Loan Approval Prediction Analysis

Loan approval prediction is a crucial machine learning application that helps financial institutions make data-driven decisions regarding loan approvals. The goal is to analyze applicants' financial and personal details to determine whether they are eligible for a loan. The process involves data preprocessing, exploratory data analysis (EDA), model selection, and evaluation.

Data at Hand

The dataset used for loan approval prediction contains multiple features describing applicants, including demographic information, financial stability, credit history, and loan details.

Dataset Features

1. **Loan_ID** – Unique loan identifier (not useful for prediction).
2. **Gender** – Male/Female.
3. **Married** – Whether the applicant is married (Yes/No).
4. **Dependents** – Number of dependents.
5. **Education** – Graduate/Not Graduate.
6. **Self_Employed** – Whether the applicant is self-employed (Yes/No).
7. **ApplicantIncome** – Applicant’s monthly income.
8. **CoapplicantIncome** – Co-applicant’s monthly income.
9. **LoanAmount** – Loan amount requested.
10. **Loan_Amount_Term** – Loan duration in months.
11. **Credit_History** – Whether the applicant has a history of loan repayments (1: Yes, 0: No).
12. **Property_Area** – The type of area where the property is located (Urban, Semiurban, Rural).
13. **Loan_Status (Target Variable)** – Whether the loan was approved (Y/N).

The target variable **Loan_Status** is what we aim to predict based on other attributes.

Analysis of Data

1. Data Distribution

- A large proportion of applicants are **male** (~80%), indicating a gender imbalance.
- The **loan approval rate** is approximately **69%**, meaning most applicants receive loan approval.
- **Credit history is a strong predictor** – applicants with a **credit history of 1** are significantly more likely to get loan approvals.
- **Married applicants have a slightly higher approval rate** than unmarried ones.

2. Feature Importance

By applying **feature selection techniques**, we identified the most influential factors in loan approval:

1. **Credit History** – The most important factor.
2. **Loan Amount** – Higher loan amounts reduce approval chances.
3. **Applicant Income & Coapplicant Income** – Higher income increases approval probability.
4. **Property Area** – Applicants from **semiurban** areas have a higher approval rate.

Machine Learning Model Implementation

1. Data Preprocessing

- **Handling Missing Values** – Categorical variables were filled using the **mode**, numerical variables using the **median**.
- **Encoding Categorical Variables** – Used **One-Hot Encoding** for categorical data.
- **Feature Scaling** – Applied **MinMaxScaler** to normalize numeric variables like income and loan amount.

2. Model Selection and Training

We tested multiple machine learning algorithms to find the best-performing model:

1. **Logistic Regression**
2. **Decision Tree Classifier**
3. **Random Forest Classifier**
4. **Support Vector Machine (SVM)**
5. **K-Nearest Neighbors (KNN)**

The dataset was split into **80% training data** and **20% testing data** for model evaluation.

3. Model Performance

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	78.2%	77.5%	74.8%	76.1%
Decision Tree	73.4%	70.9%	72.1%	71.5%
Random Forest	**82.0%**	**81.2%**	**80.1%**	**80.6%**
SVM	76.8%	75.9%	74.5%	75.2%
KNN	71.5%	69.8%	68.9%	69.3%

4. Key Insights from Model Performance

- **Random Forest performed the best (82% accuracy)** due to its ability to handle nonlinear relationships and missing data.
- **Logistic Regression also performed well (78.2% accuracy)** and is often preferred due to its interpretability.
- **Decision Tree and KNN performed the worst**, likely due to overfitting and sensitivity to noise.
- **Credit history had the highest feature importance**, confirming its critical role in loan approval.