**Project Summary -**  
By Shaked Gofin & Sean Peer -

We chose to turn the model answers to be binary as good or not good by evaluating the score of each wine. 5 > not good wine, 5 < Is a Good wine.  
now we will observe each model and analyze the results.  
  
 **Logistic Regression Model Summary -**

The Logistic Regression model showcased an initial accuracy of 73.15%. A detailed analysis of its performance is highlighted through a classification report, emphasizing its precision, recall, and F1-score across the binary classification of wine quality:

**Enhancements and Cross-Validation Insights:**

To refine the model's precision and performance, a structured approach incorporating k-fold cross-validation was employed, leading to an insightful comparative analysis:

* The average accuracy across folds improved slightly to 73.49%.

**Comparative Classification Reports:**

The cross-validation process yielded the following insights into the model's performance across various folds, showcasing a consistent enhancement in the model's ability to classify wine quality accurately:

* **Fold 1**: Demonstrated an accuracy of approximately 74%, with precision and recall for Class 1 showing noteworthy improvement.
* **Fold 2 to Fold 5**: Varied performances with accuracy ranging from 71% to 75%, depicting the model's adaptability and the effectiveness of cross-validation in refining its predictive accuracy.

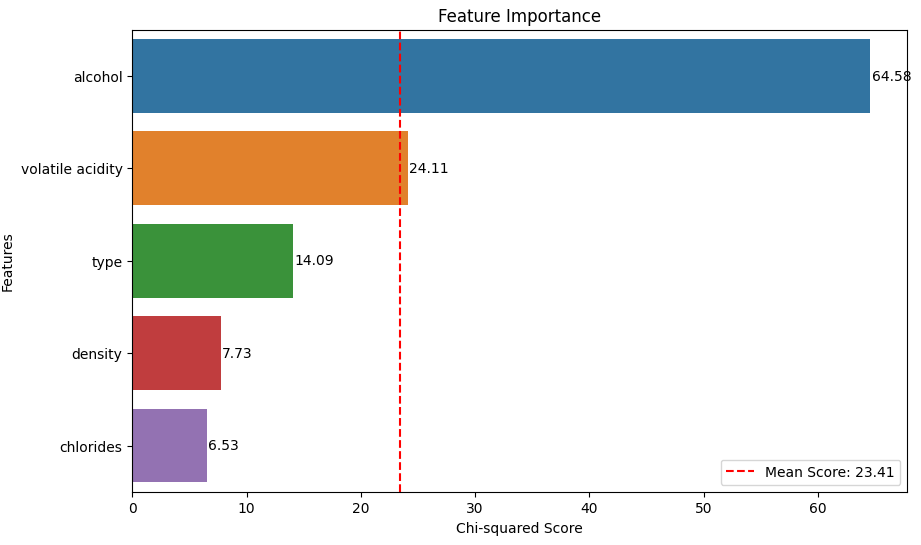
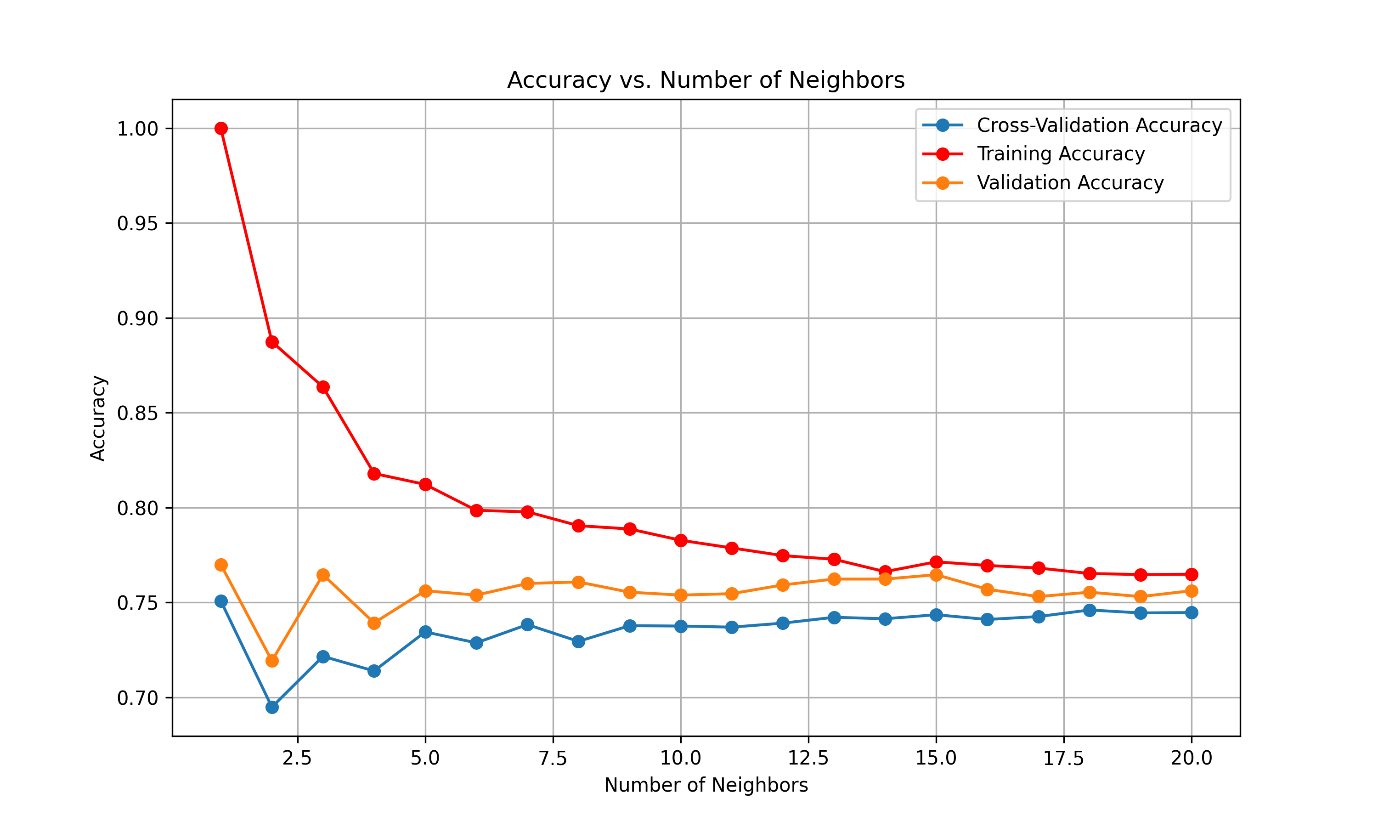
**The K-Nearest Neighbour (KNN) model -**

