Loan Approval Prediction Using Machine Learning

# Objective

The goal of this project is to build a robust machine learning model to predict whether a loan application should be approved or rejected based on applicant and financial attributes. Accurate predictions help financial institutions reduce risk and improve decision-making efficiency.

# Dataset Overview

- Source: Kaggle – Pre-Approved Loan Prediction Dataset  
- Records: 5,000+  
- Features: Demographic details (Age, Education, Marital Status), financial indicators (Income, Credit Score, Loan Amount), and categorical variables such as Property Area, Self-Employed status, and Tax Return history.  
- Target: Loan\_Status (Approved/Rejected)

# Methodology

**1. Exploratory Data Analysis (EDA):**- Checked missing values (only 1 missing in Salaried column).  
- Visualized feature distributions, detected outliers, and applied capping for numerical features (Income, Loan\_Amount, Monthly\_CC\_Usage).  
- Correlation analysis showed strong relation between Income, Loan\_Amount, and loan approval.  
  
**2. Feature Engineering:**- Encoded categorical features using One-Hot Encoding.  
- Removed irrelevant columns like ZIP Code and ID.  
- Applied Recursive Feature Elimination (RFE) for Logistic Regression, selecting top 10 features.  
  
**3. Model Building:**- Split data into Train (80%) and Test (20%) sets.  
- Trained four models:  
 \* Logistic Regression (with RFE-selected features)  
 \* Random Forest (Tuned) using GridSearchCV  
 \* AdaBoost with GridSearchCV  
 \* XGBoost with GridSearchCV  
  
**4. Evaluation Metrics:**- Accuracy, Precision, Recall, F1-score, and ROC-AUC  
- Plotted Confusion Matrix and ROC curves for each model.  
- Used SHAP analysis for XGBoost to explain feature importance.

# Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-score | ROC-AUC |
| Logistic Regression | 93.4% | 0.89 | 1.00 | 0.94 | 0.952 |
| Random Forest (Tuned) | 98.8% | 0.978 | 1.00 | 0.989 | 0.999 |
| AdaBoost | 97.9% | 0.964 | 0.998 | 0.981 | 0.998 |
| XGBoost | 99.8% | 0.996 | 1.00 | 0.998 | 0.999 |

Best Model: XGBoost, achieving near-perfect accuracy and AUC.

Key Features (SHAP Analysis):  
- Credit\_Score\_Poor had the highest negative impact on approval.  
- Property location (Property\_Area) and income significantly influenced decisions.

# Conclusion & Future Scope

XGBoost is the most effective model for this task, delivering exceptional predictive performance and interpretability. SHAP analysis enhanced transparency, making the model suitable for real-world credit risk assessment.  
  
Future improvements:  
- Implement SMOTE/ADASYN for severe imbalance scenarios.  
- Apply Bayesian Optimization for hyperparameter tuning.  
- Deploy the model as a Flask or FastAPI service for integration into financial systems.