**FULL TITLE OF THE PROJECT**

**A MINI PROJECT REPORT**

### By

### Lavanya Tetakali (21JG1A0560)

Under the esteemed guidance of

**Mrs. V GOWTAMI ANNAPURNA**

Assistant Professor

Department of CSE

## **Department of Computer Science and Engineering**

## **GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING FOR WOMEN**

[Approved by AICTE NEW DELHI, Affiliated to JNTUK Kakinada]

[Accredited by National Board of Accreditation (**NBA**) for B.Tech. CSE, ECE & IT – Valid from 2019-22 and 2022-25]

[Accredited by National Assesment and Accreditation Council(**NAAC**)– Valid from 2022-27]

Kommadi , Madhurawada, Visakhapatnam–530048

### 2021–2025

## **GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING FOR WOMEN**

## **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

## **CERTIFICATE**

This is to certify that the mini project report titled **“Credit Score Classification”** is a bonafide work of following III B.Tech. students in the Department of Computer Science and Engineering, Gayatri Vidya Parishad College of Engineering for Women affiliated to JNT University, Kakinada during the academic year 2023-2024 Semester-II.

**Lavanya Tetakali (21JG1A0560)**

#### Project Mentor

#### Mrs. V. Gowtami Annapurna

#### Assistant Professor

**ACKNOWLEDGEMENT**

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragement crown all the efforts with success.

We feel elated to extend our sincere gratitude to **Mrs. V. Gowtami Annapurna,** Assistant Professor, for encouraging all the way during project analysis. His annotations, insinuations and criticisms are the key behind the successful completion of the thesis and for providing us all the required facilities.

We express our deep sense of gratitude and thanks to **Dr. P. V. S. Lakshmi Jagadamba**, Professor and Head of the Department of Computer Science andEngineering for her guidance and for expressing her valuable and grateful opinions in the project for its development and for providing lab sessions and extra hours to complete the project.

We would like to take this opportunity to express our profound sense of gratitude to **Dr. R. K. Goswami,** Principal and **Dr. G. Sudheer,** Vice Principal for allowing us to utilize the college resources thereby facilitating the successful completion of our thesis.

We are also thankful to both teaching and non-teaching faculty of the Department of Computer Science and Engineering for giving valuable suggestions from our project.

**Lavanya Tetakali**

**(21JG1A0560)**

**TABLE OF CONTENTS**

#### TOPICS PAGENO.

|  |  |
| --- | --- |
| **Abstract** | i |

|  |  |  |
| --- | --- | --- |
| **1.** | **INTRODUCTION** | **1** |
|  | 1.1 Problem Statement | 1 |
| **2** | **LITERATURE REVIEW** | 4 |
| **3** | **TECHNOLOGY STACK** | **7** |
|  | I) Datasets/Database Used  II) Pre Processing Steps  III) Packages Used  IV) Software Requirement Specification |  |
| **4** | **METHODOLOGY** | 20 |
| **5** | 4.1 Modules Descriptions  4.2 Algorithms  4.3 Model Architecture Diagram  **IMPLEMENTATION** | 24 |
| **6** | **RESULTS / ANALYSIS** | 25 |
|  | 6.1 Output Screens **/** Results Analysis(must include Precision,Recall,F1-Score,Confusion Matrix, ROC Curve and also represent analysis in tabular format) |  |
| **7** | **CONCLUSION & FUTURE SCOPE** | 28 |
| **REFERENCES** | | 30 |

**ABSTRACT**

Credit Score is one of the most important and critical features or attributes of an individual which states their financial behavior and Market Credibility. Credit Score measures of an individual help banking and lending institutions to identify their customers as a good or a bad customer and also help them to identify the defaulters in terms of granting loans and credit cards to their customers. Credit Score may also help an individual to calculate their financial capability and eventually help them to explore various investment options so as to increase their market credibility. Machine Learning Algorithms can play important role in predicting the status of a customer as good or bad and at the same time can help them to plan a progressive financial eligibility for future prospects. The paper proposes a loan eligibility prediction system based on Credit Score calculator using machine learning algorithms. The accuracies of each algorithm can be compared and a system can be developed with the one having highest accuracy. The proposed system can help to identify an Individual’s Credit Worthiness and extend support to financial institutions in deciding whether to grant or deny loan to the customer. The system can also help an individual to understand their credit worthiness and provide various investment recommendation so as to increase their overall credit score.

Keywords—: Credit Worthiness, Credit Score, Logistic Regression, Machine Learning Algorithms.

1. **INTRODUCTION**

1.1 Problem Statement

Credit Score is one of the most important and critical attribute of an Individual to understand their Credit Worthiness and Market Credibility. There is a high rise in the financial sectors such as loan, investments, stocks which involves a lot of money, so a strategy build on the basis of just human experiences is not advisable. In money lending business there are always defaulters who may cause large losses which is not at all favorable on the part of the financial institutions. The institutions have in fact experienced many such losses historically. The money lenders have to decide some chosen set of criteria before providing loan to certain person, so as to minimize the losses and maximize the profit. Building an efficient model that can envisage the Client’s Credit Worthiness and Credit Score can aid financial institutions in order to identify the defaulters and separate out them from good borrowers. This can help to maximize their profit and to come out among the top rated financial institutions. Recently with high rise in new technologies like data science, machine learning it’s been easier to predict things and work accordingly. An ensemble strategy is combining or adding multiple classifiers that individually tests or perform to give results and get a single predictive output based on all the data. Ensemble model decrease the variance, bias and greatly improve the capability to predict with accuracy. Ensemble models are in high demands due to their impressive features and have a lot of potential to become one of the highly used methods in the financial field.

**2. LITERATURE SURVEY**

In paper [1] Customer financial Behavior is classified as a Good or a bad using algorithms of Machine Learning like Neural Network, Decision Tree, Logistic Regression, SVM is calculated. Attributes such as Age, Gender, Income, Dependents, Current Address, Married, Self Employed, City, Previous Loan, Loan Period, Job Title, House(own/Rented). Accuracies obtained were ANN-0.80 LR-0.82 Decision Tree- 0.78 SVM-0.79.

In paper[9] Customer Behavior as a Good(1) or bad(0) is classified using Logistic Discriminant Analysis, Artificial Neural Network, XGBoost, SVM was calculated using attributes such as Gender, Married, Income, Education, Dependents, Previous Loan ,House(own/rented),credit card, Number of Previous loans, Job type. Accuracies obtained were LDA-0.73 ANN-0.82 SVM-0.84.

In paper[6] Credit score of a customer is classified using its Social Networking information considering peer to peer lending, online micro lending market etc. using Random Forest, Adaboost, Light GBM. Attributes considered were Social Stability, Social Exposure, Social Quality, Gender, Age, Income, Education, Job, Years at Residence, Credit History, City, Dependents, type of loan. Accuracies obtained were Random Forest-0.74 Adaboost-0.78.

In paper[8] Credit Score is calculated to predict Good or a bad Customer using Oracle Data Miner, ROC Curves and Graphical analysis. Attributes considered were Office Details, other loans, Credit history, Nationality, Current address, income, dependents, Education, Gender, Marriage, Self Employed, DOB. Accuracies obtained were LR-0.82. SVM-0.79 ANN- 0.76.