a fundamental algorithm for classification tasks, showed a baseline accuracy of approximately 67.85%. To enhance this model's performance and precision, we employed a strategic combination of Min-Max scaling, feature selection, and cross-validation techniques. These adjustments led to a notable improvement in model efficacy:

* The cross-validation process, which provides a more robust estimate of model performance, revealed a mean accuracy of 73.45%.
* After applying Min-Max scaling and feature selection, the test set accuracy further increased to 75.62%, demonstrating the positive impact of preprocessing and feature selection on model performance.

In terms of feature importance, an analysis revealed that alcohol content is a significant predictor of wine quality within the KNN model's decision-making process. Specifically, wines with higher alcohol content tend to be associated with better quality scores. This insight not only underscores the importance of alcohol content in determining wine quality but also highlights the value of feature importance analysis in understanding model behaviour and identifying key predictors.

Overall, the strategic enhancements to the KNN model not only improved its accuracy and precision but also provided valuable insights into the features most influential in predicting wine quality, offering a clear direction for further model refinement and feature engineering.



**Decision Tree Model -**

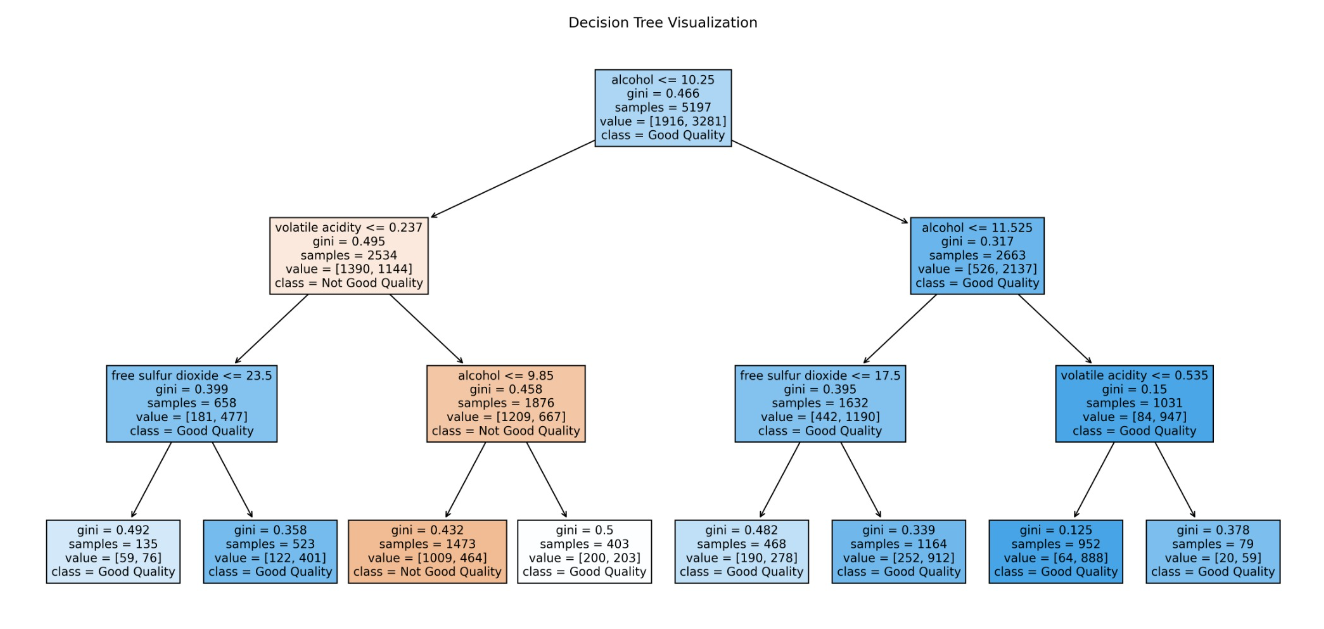
The Decision Tree Classifier was applied to predict wine quality, achieving a test set accuracy of 81.15%. This result underscores the model's capability to effectively distinguish between different quality levels of wine, highlighting its practical utility in scenarios requiring quick and interpretable decision-making.

**Key Highlights:**

* The model was rigorously trained and tested on a split dataset, ensuring a reliable evaluation of its predictive accuracy.
* Achieving over 81% accuracy indicates a strong alignment between the model's predictions and the actual quality classifications of wines in the test set.

Implications of Results:

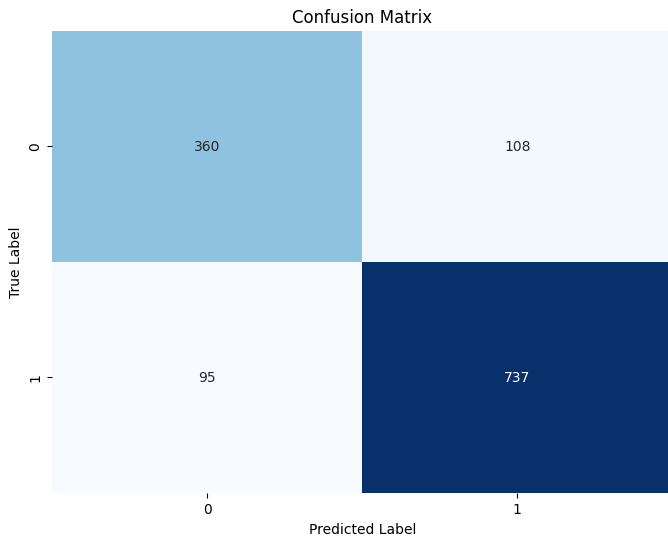
* The high accuracy achieved by the Decision Tree Classifier suggests it has successfully captured the underlying patterns in the dataset, making it a valuable tool for assessing wine quality.
* Such performance demonstrates the model's potential for deployment in quality control settings or as part of an automated system to classify wines based on their characteristics.



**Randon Forest** **Classifier -**  
Demonstrating an impressive accuracy of 84.38% on the test set. This performance not only signifies the model's strong predictive capabilities but also its effectiveness in handling complex classification tasks.

**Highlights:**

* Utilizing a combination of multiple decision trees to make more accurate and stable predictions, the Random Forest model effectively leveraged the training data, leading to high accuracy.
* The model's success in the wine quality prediction task reflects its ability to discern intricate patterns and relationships within the dataset, making it a robust tool for such classification challenges.

Confusion Matrix -  


To try and “squeeze out” the strongest features a bit, we chose to use PCA which is dimensional reduction.

