DATASCIENCE AND AI: KCBF AND ADANIAN LABS

Assignment 3: Lab - Loan Approval Predictor - Project Report

NAME: Group 6

Table of Contents

Loan Approval Predictor - Project Report

Team Members

- Vallary Ogolla
- Emmanuel Kimeu
- Fredrick Njoroge
- Nancy Kamau
- Perpetual Muthaka

1. Project Overview

This group project involved building a supervised machine learning model to predict whether a loan application should be approved or not based on various applicant features. The team completed the end-to-end ML lifecycle, including data exploration, preprocessing, feature selection, model training, and evaluation.

2. Project Goal

Predict loan approval using a binary classification model (Approved or Not Approved).

3. Dataset Information

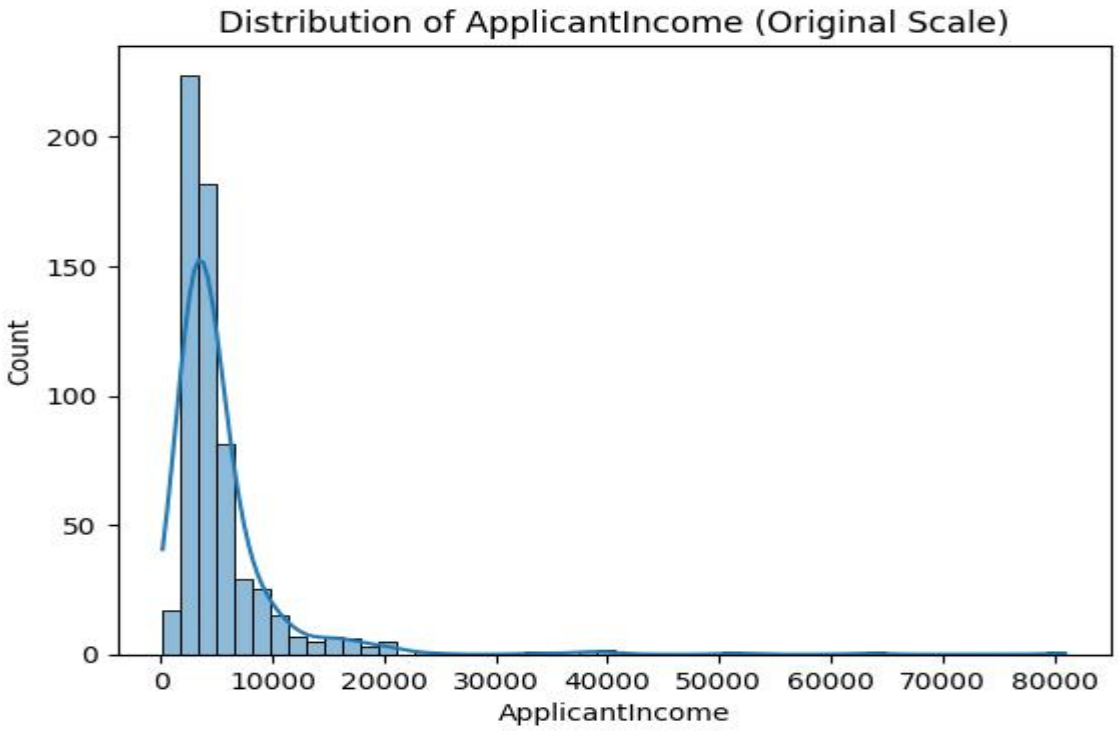
- Name: Loan Approval Dataset
- Source: https://www.kaggle.com/datasets/granjithkumar/loan-approval-data-set
- File Used: Loan_train.csv
- Records: 614 applicants
- Features: Gender, Marital Status, Education, Income, Loan Amount, Property Area, Credit History, etc.

