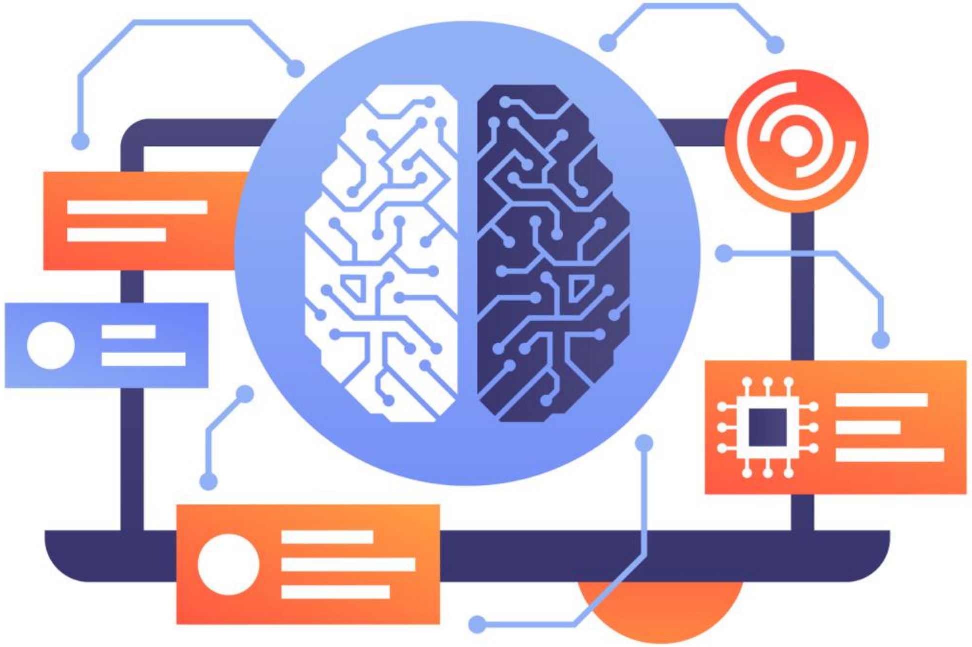
Introduction to Artificial intelligence

Loan Approvals Prediction Using Machine Learning

Done By: Group 4:

1. Piyush Sunil Borse 25PGAI0026
2. Prateek Majumder 25PGAI0027
3. Bhawana Thawarani 25PGAI0137
4. Prajwal Wagh 25PGAI0109
5. Yuvraj Singh Srinet 25PGAI0019

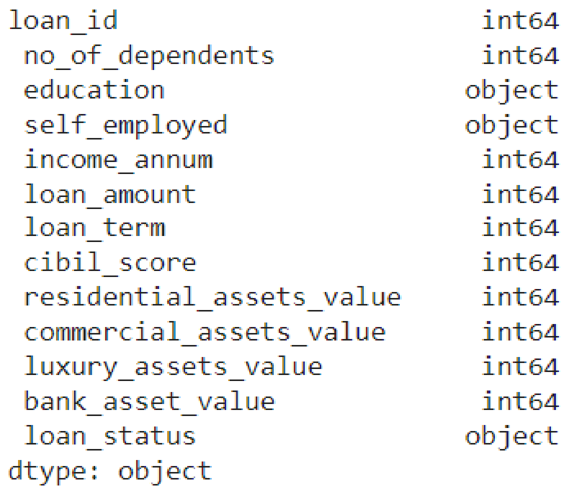


Problem Statement

- The objective of this project is to develop a machine learning model to predict loan approval status based on applicant details and financial information. The dataset includes variables such as number of dependents, education level, employment status, annual income, loan amount, loan term, credit score, and various asset values.
- The project involves preprocessing the data, performing exploratory data analysis, engineering features, selecting and evaluating classification models, and optimizing hyperparameters.
- The deliverables include a detailed report of the entire process, the trained predictive model with performance metrics, and a deployable version of the model. The implementation of this predictive model aims to enhance the loan approval process by improving decision accuracy, reducing default risks, and providing quicker responses to loan applications, thereby increasing overall customer satisfaction.

About Data SET:

- **loan_id**: Unique identifier for each loan application.
- **no_of_dependents**: Number of dependents of the applicant.
- **education**: Educational qualification of the applicant.
- **self_employed**: Employment status of the applicant (self-employed or not).
- **income_annum**: Annual income of the applicant.
- **loan_amount**: Amount of loan requested.
- **loan_term**: Term of the loan.



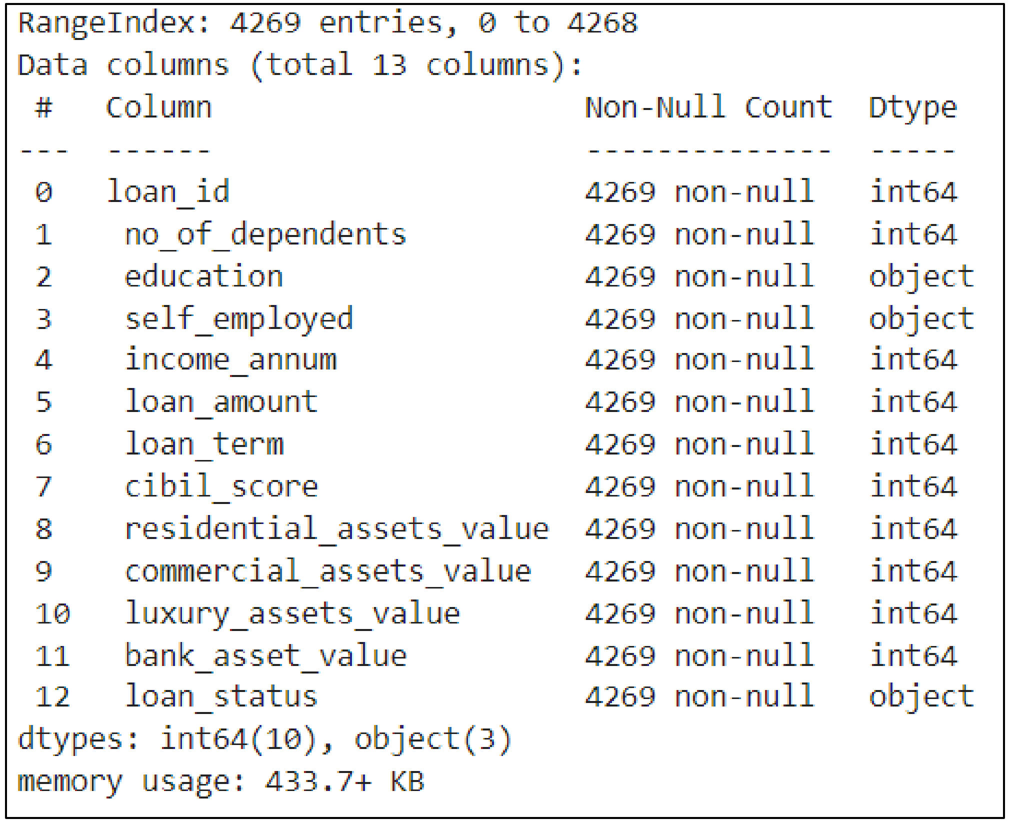
```
loan_id                        int64
 no_of_dependents              int64
 education                    object
 self_employed                object
 income_annum                  int64
 loan_amount                   int64
 loan_term                     int64
 cibil_score                   int64
 residential_assets_value      int64
 commercial_assets_value       int64
 luxury_assets_value           int64
 bank_asset_value              int64
 loan_status                  object
dtype: object
```

- **cibil_score**: Credit score of the applicant.
- **residential_assets_value**: Value of residential assets owned by the applicant.
- **commercial_assets_value**: Value of commercial assets owned by the applicant.
- **luxury_assets_value**: Value of luxury assets owned by the applicant.
- **bank_asset_value**: Total value of assets held in the applicant’s bank.
- **loan_status**: Target variable indicating loan approval status (approved or not approved).

