Loan Approval Prediction Using Machine Learning
Exploratory Data Analysis & Model Evaluation

1. Project Overview

This project is a **machine learning case study** aimed at understanding the patterns and relationships in a bank's **loan approval dataset**, and to assess the suitability of logistic regression as a predictive model. It includes **data cleaning, exploratory data analysis (EDA), correlation analysis**, and **model evaluation**.

2. Data Cleaning and Preprocessing

The dataset contains demographic and financial information such as **Gender, Education, ApplicantIncome, LoanAmount, and Loan_Amount_Term**. Missing values were found in the `LoanAmount` column.We handled this using the **median imputation technique** to avoid outlier influence:

3. Exploratory Data Analysis (EDA)

3.1 .Univariate Analysis:

- **Gender**: Male applicants dominate (477 vs. 109 females).
- **Education**: Majority (457) are **graduates**.
- **Self_Employed**: Most applicants are **not self-employed** (only 80 are self-employed).
- **Loan_Amount_Term**: 360 months is the most common loan duration (~504 cases).
- **Income**: Most **ApplicantIncome** values lie below 10,000.
- **LoanAmount**: Positively skewed; most loan amounts lie between 100 and 250.

3.2 Bivariate Analysis Insights:

- **Income vs LoanAmount vs Loan_Status**: Higher income can lead to higher loans, but **loan approval does not guarantee** with higher income alone.
- **Education vs LoanAmount**: Graduates request slightly higher and more variable loan amounts.
- **Loan_Amount_Term vs Income**: Regardless of income, most people select **360-month terms**.

1. Model Evaluation – Logistic Regression

- We built a **Logistic Regression** model for binary classification (`loan_approval` = 1 for approved, 0 for not approved). Here's the performance:

2. Correlation MatrixAnalysis

- The highest correlation is seen between:

  a. `LoanAmount` and `ApplicantIncome` → **0.52**
  b. `LoanAmount` and `CoapplicantIncome` → **0.21**

- Target variable `loan_approval` has very weak correlations:

  c. Education: **-0.07**
  d. LoanAmount: **-0.05**
  e. ApplicantIncome: **-0.0**

3.Classification Report Summary:

Metric	Class 0 (Not Approved)	Class 1 (Approved)
Precision	0.30	0.60
Recall	0.30	0.60
F1-Score	0.30	0.60

Overall Accuracy: **0.49** (or 49%)

4.Model Issues Identified:

- **Low accuracy** and **imbalanced class predictions**.
- Model fails to **generalize well** on both classes.
- **Recall for non-approved loans is very low**, which is risky for real-world credit scoring.

5..Final Conclusion and Recommendations

- The dataset shows **diverse applicant profiles**, making it suitable for ML experimentation.
- However, the **Logistic Regression model** performs poorly due to:
  - Weak feature correlations
  - Categorical complexity
  - Class imbalance
- Therefore, **Logistic Regression is not a suitable choice** for this problem.

6.Recommended Next Steps:

- Use **tree-based ensemble models** that handle non-linearity and category interactions better:
  - **Random Forest**
  - **XGBoost**
  - **LightGBM**
- Apply **SMOTE or other class balancing techniques** to improve fairness across classes.
- Perform **feature engineering** on categorical data using label encoding or one-hot encoding.