CS471 Final Project Proposal

Title: Loan Prediction Using Machine Learning

Team Members:

- Jamison Stalter
- Julian Rangel

1. Motivation

The goal of this project is to build a reliable loan prediction model, an essential tool for financial institutions in assessing loan eligibility and minimizing risk. By accurately predicting loan eligibility, banks and financial institutions can streamline the decision-making process, minimize the risk of defaults, and increase the overall efficiency of lending operations.

A similar study on Kaggle used this dataset and employed models like Logistic Regression and Decision Trees. Our approach will build on this prior work by comparing a rule-based system to multiple machine learning models, providing insights into which methods offer the highest accuracy and reliability for loan prediction.

2. Dataset

We will use the **Loan Prediction Dataset** available on Kaggle. This dataset contains key information about applicants, such as:

- **Applicant and Co-applicant Income**
- **Loan Amount and Term**
- **Credit History**
- **Gender, Marital Status, Dependents, Education, Self-Employment Status**

Given the variety of features, we anticipate needing preprocessing steps such as:

- **Handling Missing Values**: Using imputation strategies for continuous and categorical data.
- **Encoding Categorical Variables**: Converting categorical attributes (e.g., Gender, Education) into numerical values.
- **Scaling Numerical Features**: Standardizing or normalizing features like income and loan amount for better performance with machine learning models.

3. Method

To classify loan eligibility, we will employ three approaches:

- **Algorithm 1: Rule-based Approach** – This initial approach will use hand-crafted rules based on domain knowledge. For instance, applicants with high income and a positive credit history might be approved, while those with lower income and no credit history may be declined. This will provide a baseline comparison for the machine learning algorithms.
- **Algorithm 2: Decision Tree or Logistic Regression** – A simple yet effective machine learning model to classify eligibility based on loan application features. Logistic Regression, as a linear model, is often interpretable, while Decision Trees offer a structured, rule-based model.
- **Algorithm 3: Random Forest or Support Vector Machine (SVM)** – A more robust classifier that can capture complex relationships within the data. Random Forest, an ensemble method, will provide insights into feature importance, while SVM is useful for non-linear decision boundaries.

4. Intended Experiments

We plan to conduct the following experiments:

- **Model Training and Evaluation**: Train each of the three algorithms using cross-validation to ensure robustness.
- **Hyperparameter Tuning**: Experiment with different hyperparameters for each model, such as tree depth for Decision Trees or kernel selection for SVM, to optimize model performance.
- **Feature Engineering**: Investigate whether derived features (e.g., income-to-loan ratio) improve model performance.

5. Model Evaluation

Model performance will be evaluated using metrics like:

- **Accuracy**: Overall correctness of predictions.
- **Precision and Recall**: Specific focus on correctly identifying eligible vs. ineligible applicants.
- **F1-score**: To balance precision and recall, especially important if there is class imbalance in the dataset.

By comparing models using these metrics, we aim to determine the most effective approach for predicting loan eligibility.