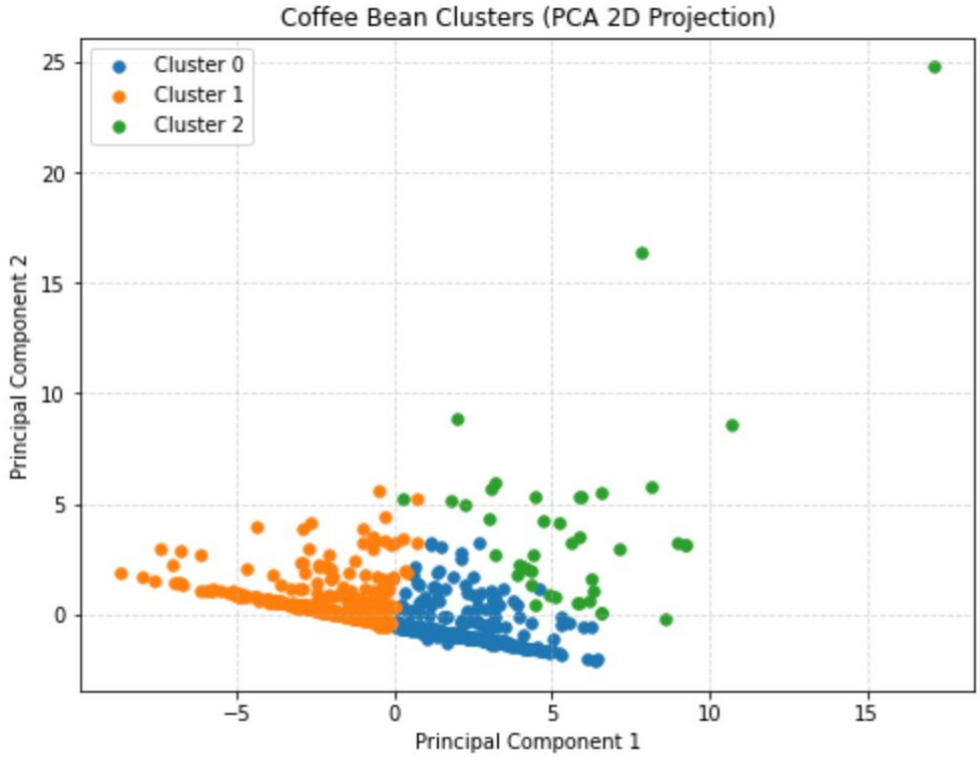
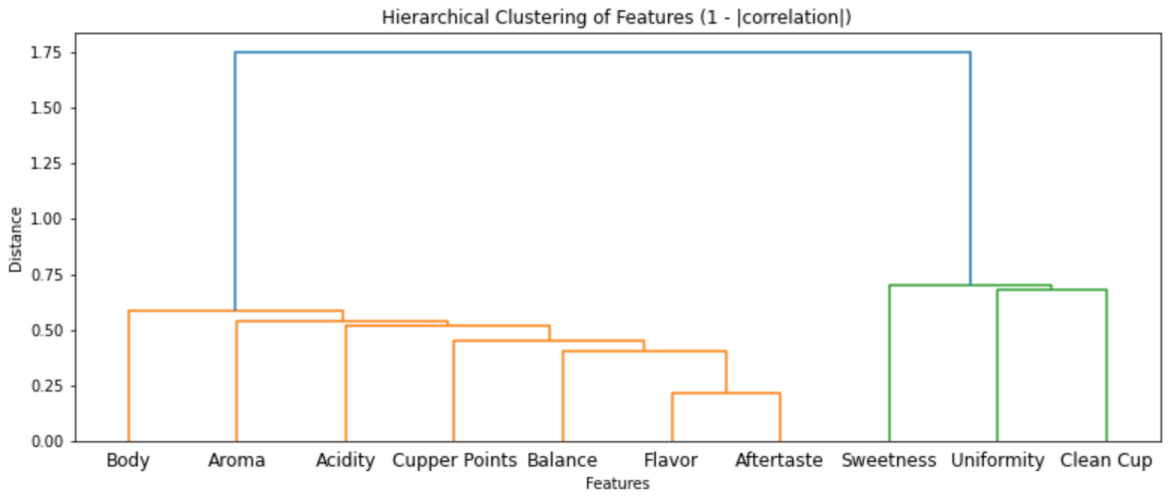
Tabular Data Modelling

The modelling methodology has started with Clustering to cluster the data about coffee beans and determine the natural groupings of features. The standardized quality attributes were subjected to k-Means (k=3). The resulting clusters reflected a moderate separation of clusters, as shown by the Silhouette Score of 0.300; (Davies-Bouldin Index=1.154), which showed that there are three quality profiles.

