Inspiring Excellence

CSE422: Artificial Intelligence

Project Title: Loan Approval Prediction

Table of Contents

Introduction	**Page 2**
Dataset Description	**Page 2-3**
Data pre-processing	**Page 3**
Models Used	**Page 3-4**
Model Comparison	**page4-8**
Conclusion	**Page 8**

Introduction

A loan request gets approved or rejected based on various criteria from the loan providers. In this project, we have used some machine learning models to predict the loan status of a borrower. Our dataset contains features such as education, loan amount, gender, marital status, etc. Based on these features our machine learning models predicted the eligibility of a person for the approval of the loan request.

Dataset description

We used the Loan Prediction dataset from Kaggle for our project.

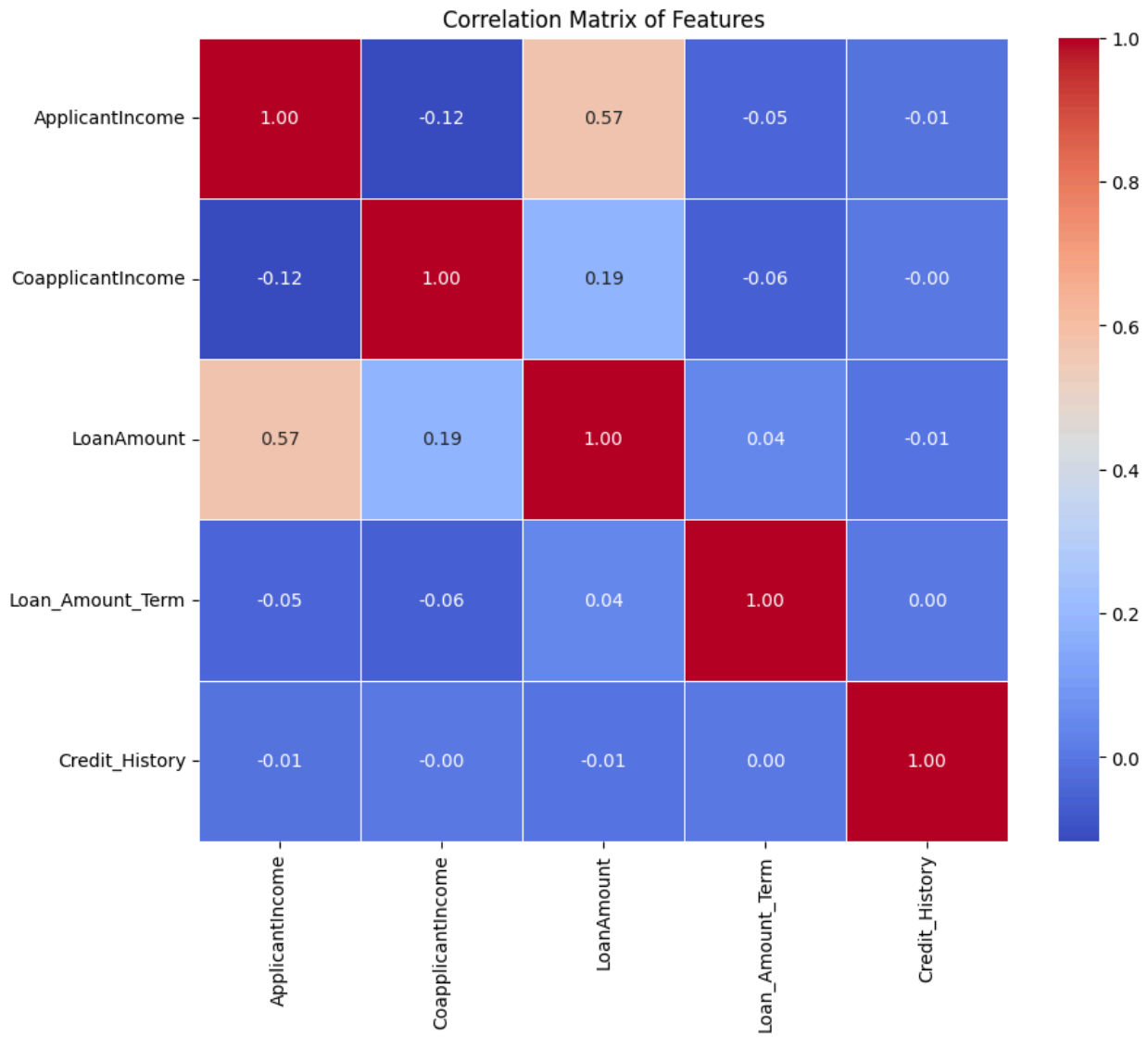
Source: https://www.kaggle.com/datasets/ninzaami/loan-predication

There are 12 features present in our dataset. Education, Property area, Applicant income, and co-applicant income are some of the key features.

