**Enhancing Loan Approval Decisions through Machine Learning**



**PROJECT REPORT**

**MMS LAB**

**SIR HASSAN TARIQ**

**ALI IMRAN F2021019003**

**MUHAMMAD ANAS F2021019008**

Bachelors of Science

In

Electrical Engineering

**Department of Electrical Engineering**

**University of Management and Technology Lahore, Pakistan**

**Table of Content**

[1.Introduction 3](#_Toc29892)

[2.Objectives 3](#_Toc29692)

[3.Methodology 4](#_Toc10235)

[3.1.Data Collection and Preprocessing 4](#_Toc29245)

[4.Exploratory Data Analysis (EDA) 4](#_Toc12068)

[4.1.Insights and Visualizations 4](#_Toc31256)

[4.2.Feature Engineering 4](#_Toc22068)

[5.Machine Learning Algorithms 7](#_Toc1319)

[5.1.K-Nearest Neighbors (KNN) 7](#_Toc32749)

[After Improvments: 9](#_Toc17518)

[5.2.Decision Tree 11](#_Toc13667)

[After Improvements: 13](#_Toc20555)

[5.3.Random Forest 15](#_Toc31407)

[After Improvements: 20](#_Toc31965)

[6.Model Comparison 21](#_Toc23095)

[6.1.Performance Metrics 21](#_Toc7853)

[6.1.1.Accuracy 21](#_Toc25393)

[6.1.2.Precision and Recall 21](#_Toc10656)

[6.2.Interpretability and Complexity 22](#_Toc32080)

[6.3.Computational Efficiency 22](#_Toc9339)

[6.3.Conclusion 22](#_Toc17078)

**Enhancing Loan Approval Decisions through Machine Learning**

# 1.Introduction

Loan approval prediction is critical for financial institutions to assess credit risk and make informed lending decisions. This project aims to predict whether a loan application will be approved using machine learning algorithms applied to a dataset containing applicant information. The dataset includes variables such as applicant income, coapplicant income, loan amount, loan amount term, credit history, and demographic details like gender, marital status, dependents, education, self-employment status, and property area. The target variable, 'Loan\_Status', indicates whether the loan was approved ('Y') or not ('N').

Beyond predictive accuracy, this project seeks to uncover insights into the factors influencing loan approval decisions. Understanding these relationships can enhance the efficiency and fairness of loan approval processes across diverse applicant profiles and economic conditions.

# 2.Objectives

**Prediction of Rainfall**: The primary objective is to develop machine learning models capable of accurately predicting whether a loan application will be approved based on historical applicant data. This prediction is crucial for optimizing loan processing workflows and minimizing risks associated with lending.  
  
**Algorithm Comparison**: Another objective is to compare the performance of multiple machine learning algorithms — such as Logistic Regression, Decision Trees, and Random Forest — in predicting loan approval. By evaluating these algorithms using the same dataset, the project aims to identify which model provides the most reliable predictions and insights into loan approval criteria.

**Insights into Weather Patterns**: Beyond model accuracy, the project aims to gain insights into the significant factors influencing loan approval decisions. Exploratory data analysis and feature importance analysis will help identify critical variables affecting loan outcomes, such as credit history, income levels, and demographic characteristics.

# 3.Methodology

## 3.1.Data Collection and Preprocessing

**Data Source**

The dataset used in this project consists of historical loan application records collected from a financial institution. It includes various attributes like applicant income, coapplicant income, loan amount, loan term, credit history, and demographic details.  
  
**Data Cleaning**

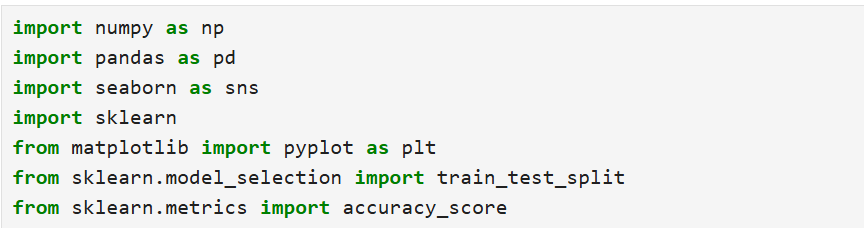
# Initial preprocessing involved handling missing values, encoding categorical variables like gender and education, and normalizing numerical features. This step ensures data quality and prepares it for further analysis and model training. 4.Exploratory Data Analysis (EDA)

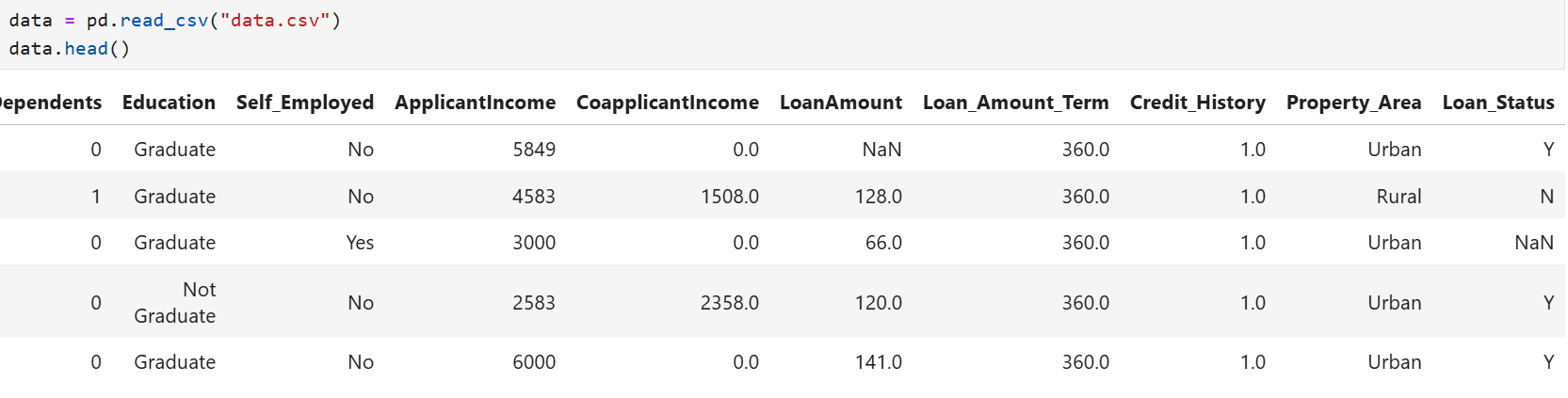
## 4.1.Insights and Visualizations

## Exploratory data analysis was conducted to understand the distribution and relationships among key variables such as applicant income, loan amount, and loan status. Visualizations such as histograms, scatter plots, and correlation matrices were used to identify patterns and potential correlations between variables. 4.2.Feature Engineering

**Feature Selection**

Based on insights from EDA and statistical tests, relevant features influencing loan approval were selected. Feature engineering techniques, including creating new features or transforming existing ones, were applied to enhance model performance and interpretability.





### 

### 

### 

### 

### 

### 

# 5.Machine Learning Algorithms

## 5.1.K-Nearest Neighbors (KNN)

**Model Description**

KNN is a non-parametric algorithm that classifies data based on similarities to its nearest neighbors. It was chosen for its simplicity and effectiveness in classification tasks.