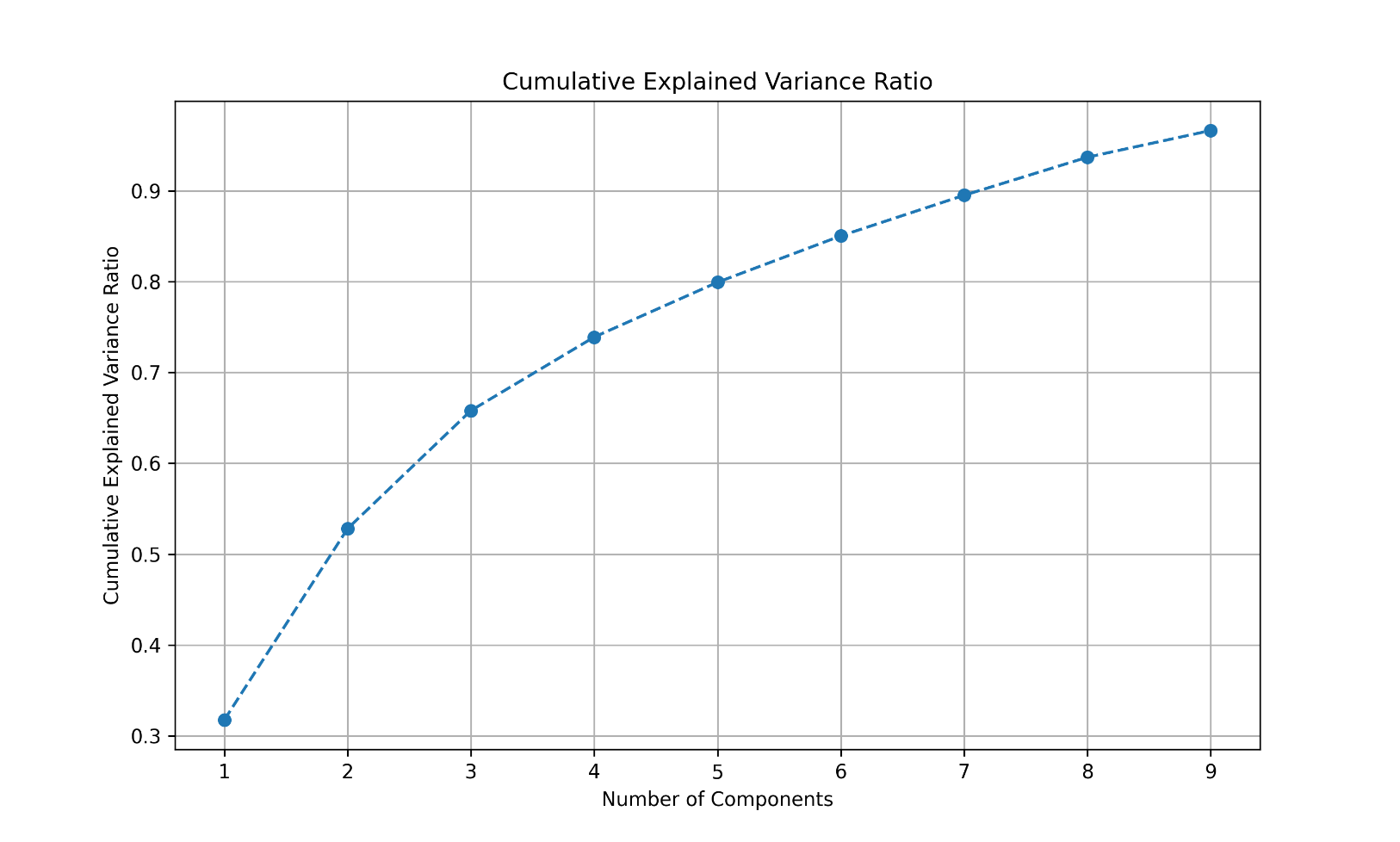
transforms a set of possibly correlated features into a smaller set of linearly uncorrelated variables called principal components.

**Random Forest PCA-**  
  
The Random Forest model combined with PCA, utilizing 6 principal components derived from the original dataset, achieved an accuracy of 82.46% on the test set. This outcome, while still high, is slightly lower than the accuracy observed with the regular Random Forest model without PCA.

**Key Points:**

* **Dimensionality Reduction**: PCA was applied to reduce the dataset to 6 principal components, focusing on capturing the most significant variance in the data with fewer features.
* **Accuracy Impact**: The reduction in accuracy compared to the standard Random Forest model can be attributed to the loss of information. While PCA aims to retain the most important variance, the elimination of dimensions may lead to the exclusion of some details that were valuable for prediction.

**Conclusion:**

The application of PCA with Random Forest showcases a trade-off between model simplicity and predictive performance. The slight decrease in accuracy emphasizes the balance between reducing complexity and maintaining the integrity of the dataset's predictive power. This approach is beneficial in scenarios where model efficiency and computational resource limitations are priorities, but it's important to consider the potential impact on prediction accuracy.

ראש הטופס

