**Loan Risk Modeling**

**Assessing Loan Default Risk with Machine Learning**

**1. Introduction**

Credit risk modeling is a critical task for financial institutions to evaluate the likelihood of loan defaults and make informed lending decisions. This project aims to build a predictive model using **decision trees** and **random forests** to classify loan applicants into "low-risk" or "high-risk" categories based on the **German Credit Dataset**. The results will help lenders optimize approval processes, reduce defaults, and improve profitability.

**2. Objectives**

1. **Primary Objective**: Develop a machine learning model to predict the probability of loan default using the German Credit Dataset.
2. **Secondary Objectives**:
   * Perform exploratory data analysis (EDA) to understand key features influencing credit risk.
   * Compare the performance of decision trees vs. random forests.
   * Identify the most significant predictors of loan default.
   * Provide actionable insights for risk management strategies.

**3. Dataset Overview**

**Dataset**: **German Credit Data**

* **Source**: UCI Machine Learning Repository or Kaggle.
* **Size**: 1,000 entries with 20 features (7 numerical, 13 categorical).
* **Target Variable**: Binary classification (Good Credit Risk = 1, Bad Credit Risk = 2).
* **Key Features**:
  + Credit\_history: Previous credit behavior (e.g., "existing credits paid back duly").
  + loan\_amount: Disbursed credit amount.
  + employment\_duration: Current employment status.
  + purpose: Loan purpose (e.g., car, education, furniture).
  + age, sex, savings\_account: Demographic and financial attributes.

**4. Methodology**

**4.1 Data Preprocessing**

* **Handling Missing Values**: Impute or remove missing data.
* **Categorical Encoding**: Convert categorical variables (e.g., purpose, credit\_history) using one-hot or label encoding.
* **Feature Scaling**: Normalize numerical features (e.g., loan\_amount, age).
* **Class Balancing**: Address class imbalance (70% "good" vs. 30% "bad" credit risks) using techniques like SMOTE or class weighting.

# 4.2 Exploratory Data Analysis (EDA)

As part of this project, we will conduct a comprehensive exploratory data analysis using Python visualizations and graph-based representations. Code will be implemented to dynamically build visual graphs (e.g., bar plots, heatmaps, histograms, box plots), allowing deeper exploration of: -

Distributions of numerical and categorical features.  
Correlations between variables and the target label.  
Anomalies or biases in demographics (e.g., age, gender).  
Network-style graphs to visualize interconnected attributes if applicable.  
  
**This section is applying graph theory concepts where relevant, such as using graphs to model feature interdependencies or decision pathways.**

**4.3 Model Development**

1. **Decision Trees**:
   * Build a baseline model using entropy/Gini impurity for splits.
   * Prune the tree to avoid overfitting (e.g., max\_depth, min\_samples\_split).
2. **Random Forests**:
   * Train an ensemble of decision trees with bootstrapping.
   * Optimize hyperparameters (e.g., n\_estimators, max\_features) via grid search.

**4.4 Model Evaluation**

* **Metrics**: Accuracy, precision, recall, F1-score, ROC-AUC.
* **Cross-Validation**: Use k-fold cross-validation (k=5/10) to ensure robustness.
* **Confusion Matrix**: Visualize false positives/negatives.

**4.5 Feature Importance Analysis**

* Rank features by their contribution to predictions (e.g., Gini importance in random forests).
* Interpret results to guide risk assessment policies.

**5. Implementation Plan**

|  |  |
| --- | --- |
| **Phase** | **Tasks** |
| **Week 1-2** | Data preprocessing, EDA, and feature engineering. |
| **Week 3** | Train and tune decision tree and random forest models. |
| **Week 4** | Evaluate models, compare performance, and analyze feature importance. |
| **Week 5** | Finalize report and present insights. |

**6. Expected Outcomes**

1. A comparative analysis of decision trees vs. random forests for credit risk prediction.
2. Identification of top features influencing loan defaults (e.g., credit\_history, loan\_amount).
3. A deployable model with actionable recommendations for lenders (e.g., stricter criteria for high-risk groups).

**7. Tools and Technologies**

* **Programming Language**: Python
* **Libraries**: Scikit-learn, Pandas, NumPy, Matplotlib/Seaborn, Imbalanced-Learn.
* **Environment**: Jupyter Notebook, Google Colab.

**8. Ethical Considerations**

* Ensure the model does not discriminate based on sensitive attributes like sex or age.
* Highlight potential biases in historical lending data during EDA.

**9. Deliverables**

1. **Technical Report**: Code, visualizations, and model performance summaries.
2. **Presentation**: Stakeholder-friendly summary of findings and recommendations.
3. **Deployment Guidelines**: Documentation for integrating the model into a loan approval system.

**10. References**

1. German Credit Dataset: UCI Repository Link
2. Bierman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.
3. Scikit-learn Documentation: [Decision Trees](https://scikit-learn.org/stable/modules/tree.html), [Random Forests](https://scikit-learn.org/stable/modules/ensemble.html#forest).

**Prepared by**: Muhammad ALBER, Zain Bin Khalid  
**Date**: 1/31/2025  
**Contact**: su92-bscsm-f23-353@superior.edu.pk