Loan Eligibility Prediction Project Report

1. Introduction

1.1. Background

Loan eligibility prediction is crucial for financial institutions to assess the risk associated with loan applicants. By analyzing applicant data, banks can predict whether a loan should be approved or not, minimizing the risk of default.

1.2. Objective

The objective of this project is to build a model that predicts whether a loan will be approved based on various features such as applicant income, education, loan amount, credit history, etc.

2. Data Overview

2.1. Data Collection

The dataset used in this project was obtained from the "loan.csv" file. It contains information about loan applicants, including demographic details, income, loan amount, and whether the loan was approved or not.

2.2. Data Description

The dataset contains the following columns:

- **Loan_ID**: Unique Loan ID
- **Gender**: Male/Female
- **Married**: Applicant marital status
- **Dependents**: Number of dependents
- **Education**: Applicant education level
- **Self_Employed**: Whether the applicant is self-employed
- **ApplicantIncome**: Applicant's income
- **CoapplicantIncome**: Co-applicant's income
- **LoanAmount**: Loan amount requested
- **Loan_Amount_Term**: Term of the loan in months
- **Credit_History**: Whether the applicant has a credit history
- **Property_Area**: The area where the property is located (Urban/Semiurban/Rural)
- **Loan_Status**: Whether the loan was approved or not (Target variable)

2.3. Data Preprocessing

Before building the model, the data underwent preprocessing:

- **Missing Values**: Several columns had missing values, which were imputed appropriately.
- **Feature Transformation**: The `LoanAmount` feature was log-transformed to normalize its distribution.
- **Encoding**: Categorical variables were encoded using one-hot encoding and label encoding as necessary.
- **Outliers**: Outliers in numerical features were detected and treated.

3. Exploratory Data Analysis (EDA)

3.1. Univariate Analysis

Each feature was analyzed individually:

- **Loan_Status**: The target variable was imbalanced, with a higher proportion of approved loans.
- **ApplicantIncome**: Most applicants had a relatively low income, with a few high-income outliers.
- **LoanAmount**: Loan amounts were mostly concentrated in the lower range, with a few large loans.

3.2. Bivariate Analysis

Relationships between the target variable and other features were explored:

- **ApplicantIncome vs. Loan_Status**: Higher income did not necessarily lead to loan approval.
- **Credit_History vs. Loan_Status**: Applicants with a credit history had a significantly higher chance of loan approval.

3.3. Multivariate Analysis

The combined effect of multiple features on loan approval was analyzed. For example, the combination of `Credit_History` and `Property_Area` revealed insights into the likelihood of loan approval.

4. Model Building

4.1. Model Selection

Several machine learning models were considered, including:

- **Logistic Regression**
- **Decision Tree Classifier**
- **Support Vector Machine (SVM)**
- **Random Forest Classifier**

4.2. Model Training

The dataset was split into training and testing sets. Each model was trained on the training set and validated on the testing set. Hyperparameter tuning was performed using cross-validation to optimize model performance.

4.3. Model Evaluation

Models were evaluated based on accuracy, precision, recall, F1-score, and AUC-ROC curve. The Support Vector Machine (SVM) model showed the best performance with an accuracy of **X%** and an AUC of **Y**.

5. Results and Discussion

5.1. Key Findings

- **Credit History** was the most significant predictor of loan approval.
- **Applicant Income** and **Loan Amount** also played crucial roles, though their impact was less than Credit History.

5.2. Model Performance

The SVM model, with its superior performance metrics, was selected as the final model for predicting loan eligibility.

5.3. Limitations

- The dataset was imbalanced, which may have impacted model performance.
- Certain features, such as `Self_Employed`, did not significantly influence the outcome, potentially due to low variance in the data.

6. Conclusion

This project successfully developed a model to predict loan eligibility with a high degree of accuracy. By leveraging machine learning techniques, the model can assist financial institutions in making informed lending decisions. Future work could involve refining the model with more data and exploring other advanced machine learning algorithms.