4. Exploratory Data Analysis (EDA)

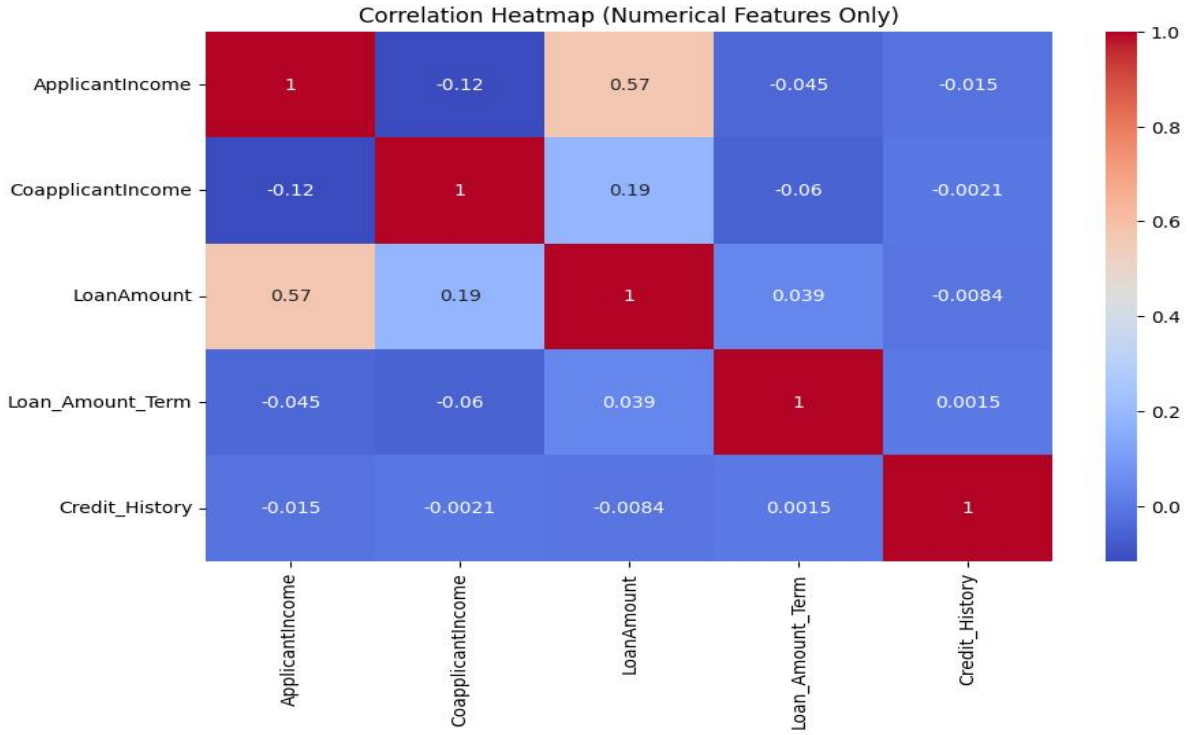
The dataset was inspected for missing values and data types. Distributions of both categorical and numerical features were visualized to understand the structure and patterns in the data. The target variable `Loan_Status` was moderately imbalanced, with 69% approved and 31% not approved but within the accepted rate we had set.

Key observations:

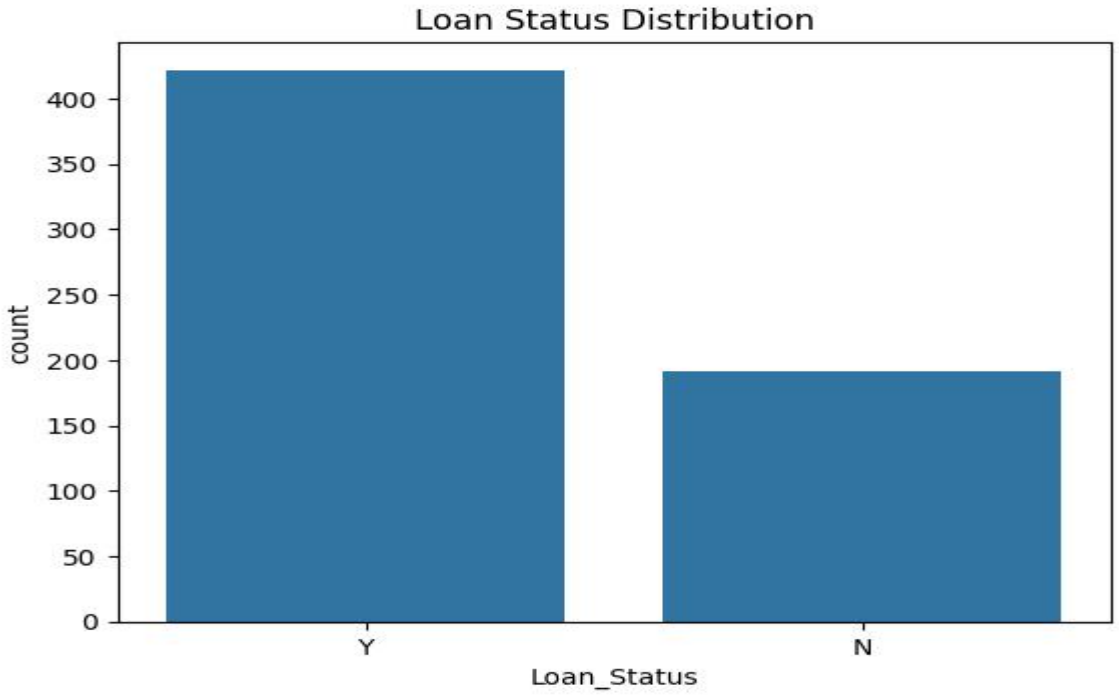
- **Applicant income and loan amount distributions were skewed.**



