ASSIGNMENT REPORT

This report presents a detailed analysis of seed classification using various machine learning models, including Decision Trees, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forests. The objective was to classify seeds into distinct categories based on their features and evaluate the performance of each model.

Dataset Overview:

The dataset comprises seven features: Area, Perimeter, Compactness, Kernel Length, Kernel Width, Asymmetry Coefficient, and Kernel_Groove_Length. The target variable, Class Label, categorizes the seeds into three classes.

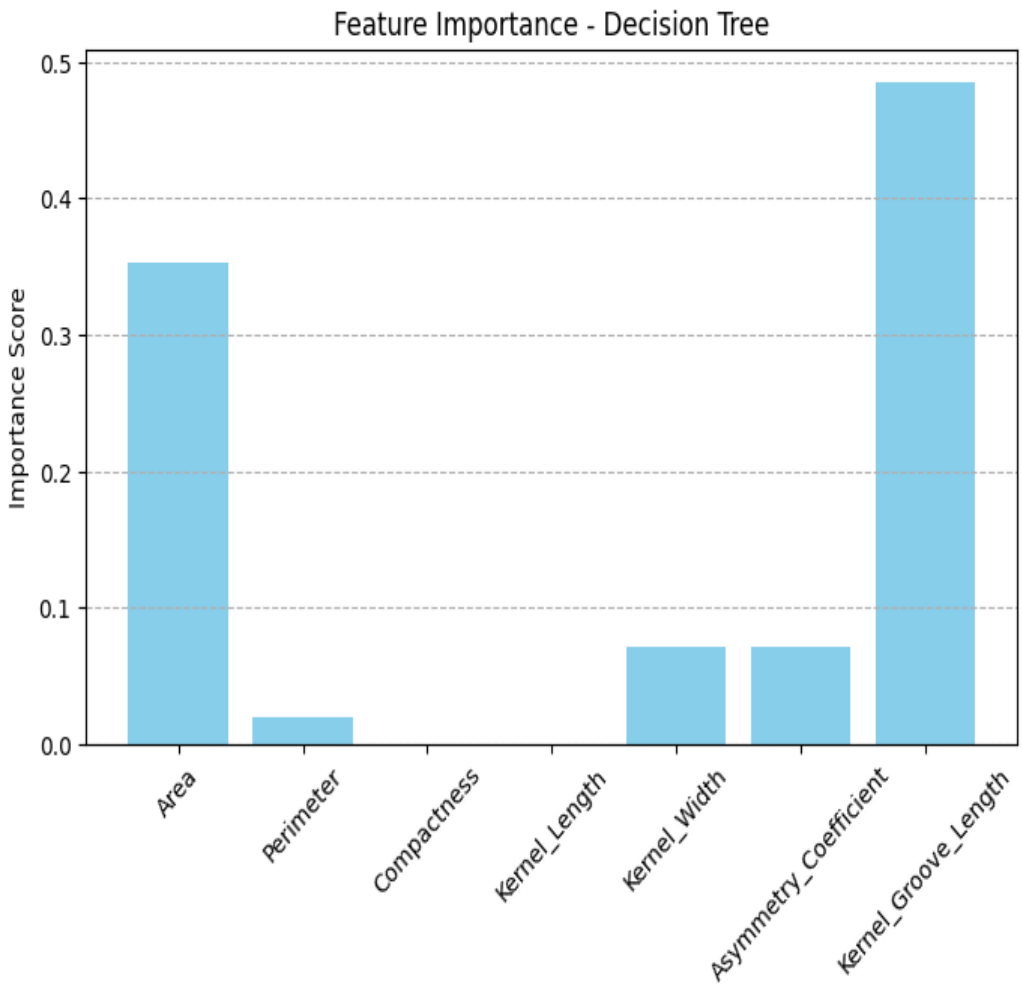
Preprocessing:

The dataset was preprocessed by splitting it into training and testing sets (70% training, 30% testing) and applying feature scaling using Standard Scaler to normalize the data.

1. ***Decision Tree Classifier***

   The Decision Tree model was able to classify the dataset with an accuracy of 93.65%.

   - *Classification Report:*
     Precision: 0.94 (macro avg)
     Recall: 0.94 (macro avg)
     F1-Score: 0.94 (macro avg)



From the feature importance plot, we observed that:

* Kernel Groove Length and Area were the most important features.

* Perimeter and Compactness had a minimal effect on classification.

