**Report on Model Performance: Decision Tree vs. Random Forest**

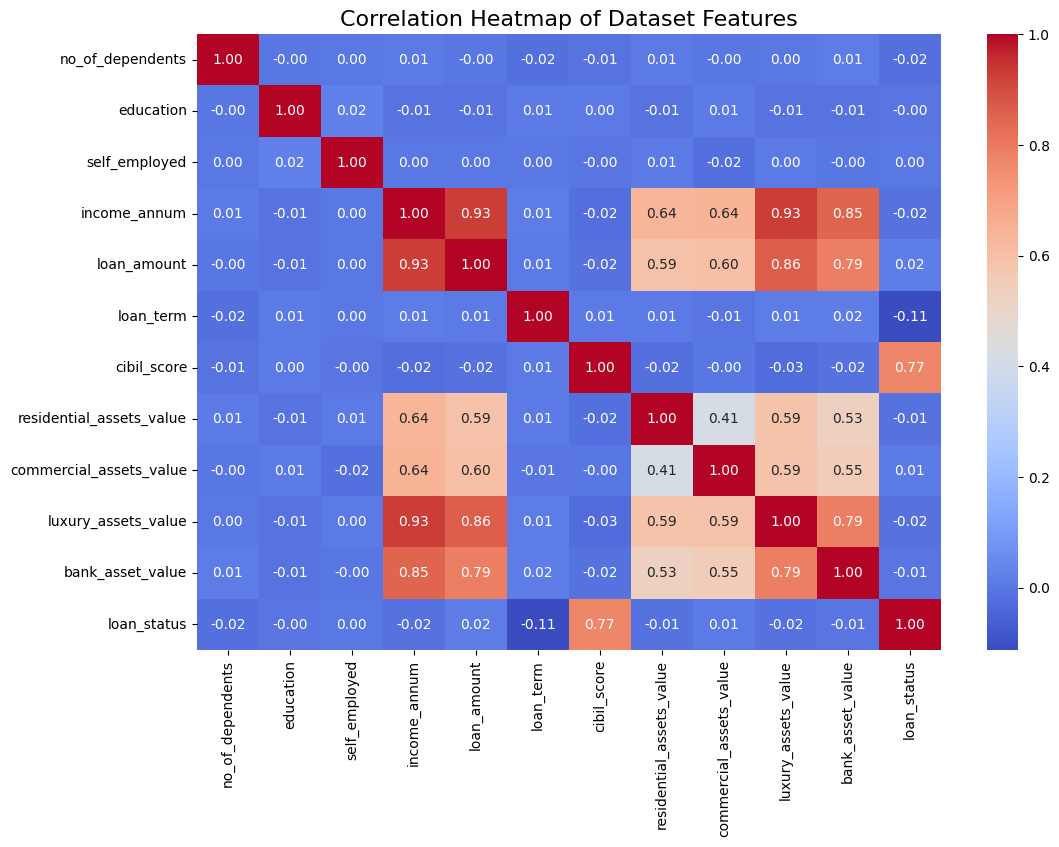
#### **Objective**

The goal was to evaluate the performance of two machine learning models, Decision Tree and Random Forest, on a binary classification task (predicting loan approval status). The evaluation metrics included **precision**, **recall**, **F1-score**, and **accuracy**.

The first step after cleaning and data processing was that I ran a correlation analysis in order to get a better understanding of the relationships between the data points.

#### 

#### **Correlation Test:**



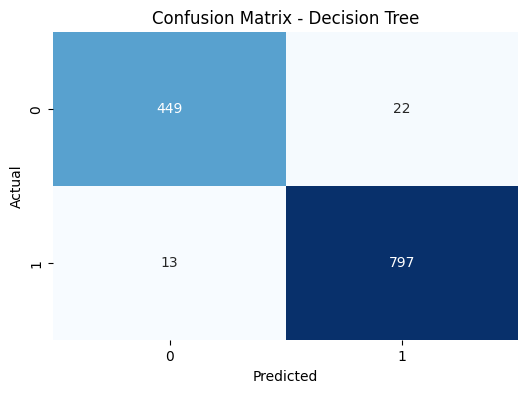
#### 

#### 

#### **Model 1: Decision Tree**

#### **Performance Metrics**

* **Precision**:
  + Class 0 (Rejected): 0.97
  + Class 1 (Approved): 0.97
* **Recall**:
  + Class 0 (Rejected): 0.95
  + Class 1 (Approved): 0.98
* **F1-Score**:
  + Class 0 (Rejected): 0.96
  + Class 1 (Approved): 0.98
* **Overall Accuracy**: 97.27%



