

## **1. Project Title**

Credit Default Risk Prediction Using Machine Learning

## **2. Problem Statement**

Loan defaults pose a major financial risk for banks and lending institutions. Approving loans without accurately assessing a customer's repayment ability can lead to increased non-performing assets and financial losses. Traditional credit evaluation methods rely on manual rules and limited indicators, which often fail to capture complex patterns in customer financial behavior.

This project aims to build a machine learning-based system that predicts whether a customer is likely to default on a loan using demographic, financial, and credit-related features.

## **3. Objective**

The main objective of this project is to predict whether a customer is likely to default on a loan.

This system helps financial institutions to:

Reduce credit risk

Identify high-risk customers early

Make accurate and data-driven loan approval decisions

## **4. Dataset Description**

The dataset contains customer demographic and financial information such as:

Income

Savings

Monthly expenses

Loan amount

Credit score

Employment details

Loan history

The target variable is Loan Default Status, where:

0 = No Default

1 = Default

## **5. Methodology**

The project follows a complete end-to-end machine learning workflow:

Data loading and inspection

Exploratory Data Analysis (EDA)

Handling missing values using imputation

Feature scaling and encoding

Train-test split

Model training

Model evaluation and comparison

## **6. Models Implemented**

The following classification models were implemented and compared:

Logistic Regression

Decision Tree

Random Forest

Support Vector Machine (SVM)

XGBoost

## **7. Evaluation**

The models were evaluated using:

Accuracy

Precision

Recall

F1-Score

Confusion Matrix

## **Model Performance Summary**

Logistic Regression → ~93%

Decision Tree → ~94%

Random Forest → ~93%

SVM → ~89–90%

XGBoost → ~95–96% (Best Model)

Among all models, XGBoost achieved the highest accuracy with balanced precision and recall, making it the most reliable model for this dataset.

## 8. Conclusion

In this project, an end-to-end Credit Default Risk Prediction system was successfully developed using machine learning techniques. The complete data science workflow—from data preprocessing and EDA to model training and evaluation—was implemented.

Among all models, XGBoost delivered the best performance with an accuracy of approximately 95–96%, indicating strong predictive capability. This demonstrates that machine learning can effectively support financial institutions in predicting credit default risk and improving lending decisions.

## 9. Business Impact

This model can help financial institutions to:

Identify high-risk customers in advance

Reduce loan default rates

Improve credit approval decisions

Minimize financial losses

Ensure fair and consistent lending

## 10. Future Improvements

Although the current model achieved strong performance, it can be further improved in the future.

Hyperparameter tuning using GridSearchCV was explored; however, due to computational resource limitations, the execution was time-consuming and complete output could not be generated. With access to higher computational resources, extensive hyperparameter tuning can be performed to further improve model performance.

Additional future enhancements include:

Handling class imbalance using **SMOTE**

Adding **model explainability** using SHAP values

Evaluating performance using **ROC-AUC and Precision-Recall curves**

Deploying the model as a **web application** using Flask or FastAPI (Hyperparameter tuning was attempted, but full results could not be included due to computational limitations.)

## 11. Final Note

This project demonstrates the ability to:

Handle real-world structured data

Apply multiple machine learning algorithms

Evaluate and compare model performance

Build a complete prediction-ready machine learning pipeline