Loan Approval Prediction – Final Project Report

# 1. Executive Summary

This project aims to develop a machine learning model that predicts whether a loan application should be approved or rejected. By analyzing historical loan data, we implemented two classification models – Logistic Regression and Decision Tree – to support automated, data-driven decision-making for financial institutions. The models were trained, evaluated, and compared based on multiple metrics, with Logistic Regression offering better generalization performance.

# 2. Project Objective

To build a binary classification model that predicts loan approval status ('Y' for Approved, 'N' for Not Approved) based on applicant data. This contributes to automating and improving consistency in the loan screening process.

# 3. Dataset Information

Source: Kaggle  
Dataset Name: Loan Approval Dataset  
File Used: Loan\_train.csv  
Size: 614 records, 13 columns  
Target Variable: Loan\_Status

# 4. Tech Stack

• Language: Python  
• Environment: Jupyter Notebook  
• Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn

# 5. Exploratory Data Analysis (EDA)

We explored patterns in the dataset using visualizations and summary statistics. Key insights include:  
• Applicants with credit history had higher approval rates.  
• Graduate applicants had slightly higher approval rates.  
• Self-employed applicants were less likely to be approved.

# 6. Data Cleaning and Preprocessing

• Missing values were handled using mode imputation for categorical and numerical features.  
• Label Encoding and One-Hot Encoding were used for categorical variables.  
• Numeric features were scaled using StandardScaler (for Logistic Regression).  
• The dataset was split into 80% training and 20% testing.

# 7. Model Development

Two machine learning models were developed and trained:  
• Logistic Regression  
• Decision Tree Classifier

# 8. Model Evaluation

Evaluation metrics used include Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.

Performance comparison:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Logistic Regression | Decision Tree | Best Performing | Comments |
| Accuracy | 0.8618 |  | Logistic Regression | Higher overall accuracy |
| Precision | 0.8400 |  | Logistic Regression | More precise positive predictions |
| Recall | 0.9882 |  | Logistic Regression | Captured more true positives |
| F1-score | 0.9081 |  | Logistic Regression | Balance between precision and recall |

# 9. Additional Analysis

• Feature importance highlighted 'Credit History' as the most influential feature.  
• Decision Tree showed signs of overfitting and may benefit from pruning or regularization.  
• No explicit feature selection step was performed as models performed well with all inputs.

# 10. Conclusions and Recommendations

• Logistic Regression is recommended for deployment due to its consistent performance and simplicity.  
• Future improvements could include using ensemble models like Random Forest or XGBoost.  
• Model should be regularly updated with new data and fairness audits to ensure unbiased decisions.

# 11. Project Members

• Eric Otuoma  
• Susan Muthoki  
• Belinder Otieno  
• Imani Lunjala  
• Martsellah Osachi  
• Sharon Ngala

# 12. GitHub Repository

Link: https://github.com/devotuoma/loan\_approval\_prediction