PROJECT REPORT

Project submitted to the
SRM University – AP, Andhra Pradesh

SRM UNIVERSITY AP

Neerukonda,Mangalagiri,Guntur
Andhra Pradesh-522502

Bachelor of Technology In
Department of computer science and engineering
Machine Learning
Project Report on

Loan Prediction System

Sai Akash	AP21110010945
Dheeraj	AP21110010935
Lokesh	AP21110010956

Under The Guidance of

Dr Mahesh Kumar Morampudi

CSE-O

Certificate:

Date:

This is to certify that the work present in this Project entitled "**LOAN PREDICTION SYSTEM**" has been carried out by Dheeraj, Sai Akash, Lokesh under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in School of Engineering and Sciences.

Supervisor

Dr. Mahesh Kumar Morampudi

Designation- Asst.Professor dept of CSE Affiliation.

Acknowledgement

Foremost, we express our deepest appreciation to our mentor, Dr. Mahesh Kumar Morampudi, whose wisdom, insights, and steadfast assistance has been crucial in forming our project. Your guidance has illuminated our path and instilled in us a profound appreciation for the intricacies of machine learning and exploratory data analysis.

We extend our gratitude to our guide, Dr. Mahesh Kumar Morampudi, who has generously shared their expertise, serving as our guiding light through this intricate journey. Your mentorship has been invaluable, and your patience in addressing our queries has been greatly appreciated.

Table of contents:

Abstract

As the needs of people are increasing, the demand for loans in banks is also frequently getting higher every day. Banks typically process an applicant's loan after screening and verifying the applicant's eligibility, which is a difficult and time-consuming process. In some cases, some applicants default and banks lose capital. The machine learning approach is ideal for reducing human effort and effective decision making in the loan approval process by implementing machine learning tools that use classification algorithms to predict eligible loan applicants. Banks desire to deliver the loan toan individual who can be recompensate the loan on time and can afford maximum profit to the bank. So, there is a need of a system which could do this analysis and save the banks time and resources. This can be done using Machine Learning. Using these records variables and bank loan rules we will train a machine learning model which will predict that a person is eligible for loan or not. We will use sklearn for our model and train_test_split for splitting the data set into train dataset and test dataset. Here we are going to use various models like Decision Tree(DT) and Random Forest(RF) so as to fetch more accurate results as the given problem is a supervised classification problem.

Introduction

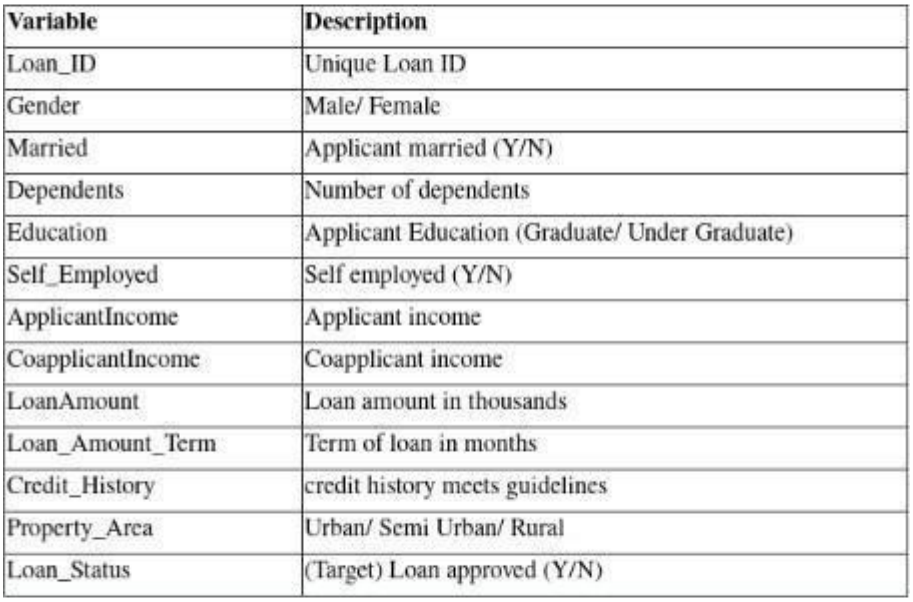
Loans are the core business of banks. The main profit comes directly from the loan's interest. The loan companies grant a loan after an intensive process of verification and validation. However, they still don't have assurance if the applicant can repay the loan with no difficulties. In our Project, we built a predictive model to predict if an applicant is eligible for the loan or not. We prepared a model to predict the target variable. And made a User Interface. Customer first apply for loan after that company or bank validates the customer eligibility for loan. Company or bank wants to automate the loan eligibility process (real time) based on customer details provided while filling application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and other. This project has taken the data of previous customers of various banks to whom on a set of parameters loan were approved. So, the machine learning model is trained on that record to get accurate results. Our main objective of this project is to predict the safety of loan. To predict loan safety, the Random forest and Naïve bayes algorithm and Decision Tree tare used.