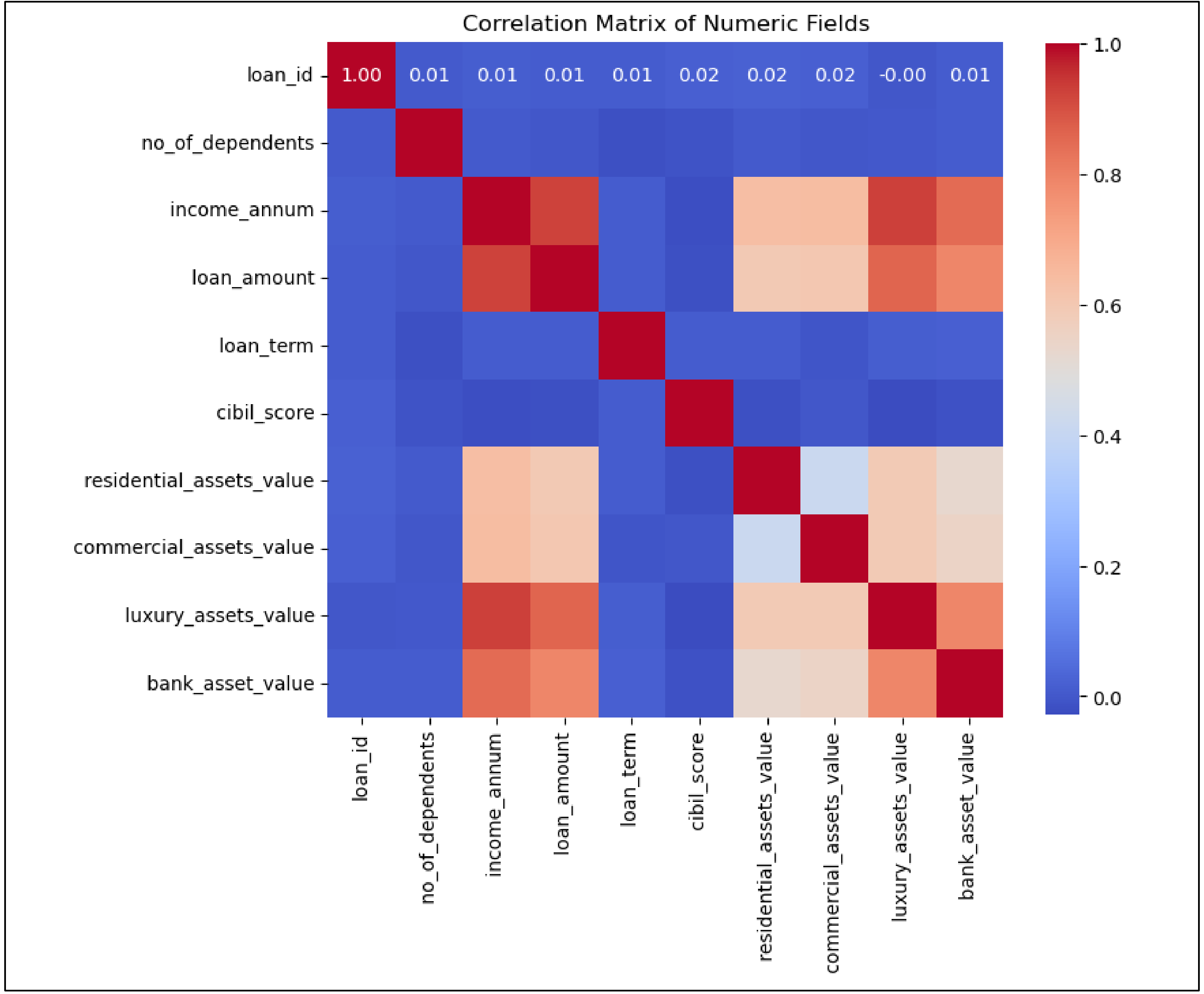
Project WorkFlow:

1. Data Reading
2. Data Exploration
3. Data Visualisation and Analysis
4. Data Preparation and Data Scaling
5. Train Test Split of Data
6. Model Training
7. Model Prediction and Accuracy Metrics
8. Building a GUI Application

Read Data and Analyse



```
RangeIndex: 4269 entries, 0 to 4268
Data columns (total 13 columns):
 #   Column                     Non-Null Count  Dtype 
---  ------                     --------------  ----- 
 0   loan_id                    4269 non-null   int64 
 1    no_of_dependents          4269 non-null   int64 
 2    education                 4269 non-null   object
 3    self_employed             4269 non-null   object
 4    income_annum              4269 non-null   int64 
 5    loan_amount               4269 non-null   int64 
 6    loan_term                 4269 non-null   int64 
 7    cibil_score               4269 non-null   int64 
 8    residential_assets_value  4269 non-null   int64 
 9    commercial_assets_value   4269 non-null   int64 
 10   luxury_assets_value       4269 non-null   int64 
 11   bank_asset_value          4269 non-null   int64 
 12   loan_status               4269 non-null   object
dtypes: int64(10), object(3)
memory usage: 433.7+ KB
```