This is a classification problem we've worked on. It is a classification problem because we are predicting whether the data falls under the "yes" or "no" class in the label loan status.

We've got 614 data points present in this dataset.

Features present in the dataset are mainly categorical. 7 features are categorical and the rest 5 are numerical.



This is the correlation graph of our dataset, indicating that the features are linearly dependent.

Dataset pre-processing

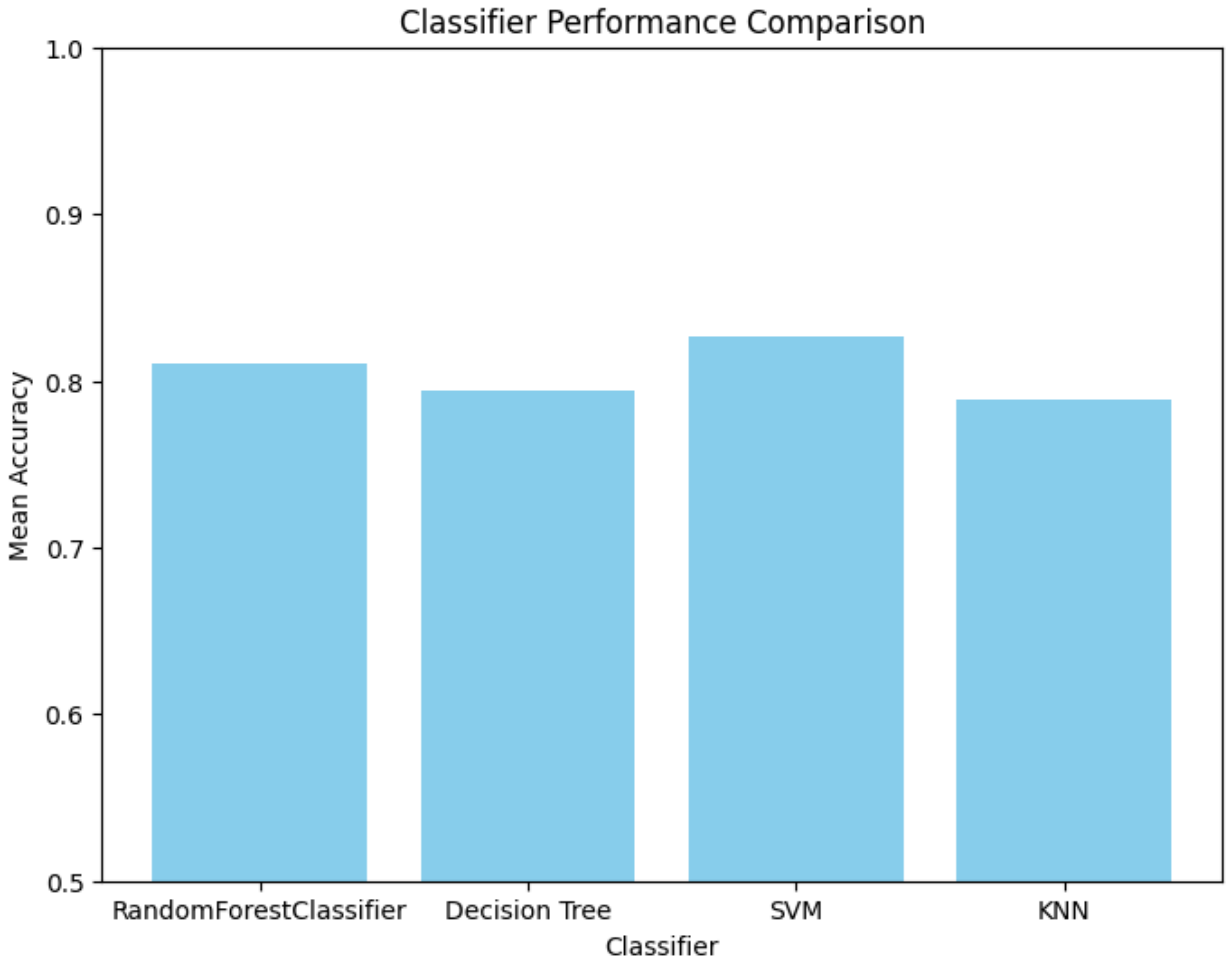
For null values, we imputed them with the mean value of that specific column or feature. And for the categorical missing values, we imputed them with the most frequent value of that feature column and applied one-hot encoder to represent them in numeric format.