- **Strong correlation observed with Credit_History.**

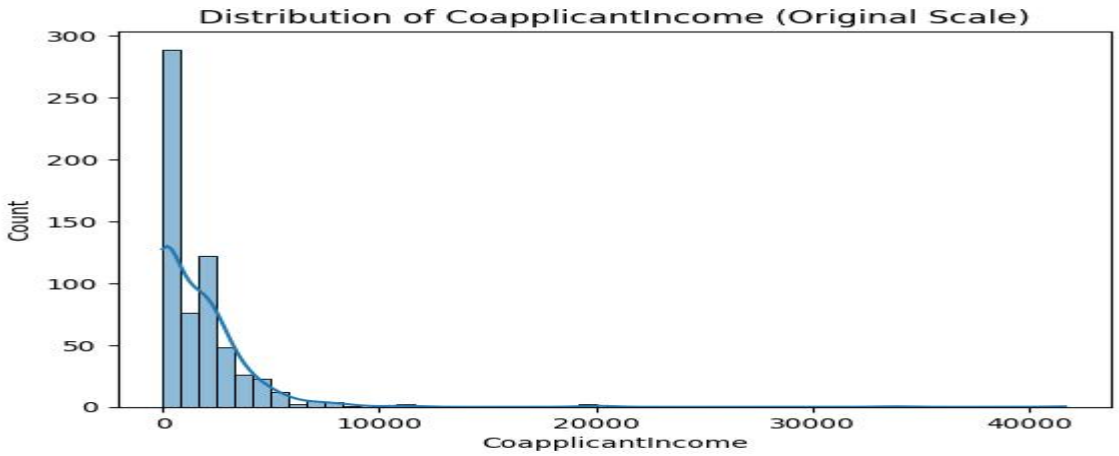


Bar chart of Loan_Status distribution



:

Histograms for numerical features



- Some features had missing values (e.g., Gender, Married, LoanAmount) and we filled them and dropped some.