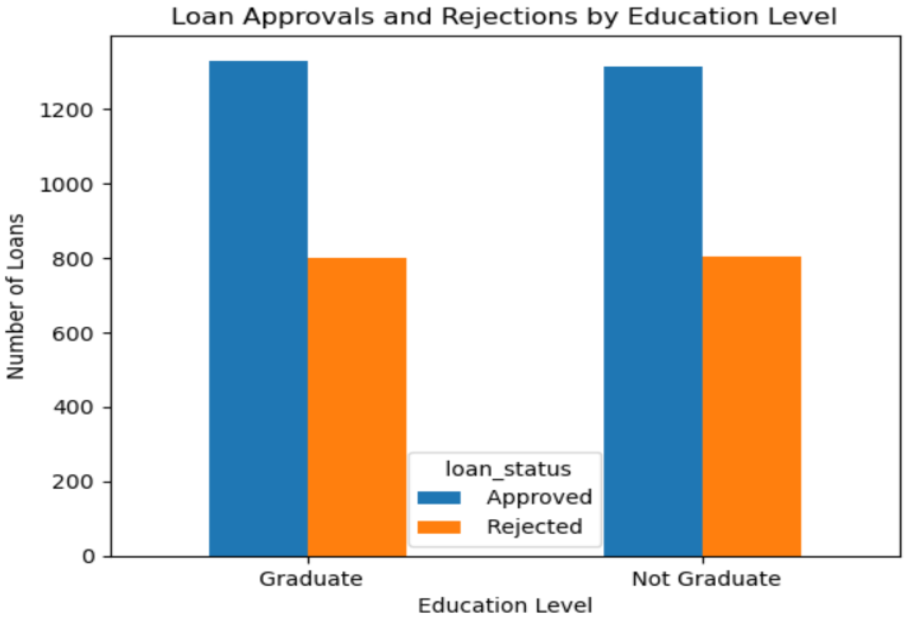
Correlation Matrix of Numeric Fields

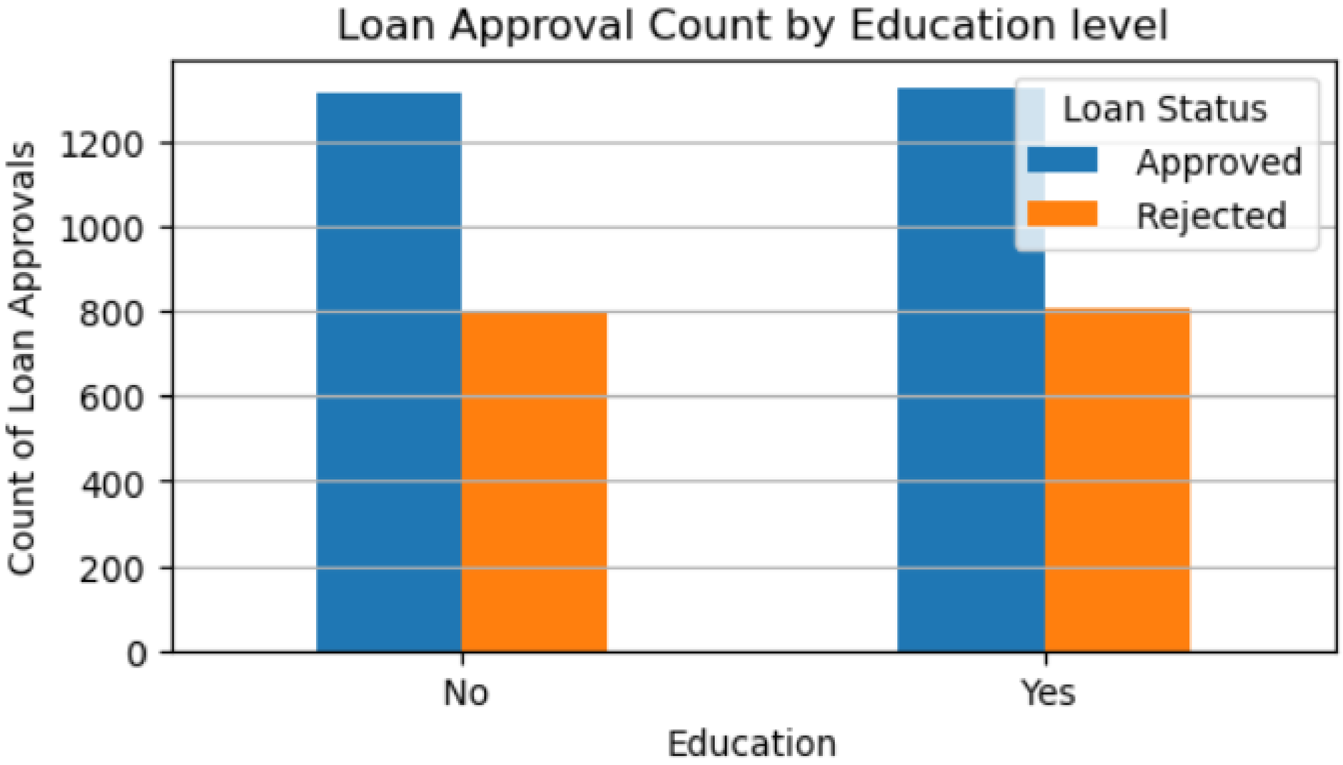
Data Exploration and Data Correction

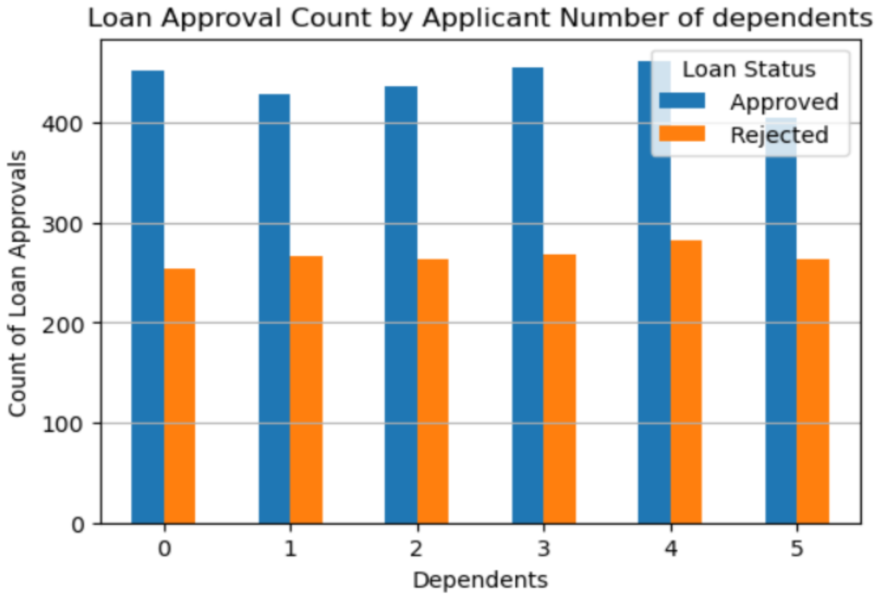
Important Steps:

1. Checking Data Types of columns.
2. Checking for null values.
3. Correlation among data.
4. Getting discriptive statistics of the data.
5. Removing some
   negative values in residential_assets_value field.