In paper [7] Customer behavior as a Good or a bad is calculated using machine learning tools like Neural Network, Decision tree, Adaptive boosting, SVM, Logistic Regression. Gender, Marriage, Income, Dependents, Property Details, House(own/rented), type of loan, City, Monthly expense, previous loans, dependent income, savings, credit history of an individual were taken into consideration. Accuracy obtained were ANN-0.82 Decision tree-0.78 Adaptive boosting-0.76 SVM-0.84 LR-0.88.

**4. METHODOLOGY**

**4.1. Modules Description**

1. NumPy: It is a fundamental package for scientific computing in Python. It provides support for arrays, matrices, and mathematical functions, which can be useful for data manipulation and calculations.

2. Pandas: Pandas is a powerful data manipulation library. It offers data structures like DataFrame and Series, which are highly useful for data preprocessing, cleaning, and analysis tasks.

3. Seaborn: Seaborn is a statistical data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

4. Matplotlib: Matplotlib is a plotting library for Python. Pyplot is a Matplotlib module that provides a MATLAB-like interface for creating plots and visualizations.

5.Tree: This module contains the DecisionTreeClassifier class, which implements decision tree algorithms for classification tasks. Decision trees are used for both classification and regression tasks.

6.LabelEncoder: LabelEncoder is a utility class in scikit-learn used to encode categorical labels with numerical values. It's commonly used to preprocess categorical features before feeding them into machine learning models.

7. Train\_test\_split: This function from the model\_selection module splits datasets into random train and test subsets. It's essential for evaluating machine learning models' performance on unseen data.

8. RobustScaler: RobustScaler is a preprocessing technique used to scale features to a robust range, less sensitive to outliers compared to standard scaling methods like MinMaxScaler or StandardScaler.

9. VotingClassifier: VotingClassifier is an ensemble learning method in scikit-learn that combines multiple individual machine learning models and aggregates their predictions using a voting strategy (e.g., majority voting).

10.DecisionTreeClassifier: This class implements decision tree algorithms specifically for classification tasks. Decision trees partition the feature space into a set of rectangular regions and assign a class label to each region.

11.KNeighborsClassifier: KNeighborsClassifier is a classification algorithm based on the k-nearest neighbors principle. It predicts the class of a data point by a majority vote of its k nearest neighbors in the feature space.

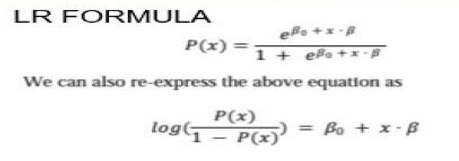
12.RandomForestClassifier: RandomForestClassifier is an ensemble learning method that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting.

13.LogisticRegression: Logistic regression is a linear model used for binary classification tasks. It estimates the probability that a given input belongs to a particular class using a logistic function.

**4.2. Algorithms:**

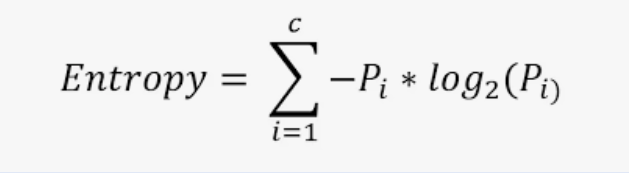
1) Logistic Regression

The model that will be used for designing the credit score system is Logistic Regression Model. It is one of the most commonly used Machine Learning Algorithm utilized in Credit Scoring Model. While comparing Logistic regression and linear regression it is observed that both alter in their results and outcome, as the outcome of latter is continuous and the outcome of former one is discrete. Logistic regression establishes the relationship between the independent variables where the number of variables can be single or multiple.



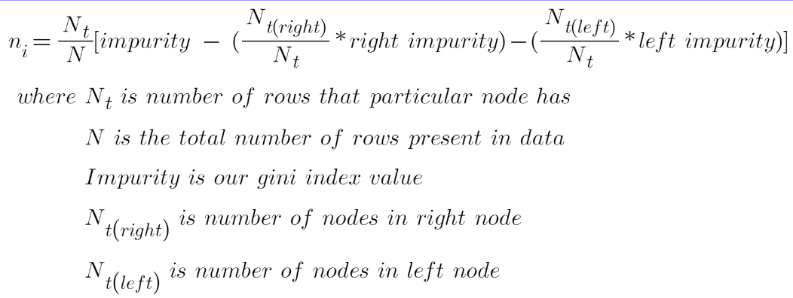
2)Decision Tree

Decision Trees is a tree based learning algorithms and has high efficiency and used mostly in the field of machine learning techniques. Tree based techniques have higher accuracy, good stability and ease of research. Decision Tree algorithms establishes non-linear relationships between attributes such that the accuracy obtained is good. Any conversion in the input do not directly change the outputs as Decision trees are non-linear. Here, we divide the dataset into small subsets and decision tree is designed in an incremental manner. These steps lead to creation of decision nodes and tree nodes.



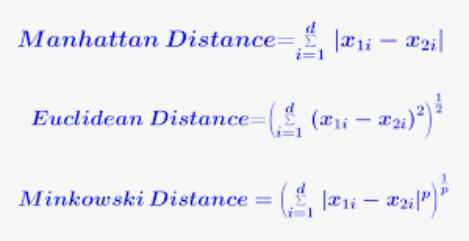
3)Random Forest Algorithm

Random Forest is an all-round Algorithm which has features regression as well as classification tasks. It also deal with missing data, removing outliers, commence dimensional reduction technique and other learning methods. It designs a strong model by combining weak models. Each tree assigns a classification to classify a new object based on its attributes, i.e. voting is carried out for that specific selected class. This Algorithm then selects the tree having the most votes (from the list of all trees in forest) and then takes average of different outputs to compute the results in the case of regression.



4) KNN

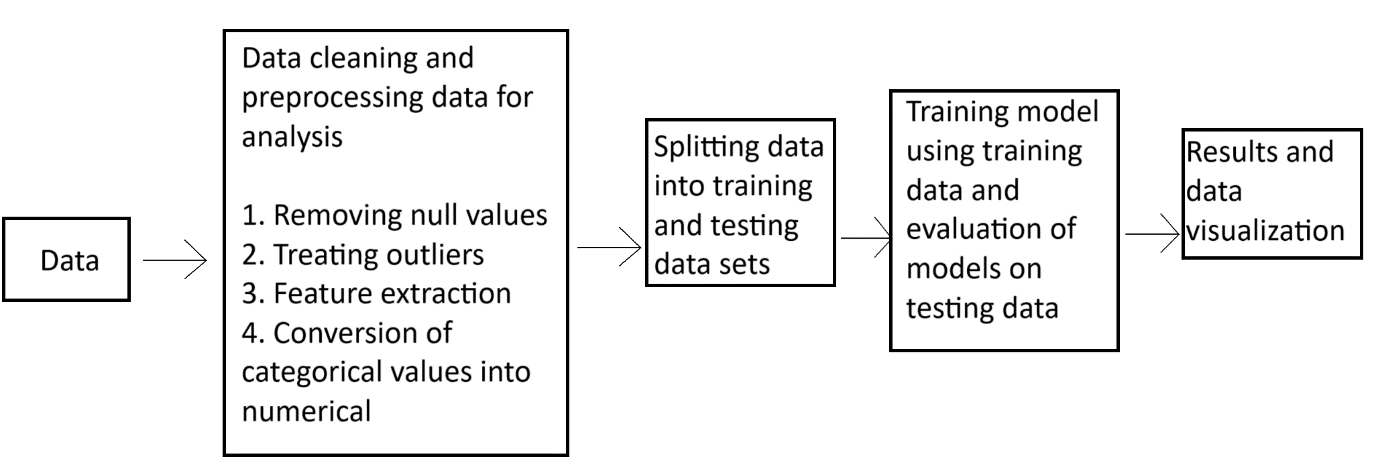
The K-Nearest Neighbors (KNN) algorithm is a non-parametric method used for classification and regression tasks. During training, KNN memorizes the entire dataset, and for each new data point, it calculates distances to all training instances. It then selects the K nearest neighbors based on a distance metric, such as Euclidean distance. For classification, KNN assigns the most common class label among its K nearest neighbors to the new data point. However, the choice of K is crucial and can significantly impact performance. KNN is sensitive to feature scale and can be computationally expensive for large datasets.



