Credit Risk Prediction Project Report

# 1. Problem Statement

Financial institutions face a critical challenge in assessing the creditworthiness of loan applicants. An accurate credit risk prediction system helps minimize defaults and maintain financial stability. This project uses the German Credit dataset to develop a machine learning model that predicts whether an applicant is a good or bad credit risk. The model is designed to support credit officers by providing reliable predictions and transparency into the decision-making process.

# 2. Dataset Overview

The dataset contains information for 1000 loan applicants, with a mix of numerical and categorical features:  
  
- Numerical Features: Age, Duration, Credit amount  
- Categorical Features: Sex, Job, Housing, Purpose, Saving accounts, Checking account  
  
Target variable: CreditRisk (0 = Good credit risk, 1 = Bad credit risk)

# 3. Data Preprocessing

The preprocessing pipeline included:  
- Imputing missing values (mean for numeric, most frequent for categorical)  
- One-hot encoding of categorical features  
- Scaling numerical features using ColumnTransformer  
- Integration of all steps into a unified Pipeline to ensure repeatability and avoid data leakage

# 4. Feature Engineering

A domain-informed rule-based target was created using:  
- High credit amount and long duration  
- Low savings or checking balances  
- Low job category (0 or 1)  
  
These rules align with common credit risk evaluation policies used by financial institutions.

# 5. Modeling Approach

Model used: RandomForestClassifier  
  
- Robust to overfitting  
- Handles mixed data types well  
- Provides feature importance metrics  
  
The entire workflow was encapsulated in a Pipeline for simplicity and reproducibility.  
Hyperparameters were tuned using GridSearchCV with 5-fold cross-validation.  
  
Best Parameters: {'classifier\_\_max\_depth': None, 'classifier\_\_min\_samples\_split': 2, 'classifier\_\_n\_estimators': 200}

# 6. Evaluation Results

Classification Report on Test Set:  
  
Class Precision Recall F1-Score  
0 (Good) 0.93 0.97 0.95  
1 (Bad) 0.98 0.96 0.97  
  
Accuracy: 0.96 overall  
  
These results show that the model is highly accurate and performs well across both classes.

# 7. Top Influential Features

Based on feature importance from the Random Forest model, the top features influencing credit risk are:  
- Duration  
- Credit amount  
- Saving accounts\_little  
- Job  
- Checking account\_no checking  
  
These align with financial intuition: higher loan amounts and durations indicate risk, while low savings and job stability reduce creditworthiness.

# 8. Streamlit User Interface

A Streamlit web application was developed to:  
- Collect applicant details via input fields  
- Predict credit risk in real time  
- Display predictions with confidence levels  
  
The app makes the model accessible to non-technical users and suitable for demonstration.

# 9. Conclusion & Recommendations

The model accurately predicts credit risk with 96% accuracy and identifies important financial indicators.  
  
Recommendations:  
- Use the model in a decision-support role  
- Integrate SHAP for explainability  
- Include behavioral data for enhanced prediction  
- Retrain periodically to maintain accuracy over time