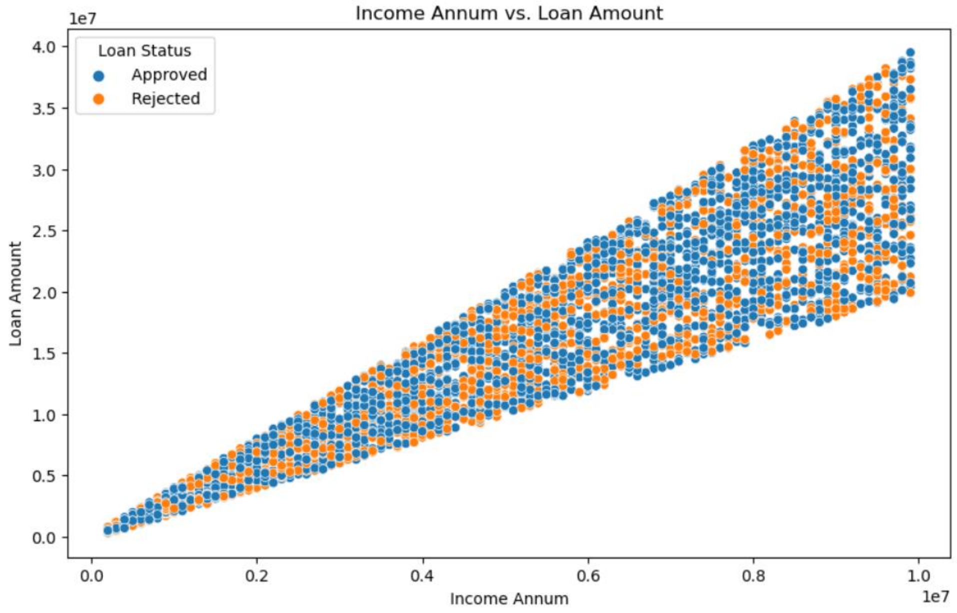
Data Analysis : Plotting and Charting

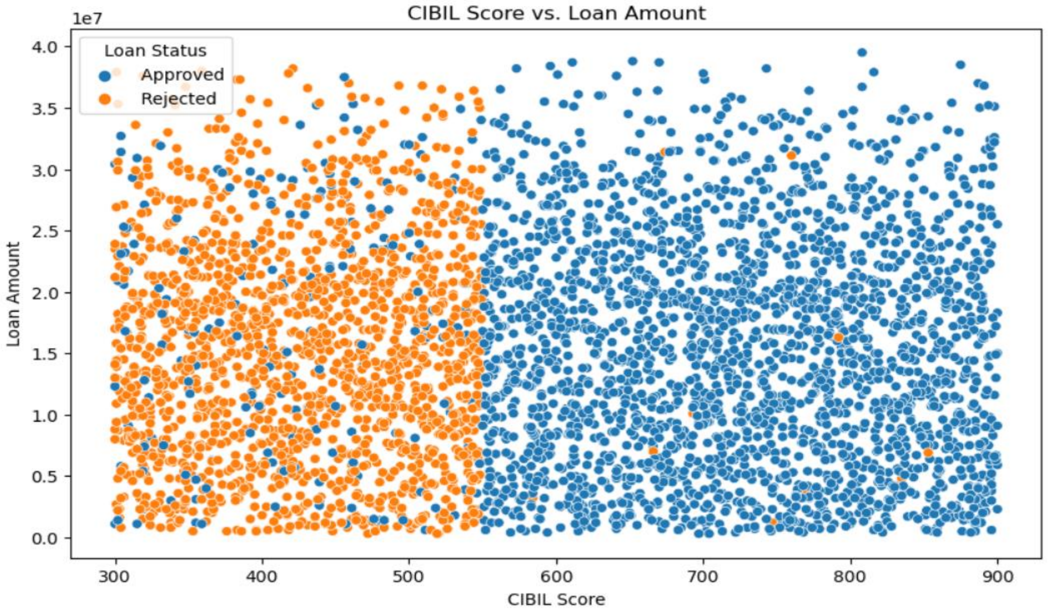


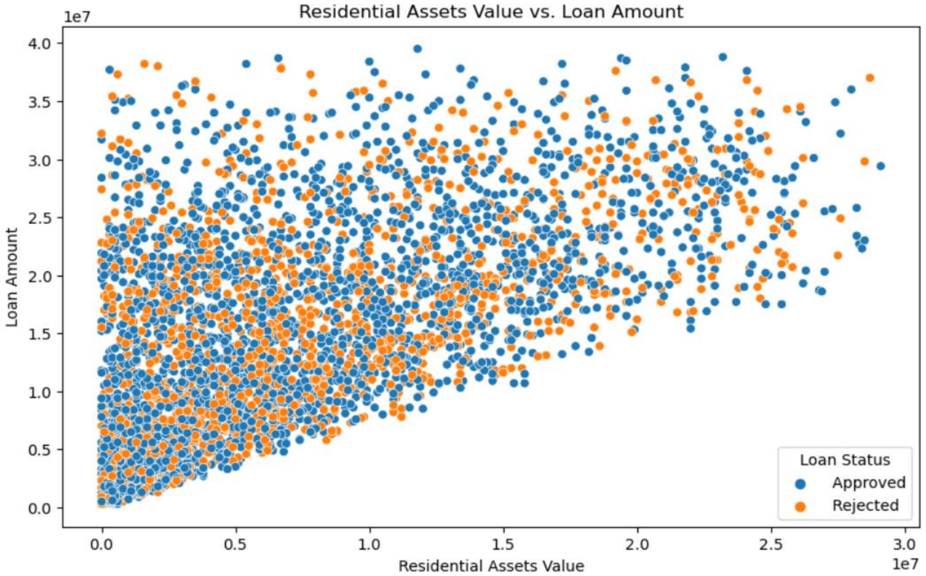




Data Analysis







Convert Categorical Variables To Numeric

'Graduate': 1, 'Not Graduate': 0
'Yes': 1, ' No': 0
'Approved': 1, 'Rejected': 0

- Categorical features refer to string data types and can be easily understood by human beings. However, machines cannot interpret the categorical data directly. Therefore, the categorical data must be converted into numerical data for further processing.
- We mapped categorical variables to numerical values for better processing by machine learning algorithms.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Min-max Scaling

- Min-max scaling, also known as normalization, is a technique commonly used in data preprocessing. It is used to transform numerical features into a specific range, typically between 0 and 1.
- Many machine learning algorithms perform better when the input features are normalized. By scaling the features to a specific range, you can prevent any particular feature from dominating the learning process. This is especially important when working with algorithms that are sensitive to the scale of the data.

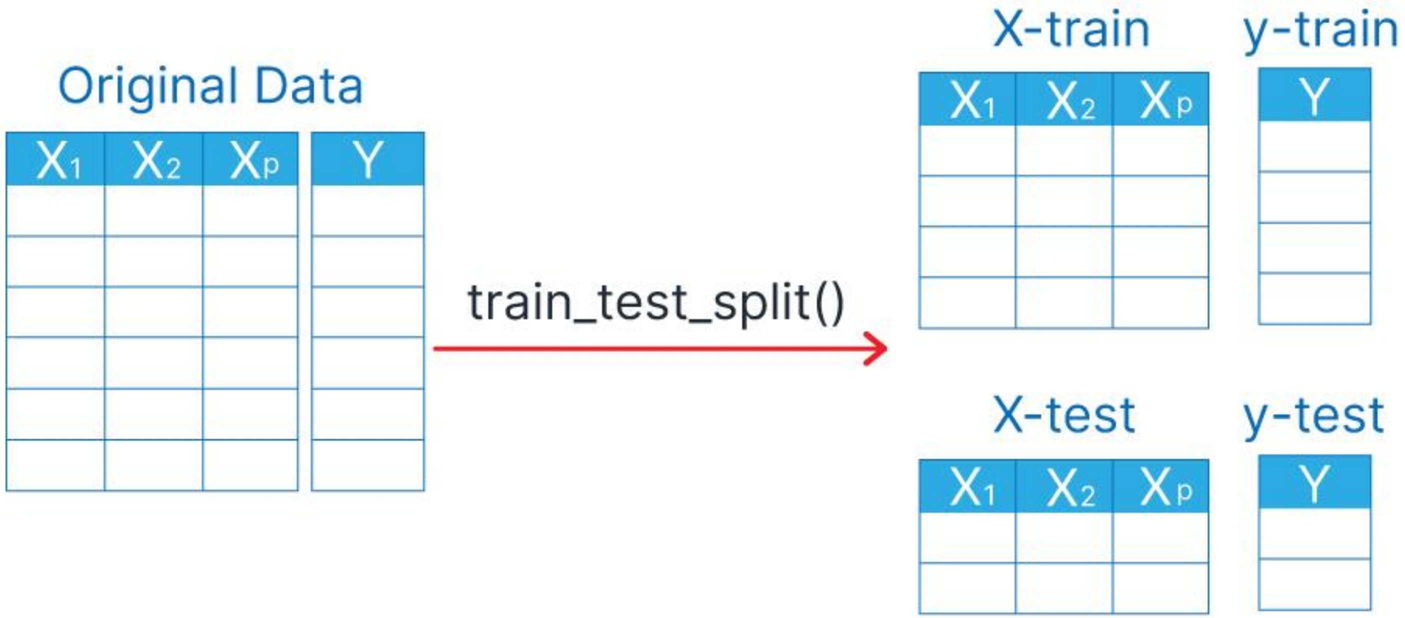
Data Preparation

Input Features: X
Output: y

Supervised machine learning is a type of machine learning that learns the relationship between input and output. The inputs are known as features or X variables and output is generally referred to as the target or y variable. The type of data which contains both the features, and the target is known as labeled data.

