Loan Approval Status Classification Models

ML Assignment 2 - Submission

Submitted: February 14, 2026

1. GitHub Repository

https://github.com/powarsg/ml-classification-models

Complete source code with app.py, requirements.txt, model training scripts, and comprehensive README

2. Live Streamlit App

https://powarsg-ml-classification-models.streamlit.app

Interactive frontend deployed on Streamlit Community Cloud with model selection, CSV upload, and real-time predictions

3. BITS Virtual Lab Execution

Screenshot: Place bits_lab_screenshot.png in project folder

4. Project Documentation

a. Problem Statement

Predict loan approval status based on applicant demographics and financial information. Binary classification: 0 = Not Approved, 1 = Approved

b. Dataset Description

Dataset: 45,000 records (44,995 after outlier removal) | **Features:** 13 input features + target | **Split:** 80% train, 10% val, 10% test

Class Distribution: 77.8% Not Approved (35,000) | 22.2% Approved (10,000) - Imbalanced

Preprocessing: StandardScaler (numeric), LabelEncoder (categorical), Stratified split

c. Models Used

6 Classification Models: Logistic Regression, Decision Tree, KNN, Naive Bayes, Random Forest, XGBoost

Model	Accuracy	AUC	F1	MCC
Logistic Regression	0.8927	0.9507	0.7564	0.6876
Decision Tree	0.9204	0.9605	0.8042	0.7601
KNN	0.8884	0.9292	0.7402	0.6699
Naive Bayes	0.7282	0.9403	0.6205	0.5410
Random Forest	0.9211	0.9742	0.8084	0.7629
XGBoost	0.9311	0.9788	0.8362	0.7948

d. Model Performance Observations

Logistic Regression: Moderate baseline (89.27% accuracy, 0.9507 AUC). Interpretable but limited to linear decision boundaries.

Decision Tree: Good performance (92.04% accuracy, 0.9605 AUC). Natural feature interaction handling with interpretable rules.

KNN: Solid performance (88.84% accuracy, 0.9292 AUC). Requires feature scaling; computationally expensive at scale.

Naive Bayes: Poor practical performance (72.82% accuracy). Critical flaw: 100% recall but 0.45% precision - unacceptable for production.

Random Forest: Excellent (92.11% accuracy, 0.9742 AUC). Robust ensemble with feature importance insights. Strong candidate.

XGBoost ■: BEST MODEL (93.11% accuracy, 0.9788 AUC, 0.8362 F1). Sequential boosting, native imbalance handling. Recommended for production.

Key Insights on Imbalanced Data:

• Accuracy misleading (77.8% vs 22.2% class ratio) • AUC, F1, MCC more informative • Ensemble methods handle imbalance naturally • Precision-recall trade-off critical for loan decisions