Problem Statement

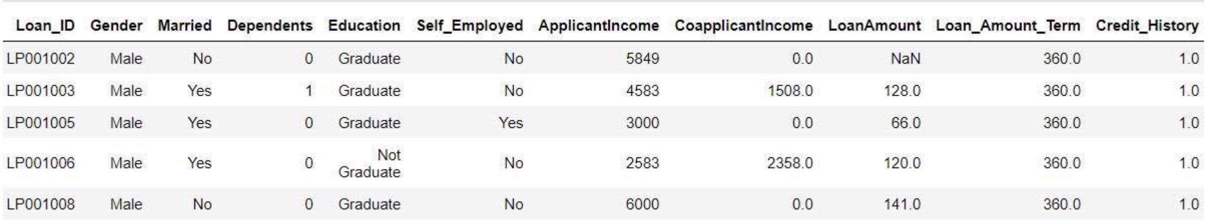
The current loan verification and validation process is highly time-consuming, often requiring extensive manual efforts. This labor-intensive approach not only consumes valuable resources but also increases the likelihood of human errors creeping into the validation process. Additionally, the lack of cross- referencing previous loan records further exacerbates the inefficiencies, leading to a fragmented and less reliable assessment of borrower creditworthiness. Consequently, this outdated system demands a significant allocation of human resources, hindering efficiency and scalability within financial institutions. For this Our Machine learning model calculates all the parameters given and predicts if the applicant is eligible for loan or not in very less time.

Dataset

This is taken from the Kaggle which contains 13 features and 614 instances



Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	(Target) Loan approved (Y/N)



Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0
LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0
LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0
LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0
LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0

Project Workflow

1) **Reading the Dataset**:

- The dataset is read from a CSV file using loan_dataset=pd.read_csv('dataset1.csv')

2) **Splitting into Train and Test Sets**:

- The dataset is split into training and test sets using train_test_split from sklearn. This is a common practice to evaluate the model's performance on unseen data.

3) **Handling Missing Values**:

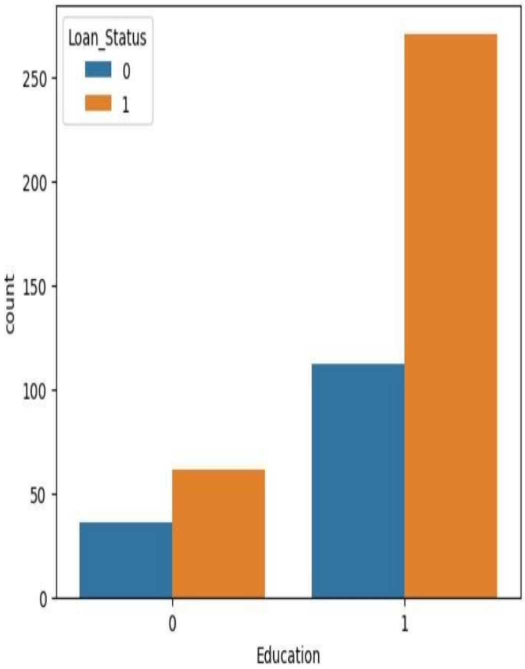
- Missing values are handled in multiple columns such as Gender, Married,Dependents,Self_Employed,Loan_Amount,Loan_Amount _Term, Credit_ History