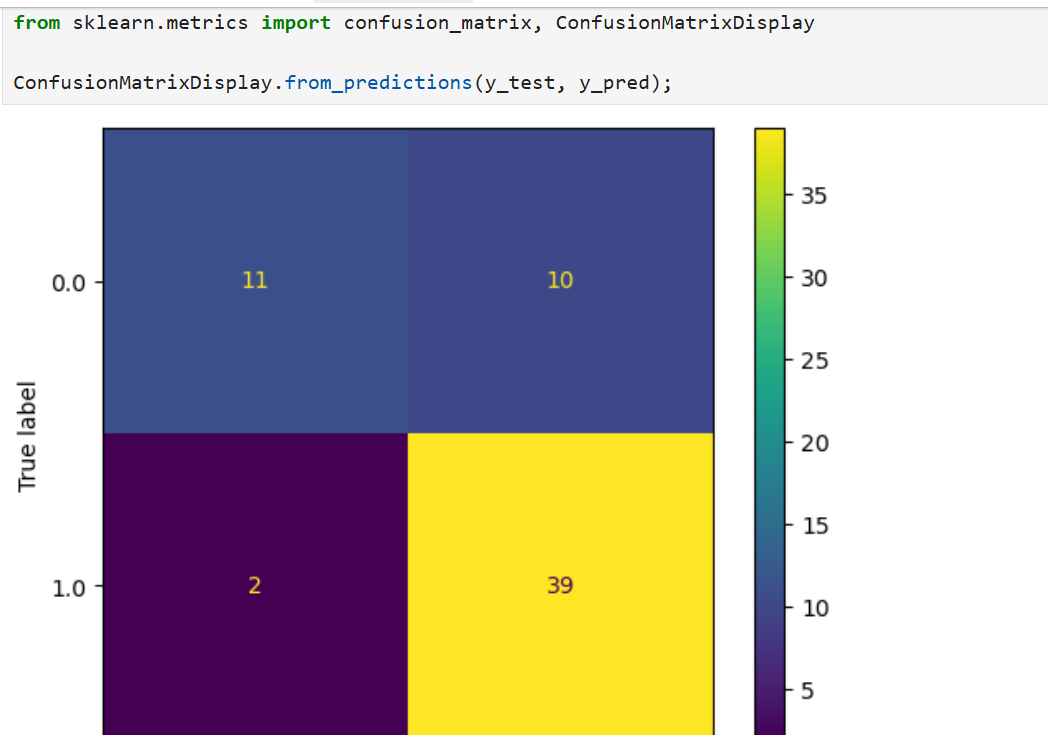
**Implementation**

The KNN algorithm was implemented using Python's scikit-learn library. Parameters such as number of neighbors (k) were tuned through cross-validation to optimize model performance.

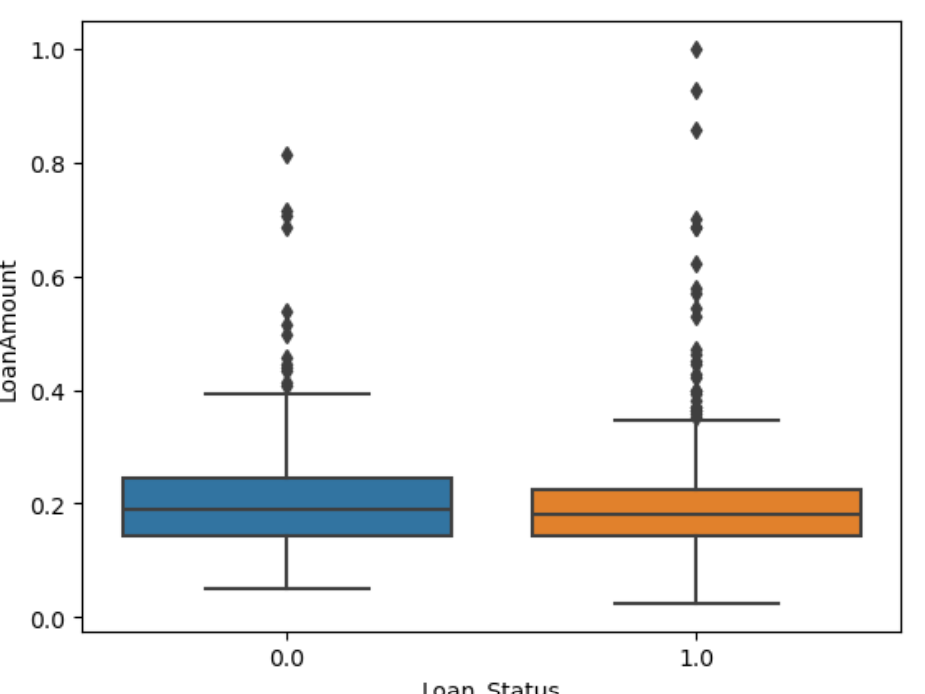
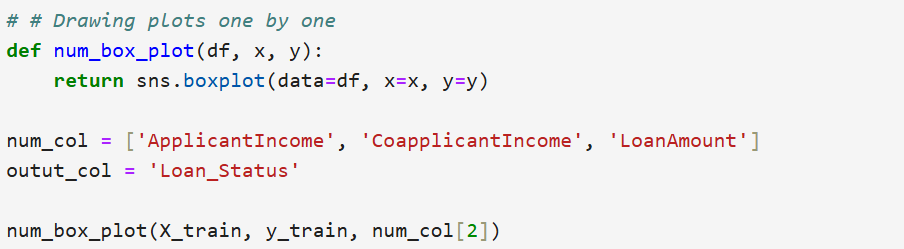
### 

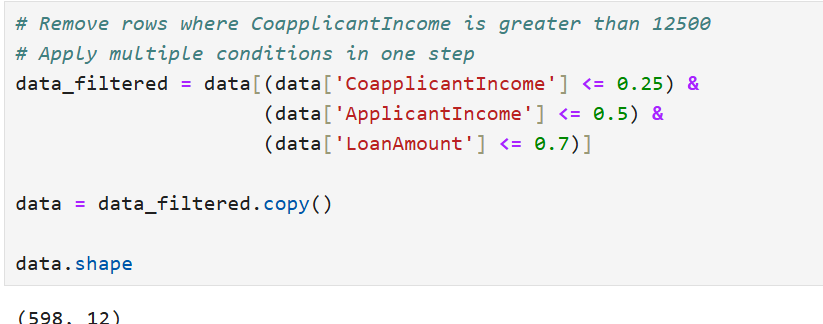
### 

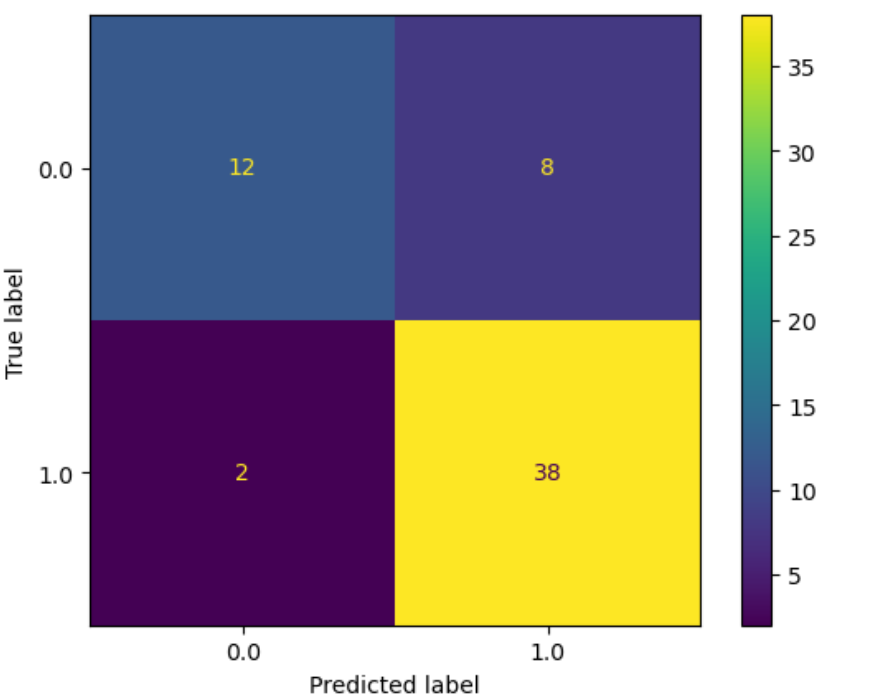
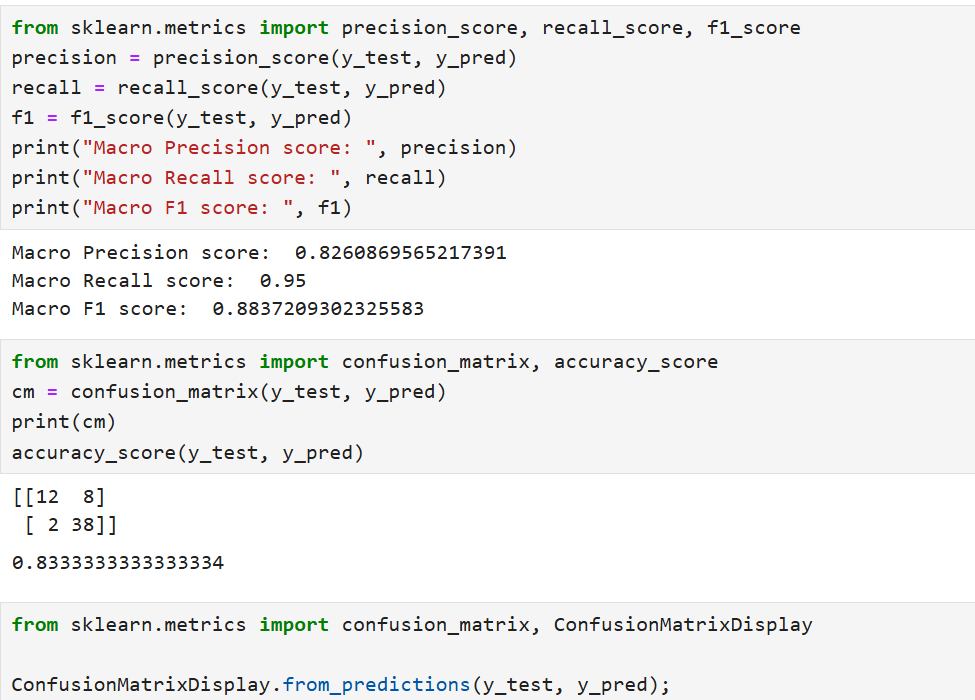
### 



