Project Report: Random Forest Classification on Loan Approval Dataset

Author: Usman Turajo | **Date**: August 15, 2025 | **Toolset**: Python, scikit-learn, pandas, matplotlib

1. Abstract

This project explores the application of a Random Forest Classifier to a supervised learning problem using a structured dataset. The goal was to build a robust model capable of generalizing well to unseen data. After preprocessing and tuning, the model achieved a training accuracy of 100% and a cross-validation accuracy of approximately 95.7%. While the training accuracy suggests potential overfitting, cross-validation results indicate strong generalization. Feature importance analysis and fold performance variability were also examined to assess model stability.

2. Introduction

Machine learning models are increasingly used to solve classification problems across domains. This project focuses on building a Random Forest Classifier to predict a target variable from structured input features. The objective was to evaluate model performance, identify signs of overfitting, and ensure generalization through cross-validation.

3. Dataset Description

- **Source**: Kaggle.com
- **Samples**: 4269
- **Features**: `education, self_employed, loan_status, loan_id, no_of_dependents, income_annum, loan_amount, loan_term, cibil_score, residential_assets_value, commercial_assets_value, luxury_assets_value, bank_asset_value'`
- **Target Variable**: Loan_status
- **Preprocessing**:
    - Missing values handled
    - Categorical features encoded

4. Methodology

- **Model**: Random Forest Classifier
- **Hyperparameters**:
    - max_depth=10
    - n_estimators=100
    - min_samples_split=2
- **Validation Strategy**:
    - 5-Fold Cross-Validation

- **Libraries Used**:
  - scikit-learn
  - pandas
  - matplotlib
  - seaborn

5. Results

Accuracy Scores

Metric	Value
Training Accuracy	100%
CV Accuracy (mean)	95.7%
CV Std Deviation	3.8%

Fold Scores

Fold	Accuracy
1	96.6%
2	97.5%
3	98.2%
4	98.0%
5	88.2%

6. Discussion

The model achieved perfect accuracy on the training set, which is a strong indicator of overfitting. However, the cross-validation score of 95.7% suggests that the model generalizes well to unseen data. Fold 5 showed a noticeable drop in accuracy (88.2%), which may be due to class imbalance or outliers. Feature importance analysis helped identify the most influential predictors, guiding future feature selection.

7. Conclusion

The Random Forest model performed well overall, with high accuracy and reasonable stability across folds. While overfitting was observed in training, cross-validation helped validate the model's generalization ability. Future work could include:

- Testing on a separate holdout set
- Trying other models (e.g., XGBoost, SVM)
- Performing hyperparameter optimization
- Investigating Fold 5 anomalies

8. References

- Scikit-learn documentation: https://scikit-learn.org/
- Random Forest algorithm overview: https://en.wikipedia.org/wiki/Random_forest