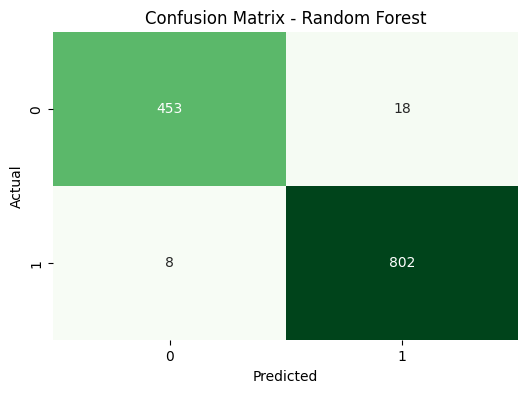
#### **Some Observations**

* The model performed well overall, achieving high precision and recall for both classes.
* It had slightly lower recall for rejected loans (0.95), which means it missed a small proportion of rejected cases.
* The weighted average of F1-scores (0.97) demonstrates strong overall balance between precision and recall.

### **Model 2: Random Forest**

#### **Performance Metrics**

* **Precision**:
  + Class 0 (Rejected): 0.98
  + Class 1 (Approved): 0.98
* **Recall**:
  + Class 0 (Rejected): 0.96
  + Class 1 (Approved): 0.99
* **F1-Score**:
  + Class 0 (Rejected): 0.97
  + Class 1 (Approved): 0.98
* **Overall Accuracy**: 97.97%



#### **Some Observations**

* Random Forest outperformed the Decision Tree in overall accuracy (97.97% vs. 97.27%).
* It exhibited slightly better precision and recall for the approved loans class (Class 1), indicating a better ability to correctly classify loans as approved.
* The model is robust and reduces overfitting by combining predictions from multiple trees.

**Comparison and insights on performance:**

Both models performed very well, with the Random Forest slightly better than the Decision Tree across all metrics.

Random Forest demonstrated higher overall accuracy and F1-score due to its ensemble approach, resulting in lower variance when compared to a just one decision tree.

**Precision versus Recall:**

Both models showed a balanced distinction between precision and recall.

Random Forest had better recall for approved loans (Class 1), which reduced the likelihood of missing true positives in this category.

**Model Selection:**

Random Forest is the best option in this situation because of its high accuracy and ability to make assumptions to previously unseen data.

### **Insights from Decision Tree Hyperparameter Tuning**

#### **Best Parameters**

The most effective parameters found through hyperparameter tuning for the Decision Tree model using GridSearchCV are listed below:

Criterion: "entropy"

The entropy criterion is used by the decision tree to evaluate the quality of splits, and the best split is determined based on information gain.

max\_depth: None.

The decision tree is allowed to grow indefinitely with no depth restrictions, which means it grows until all leaves are pure or contain fewer than min\_samples\_split samples.

Minimum number of leaf samples: 2.

Each leaf node must contain at least two samples. This prevents overfitting by ensuring that no leaf contains a single sample.

Minimum sample split: 10.

A node must have at least ten samples before considering further splitting. This parameter aids in tree complexity management and prevents overfitting.

Performance with best parameters

Best cross-validated accuracy: 98.16%.

This represents the average accuracy achieved during cross-validation. It indicates that the model is performing really well on the training data, with high generalizability across all folds.

### **Insights from Random Forest Hyperparameter Tuning**

#### **Best Parameters**

The best parameters found through hyperparameter tuning for the Random Forest model using GridSearchCV are listed below:

Number of estimators: 100.

The number of decision trees in a random forest. A higher value improves averaging and reduces variance, but at the expense of increased computational time.

max\_depth: None.

This allows the trees to grow fully, without regard for depth, ensuring that each tree captures as much complexity as required.

Minimum number of leaf samples: two.

Ensures that every leaf node has at least two samples. This helps to prevent overfitting by avoiding overly specific splits.

Minimum number of samples split: two.

A node requires at least two samples to consider splitting further, giving the model flexibility in constructing splits.

#### **Performance with Best Parameters**

* **Best Cross-Validated Accuracy:** 98.03%  
  This represents the mean accuracy across folds during cross-validation, indicating strong performance and generalization of the Random Forest model.