5) Voting classifier

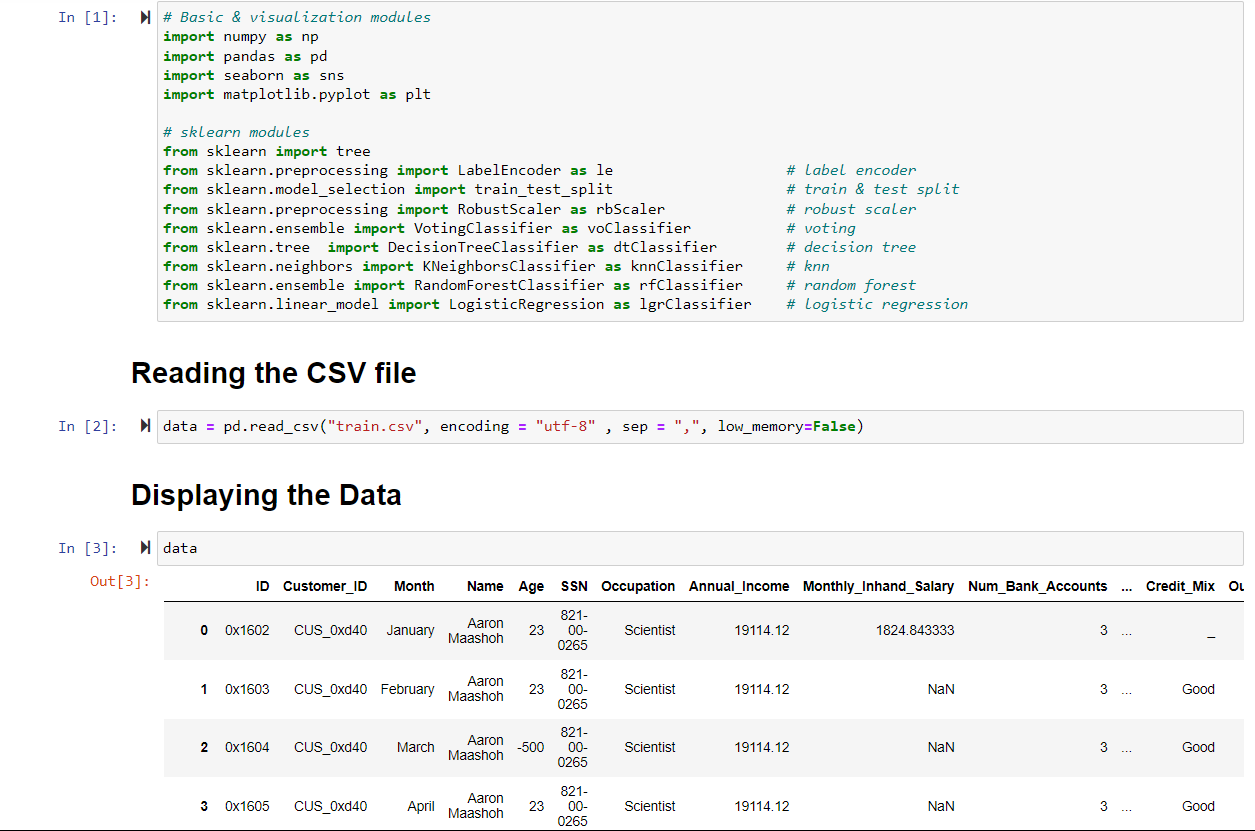
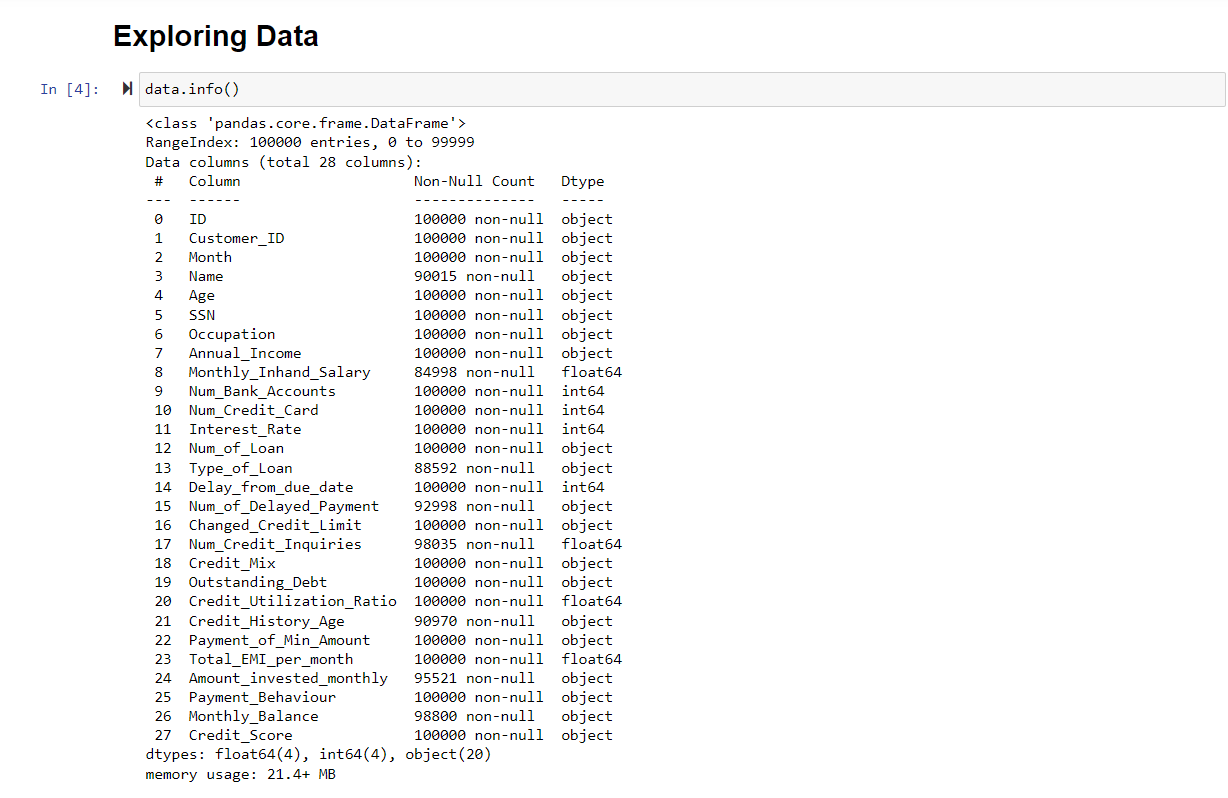
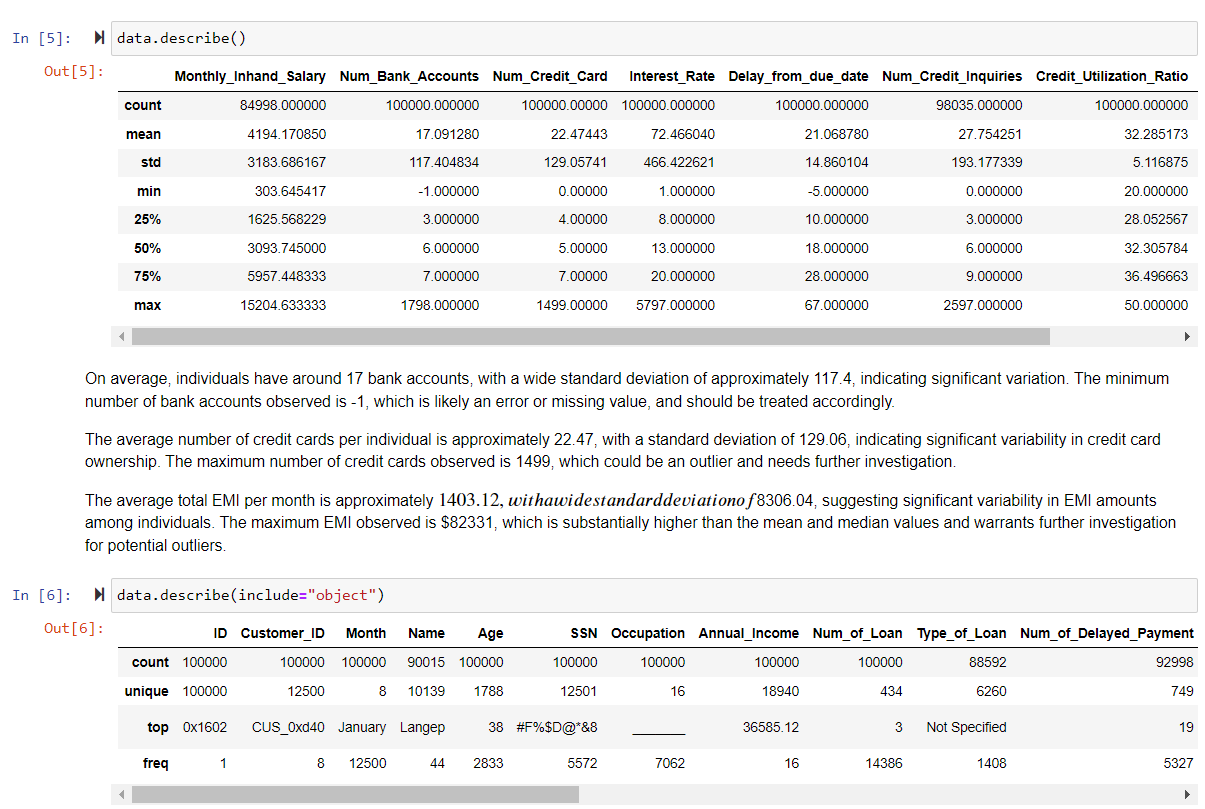
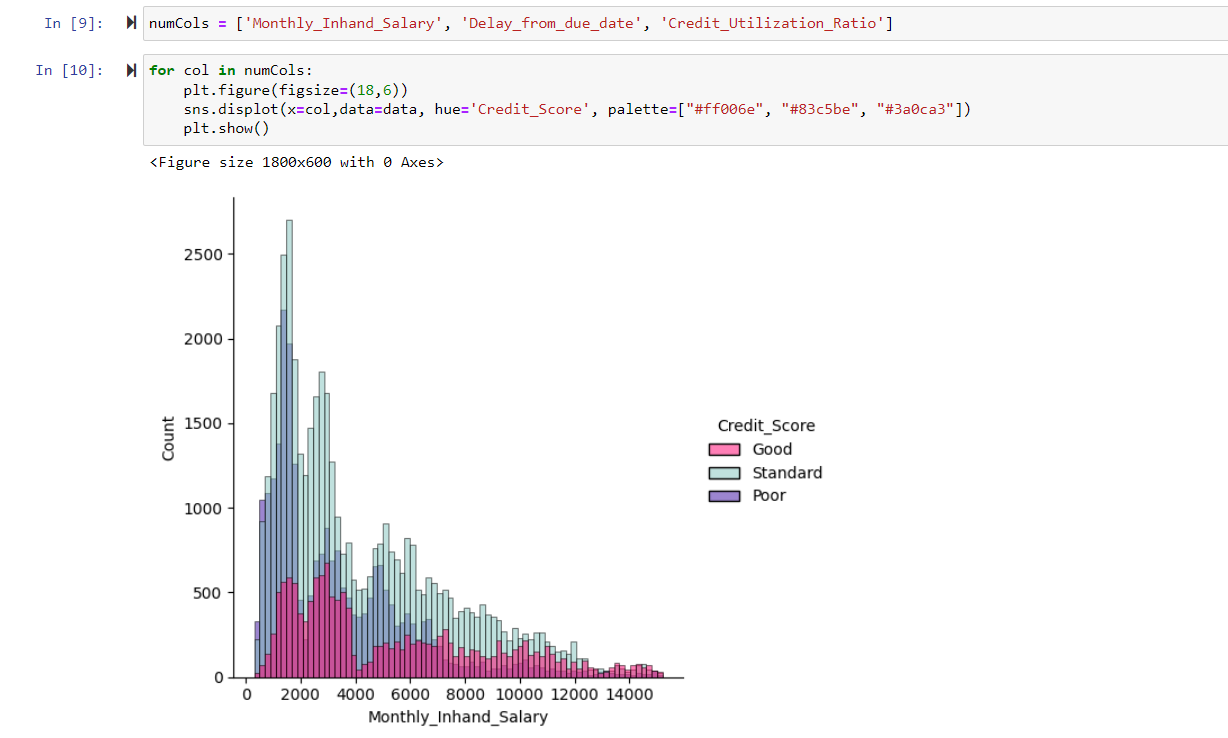
Voting Classifier is an ensemble learning method that combines multiple classifiers' predictions using a majority voting strategy. It supports both hard and soft voting, where classifiers' predictions are weighted by their confidence. This ensemble technique often leads to improved performance and robustness by leveraging diverse classifiers' collective decision-making.

