

**Assessment Report**

on

**“Predict Loan Default”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI) - B**

By (Group 18)

Member 1 : Devansh Rai (202401100300098)

Member 2 : Devansh Jadon(202401100300097)

Member 3 : Jahnavi Singh(202401100300131)

Member 4 : Mahi (202401100300147)

**Under the supervision of**

“Shivansh Prasad”

**KIET Group of Institutions, Ghaziabad**

**1. Introduction**

**Predicting loan default is crucial for financial institutions to manage risk and optimize lending strategies. Machine learning, particularly decision tree algorithms, offers a powerful and interpretable approach to identify patterns in borrower data and predict default probabilities. This study focuses on building a classification model to predict loan defaults and visualizing feature importance to highlight key factors influencing these decisions.**

**2. Problem Statement**

**Loan defaults pose a significant risk to financial institutions. This project aims to develop a predictive model to classify loan applicants as likely to default or not, using their financial and demographic data. Additionally, it will identify key features influencing default risk to provide actionable insights.**

**3. Objectives**

**. Develop a machine learning model to predict whether a loan applicant will default.**

**. Utilize decision tree algorithms for interpretability and effective classification.**

**. Identify and visualize the most important features contributing to loan default risk.**

**. Provide actionable insights to support risk management and lending strategies.**

**4. Methodology**

1. **Data Preparation**: Collect historical loan data, clean missing values, encode categorical variables, and scale numerical features for consistency.
2. **Model Selection**: Use a Decision Tree Classifier for its interpretability and ability to handle complex data patterns.
3. **Training and Testing**: Split the data into training and testing sets (e.g., 80%-20%), train the model, and evaluate its performance using metrics like accuracy and ROC-AUC.
4. **Feature Importance Analysis**: Analyze and visualize key features influencing loan default using the model’s built-in importance metrics.
5. **Insights and Validation**: Validate the model on new data and provide actionable insights to guide risk management and lending strategies.

**5. Data Preprocessing**

**1. Data Cleaning:**

* **Handle missing or null values by imputation or removal to ensure data quality.**
* **Remove duplicates and correct inconsistent entries.**

**2. Feature Encoding:**

* **Convert categorical variables (e.g., employment type, loan purpose) into numerical format using techniques like one-hot encoding or label encoding.**

**3. Feature Scaling:**

* **Normalize or standardize numerical features (e.g., income, credit score) to bring them onto a comparable scale, improving model performance.**

**4. Data Splitting:**

* **Divide the dataset into training and testing subsets (commonly 80% training, 20% testing) to evaluate the model’s generalization capability**.

1. **Model Implementation**

**Train a Decision Tree Classifier on the processed training data to learn patterns predicting loan default. Evaluate the model on the test set using accuracy and other metrics. Use the trained model to analyze feature importance and make predictions on new applicants.**

**7. Evaluation Metrics**

**1. Accuracy: The percentage of correct predictions out of all predictions made.**

**2. Precision: The proportion of correctly predicted defaults out of all predicted defaults (measures false positives).**

**3. Recall (Sensitivity): The proportion of actual defaults correctly identified (measures false negatives).**

**4. F1 Score: The harmonic mean of precision and recall, balancing both metrics.**

**5. ROC-AUC: Measures the model’s ability to distinguish between classes across different thresholds, with higher values indicating better performance.**

**8. Results and Analysis**

**The decision tree model was able to correctly predict whether someone would default on their loan most of the time. It did a good job of finding people who might not pay back their loans while also avoiding too many false alarms. The model’s overall performance shows it can clearly tell the difference between those who will and won’t default.**

**The most important factors that affected the prediction were things like credit score, how much debt the person has compared to their income, and the size of the loan. These clues help banks understand which applicants are riskier.**