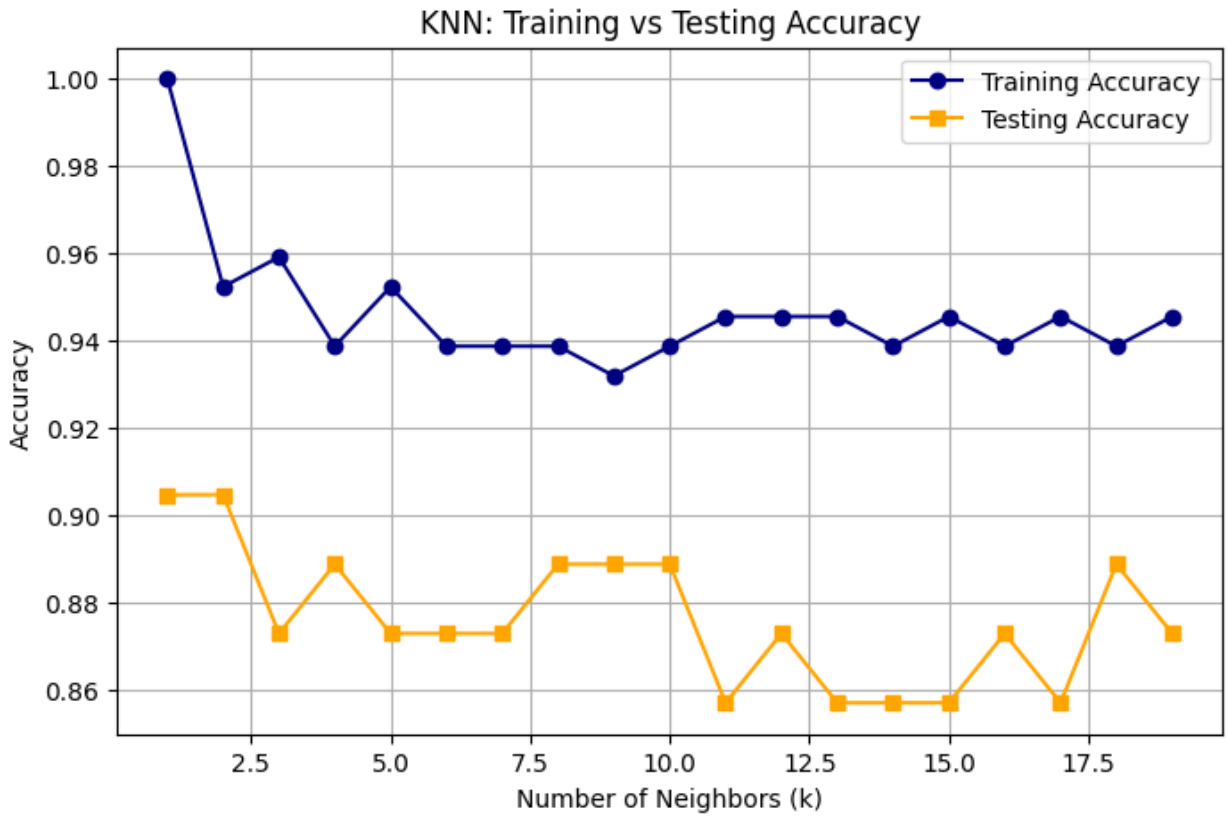
The classification report showed high precision and recall values across all three classes, suggesting that the decision tree effectively separated the classes based on feature importance.

2. K-Nearest Neighbors (KNN)

The KNN model achieved an accuracy of 87.30%.

- classification Report:
  Precision: 0.87 (macro avg)
  Recall: 0.87 (macro avg)
  F1-Score: 0.87 (macro avg)

The accuracy vs. number of neighbors plot showed:



*A sharp accuracy increase from 1 to 2 neighbors.

* Accuracy stabilized between 90-95% after selecting more features, but adding more neighbors beyond 7-10 led to potential overfitting.