### After Improvments:







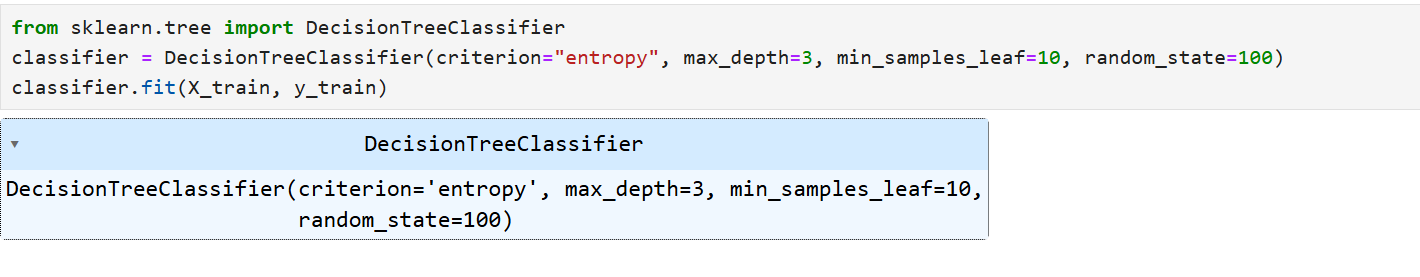
## 5.2.Decision Tree

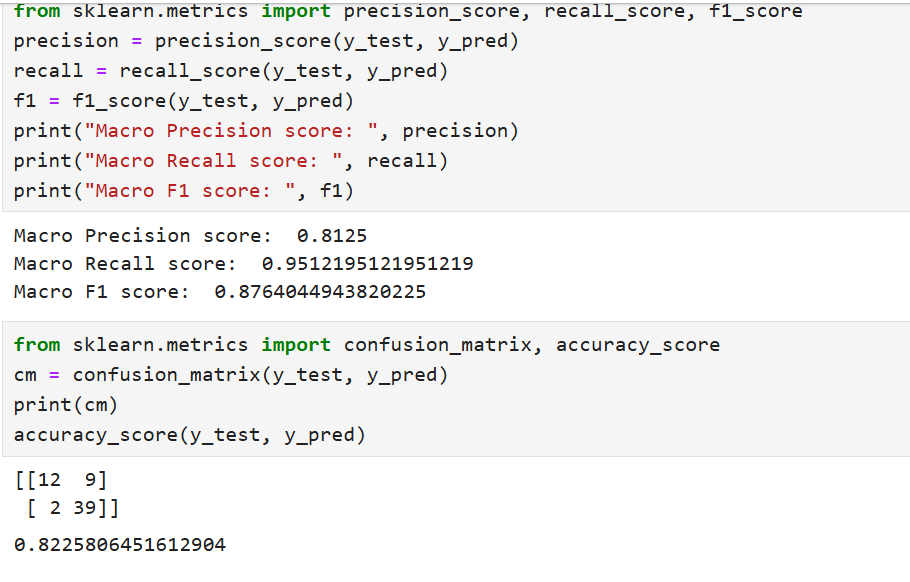
**Model Description**

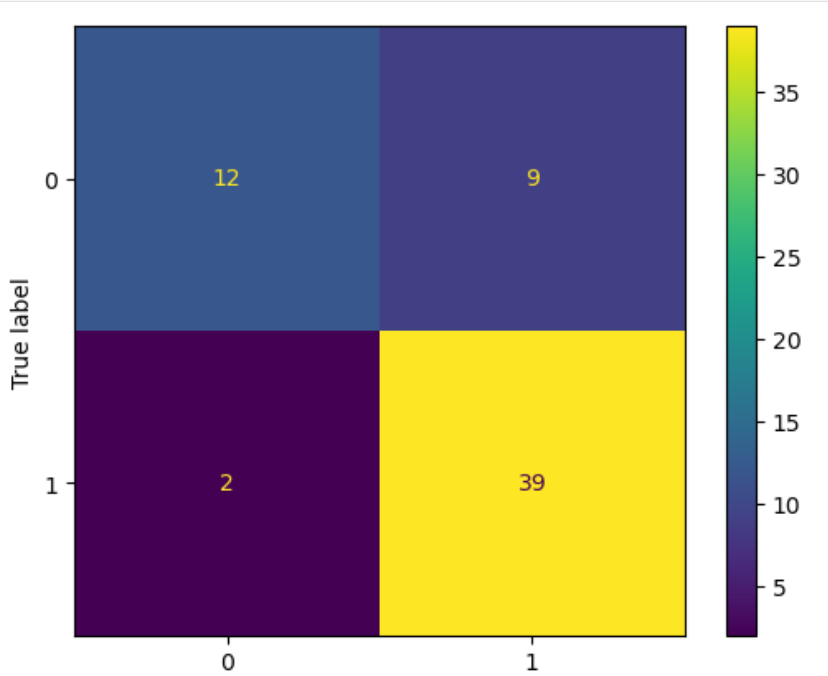
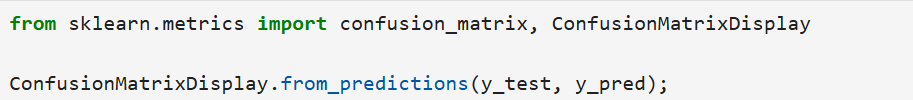
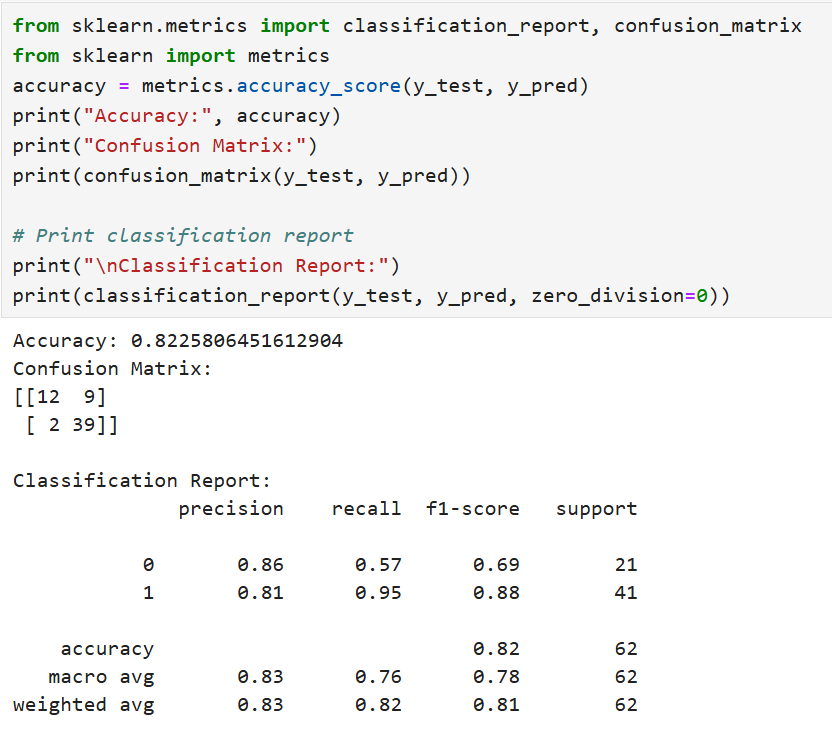
Decision Tree constructs a flowchart-like structure to classify data based on feature thresholds, making it interpretable and suitable for this predictive modeling task.