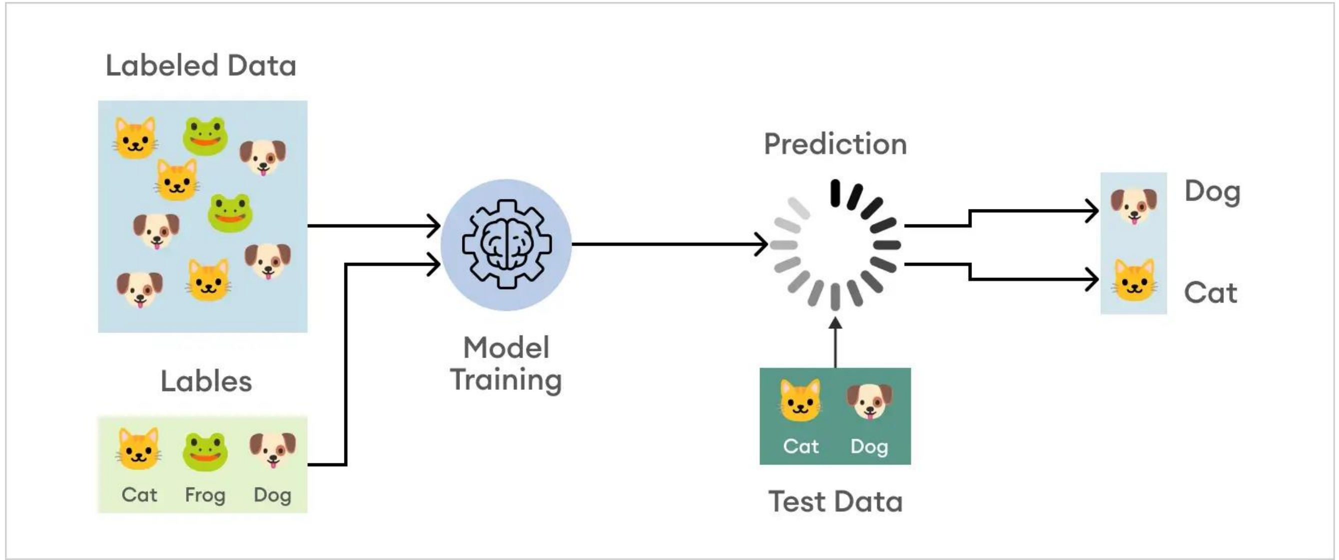
Train Test Split

Train-test split divides the data once into distinct training and test sets used for model evaluation.



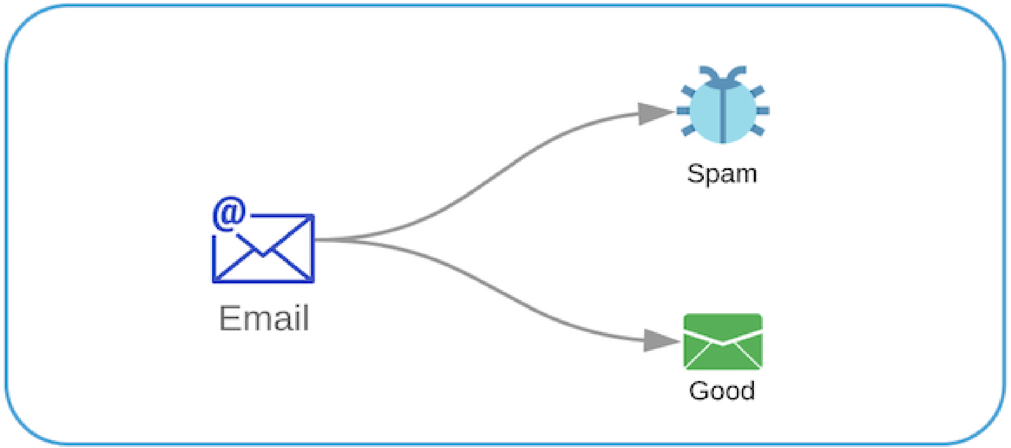
Machine Learning : Classification

Classification is a supervised machine learning method where the model tries to predict the correct label of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data.



Binary Classification:

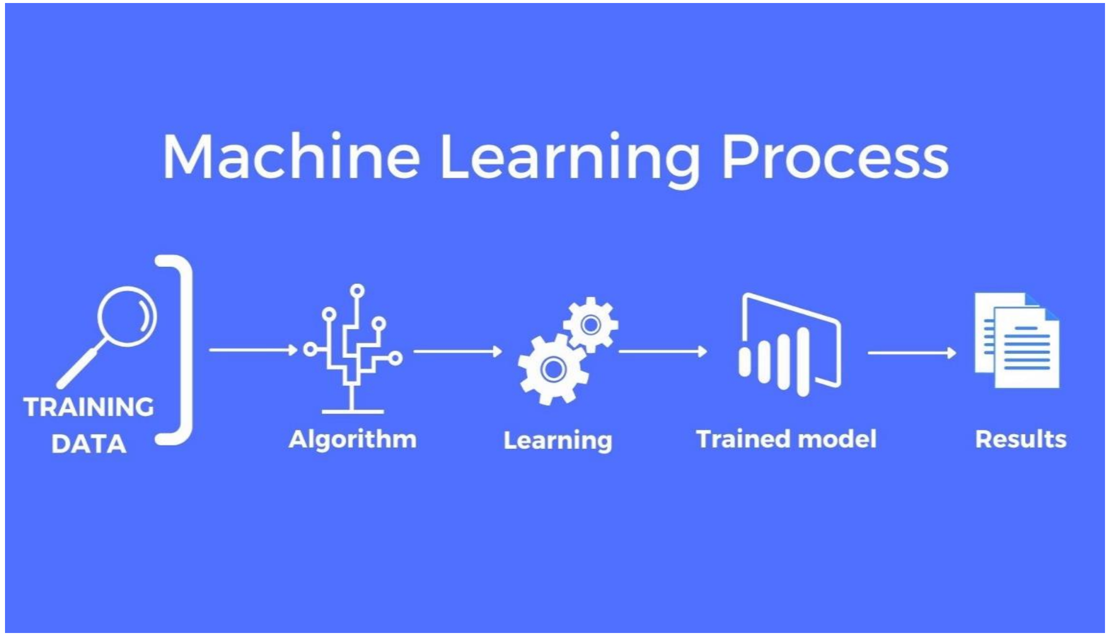
In a binary classification task, the goal is to classify the input data into two mutually exclusive categories. The training data in such a situation is labeled in a binary format: true and false; positive and negative; O and 1; spam and not spam, etc. depending on the problem being tackled. The loan approvals prediction is a binary classification problem.



Binary classification

Model Training

A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output. The training model is used to run the input data through the algorithm to correlate the processed output against the sample output.