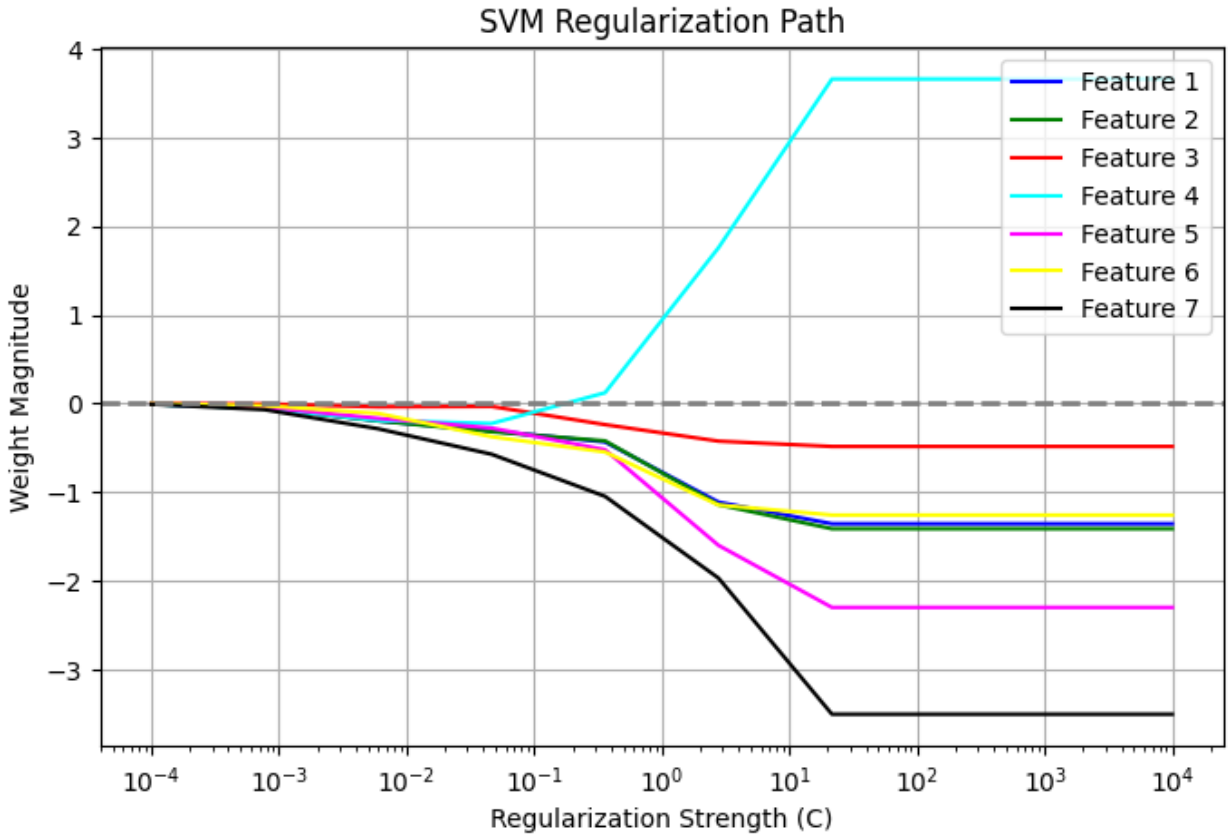
This suggests that using fewer, more significant features improves model accuracy, but increasing neighbors too many leads to diminishing returns.

3. Support Vector Machine (SVM)

The SVM model also achieved an accuracy of 87.30%.

- Classification Report:
  Precision: 0.87 (macro avg)
  Recall: 0.87 (macro avg)
  F1-Score: 0.87 (macro avg)

From the SVM Regularization Path Plot, we observed:



* Weight coefficients change as regularization strength (C) increases.

* Some features dominate at low regularization strengths, while others stabilize as C increases.

* The black and cyan features had the highest coefficient variation, indicating strong feature importance. And

* The SVM model demonstrated the effect of regularization strength (C) on the weight coefficients, showing that a non-linear SVM was necessary due to the overlapping data between classes.

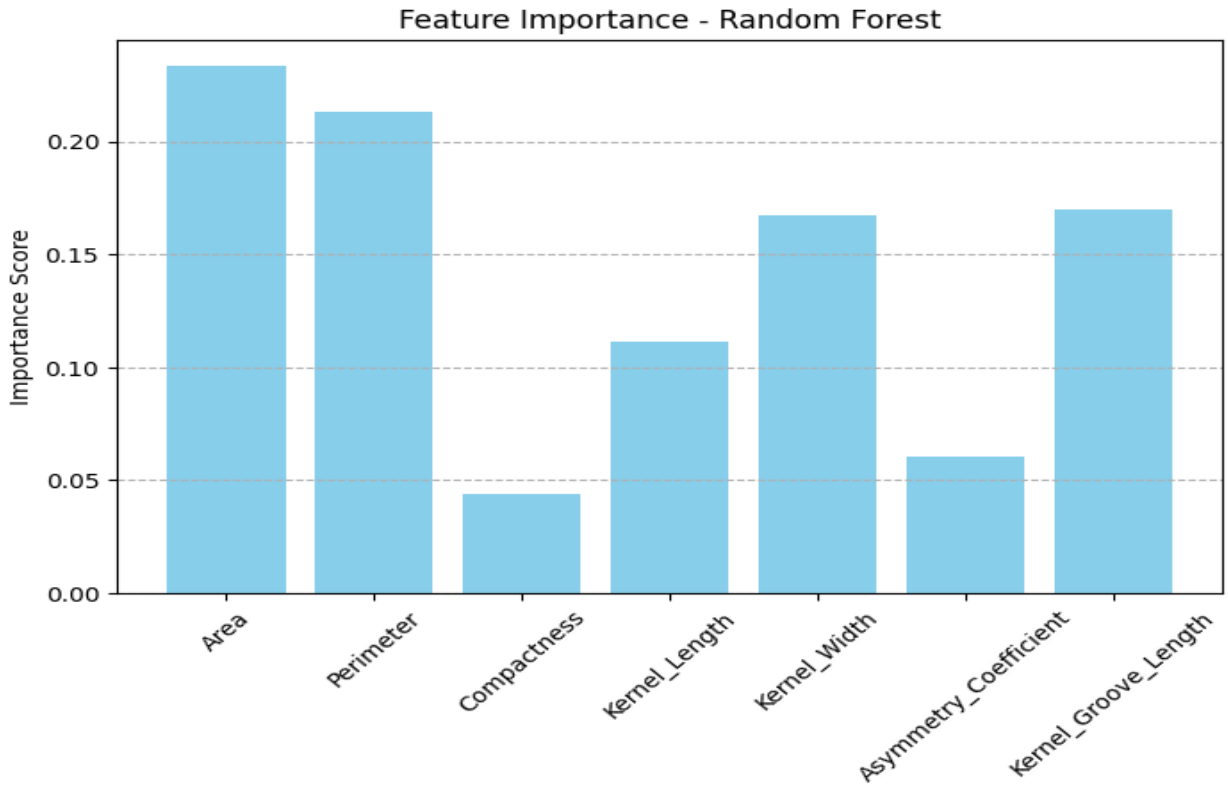
This confirms that SVM requires proper tuning of the regularization parameter C to achieve the best classification results.