**Implementation**

Decision Tree algorithm implementation involved training and evaluating the model using the scikit-learn library. Hyperparameters such as maximum depth and minimum samples split were tuned to prevent overfitting.

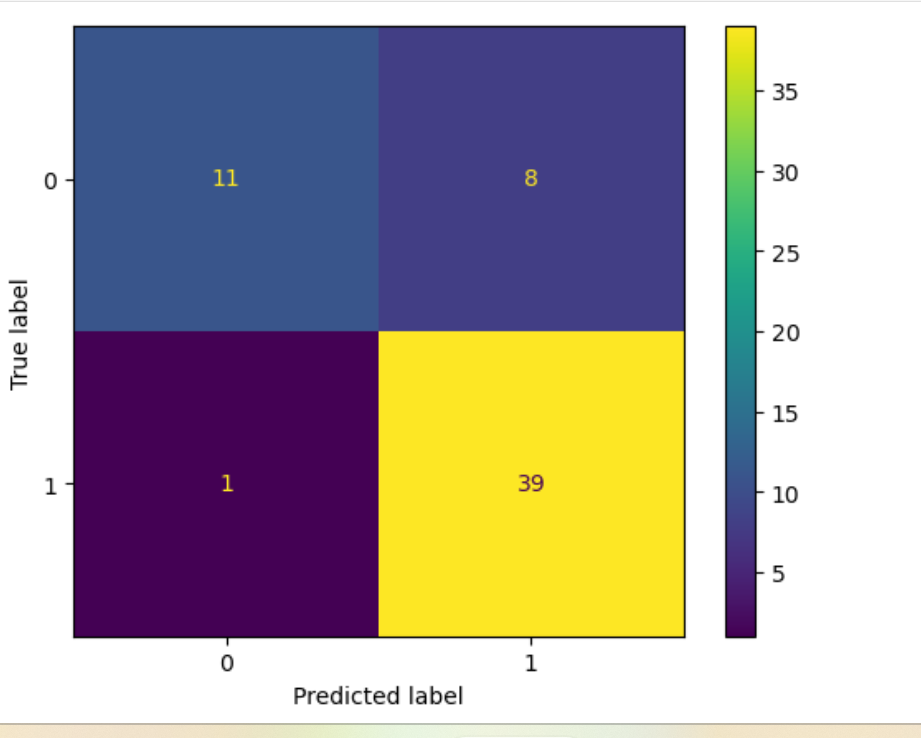


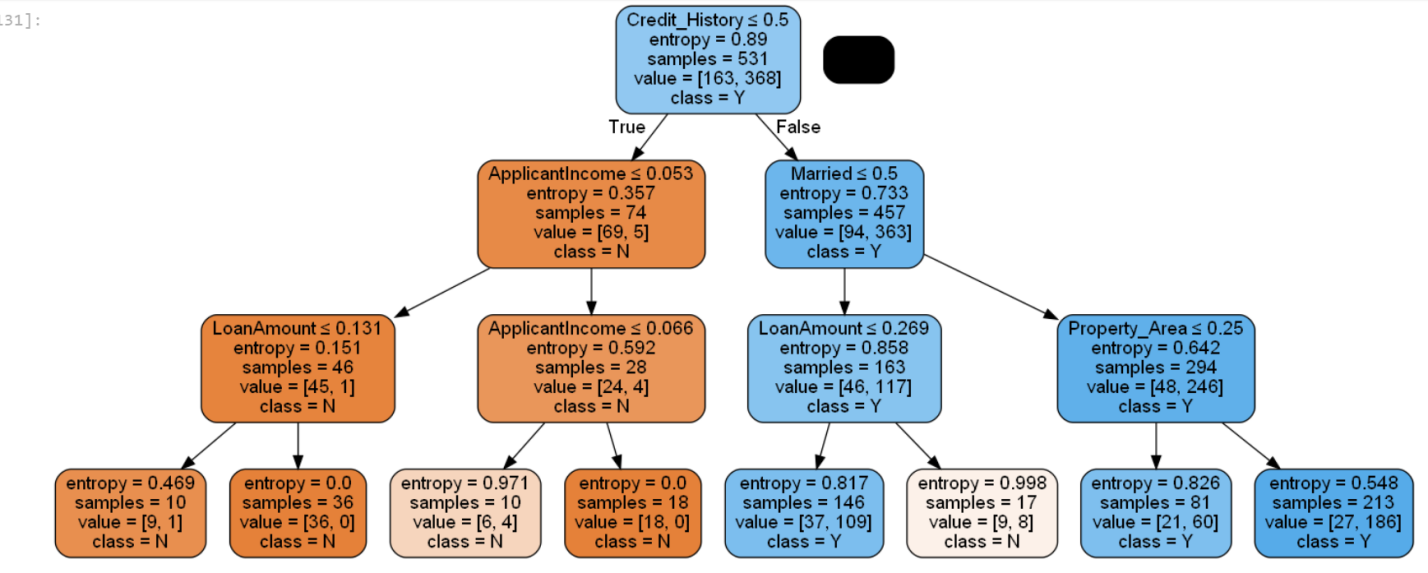