Models Used

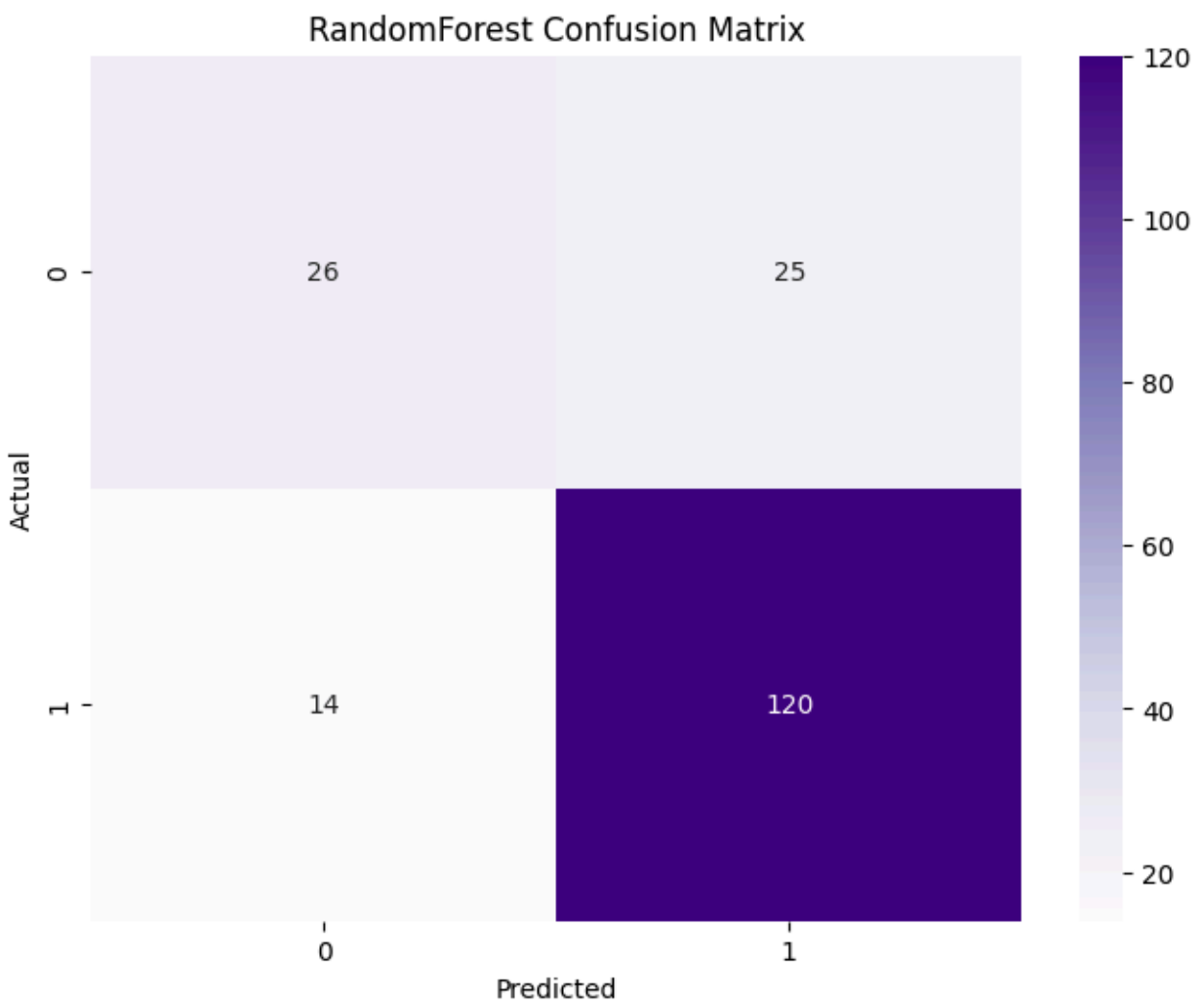
We divided our dataset into two parts, test data and train data. We splitted 70 percent of the data into train data and remaining thirty percent into test data, and splitted them randomly. We have used 4 machine learning models to predict the result. We have used Random Forest Classifier, SVM, KNN, and Decision Tree algorithms to predict our outcomes.