### **Comparison Between Decision Tree and Random Forest**

#### **Accuracy**

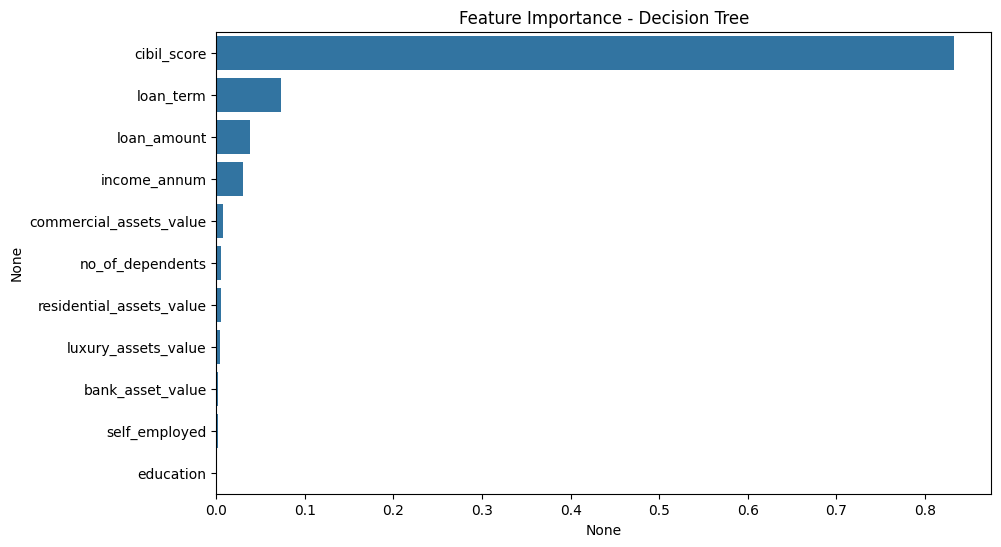
Both models performed similarly during cross-validation, achieving high accuracy scores. However, **Random Forest** is typically more robust in real-world applications due to its ensemble nature, which averages predictions across multiple decision trees, reducing the risk of overfitting and improving generalization.

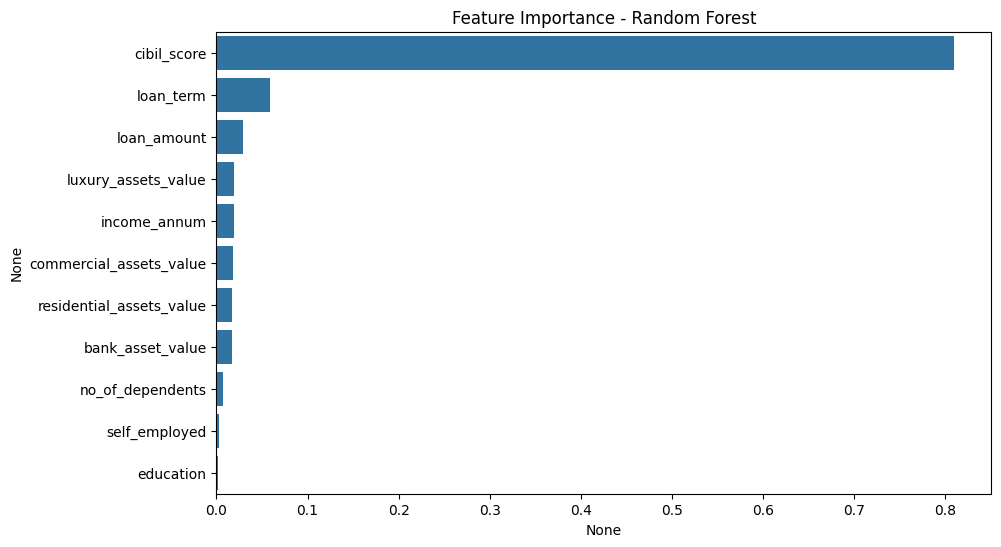
#### **Complexity**

* **Decision Tree**: Simple and interpretable but prone to overfitting, especially with complex datasets.
* **Random Forest**: Inherently more complex due to the ensemble of multiple trees. This increases computational overhead but significantly improves stability and predictive performance.

#### **Feature Importance**

Random Forest provides more reliable and nuanced feature importance scores due to its averaging mechanism across multiple trees. In this analysis, the **most important feature was credit\_score**, highlighting its critical role in predicting loan approval outcomes.\





### **Conclusion**

The tuned Random Forest model is an excellent candidate for deployment. It combines **high accuracy** with **robustness**, making it more reliable than a standalone Decision Tree for predicting loan approval outcomes. Additionally, its identification of credit\_score as the most influential feature provides actionable insights for focusing future efforts on refining credit evaluation processes.