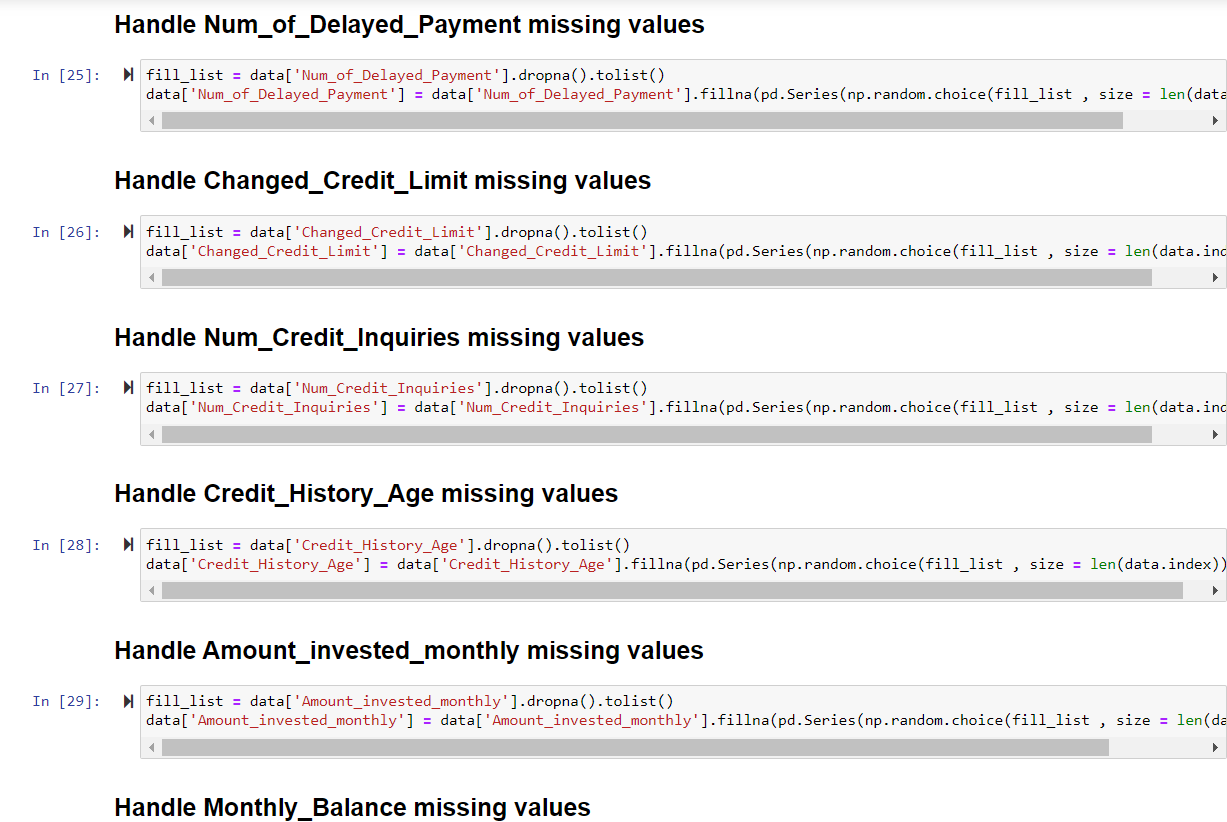
4.3. Model Architecture diagram

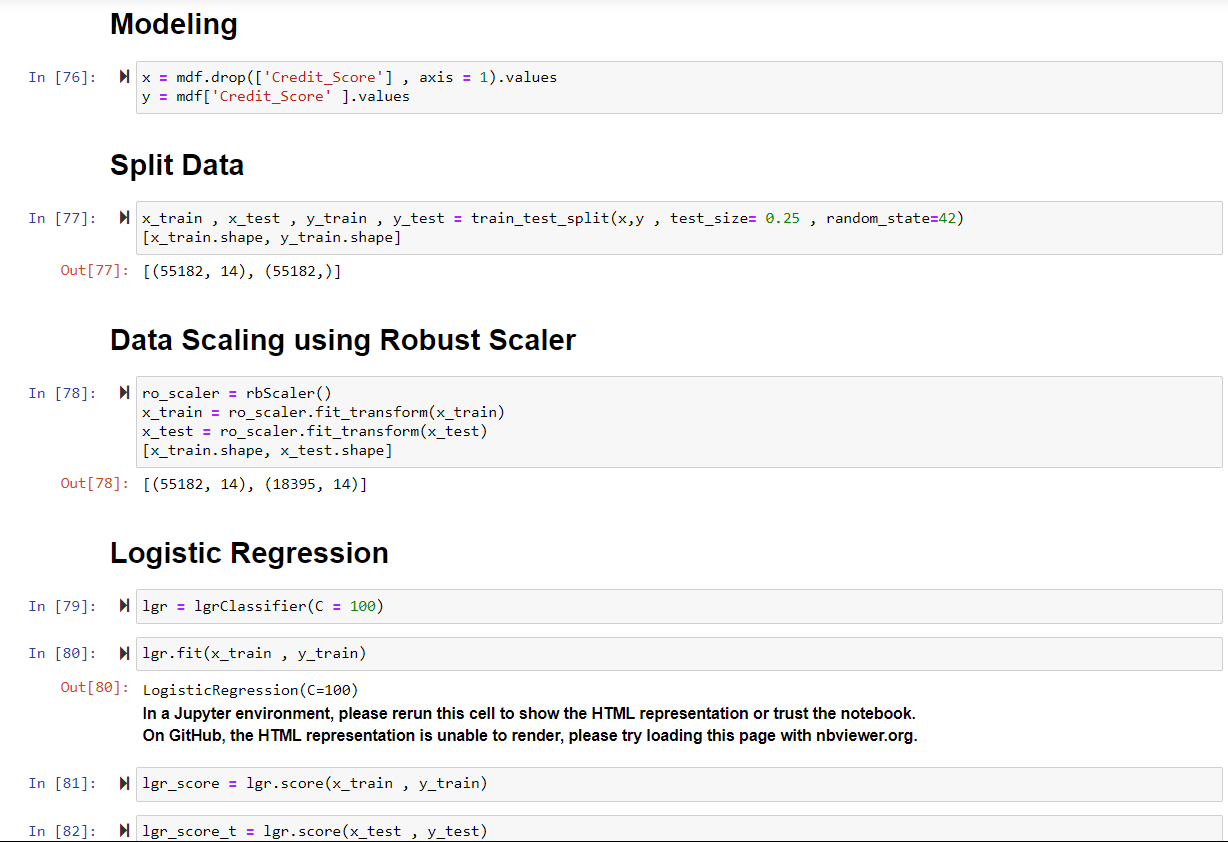


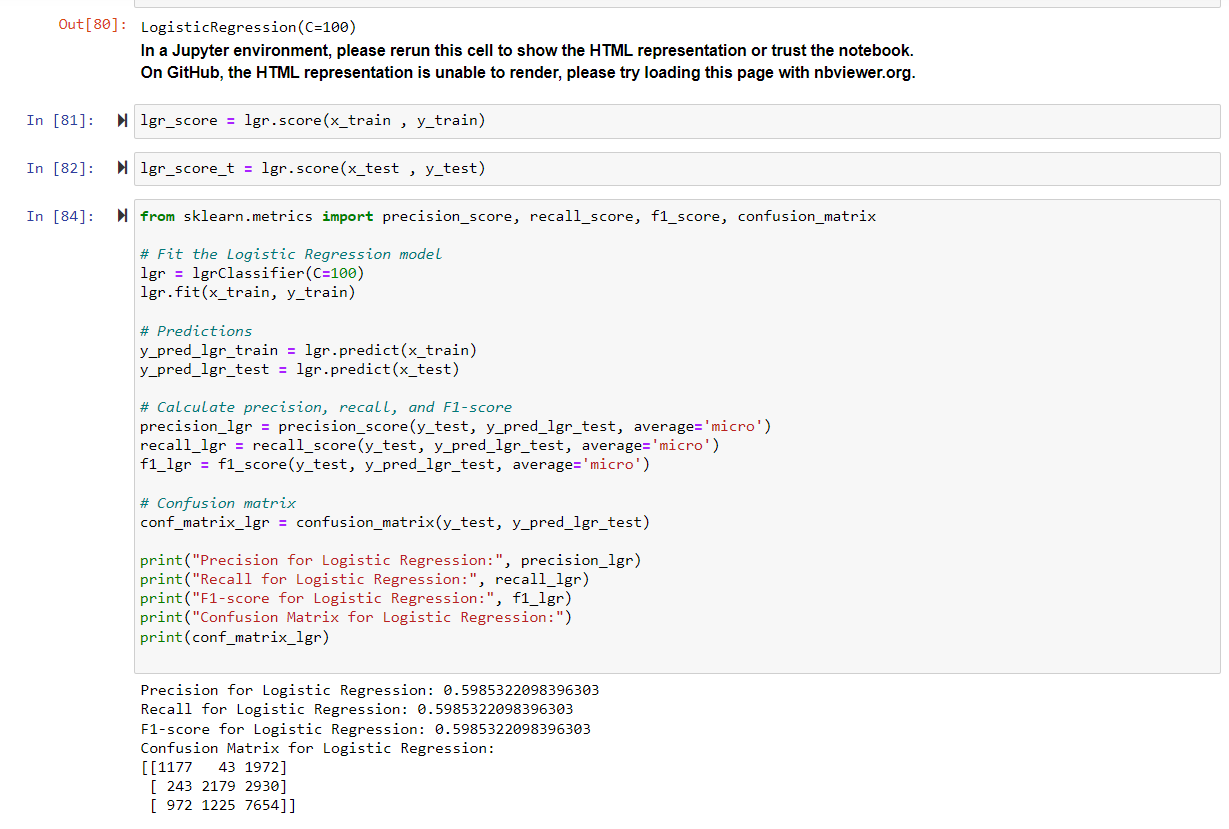
**Fig 4.1. Architecture diagram**