4) **Data Encoding:**

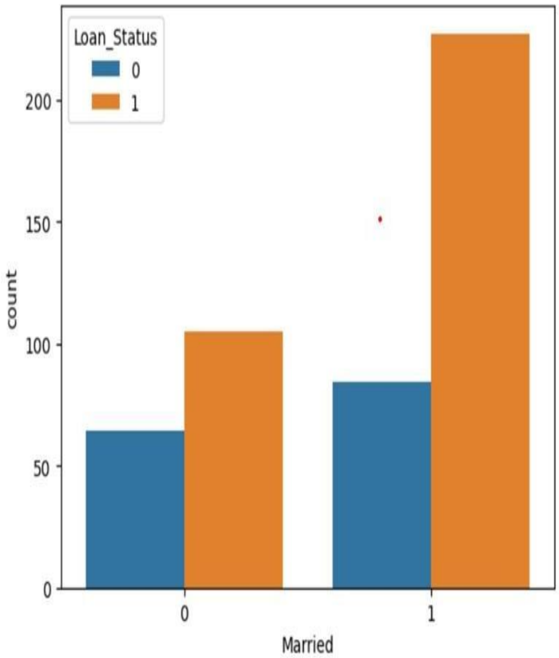
- Categorical columns such as 'Married', 'Gender', 'Self_employed', 'Property_area', 'Education' and 'Loan Status' are converted into numericals. This step is crucial because machine learning models cannot directly handle categorical data. (Label Encoding)

5) **Dataset Analysis:**

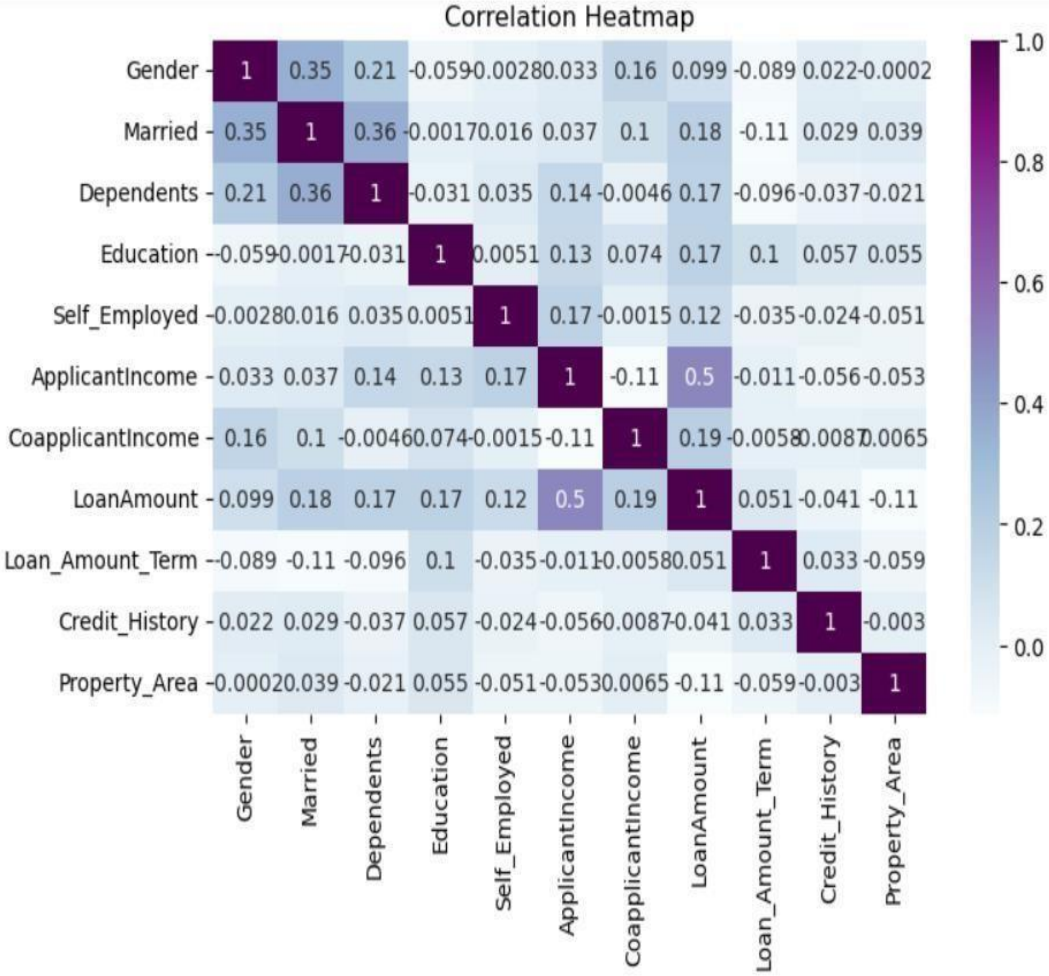
- Understanding the dependencies between the features and individual importance of the features with some visualizations.
- Exploring correlations between features to identify potential multicollinearity issues and assess the strength and direction of relationships between variables.