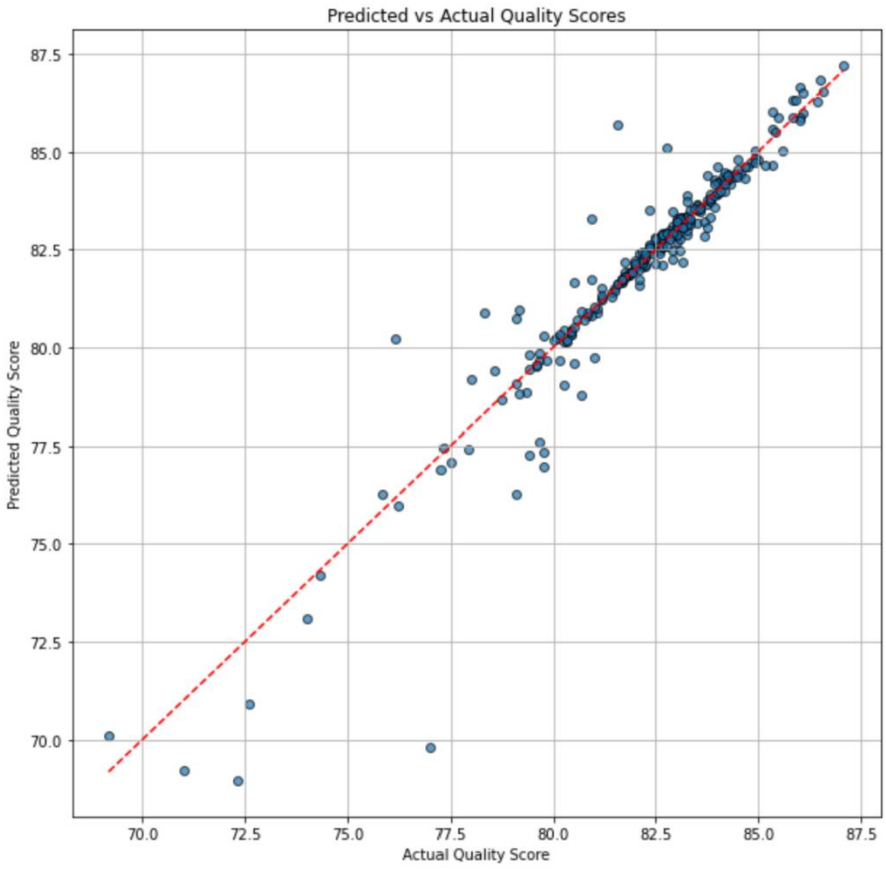


In selecting the features, two approaches were utilized in actualizing the variance of the feature means amongst the three clusters and hierarchical clustering of features to seek correlated feature. Sweetness, Uniformity and Clean Cup were the most discriminative features found in the variance analysis. Five refined features (Body, Aroma, Uniformity, Clean Cup, Sweetness) were then chosen, and the clusters began to separate with a great deal more clarity (Silhouette =0.7081). These attributes, and one-hot coded cluster clusters were very important Engineered Features to the predictive model.

```
Feature variance across clusters:
Clean Cup         6.119640
Uniformity        2.556011
Sweetness         1.985676
Flavor            1.705316
Aftertaste        1.693436
Balance           1.586376
Cupper Points     1.470224
Aroma             1.081433
Acidity           0.888149
Body              0.835634
```



Lastly, Final Prediction of the quality score was done by training a Multi-layer Perceptron (MLP Regressor) on this clean dataset. The model was very generalizable as it attained an R^2 of 0.886 in the test set which implies that it can explain more than 82 percent of the variance in quality score. The model had a Mean Absolute Error (MAE) of 0.758 and an approximate relative accuracy of 99.53% at the final score prediction.



Project Overview: Anticipated Coffee Quality.

This project was aimed at building a high-precision predictive model which would be able to predict the continuous Quality Score (regression), as well as categorize the Quality Tier (High, Medium, Low) of Arabica coffee beans based on the detailed data of sensory attributes. The strategy was focused on resistance training of the predictive pipeline via strategic feature engineering, specifically, by using unsupervised learning methods like the K-Means clustering to generate more meaningful and structured inputs to the downstream models.

Methodology Overview

K-Means clustering was used as the feature engineering process, with 10 standardised Features and divided in 3 clusters. This has enabled finding natural groupings among the set of data and selecting the most discriminative features according to inter-cluster variance. The resulting engineered feature space consisted of a refined set of five major sensory features and a one-hot encoded cluster membership feature, which also acted as other high-value predictive features.

Various regression and classification algorithms have been tested, such as simple algorithms (Linear Regression, Decision Tree) and more advanced algorithms (MLP Regressor to regression and KNN Classifier, to predict a tier). These experiments allowed making a strong comparison of model capabilities in both predictive tasks.

Key Outcomes

The models based on their predictive performance were very effective which proved the efficiency of the pipeline:

Regression Task:

The MLP Regressor performed the highest with a Mean Absolute Error of 0.5186 and an R^2 score of 0.8865 which is far better than the desired performance (MAC < 1.0).

Classification Task:

KNN Classifier proved to be the best model reaching a staggering 98 percent accuracy indicating that samples of coffee of close quality levels naturally fall in the same sensory space.

Meaning and Understandings.

The excellent performance of the MLP Regressor supports the fact that it can be applied in the predictive application in real-world, as it is precise and can predict on diverse sensory profiles. On the same note, the high-quality KNN classifier accuracy indicates the hypothesis that there is an innate quality division between tiers in the multidimensional sensory domain.

The feature engineering pipeline, especially the inclusion of clustering, was very useful in increasing the overall structure and clarity of the data. This was reflected by the fact that the Silhouette Score saw an increase in the score of 0.300 to 0.3824 which shows tighter and more significant cluster formation. By and large the joint methodology was able to yield a strong and predictable system of coffee quality prediction both in terms of quality scores and in categorical quality levels.