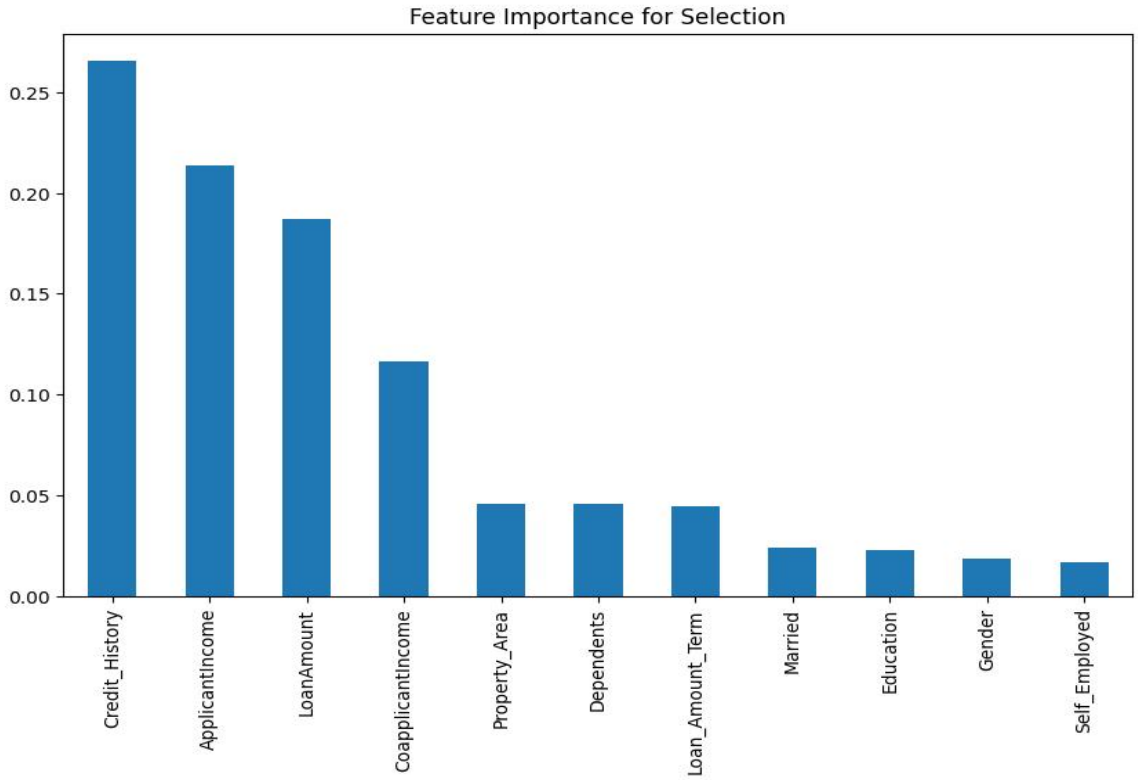
5. Data Preprocessing

- Missing values were filled using the mode or mean.
- '3+' values in Dependents were cleaned to numeric values.
- Label Encoding was applied to categorical variables.
- Numerical features (ApplicantIncome, CoapplicantIncome, LoanAmount) were scaled using StandardScaler.

6. Feature Selection

RandomForestClassifier was used to determine feature importance. Using SelectFromModel, we retained the most impactful features for model training. This improved performance and interpretability.

Bar chart of feature importance from Random Forest



7. Model Building

Two models were built and trained:
- Logistic Regression
- Decision Tree Classifier

An 80/20 split was used for training and testing sets.

8. Model Evaluation

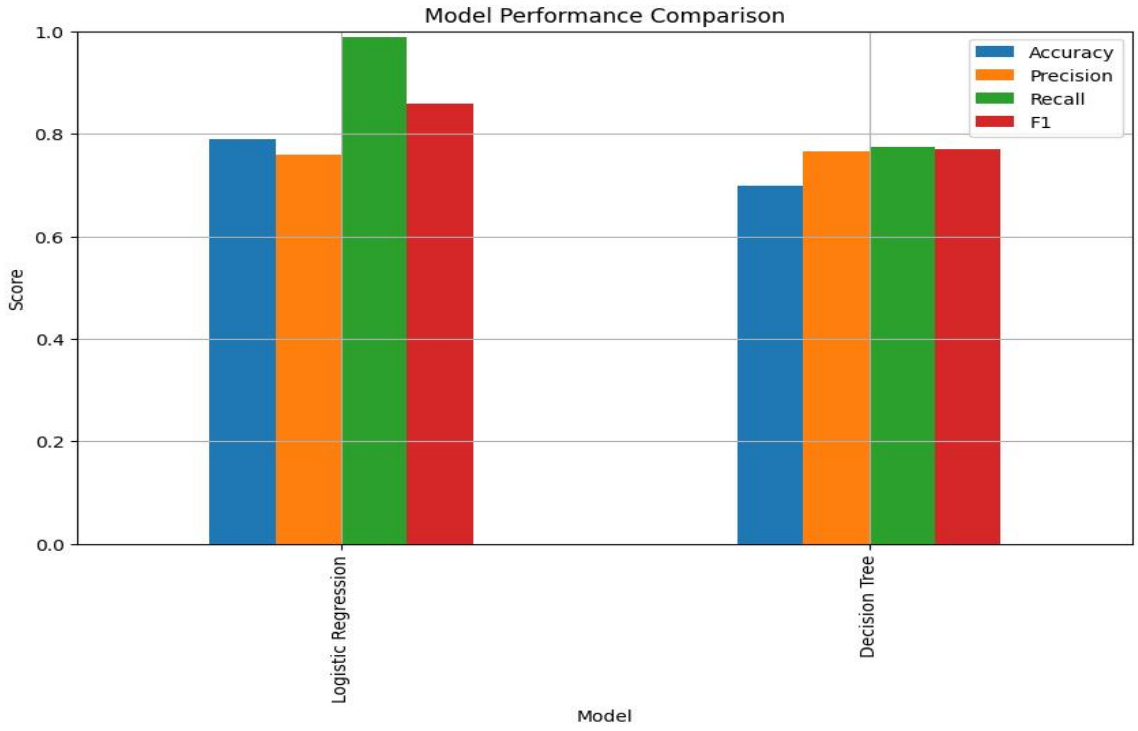
The models were evaluated using the following metrics:
- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

Logistic Regression outperformed the Decision Tree in all key metrics.

Logistic Regression Results:
Accuracy: 78.9%
Recall: 98.75%
F1 Score: 0.86

Decision Tree Results:
Accuracy: 69.9%
Recall: 77.5%
F1 Score: 0.77

Bar chart comparing model performance



9. Conclusion & Recommendation

Based on the evaluation metrics, Logistic Regression was chosen as the final model by our team for this project. It had the highest recall and F1-score, ensuring minimal rejection of eligible applicants. The Decision Tree, although interpretable, did not generalize as well.

This end-to-end ML pipeline demonstrates how data-driven decision-making can be applied to financial services, ensuring both fairness and efficiency.