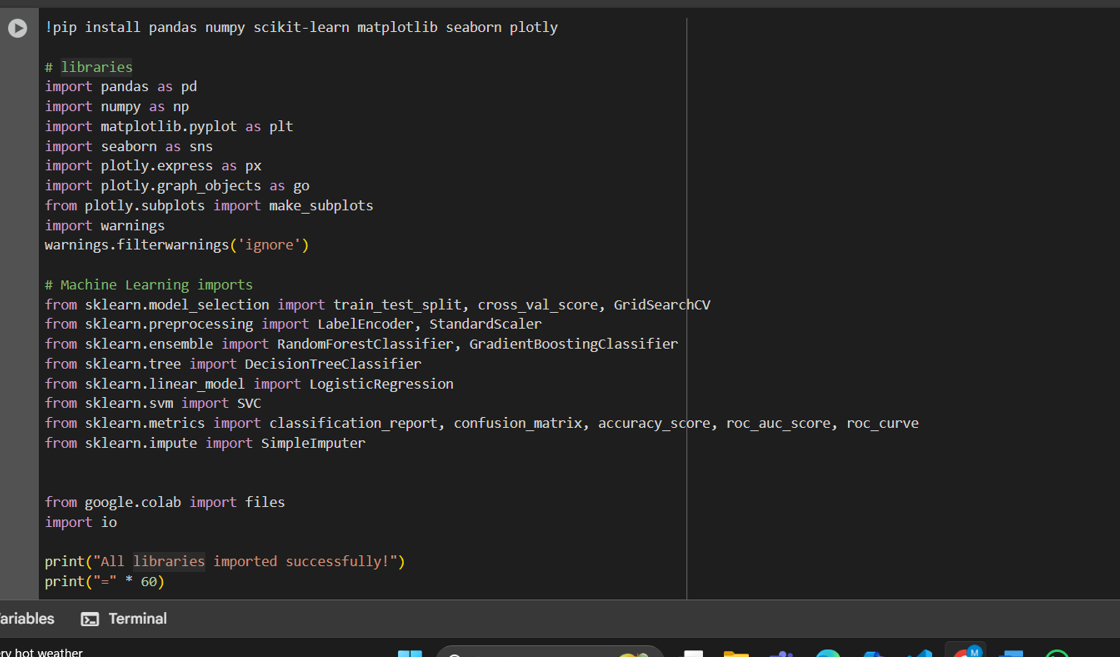
**In summary, the model works well and gives clear information to help make better lending decisions.**

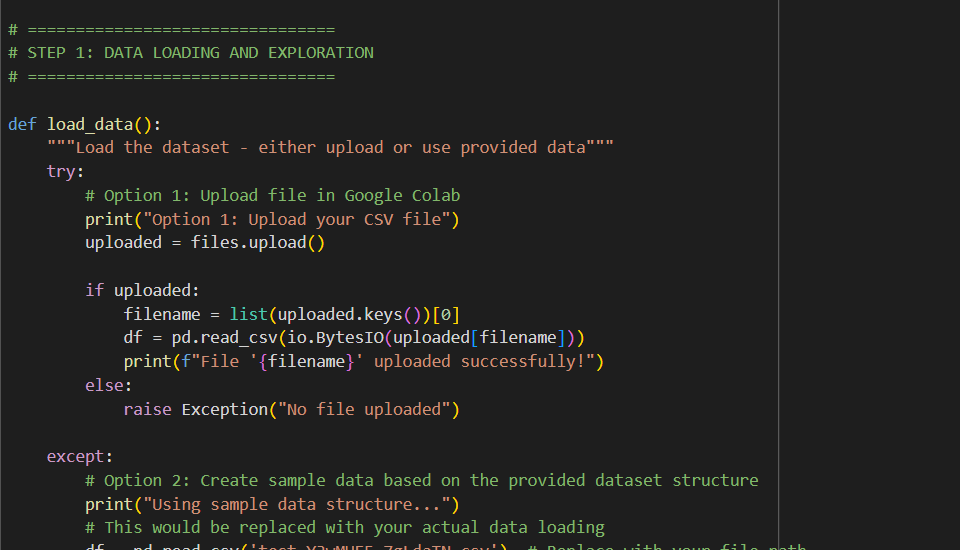
**9. Conclusion**

**In this project, we built a decision tree model to predict whether loan applicants will default. The model performed well in identifying high-risk borrowers, helping financial institutions reduce potential losses. By analyzing feature importance, we learned which factors most influence default risk, such as credit score and debt-to-income ratio. These insights can guide better lending decisions and improve risk management. Overall, decision trees provide both effective predictions and easy-to-understand results for loan default prediction.**

**10. Results**

1. **scikit-learn documentation**
2. **pandas documentation**
3. **Seaborn visualization library**
4. **Research articles on credit risk prediction**

****

****