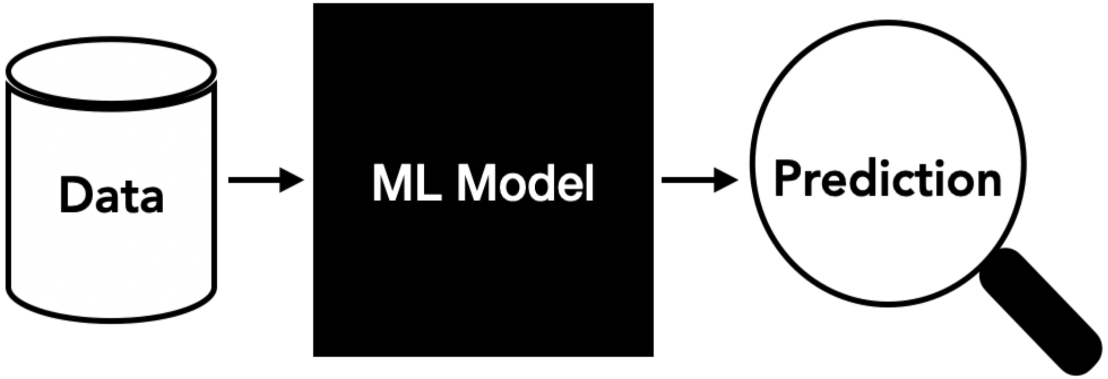
Models Used:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)

SCIKIT-LEARN FOR CLASSIFICATION

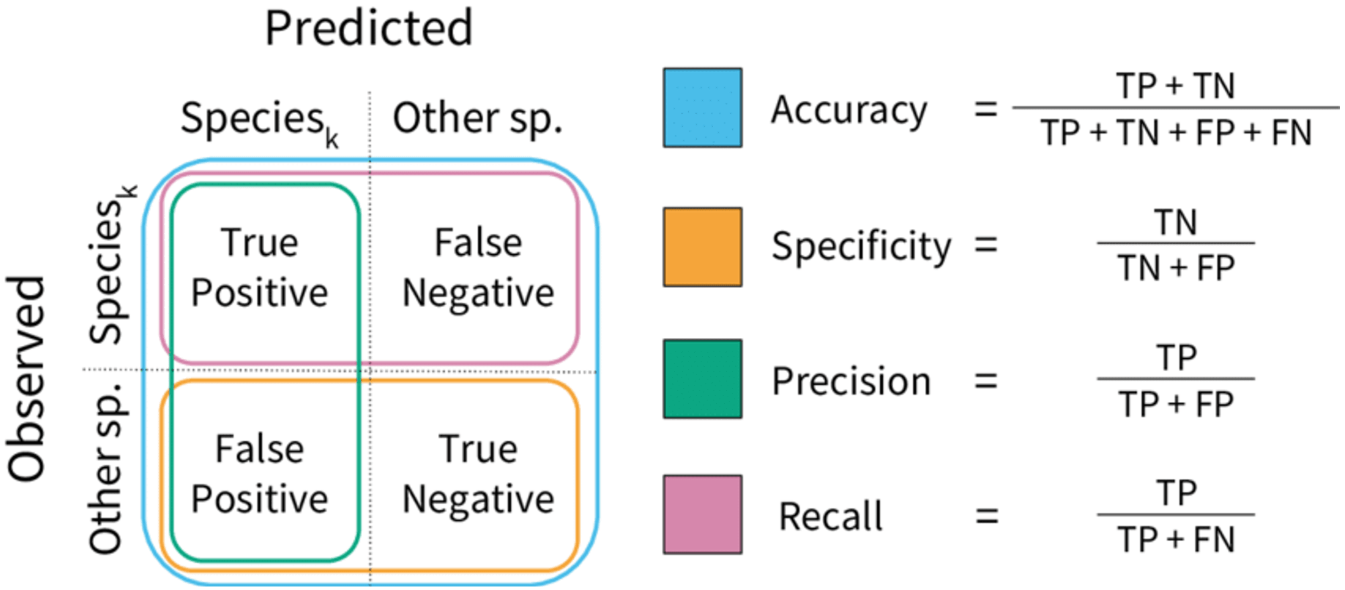
Model Prediction and Accuracy

Each input variable gets a label marking a category. In other words, the classification technique is used to map the input data to one of the categorial output labels.



Accuracy:

Evaluating the performance of your classificatio model is crucial to ensure its accuracy and effectiveness.



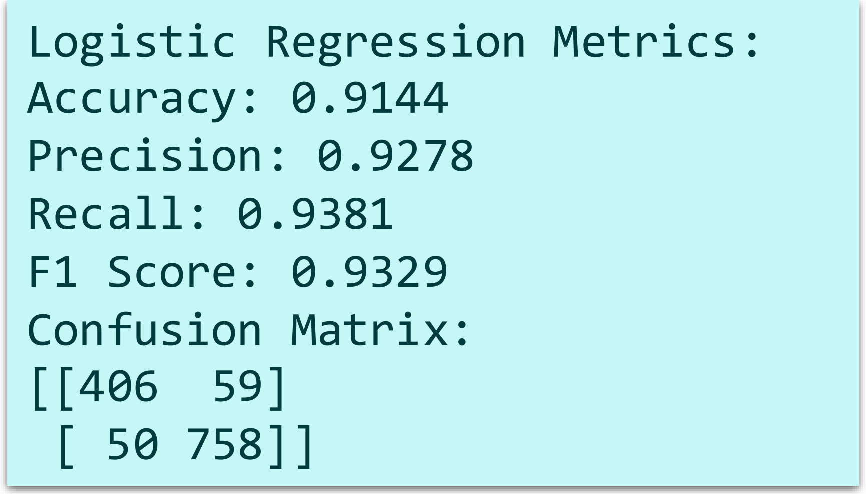
🟦 Accuracy $= \frac{TP + TN}{TP + TN + FP + FN}$