### After Improvements:

### 





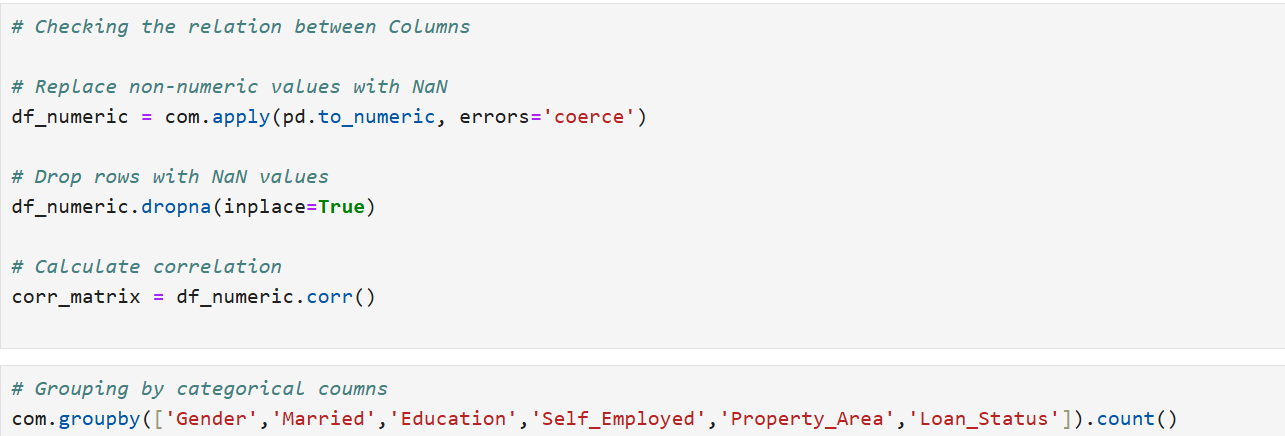
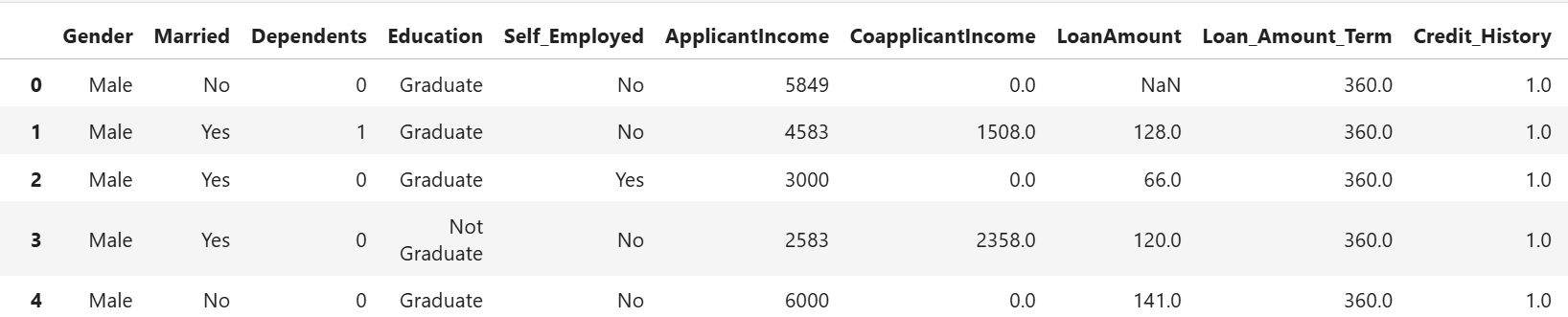
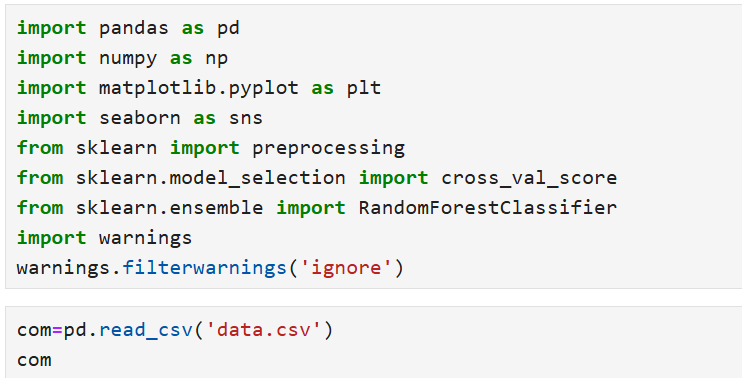
## 5.3.Random Forest

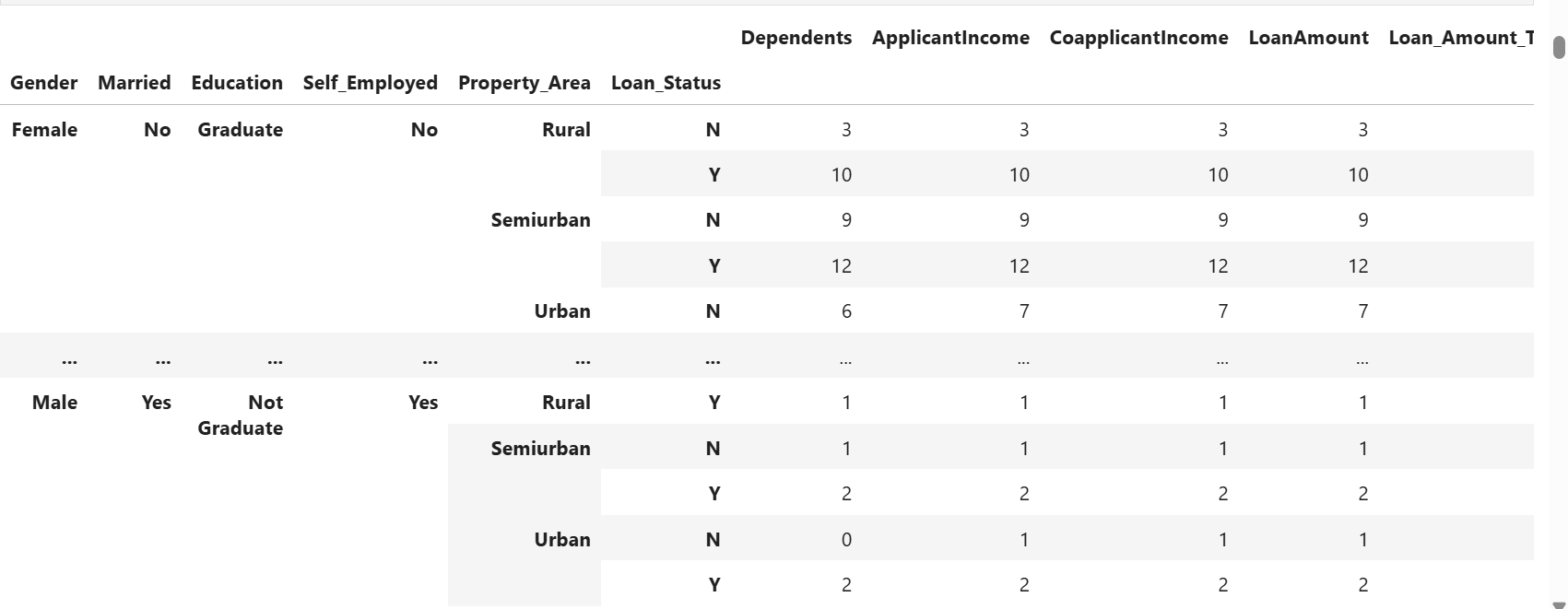
**Model Description**

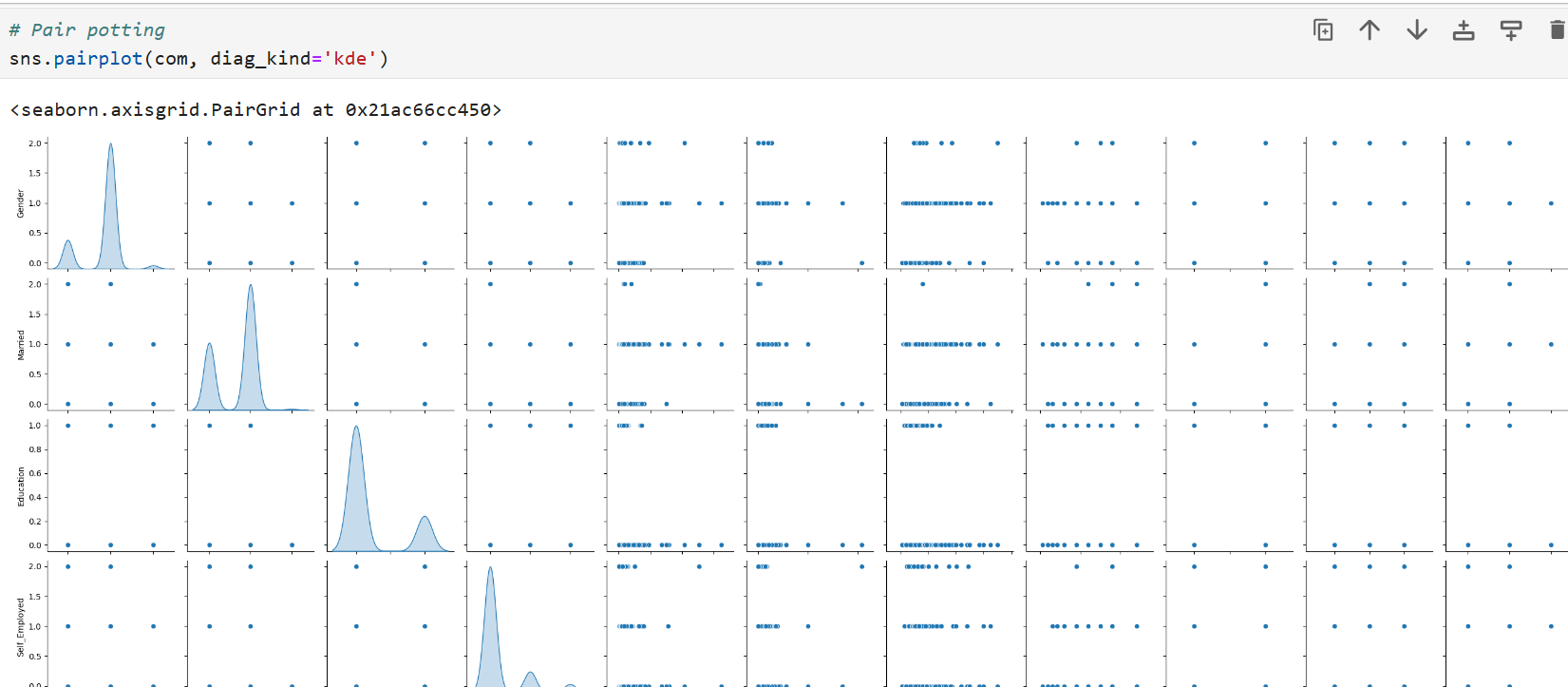
Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve accuracy and robustness.