Loan Status and Education



Loan Status and Marriage

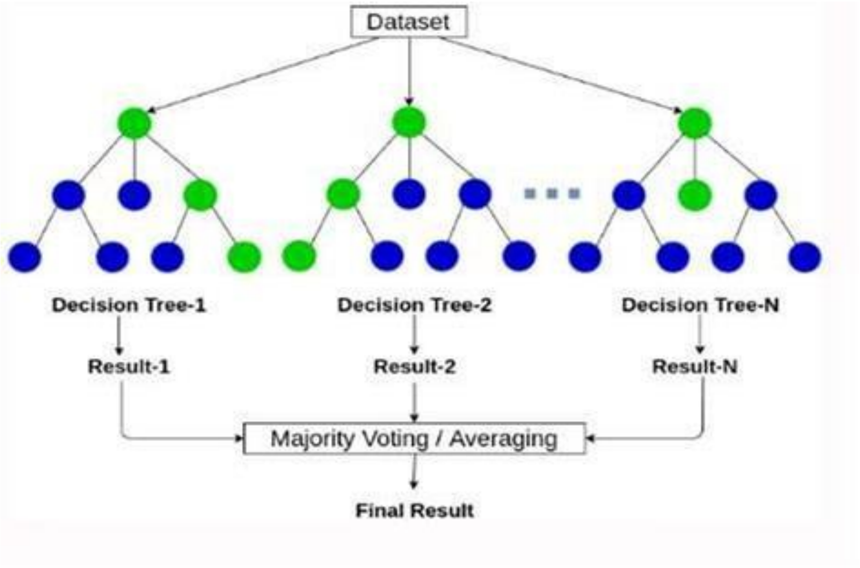


6) **Models Used**

1.Random Forest:

Random Forest is a machine learning algorithm that belongs to the ensemble learning family. It works by constructing a multitude of decision trees during training and outputting the mode (for classification) or mean prediction (for regression) of the individual trees.

Advantages:
1. Reduces overfitting by combining multiple trees.
2. Handles noisy data well.
3. Provides feature importance information.



2.Navie-Bayes

Naive Bayes is a classification algorithm based on Bayes' theorem with the "naive" assumption of independence between features. It's widely used in machine learning for various classification tasks, especially when dealing with text data like spam detection or document categorization.