🟧 Specificity $= \frac{TN}{TN + FP}$

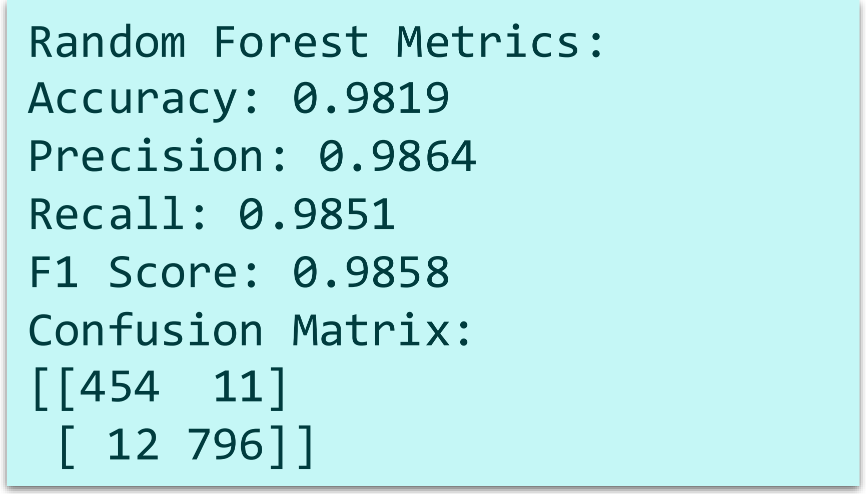
🟩 Precision $= \frac{TP}{TP + FP}$

🟪 Recall $= \frac{TP}{TP + FN}$

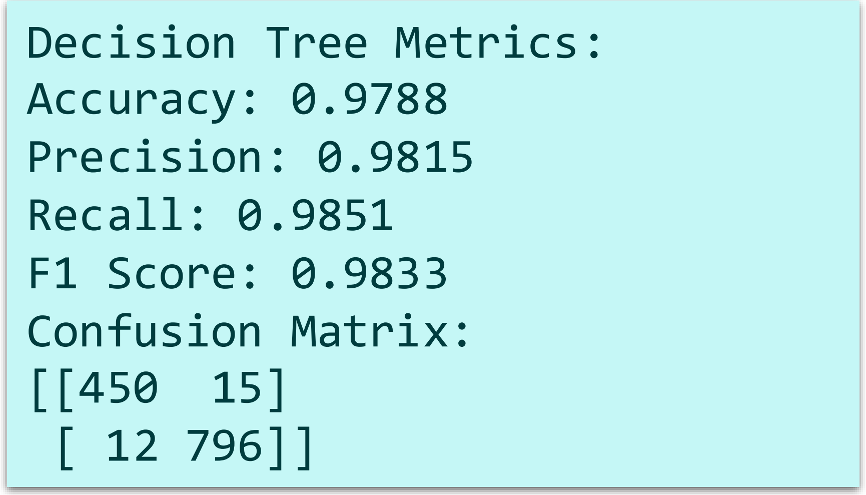
Results



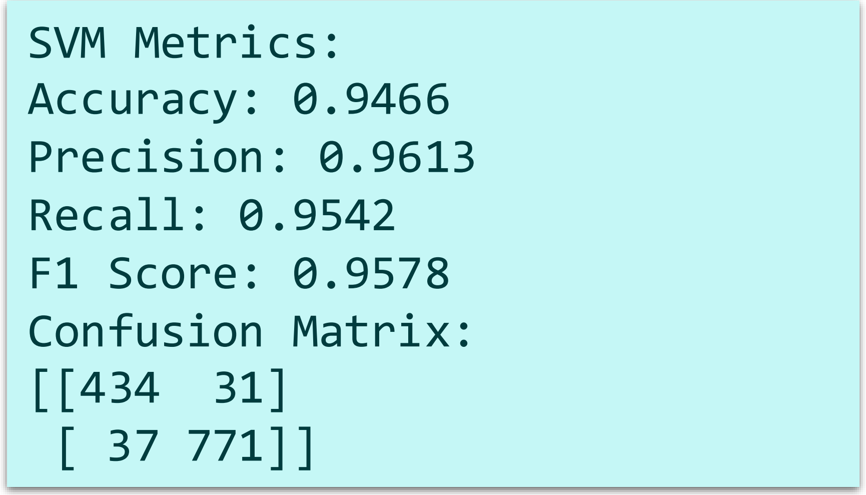
```
Logistic Regression Metrics:
Accuracy: 0.9144
Precision: 0.9278
Recall: 0.9381
F1 Score: 0.9329
Confusion Matrix:
[[406  59]
 [ 50 758]]
```



```
Random Forest Metrics:
Accuracy: 0.9819
Precision: 0.9864
Recall: 0.9851
F1 Score: 0.9858
Confusion Matrix:
[[454  11]
 [ 12 796]]
```



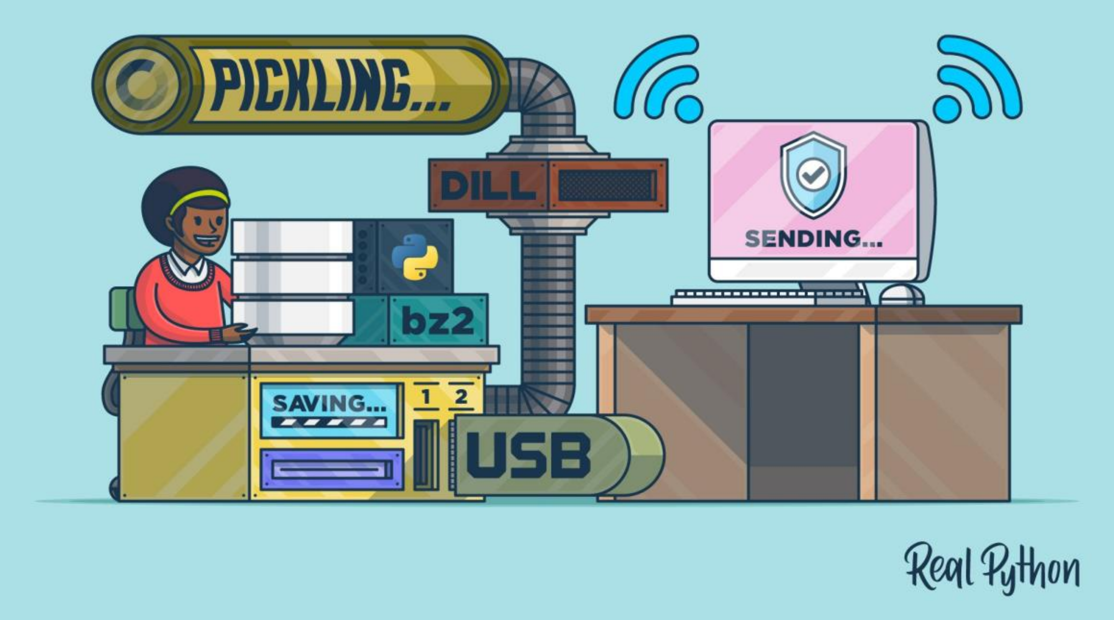
```
Decision Tree Metrics:
Accuracy: 0.9788
Precision: 0.9815
Recall: 0.9851
F1 Score: 0.9833
Confusion Matrix:
[[450  15]
 [ 12 796]]
```

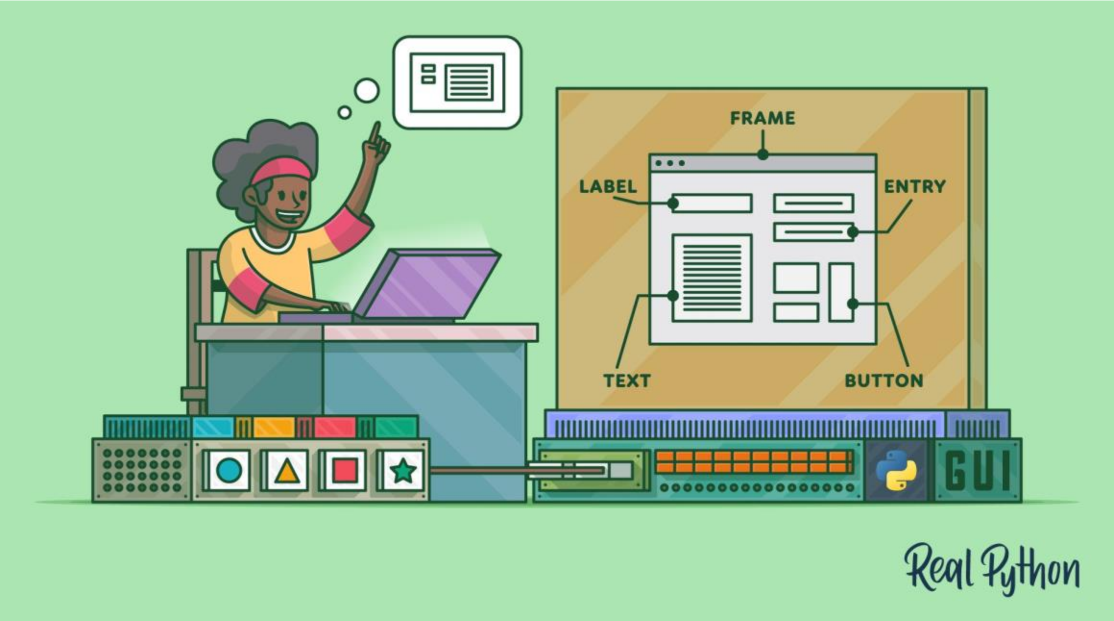


```
SVM Metrics:
Accuracy: 0.9466
Precision: 0.9613
Recall: 0.9542
F1 Score: 0.9578
Confusion Matrix:
[[434  31]
 [ 37 771]]
```