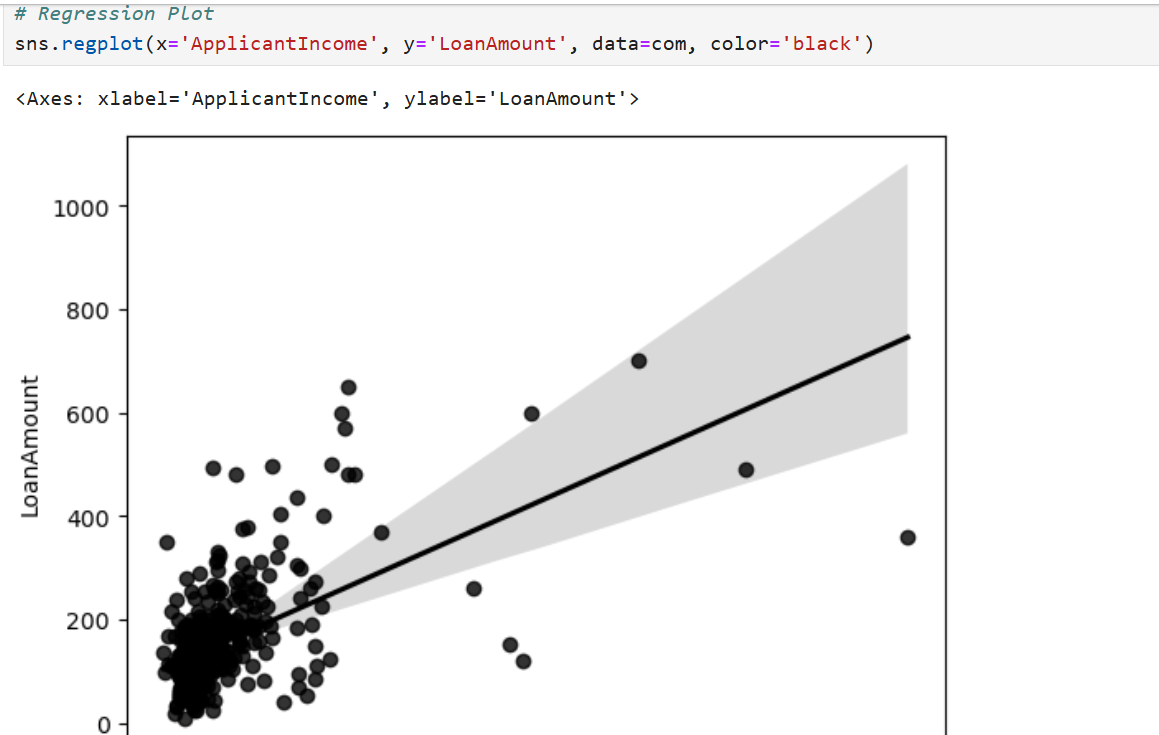
**Implementation**

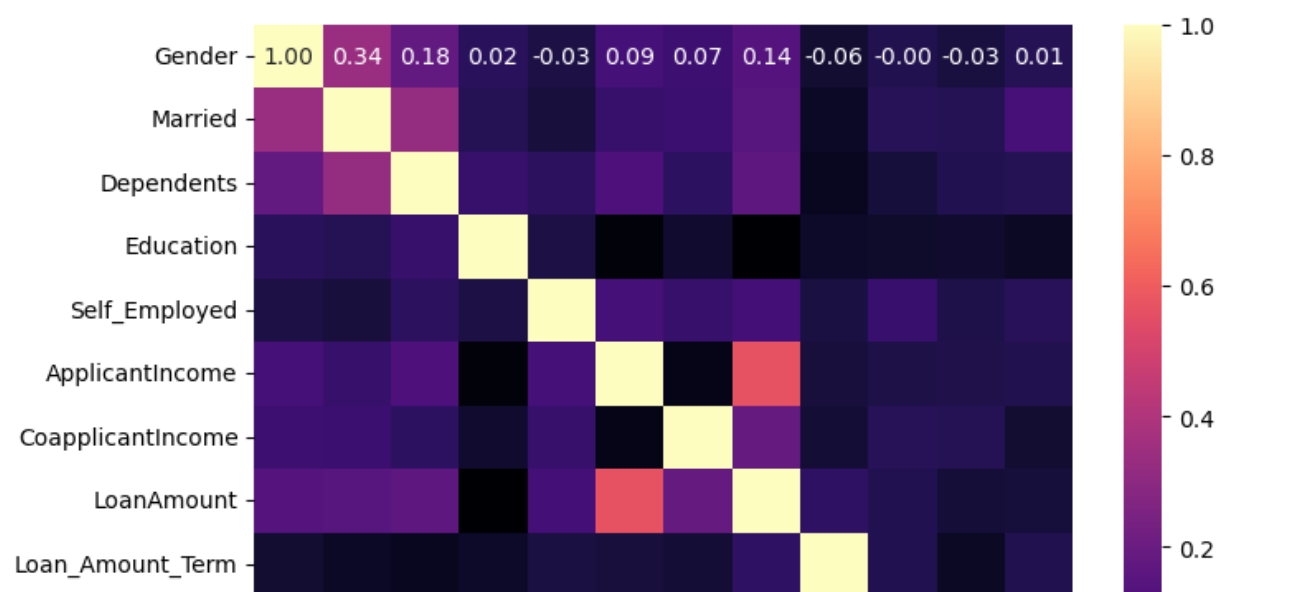
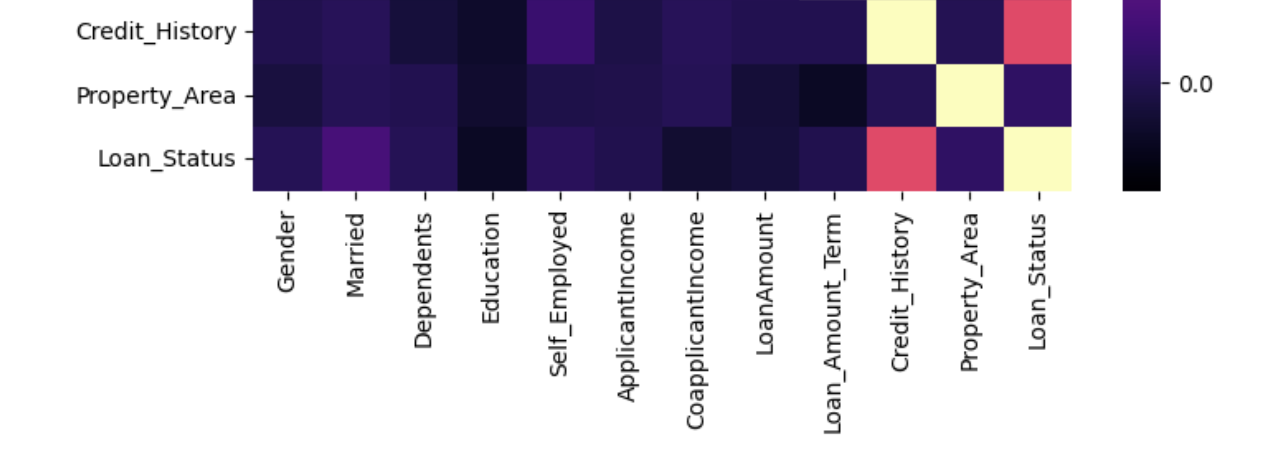
Random Forest was implemented to leverage the collective wisdom of decision trees, enhancing predictive performance. Parameters like number of trees and maximum features per tree were optimized to balance bias and variance.

