Final Report

Title: Credit Risk & Loan Default Prediction System

Financial institutions face major losses due to loan defaults. This project develops an **ML-driven credit risk assessment system** that predicts whether a loan applicant is likely to default or repay. Using historical financial and demographic data, multiple machine learning models were trained and evaluated with emphasis on **risk-sensitive metrics such as recall and precision**.

The final solution not only predicts default risk but also provides **clear explanations of why an applicant is risky**, enabling banks and fintech companies to make **transparent, compliant, and data-driven lending decisions**.

2. BUSINESS OBJECTIVE

- Predict loan default probability
- Reduce financial risk exposure
- Improve loan approval accuracy
- Support automated & manual decision-making
- Enhance regulatory explainability

3. DATASET OVERVIEW

Source: Public Loan / Credit Risk Dataset (Kaggle-style)

Features Used:

- Applicant Income
- Co-applicant Income
- Loan Amount
- Loan Term
- Credit History
- Employment Type
- Education
- Marital Status
- Property Area
- Existing EMIs

Target Variable:
Loan_Status → Default (1) / Non-Default (0)

4. DATA CLEANING & PREPROCESSING

- Missing values imputed using statistical methods
- Outliers treated for income & loan amount
- Categorical features encoded
- Numerical features scaled
- Final dataset prepared for modeling

5. EXPLORATORY DATA ANALYSIS (EDA)

Key Insights:

- Strong credit history drastically reduces default risk
- Higher loan amount with lower income increases default probability
- Salaried individuals show lower default rates
- Urban property areas have relatively lower risk

Visualizations:

- Distribution plots
- Box plots
- Bar charts
- Correlation heatmap

6. FEATURE ENGINEERING

Engineered features for stronger risk detection:

- Debt-to-Income Ratio
- EMI-to-Income Ratio
- Income buckets
- Loan burden indicators

These significantly improved model performance.

7. MODEL BUILDING

Models trained:

- Logistic Regression
- Decision Tree
- Random Forest

Evaluation Metrics:

- Accuracy
- Precision
- Recall
- ROC-AUC
- Confusion Matrix

8. MODEL COMPARISON TABLE

Model	**Accuracy**	**Precision**	**Recall**	**ROC-AUC**
Logistic Regression	High	High	Medium	High
Decision Tree	Medium	Medium	Medium	Medium
Random Forest	**Highest**	**Highest**	**Highest**	**Highest**

Selected Model: Random Forest

Reason: Best performance on risk-sensitive metrics and strong generalization.

9. MODEL EXPLAINABILITY

High-Risk Indicators:

- Poor credit history
- High EMI-to-income ratio
- Low applicant income
- High loan amount

Low-Risk Indicators:

- Stable income
- Strong credit history
- Low debt burden

Feature importance allows **transparent and regulator-friendly decisions**.