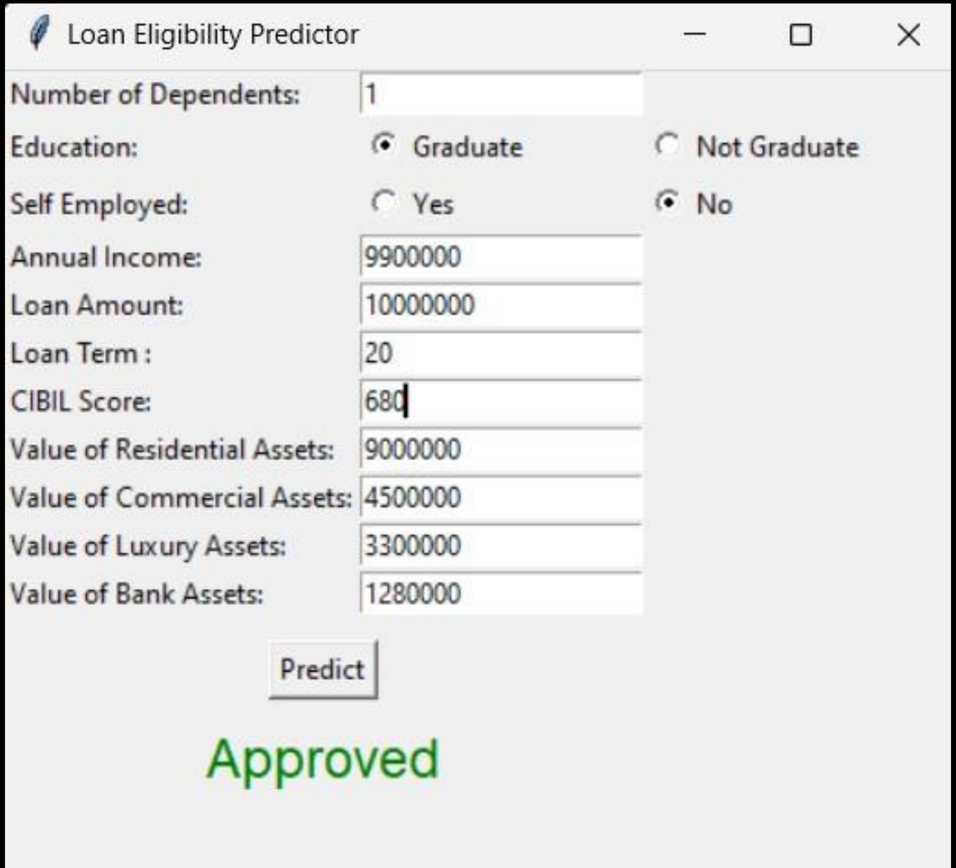
Building an End Product

1. Saving the Model and Scaler
2. Taking test Inputs
3. Scaling the Inputs
4. Passing the inputs to the model
5. Getting the Output
6. Displaying if Loan will be Approved or Rejected
7. Building a GUI





End Product



Loan Eligibility Predictor

Number of Dependents: 1

Education: ● Graduate ○ Not Graduate

Self Employed: ○ Yes ● No

Annual Income: 9900000

Loan Amount: 10000000

Loan Term : 20

CIBIL Score: 680

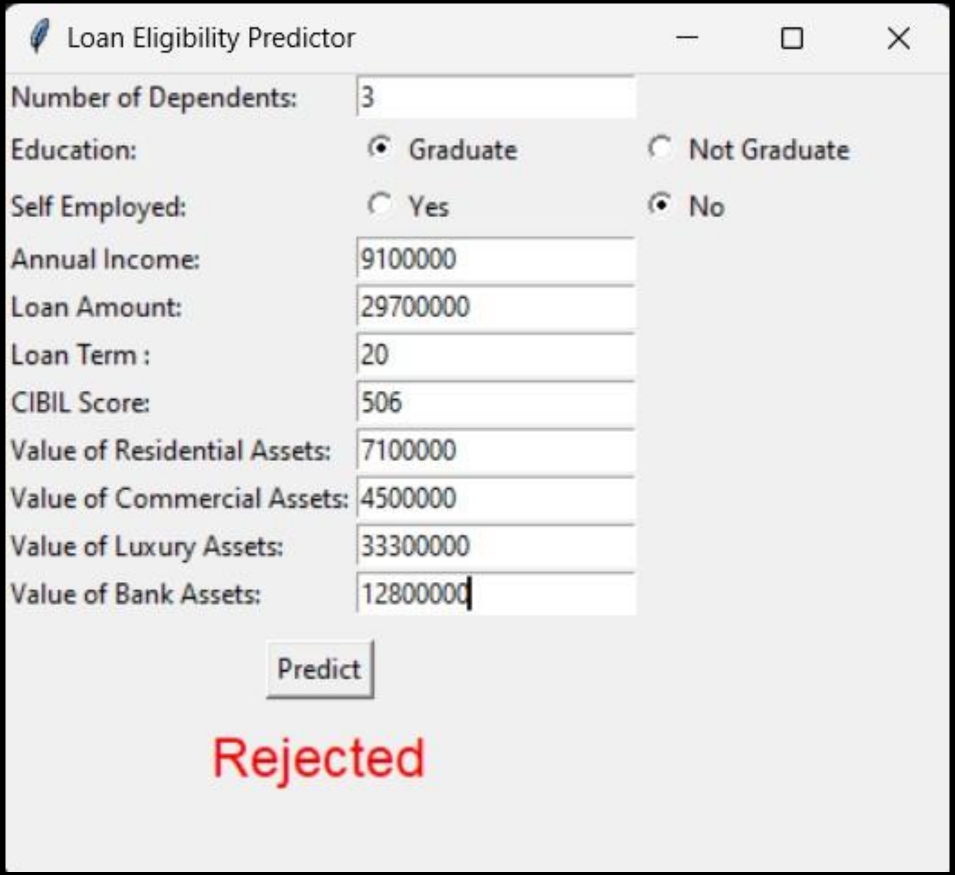
Value of Residential Assets: 9000000

Value of Commercial Assets: 4500000

Value of Luxury Assets: 3300000

Value of Bank Assets: 1280000

Approved



Loan Eligibility Predictor

Number of Dependents: 3

Education: ● Graduate ○ Not Graduate

Self Employed: ○ Yes ● No

Annual Income: 9100000

Loan Amount: 29700000

Loan Term : 20

CIBIL Score: 506

Value of Residential Assets: 7100000

Value of Commercial Assets: 4500000

Value of Luxury Assets: 33300000

Value of Bank Assets: 12800000

Rejected

End Remarks

1. The project involved assessing the performance of different machine learning models on a dataset.
2. The models used were Decision Tree, Random Forest, Logistic Regression, and SVC.
3. Among the models examined, the Random Forest Classifier had the most accuracy in the project.

Thank You