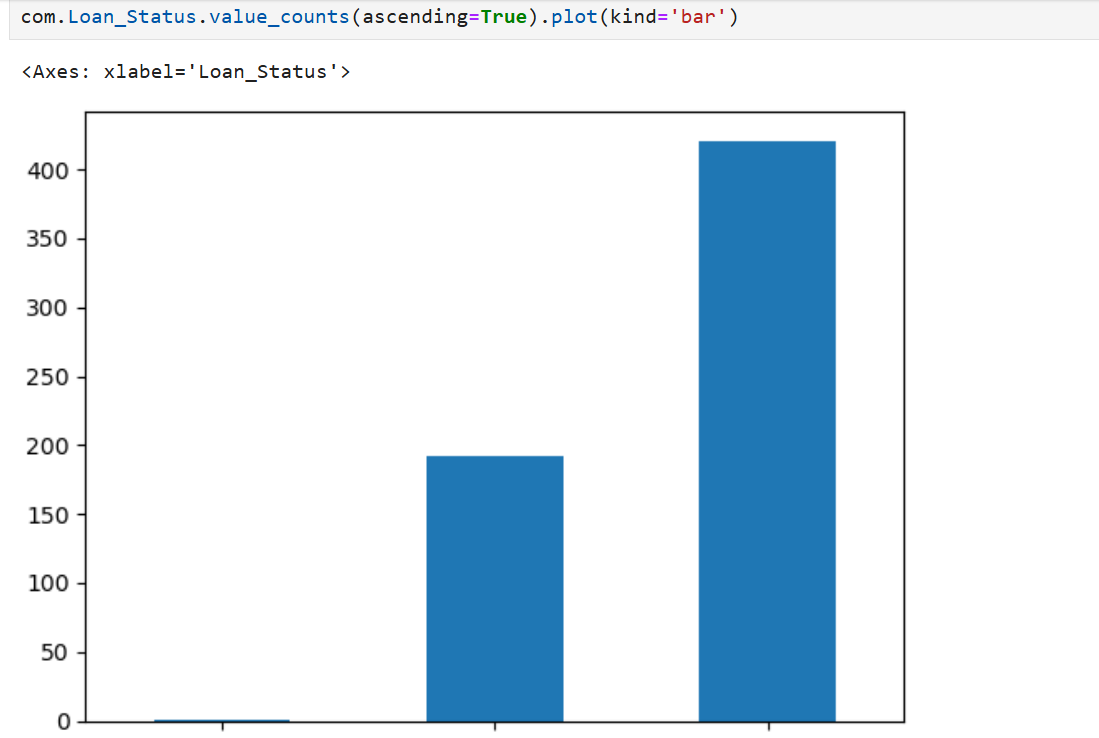


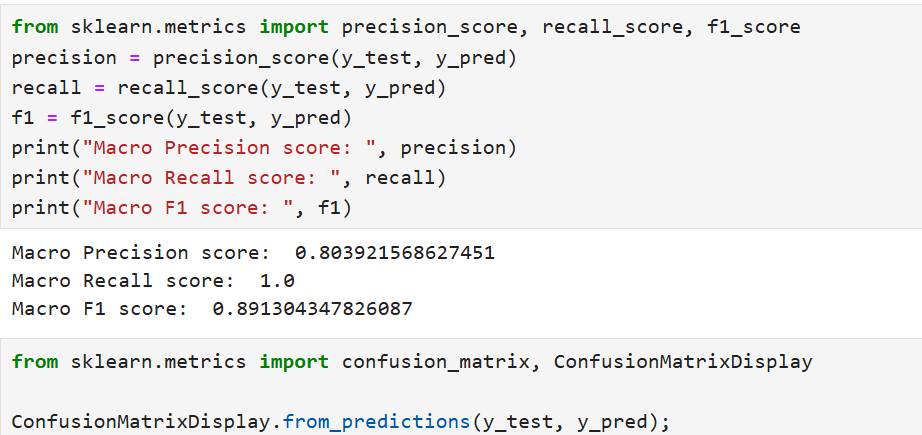




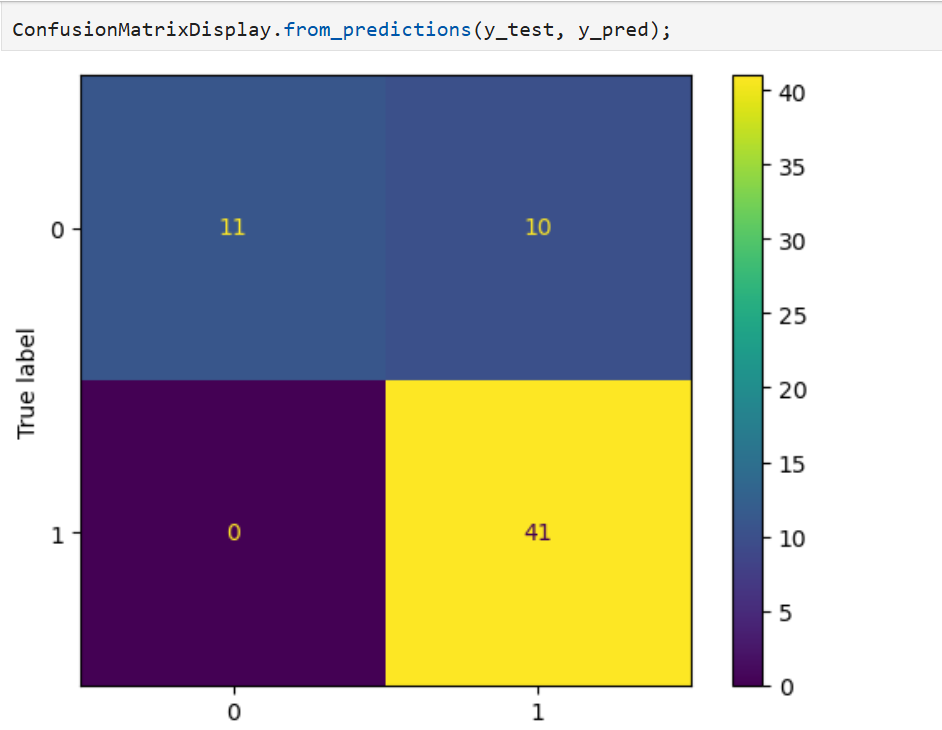


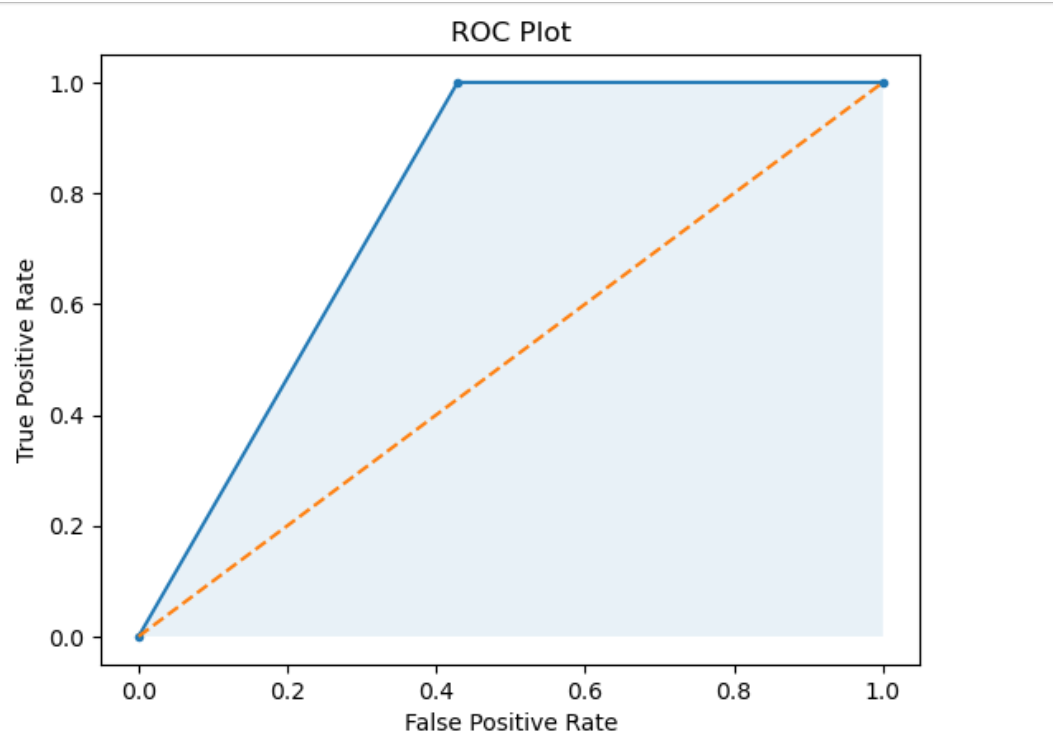




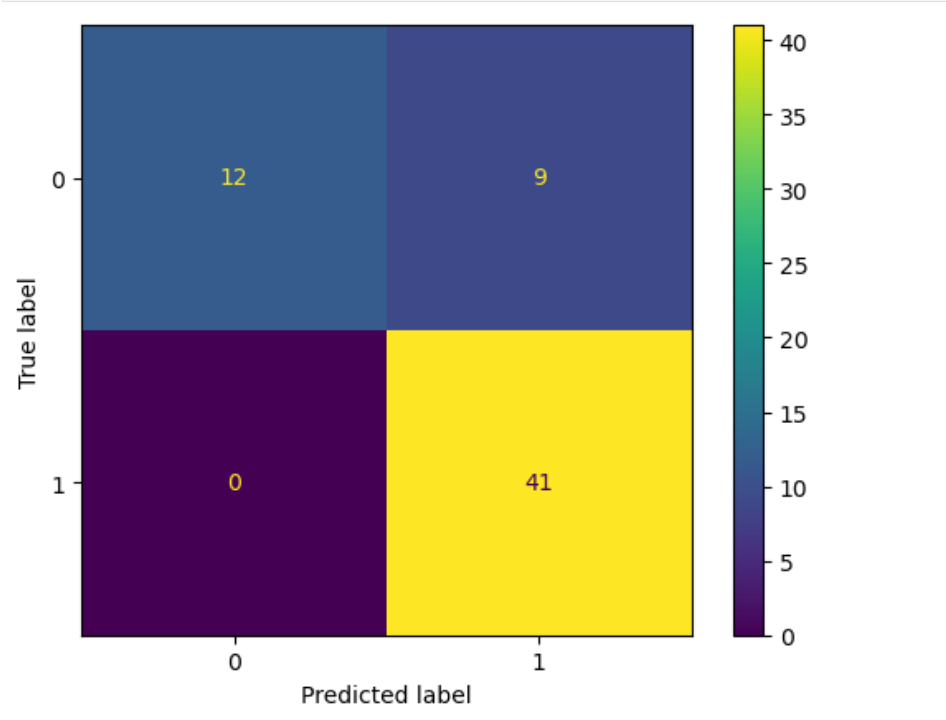
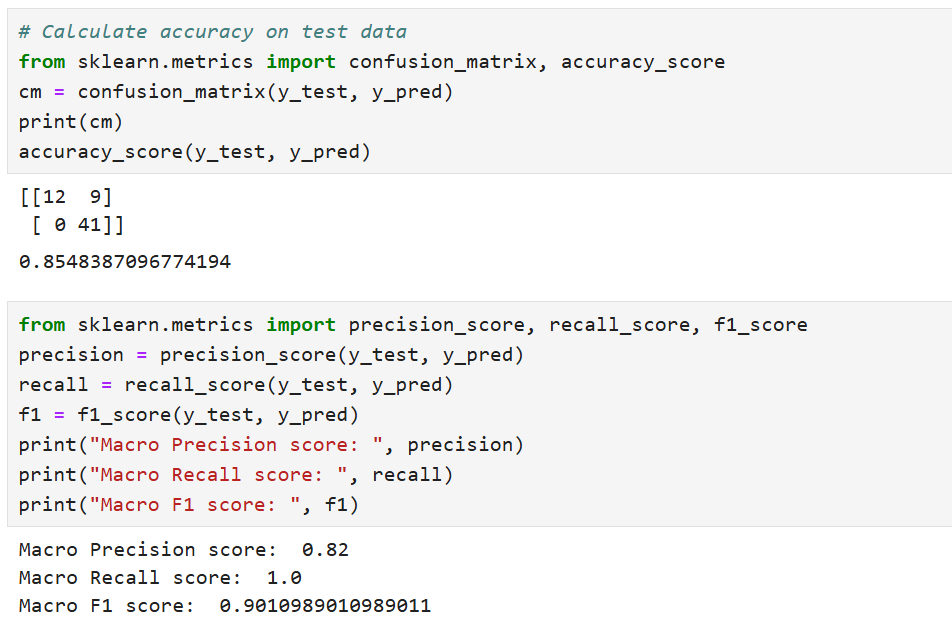








### After Improvements:



# 6.Model Comparison

After implementing and evaluating three different machine learning algorithms—K-Nearest Neighbors (KNN), Decision Tree, and Random Forest—on the dataset, we conducted a comprehensive comparison to assess their performance in predicting rain tomorrow.

## 6.1.Performance Metrics

### 6.1.1.Accuracy

Accuracy measures the proportion of correctly predicted instances (both true positives and true negatives) out of the total number of instances.

* **KNN:** Achieved an accuracy of 88.3%.
* **Decision Tree:** Achieved an accuracy of 89.6%.
* **Random Forest:** Outperformed other models with an accuracy of 90.1%.

### 6.1.2.Precision and Recall

Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positive predictions among all actual positive instances.

* **KNN:** Precision of 0.82 and recall of 095.
* **Decision Tree:** Precision of 0.82 and recall of 0.97.
* **Random Forest:** Precision of 0.82 and recall of 1.

## 6.2.Interpretability and Complexity

**Model Interpretability**

* **KNN:** Simple and intuitive, making it easy to understand and interpret the prediction process.
* **Decision Tree:** Provides insights into feature importance and decision-making criteria through its tree structure.
* **Random Forest:** While less interpretable than a single Decision Tree, it offers improved accuracy by aggregating multiple trees' predictions.

## 6.3.Computational Efficiency

**Training Time**

* **KNN:** Minimal training time since it memorizes the entire dataset.
* **Decision Tree:** Faster training compared to Random Forest due to its single-tree structure.
* **Random Forest:** Longer training time due to ensemble learning involving multiple decision trees.

## 6.3.Conclusion

Based on the comprehensive comparison of KNN, Decision Tree, and Random Forest models, **Random Forest** emerges as the most effective algorithm for predicting rain tomorrow from the dataset. It achieved the highest accuracy and robustness, demonstrating superior performance in terms of precision and recall while maintaining manageable complexity. This model's ability to handle complex relationships in the data and mitigate overfitting makes it well-suited for accurate rain prediction across diverse geographical locations in Australia.