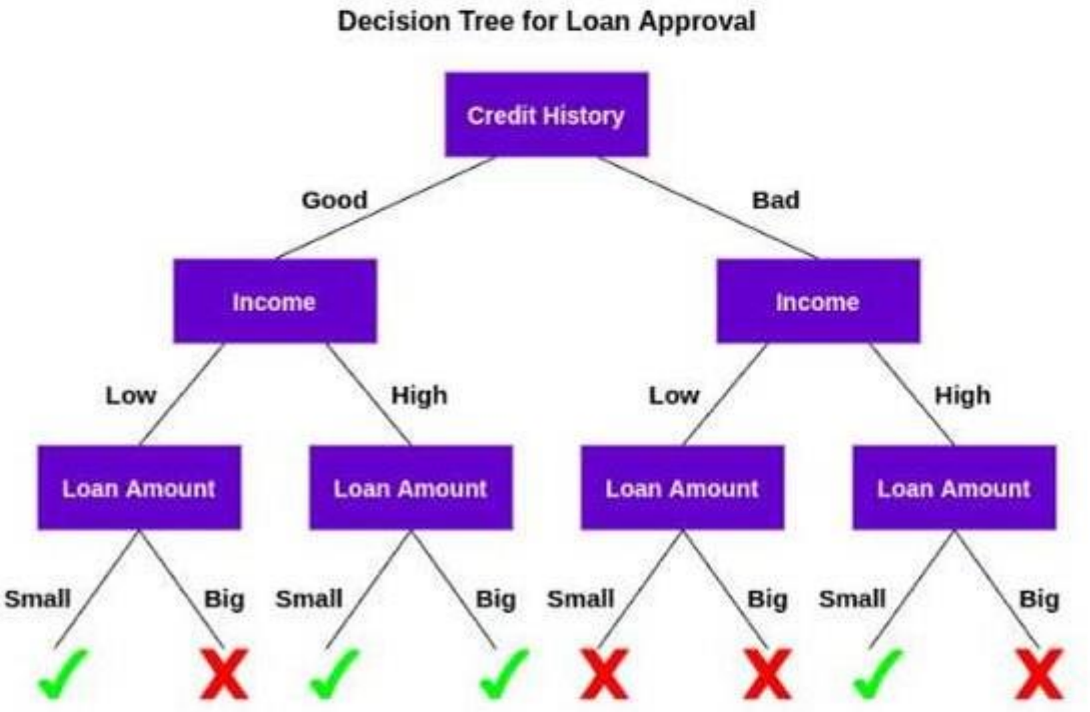
Advantages:
It is simple and easy to implement.
It doesn't require as much training data.
It handles both continuous and discrete data.
It is highly scalable with the number of predictors and data points.
It is fast and can be used to make real-time predictions.

3. DecisionTree:

A decision tree is a popular supervised learning algorithm used for both classification and regression tasks in machine learning. It creates a tree-like structure where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents the outcome or class label.

Advantages:

1. Simple to understand and to interpret. Trees can be visualized.
2. Able to handle multi-output problems.
3. Even if the underlying model from which the data were created violates some



6) Model Training and Prediction:

- Three models Random Forest, Navie-bayes and Decisson Tree are trained on the pre-processed data.
- Random Forest ,Navie-bayes and Decission Tree are used due to their simplicity and effectiveness in regression tasks.
- After training, the models are used to make predictions on the test set.

7) **Saving the Model:**

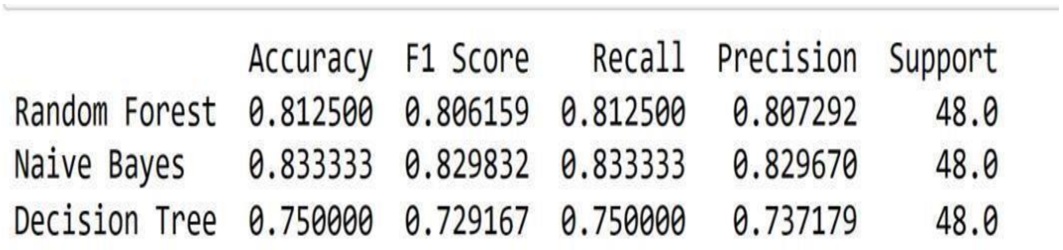
- The Random Forest Model which has secured the highest accuracy have saved using the joblib library.
- Once a model is trained, saving it allows for easy reuse of new data without the need to retrain the model each time.