4. Random Forest Classifier

The Random Forest model classified the data with 88.89% accuracy.

- Classification Report:
  Precision: 0.89 (macro avg)
  Recall: 0.89 (macro avg)
  F1-Score: 0.89 (macro avg)

From the feature importance plot, we noted:



* Perimeter and Area were the most influential features.

* Compactness had the least impact.

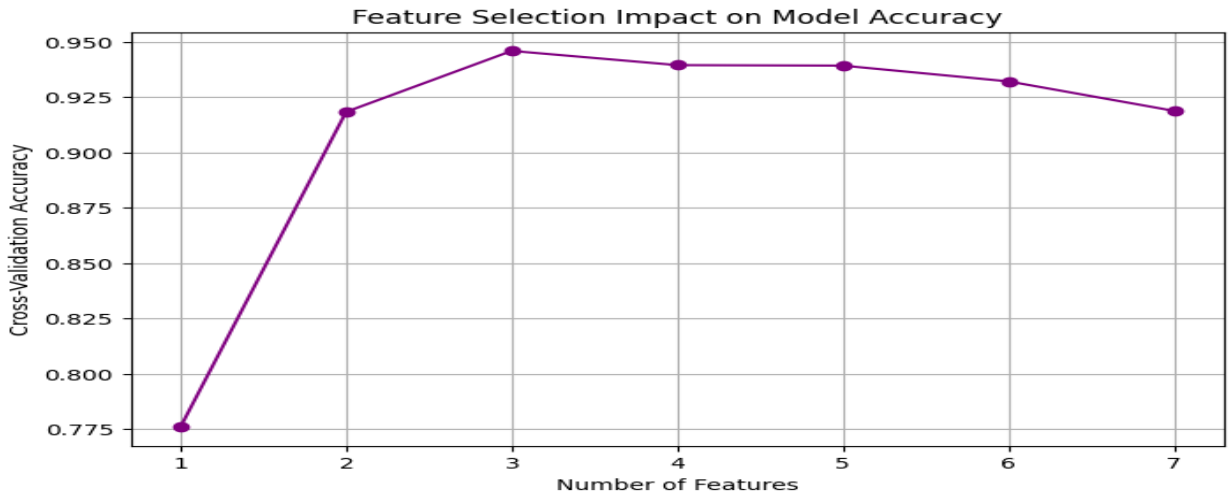
* Feature importance was evenly distributed, showing that multiple features contributed to classification.

Since Random Forest averages predictions from multiple decision trees, it provided more stable and generalized results compared to the single decision tree model.

5. Feature Selection Analysis

We performed feature selection using Sequential Feature Selector (SFS) to analyze how the number of features impacts model accuracy.

From the Feature Selection Impact Plot, we found that:



* Using only 2-3 features resulted in the highest accuracy (~95%).

* Adding more features beyond 4 did not significantly improve accuracy.

* After 7 features, accuracy slightly dropped, indicating potential overfitting.

This suggests that reducing the number of features can improve model performance by removing unnecessary complexity.

Conclusion: The analytical insights reveal that Decision Trees and Random Forests achieved the highest accuracy among the models. Feature selection played a critical role in model performance, with fewer features often leading to better accuracy. The SVM model highlighted the necessity of non-linear decision boundaries due to the overlapping nature of the data. Overall, the analysis underscores the importance of feature selection and regularization in achieving accurate classifications.