Loan Approval Predictor – Group Report

□ Introduction

As part of our machine learning group assignment, we were tasked with building a predictive model to determine whether a loan application is likely to be approved or rejected. We used the Loan Prediction dataset from Kaggle, which includes information about applicants such as income, credit history, employment status, loan amount, etc.

Our goal was to explore the data, prepare it for modeling, build and evaluate different classifiers (Logistic Regression and Decision Tree), and then reflect on which model performs best. We also added visualizations and provided recommendations for improvement.

□ Dataset Overview

Rows: 614 loan applications

Target column: Loan_Status (Y/N – approved or not)

Features: Gender, Married, Education, ApplicantIncome, LoanAmount, Credit_History, and more

Data Source: Kaggle - Loan Prediction Practice Dataset

This is a binary classification problem.

□ Exploratory Data Analysis (EDA)

We started by getting familiar with the data:

Checked basic structure and types using .info() and .describe()

Visualized distributions for key features like ApplicantIncome, LoanAmount, and Credit_History

Plotted a bar chart of loan approvals — around 69% of applicants were approved

Looked at how variables like education, gender, and property area relate to loan status

□ Data Cleaning & Preprocessing
We found missing values in several columns:

Categorical columns: Filled with mode (most common value)

Numerical columns: Filled with the median

Label encoding for binary columns

One-hot encoding for multi-category features

StandardScaler used for Logistic Regression to normalize numeric features

Split dataset: 80% train / 20% test

□ Model Building
We trained and compared two models:

Logistic Regression

Baseline classifier

Simple and interpretable

Decision Tree Classifier

Learns decision rules

Easier to visualize but prone to overfitting

Both models used the same train-test split.

□ Evaluation & Results

MetricLogistic RegressionDecision Tree

Accuracy78.9%67.5%

Precision75.9%73.3%

Recall98.7%78.7%

F1 Score85.9%75.9%

Logistic Regression performed better across most metrics

Especially strong in recall – correctly identified nearly all approvals

Decision Tree had slightly better precision but worse recall

We also reviewed confusion matrices for both the Logistic Regression and Decision Tree models. These matrices allowed us to understand where each model excelled and where it made errors in classifying loan applications.

□Visualizations Used in the Report

Based on the analysis performed, the following visualizations were used:

1. Loan Approval Class Distribution

The bar chart shows the distribution of loan approval statuses ('Y' for approved, 'N' for not approved) in the dataset. It helps visualize the class balance, showing whether one class is significantly more represented than the other.

2. Applicant Income Distribution
The histogram shows the distribution of applicant incomes. It provides insights into the range and skewness of income levels among applicants.

3. Loan Amount by Approval Status
The boxplot compares the distribution of loan amounts for approved ('Y') and not approved ('N') loans. It helps to see if there is a noticeable difference in loan amounts between the two groups.

4. Number of Columns Before and After Encoding
The bar chart illustrates the change in the number of features (columns) in the dataset before and after applying encoding techniques (one-hot encoding) to categorical variables.

5. Loan Status Distribution in Train vs Test Sets
The stacked bar chart compares the distribution of loan approval statuses in the training and test sets after splitting the data. It helps verify that the split maintained a similar class distribution in both sets.

6. Confusion Matrix for Logistic Regression and Decision Tree
These confusion matrices show the performance of the Logistic Regression and Decision Tree models on the test set. They provide a detailed view of each model's classification accuracy.

7. Comparison of Evaluation Metrics

The bar chart compares the key evaluation metrics (Accuracy, Precision, Recall, and F1 Score) for the Logistic Regression and Decision Tree models. It allows for a quick comparison of how each model performed across different metrics.

□ Key Takeaways
Credit history was the most powerful predictor of loan approval

Higher income helped, but wasn’t the only factor

Logistic Regression outperformed Decision Tree in this case

Dataset had slight class imbalance, but not critical

□ Suggestions for Improvement
If extended further:

Try Random Forest or XGBoost

Use GridSearchCV to tune hyperparameters

Apply SHAP values for explainability

Add Streamlit app for interactive predictions

Explore k-fold cross-validation for robustness

□ Conclusion
This project helped us understand the full machine learning pipeline:

Cleaning and preparing data

Exploring features and their impact

Selecting and evaluating models

Communicating results clearly

Logistic Regression provided better performance and interpretability for this dataset. It would be a suitable starting point in a real banking context — especially where recall is critical to avoid wrongly rejecting good applicants.

□□ Group Member
1Nicole Atieno
2.Farida Kendi
3.Desmond Mutuma
4.Sharon were

Course: Artificial intelligence/ Data Science
School:Adanian labs
Supervisor :Sharon Cherop