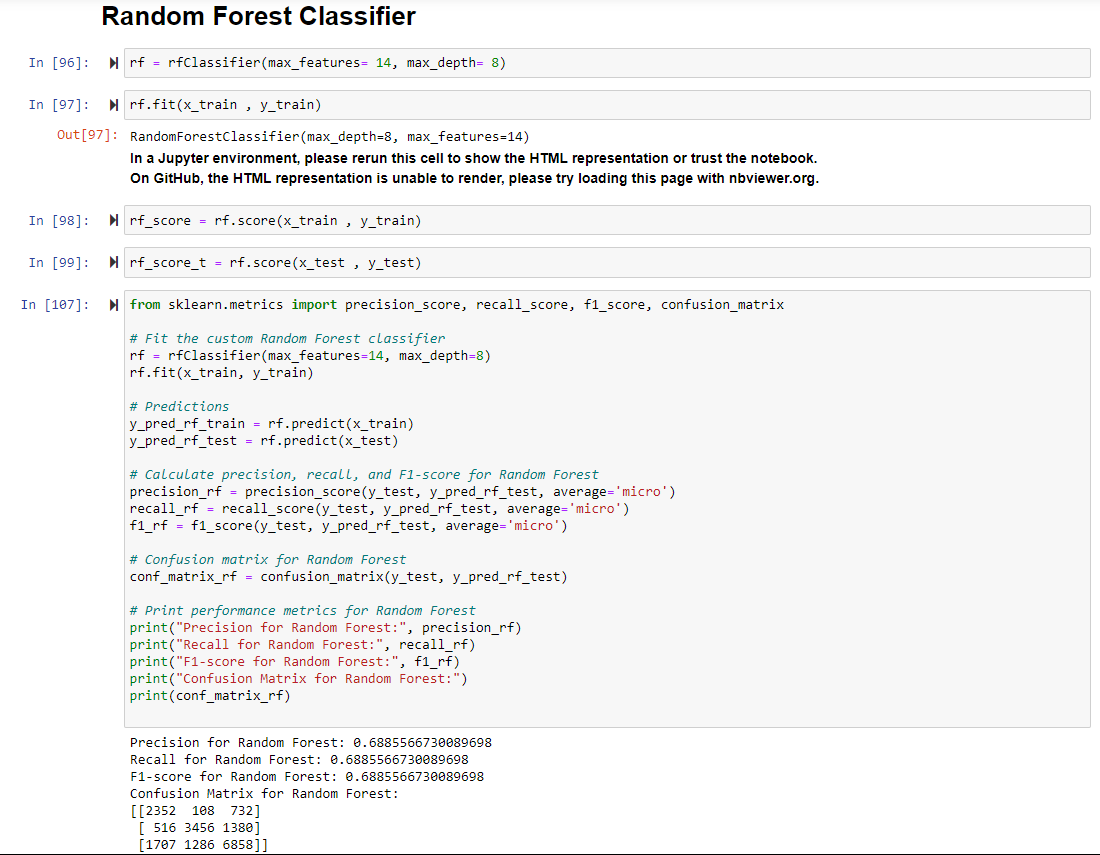
**6. Implementation**

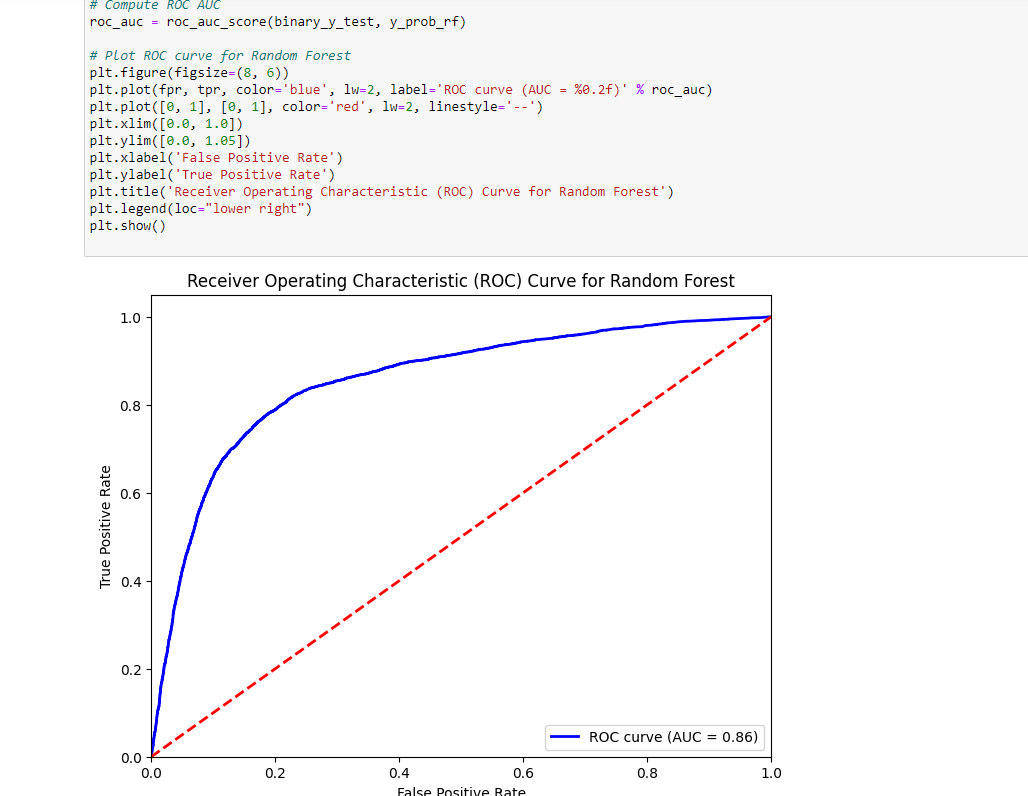
   



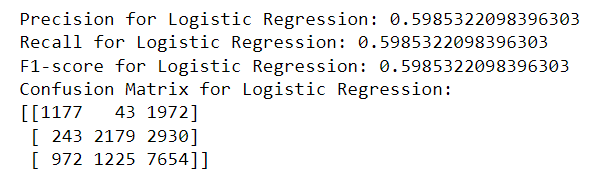


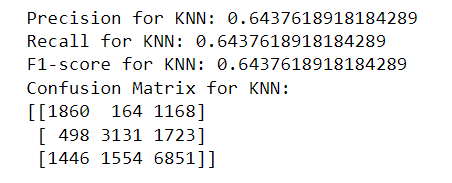
