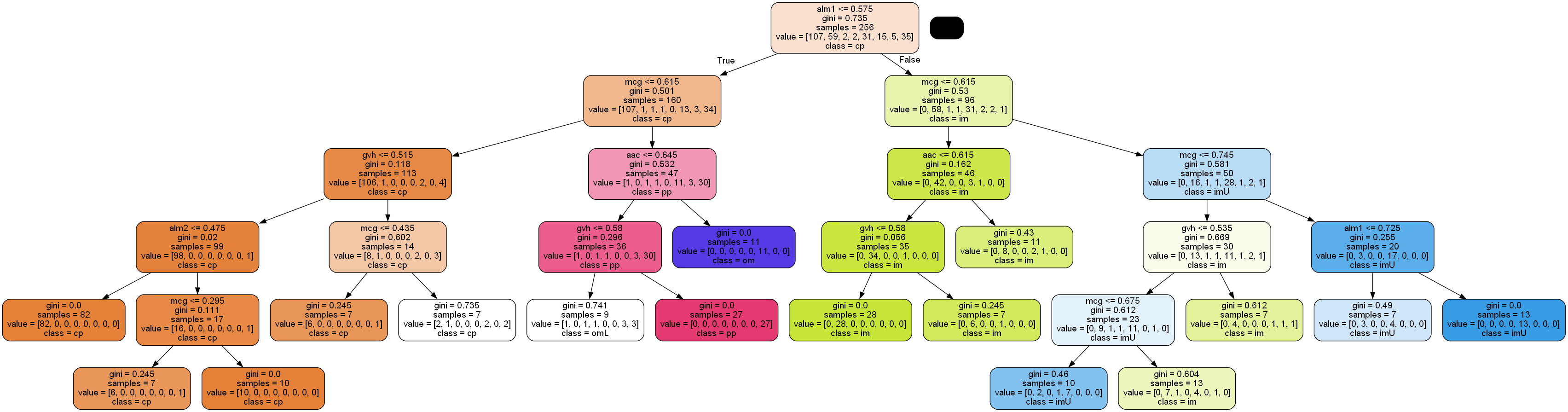
Agglomerative Hierarchical clustering method appears to perform better than the K-means method, as it has a higher ARI score (0.731) compared to K-means (0.69).

**E**

This suggests that using too few principal components (i.e., 5 or 6) may not capture enough important information to accurately cluster the data.

**G**

**G.I What do you learn from the tree about the protein’s dataset?**



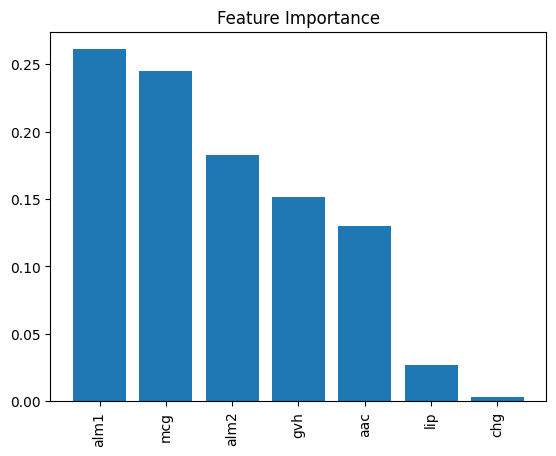
The root node of the tree splits the samples based on the value of the feature 'alm1' which is less than or equal to 0.575. The tree consists of eight levels, with the maximum depth of five.

The first split is based on the value of the feature 'mcg', which is less than or equal to 0.615. 160 samples belong to this node, and the majority of the samples belong to the class 'cp'. The second level split is based on the feature 'gvh', which is less than or equal to 0.515. 113 samples belong to this node. Among these, 106 samples belong to the class 'cp', and the majority of samples belong to the class 'cp' at this node.

As the conclusion, The tree can be used to predict the protein class for a new set of protein features by following the decision path from the root node to a leaf node. The predicted class is the class associated with the leaf node.

**G.II** **How does this accuracy compare with the precision of the clustering methods from part D?**  
Accuracy of Random Forest is 0.81538 and it suggests that RandomForest performs better compared to clustering method.

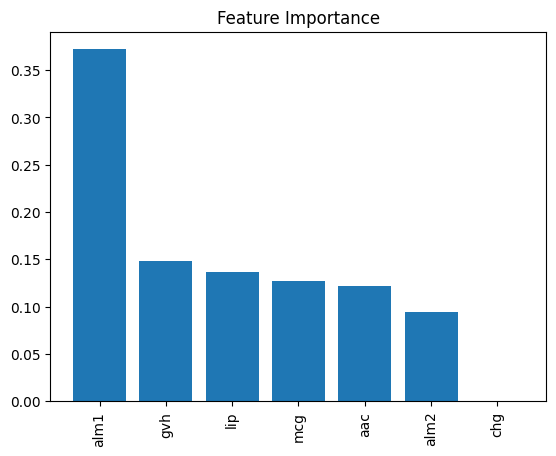
**G.III Using RandomForest to determine the feature importance. Are all features in the proteins dataset important?**

****

No, not all features in the proteins dataset are equally important for predicting protein localization sites in E.coli. The feature importance values obtained from the Random Forest model indicate that some features are more important than others.

These feature importance scores suggest that alm1 and mcg are the most important features, while chg is the least important feature. Therefore, we can conclude that not all features in the dataset are equally important and that some features are more important than others for predicting the protein localization sites in E.coli.

**G.IV**



In the proteins dataset, not all features are equally important when it comes to predicting protein localization sites in E.coli, as is evident from the feature importance data of RandomForest. The plot indicates that alm1 is the most important feature, while gvh, lip, mcg, and alm2 are of lesser but similar importance. On the other hand, chg does not appear to have any feature importance in this context. Therefore, it can be inferred that certain features play a more significant role than others in predicting protein localization sites in E.coli.

**G.V**

RandomForest and Xgboost performs better compared to clustering method.

**G.VI Which approach (clustering versus the decision trees-based methods) do you think is best to use for this prediction and why?**

I would recommend using decision tree-based methods for predicting the localization sites of E.coli proteins. This is because the task involves predicting the class labels of new, unseen instances based on the values of several predictor variables. Decision tree-based methods are well-suited for such classification tasks and can provide clear rules for predicting the class labels of new instances. Additionally, decision trees can easily incorporate new data, making them suitable for updating the model as new proteins are identified.