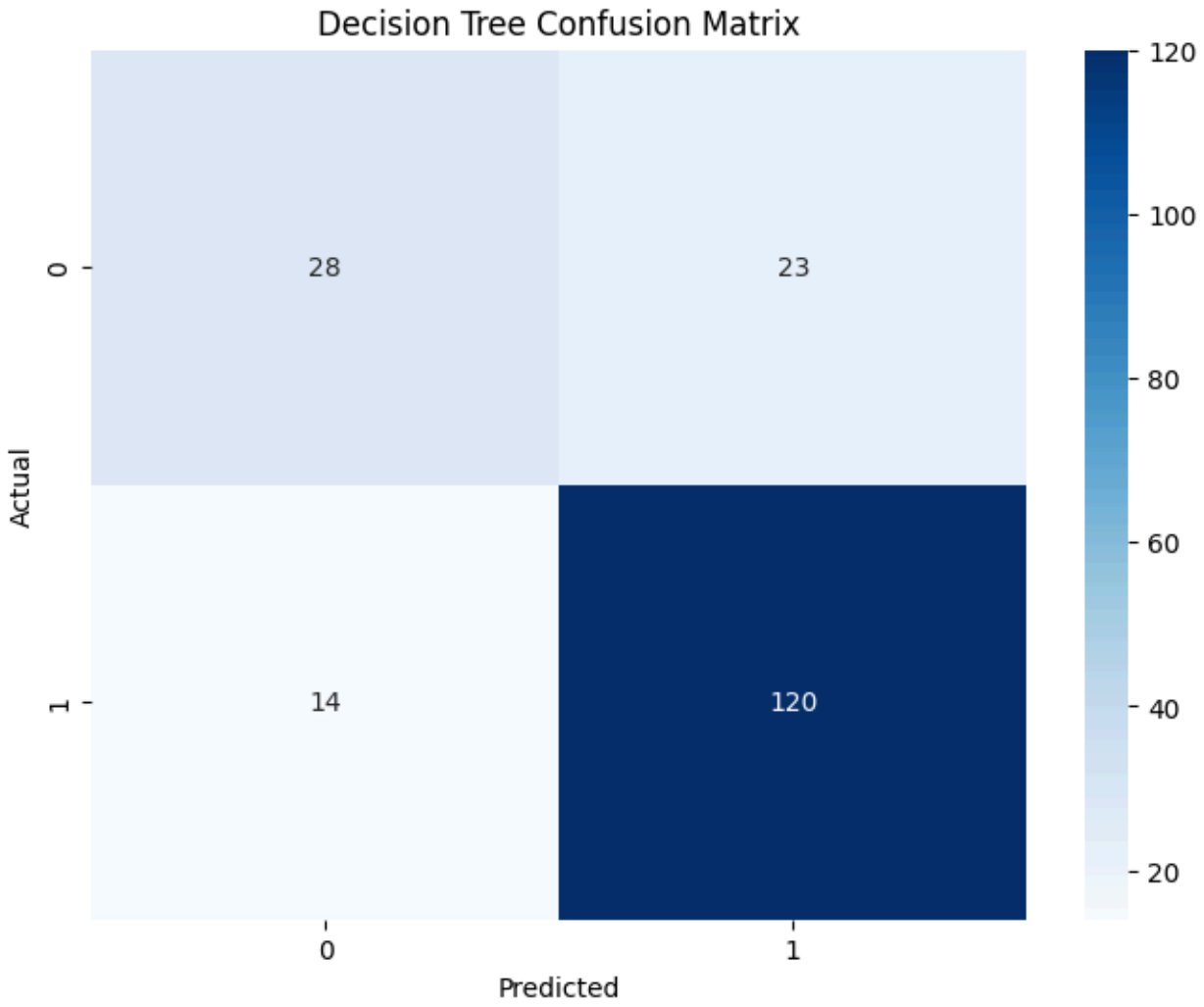
For Random forest classifier, out estimator value was 10, indicating that it used 10 decision trees to get its prediction value. For svm, we used svc because it is a classification problem and svc is used for creating the best separating hyperplane between classes.