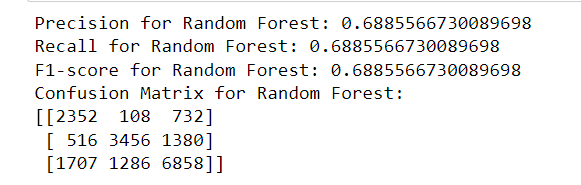
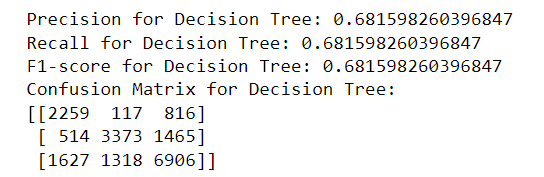


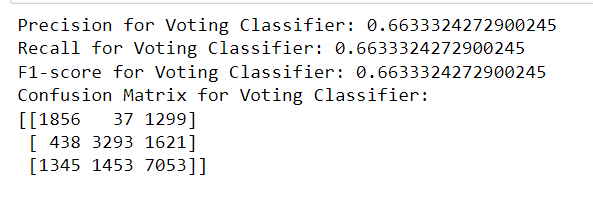


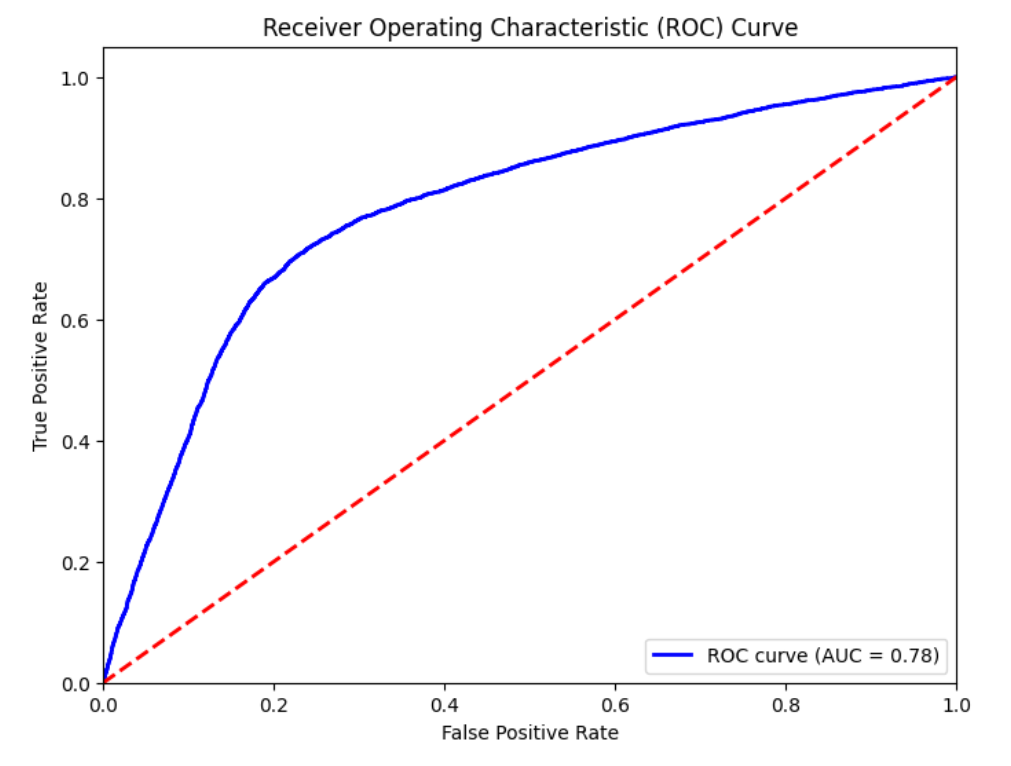
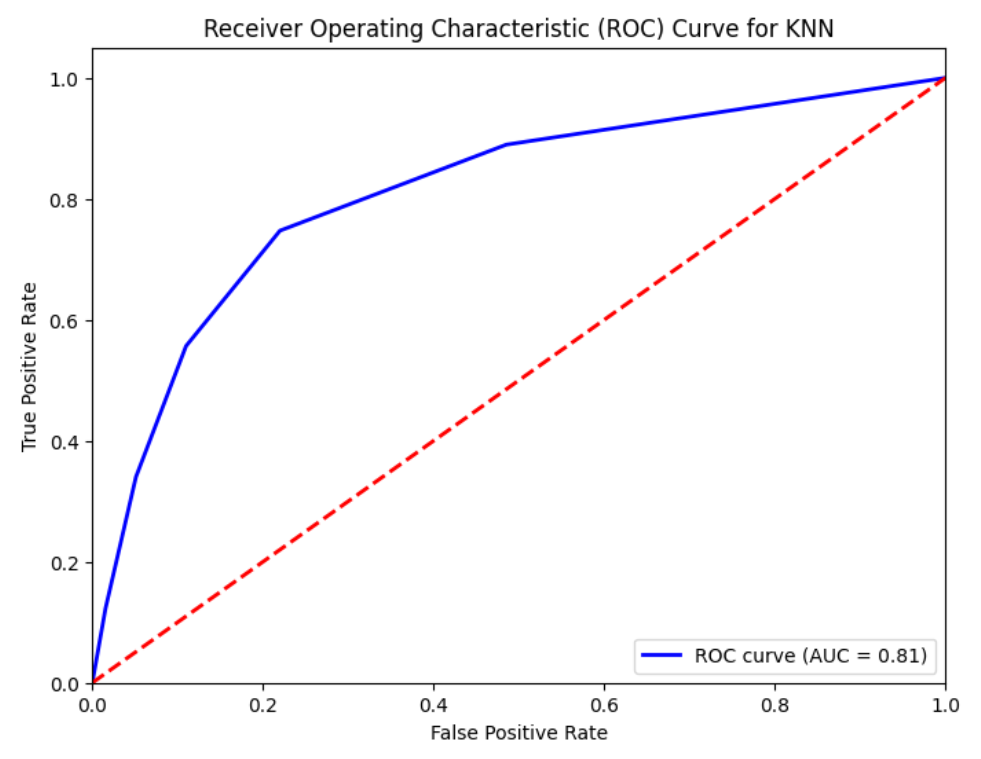