8) Interface Creation:

- Creating the Graphical User Interface(GUI) with the Tkinter library which is the standard GUI library enhance the accessibility of the model to enter the input values in order to get the prediction.
- GUI has made in which the saved Random Forest has been incorporated so that the input values are given to the Radom Forest Model by which the prediction is returned

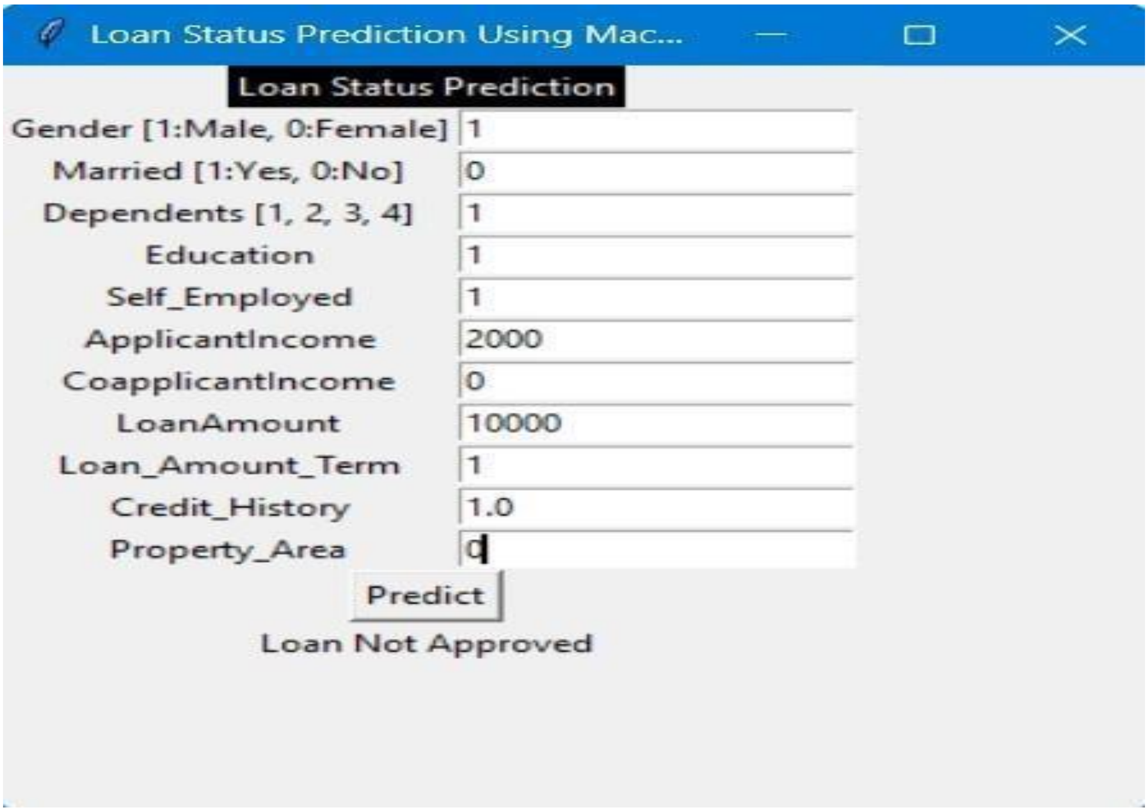
R2 Score

If 90:10 (Train: Test):



```
               Accuracy  F1 Score    Recall  Precision  Support
Random Forest  0.812500  0.806159  0.812500   0.807292     48.0
Naive Bayes    0.833333  0.829832  0.833333   0.829670     48.0
Decision Tree  0.750000  0.729167  0.750000   0.737179     48.0
```

RESULT:

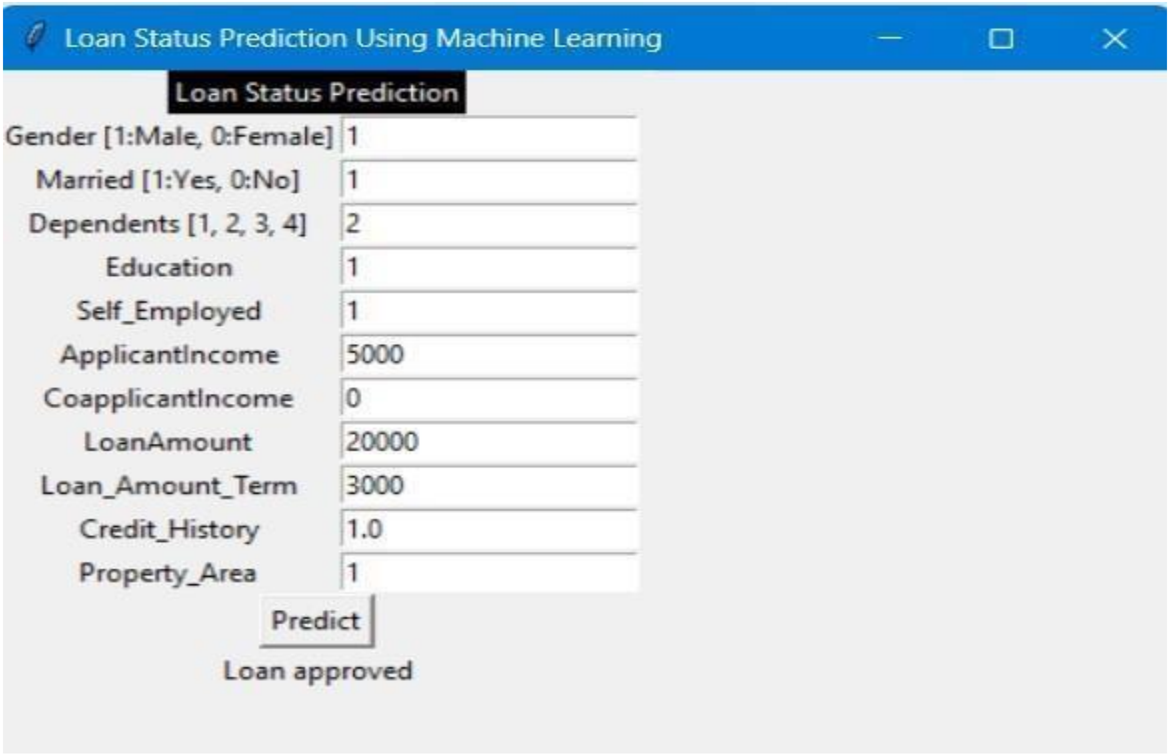


Loan Status Prediction Using Mac...

Loan Status Prediction

Gender [1:Male, 0:Female]	1
Married [1:Yes, 0:No]	0
Dependents [1, 2, 3, 4]	1
Education	1
Self_Employed	1
ApplicantIncome	2000
CoapplicantIncome	0
LoanAmount	10000
Loan_Amount_Term	1
Credit_History	1.0
Property_Area	0

Loan Not Approved



Loan Status Prediction Using Machine Learning

Loan Status Prediction

Gender [1:Male, 0:Female]	1
Married [1:Yes, 0:No]	1
Dependents [1, 2, 3, 4]	2
Education	1
Self_Employed	1
ApplicantIncome	5000
CoapplicantIncome	0
LoanAmount	20000
Loan_Amount_Term	3000
Credit_History	1.0
Property_Area	1

Loan approved

Conclusion:

The system approved or rejects the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. Machine Learning (ML) techniques are very useful in predicting outcomes for large amount of data. In our project, three machine learning algorithms, Navie Bayes(NB), Decision Tree (DT) and Random Forest (RF) are applied to predict the loan approval of customers. The experimental results conclude that the accuracy of Random Forest machine algorithm is better than compared to decision tree and Navie Bayes machine learning approaches.