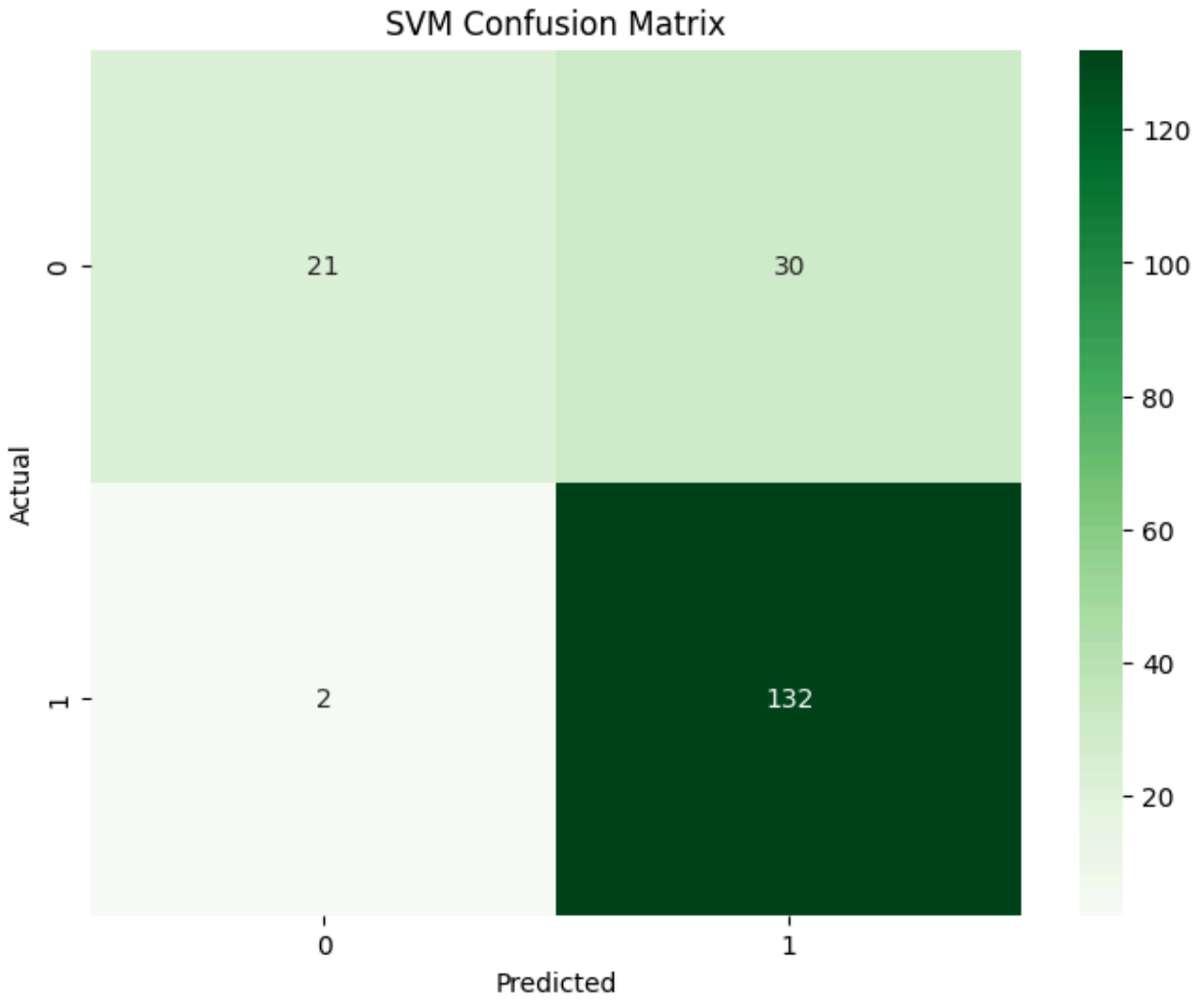
Model Comparison

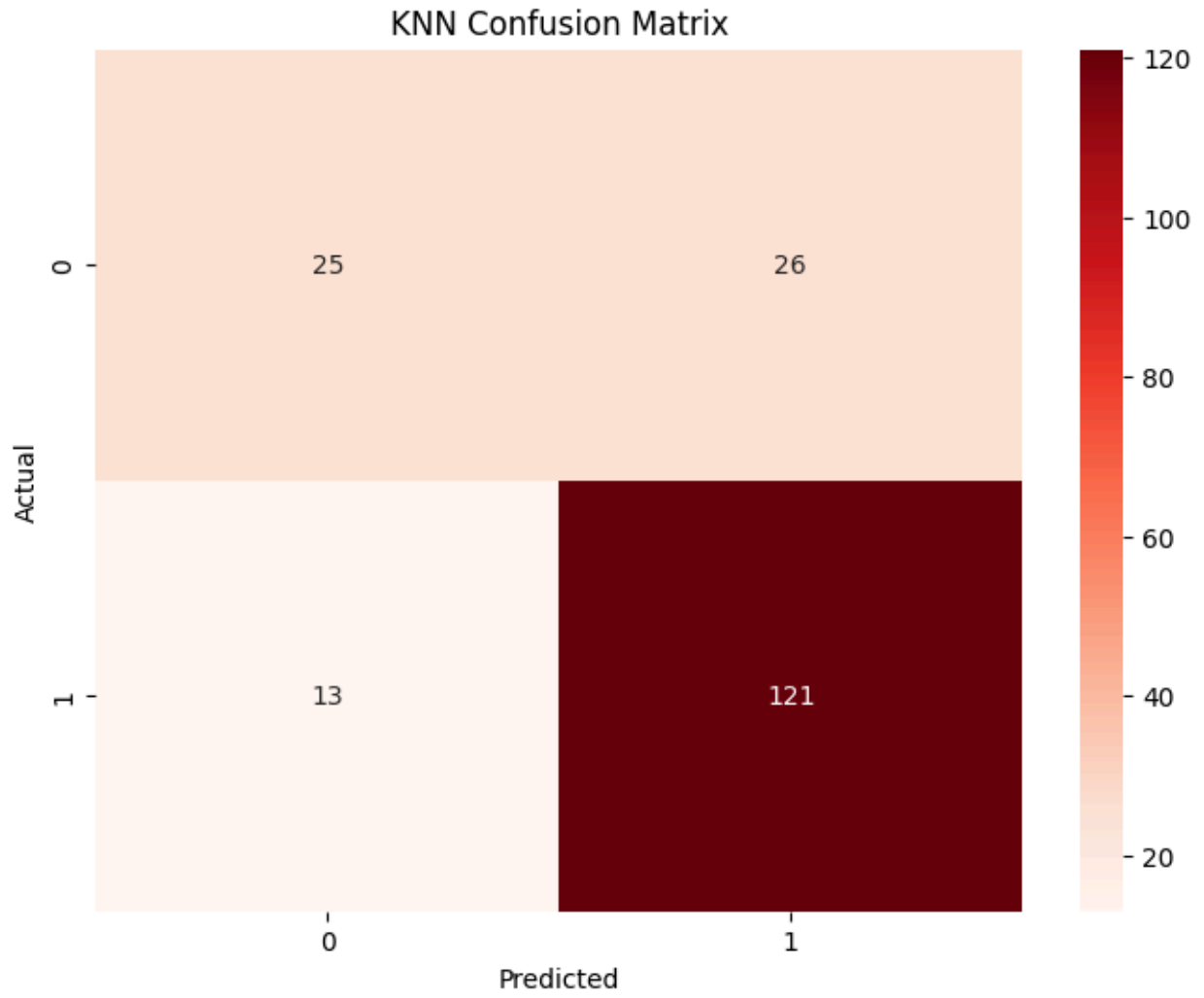


We got 81.08% accuracy from Random Forest classifier, 84.3% from SVM and 80% from KNN and Decision Tree models. Among all of the models, the SVM model has given a higher accuracy rate among all of the models during each runtime. It also bears a higher recall rate among all the four models indicating the proportion of predicted positive values to be actually positive is higher when using a SVM model.


RandomForest Confusion Matrix
Actual
Predicted
0
1
26
25
14
120
120
100
80
60
40
20


Decision Tree Confusion Matrix
28
23
14
120
Actual
0
1
Predicted
0
1
120
100
80
60
40
20


SVM Confusion Matrix
21
30
2
132
Actual
0
1
Predicted
0
1
120
100
80
60
40
20



These are the confusion matrices showing the proportion of false positive, false negative, true positive and true negative data predicted by each of the used models.

Conclusion:

Among all four models that we've used, the SVM model emerged as the most promising algorithm for loan approval prediction due to its higher accuracy and recall rates when compared to other models. Further model fine-tuning and optimization could improve prediction accuracy and assist financial institutions in making informed loan approval decisions.

The findings of this project provide a foundational understanding of the critical factors influencing loan approval, which is useful guidance for financial institutions looking to optimize their loan approval processes.