****

**Assessment Report**

on

**“Predict Loan Default”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

**GROUP 8**  
1. Aman Upadhyay - Roll No. 202401100300032  
2. Arnav Mishra - Roll No. 202401100300063   
3. Anshu Kumari - Roll No. 202401100300055   
4. Aditya Shukla - Roll No. 202401100300019   
5. Alok Kharwar - Roll No. 202401100300028

**Under the supervision of**

“Teaching Faculty Name”

**KIET Group of Institutions, Ghaziabad**

# 1. Introduction

With the growth of digital lending platforms, assessing loan applications efficiently and accurately has become a critical task. This project focuses on predicting loan defaults using supervised machine learning algorithms. By analyzing historical loan applicant data, we aim to build models that can automate the credit risk evaluation process, helping financial institutions identify high-risk borrowers and make informed decisions.

# 2. Problem Statement

To develop a classification model that predicts whether a loan applicant will default based on their financial and demographic information, aiding in minimizing loan risks for lenders.

# 3. Objectives

● Preprocess and clean the dataset for training.

● Train Decision Tree and Random Forest classifiers for predicting loan defaults.

● Evaluate models using performance metrics.

● Visualize data distributions, relationships, and feature importance for better interpretability.

# 4. Methodology

● Data Collection:

- Dataset Source: Kaggle Loan Prediction Dataset

● Data Preprocessing:

- Combined train and test data for consistent preprocessing.

- Filled missing categorical values with mode and numerical values with median.

- Label encoding was used to convert categorical variables into numeric format.

● Model Building:

- Separated features (X) and target (y) from the processed training dataset.

- Split the data into training and validation sets (80%-20%).

- Trained both Decision Tree and Random Forest classifiers.

● Model Evaluation:

- Evaluated performance using Accuracy Score and Classification Report.

- Visualized data and results using Seaborn and Matplotlib.

# 5. Data Preprocessing

● Handling Missing Values:

- Categorical: Mode imputation for features like Gender, Married, Dependents, etc.

- Numerical: Median imputation for LoanAmount.

● Encoding: Label encoding was used to convert all categorical features into numerical format.

● Data Split: 80% training and 20% validation split for model development.

# 6. Model Implementation

● Decision Tree Classifier:

- Simple and interpretable classification algorithm that splits data based on feature thresholds.

● Random Forest Classifier:

- Ensemble technique combining multiple decision trees to improve accuracy and reduce overfitting.

# 7. Evaluation Metrics

● Accuracy: Measures the proportion of correctly predicted instances.

● Precision: Indicates the proportion of correctly predicted loan defaults.

● Recall: Measures the ability of the model to detect actual defaults.

● F1 Score: Harmonic mean of precision and recall.

● Classification Report: Includes class-wise precision, recall, and F1-score.

# 8. Results and Analysis

● Decision Tree Accuracy: Achieved moderate performance with basic decision rules.

● Random Forest Accuracy: Outperformed Decision Tree with improved accuracy and balanced precision-recall.

● Visual Insights:

- Loan Status Distribution: Shows class imbalance in dataset.

- Applicant Income Distribution: Wide range of income values, slightly skewed.

- Loan Amount vs Loan Status: Boxplot shows typical loan amounts for approved vs. defaulted.

- Feature Correlation: Heatmap highlights key influencing factors like Credit History and Income.

- Feature Importance: Random Forest identifies the most influential features, including Credit History, Applicant Income, and Loan Amount.

# 9. Conclusion

The Random Forest classifier successfully predicted loan defaults with superior performance over the Decision Tree model. This project highlights the effectiveness of machine learning in risk assessment for financial institutions. Future improvements may include hyperparameter tuning, handling class imbalance using SMOTE, and trying other ensemble models like XGBoost.

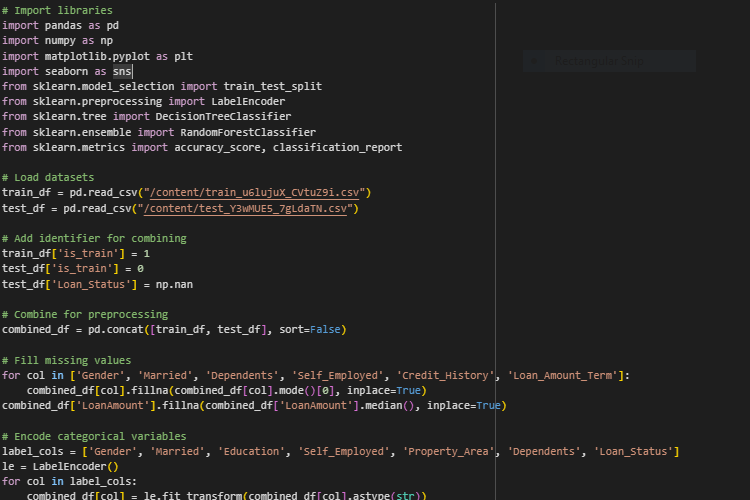
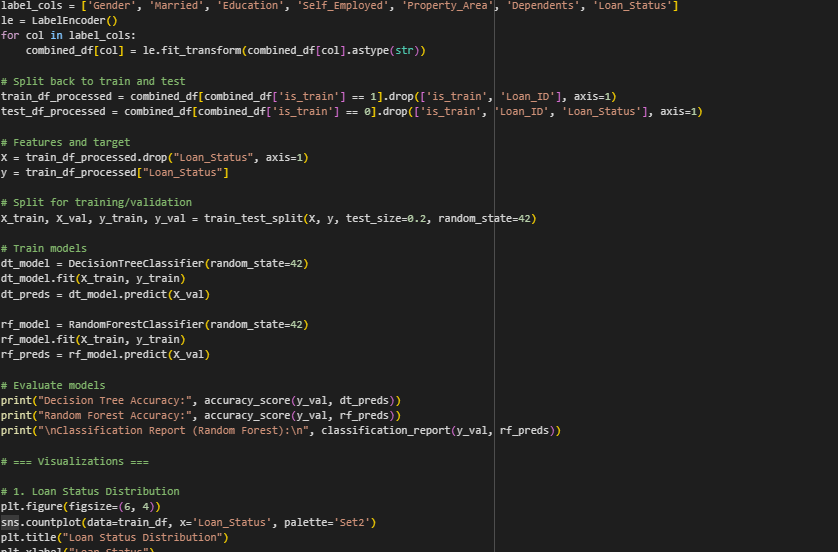
# 10. References

● Scikit-learn Documentation

● Pandas Documentation

● Seaborn Visualization Library

● Research papers and online articles related to credit risk and loan default prediction.



A screen shot of a computer program

AI-generated content may be incorrect.