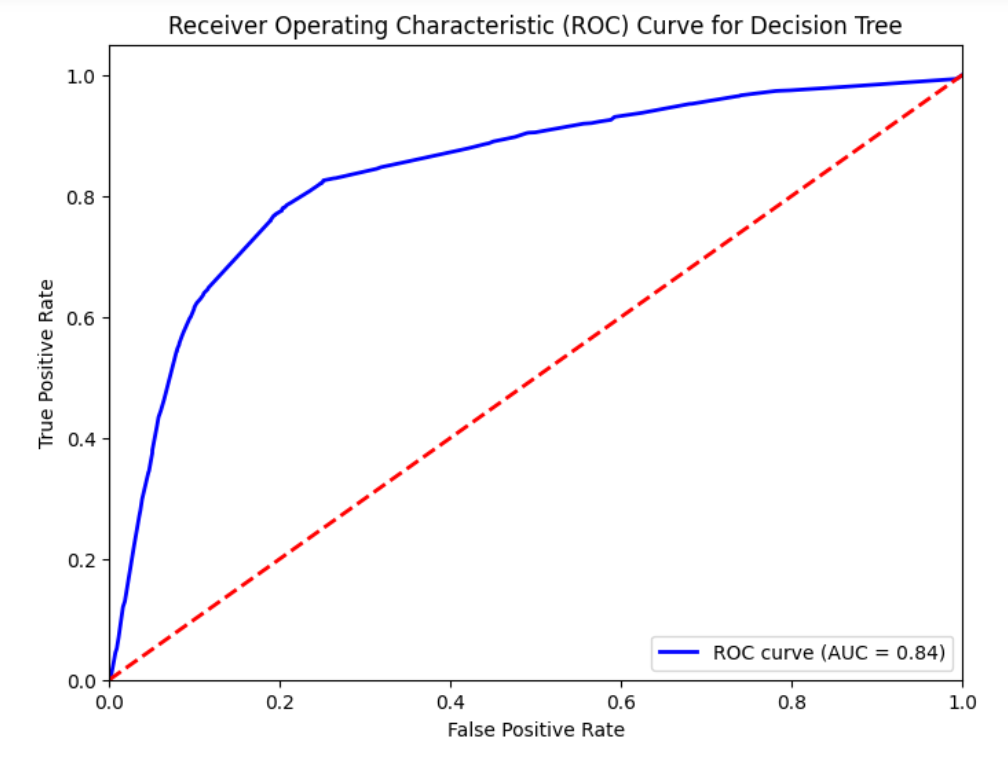
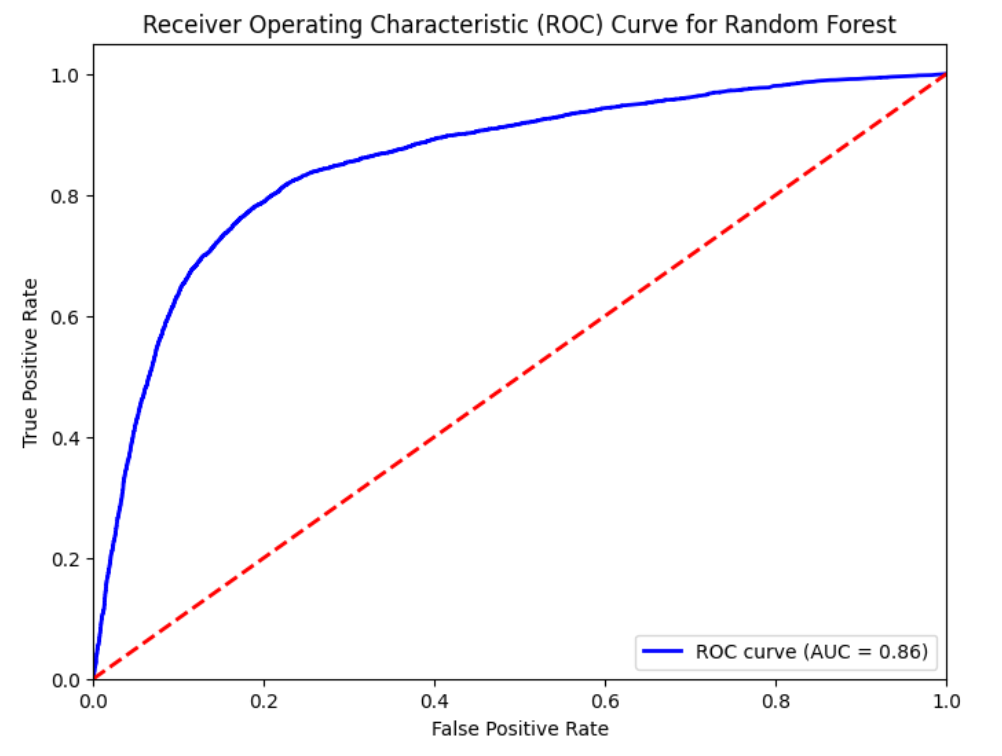
**5. Results**







**References**

[1] Huseyin Ince and Bora Aktan**,** “A comparison of data mining techniques for credit scoring in banking: A managerial perspective”, 14 Oct 2010.

[2] Jasmina Nalić and Amar Švraka , “Using Data Mining Approaches to Build Credit Scoring Model” ,17th International Symposium INFOTEH- JAHORINA, 21- 23 March 2018.

[3] Durgesh Kumar Singh and Noopur Goel**,** “Analysing Data Mining Techniques on Bank Customers for Credit Score”, 2020 8th International Conference on Reliability, Infocom Technologies and Optimization, June 2020.

[4] Archana Gahlaut, Tushar , Prince Kumar Singh **,** “Prediction analysis of risky credit using Data mining classification models”, 8th ICCCNT 2017 IIT Delhi , July 3-5, 2017.

[5] Wei Li and Jibiao Liao , “ An Empirical Study on Credit Scoring Model for Credit Card by using Data Mining Technology” , 2011 Seventh International Conference Intelligence and Security.

[6] Beibei Niu, Jinzheng Ren , Xiaotao Li , **“**Credit Scoring Using Machine Learn- ing by Combing Social Network Information: Evidence from Peer-to-Peer Lending” , 17 December 2019.

[7] Ali Al-Aradi **,** “ Credit Scoring via Logistic Regression” ,February 28, 2014.

[8] Ansen Mathew , “ Credit Scoring Using Logistic Regression” ,24 th May 2017.

[9] Pratik Sharma, Sunil Bhatia, Rohit Burman, Santosh Hazari, Rupali Hande, “Credit Scoring using Machine Learning Techniques” , March 2017.

[10] ROHIT KUMAR, MAYUKH BISWAS, SANGRAM MONDAL, “Exploratory Analysis For Credit Score With Applied Machine Learning”, May 2020.