



T.C.

MARMARA UNIVERSITY

FACULTY OF ENGINEERING

INTRODUCTION TO MACHINE LEARNING

CSE 4288

Term Project

Student

150121824 – Burak KARAYAĞLI

150121539 – Gülsüm Ece GÜNAY

150120027 – Mert MUSLU

150120051 – Erkut DÖNMEZ

150120026 – Ardacan ÖZENER

Lecturer

Assoc. Prof. Murat Can GANİZ

1.Problem Statement and Objectives

Problem Statement:

In the financial domain, assessing the eligibility of loan applicants is a critical task for banks and financial institutions. Misjudging an applicant's loan approval probability can lead to financial losses and inefficiencies. This dataset provides a comprehensive set of features, including financial and demographic attributes of applicants, as well as loan-specific details. By analyzing these features, the aim is to build a machine learning-based classification model to predict whether a loan application should be approved or rejected.

The problem involves identifying patterns and relationships in the data that influence loan approval decisions, which can support automated decision-making processes, improve efficiency, and reduce biases in the approval process.

Objectives:

1. Preprocess the Dataset:

Prepare the dataset for modeling by handling missing values, encoding categorical variables, and normalizing or scaling numerical features as needed.

2. Develop a Classification Model:

Create and evaluate a machine learning classification model capable of predicting loan approval outcomes (approved or rejected) based on the provided features.

3. Test the Model:

Evaluate the model's performance using a separate test set to assess its accuracy and reliability.

4. Analyze Results:













Interpret the model's predictions and analyze the results to gain insights into its performance and decision-making patterns.



By addressing these objectives, the project aims to enhance the efficiency and accuracy of loan approval processes while maintaining fairness and transparency.

2.Dataset Description and Source

Dataset Description:

This dataset provides a detailed view of 45,000 loan applications, containing 14 key features that encompass both applicant-specific and loan-specific attributes. These features are a mix of categorical and continuous data types, designed to support the development of a classification model for predicting loan approval status. Below is a breakdown of the dataset's features:

Column	Description	Data Type
 person_age	Applicant's age	Float
 person_gender	Applicant's gender	Categorical
 person_education	Applicant's highest level of education	Categorical
 person_income	Applicant's annual income in currency	Float
 person_emp_exp	Years of employment experience	Integer
 person_home_ownership	Home ownership status (e.g., rent, own, mortgage)	Categorical
 loan_amnt	Amount of loan requested	Float
 loan_intent	Intended purpose of loan (e.g. , personal, education)	
 loan_int_rate	Interest rate applicable to loan	Float
 loan_percent_income	Loan amount as a percentage of annual income	Float
 cb_person_cred_hist_length	Number of years of credit history	Float
 credit_score	Applicant's credit score	Integer

 previous_loan_defaults_on_file	Indicator of previous loan defaults (Yes/No)	Categorical
 loan_status	Loan status outcome (1 = approved, 0 = rejected)	Integer

The target variable, **loan_status**, indicates whether the loan application was approved (1) or rejected (0). This classification task aims to analyze the factors contributing to loan approval or rejection while developing a predictive model.

Dataset Source:

This dataset was obtained from Kaggle, a popular platform for data science and machine learning projects. While the dataset was not pre-split into training and testing sets, it will be divided into training and testing subsets as part of the preprocessing phase, using a standard split ratio such as 70/30 or 80/20 to ensure reliable model evaluation.

The dataset serves as a robust resource for building machine learning models and understanding the underlying patterns in loan approval processes.

3.Planned Methodology and Tools

Binary classification is a supervised learning task where the goal is to predict one of two classes for a given input vector.

3.1. Data Preparation

Methodologies:

- **Data Cleaning:** We need to handle missing values, outliers, and inconsistencies if they exist in the data.
- **Feature scaling:** Gradient descent-based methods like logistic regression are sensitive to the scale of features. Normalization or standardization must be done for numeric features. For algorithms like decision tree that use categorical features,

binning will be used for numeric values and feature scaling will not be used for those algorithms.

- **Encoding categorical variables:** One-hot encoding will be used for **loan_intent** feature and label encoding will be used for **person_education** feature for gradient descent-based methods.
- **Data Splitting:** Split the data into training, validation, and test sets as 70%-20%-10%.

Tools:

We will use Pandas and NumPy for data manipulation.

3.2. Model Selection

Methodologies:

We will choose several methods and create at least one model for each. Following methods will be used:

- Logistic Regression.
- Decision Tree
- Naive Bayes
- Random Forest
- Gradient Boosting

Tools:

Scikit-learn will be used for these methods and building the models.

3.3. Model Training

Methodologies:

We will train the model on the training dataset and fine-tune hyperparameters. Then we'll handle class imbalance with certain techniques.

Tools:

Scikit-learn will be used for basic training pipelines and parameter tuning.

3.4. Model Evaluation

Methodologies:

We evaluate the models and compare their metrics for binary classification. The dataset's class labels are slightly imbalanced. Instead of accuracy, F1 score will be more important in evaluation. Confusion matrix will be constructed also. Following metrics will be considered:

- Accuracy
- Precision
- Recall
- F1-Score

Tools:

- Scikit-learn: For computing metrics and visualizing confusion matrices.
- Matplotlib: For visualizing results.

4.Team Roles and Responsibilities

The roles and responsibilities within the team have not yet been assigned to individual members. However, the general outline of the plan is to use multiple machine learning algorithms on the same dataset and train several models. The division of tasks will be based on the number of models created, with the workload distributed among team members. While some members will focus on training the models, others will concentrate on preparing the dataset for training, also testing will be done by different team members who have not participated in that model's training phase. Tasks such as writing the report will be shared equally among all team members. A more detailed breakdown of responsibilities will be provided in the project report.

5.Timeline and Milestones

Week 7: Project Initiation:

- Forming teams
- Selecting dataset and project topic
- Submitting proposal

Week 9: Data Preprocessing and EDA

- Understanding dataset
- Performing data cleaning and data preprocessing

- Selecting key features through detailed examination of dataset

Week 10: Model Development

- Selecting appropriate machine learning algorithms
- Training models using dataset

Week 11: Model Evaluation and Optimization

- Evaluating model's performance with the help of test dataset
- Experimenting dataset with different algorithms

Week 12: Finalization and Presentation Preparation

- Finalizing the model and results.
- Preparing a report covering all work completed.
- Creating presentation slides and rehearsing the presentation as a team.

Milestones:

Milestone No.	Milestone	Description	Completion Date
1	Project Proposal Submission	Submit project proposal including topic and methodology.	End of Week 7
2	Data Preprocessing and EDA Completion	Complete data cleaning, EDA, and feature engineering.	End of Week 9
3	Initial Model Development	Develop and train initial model, submit progress report.	End of Week 10
4	Model Evaluation and Optimization	Evaluate models, optimize, and submit evaluation report.	End of Week 11
5	Final Report and Presentation Preparation	Prepare final report, slides, and rehearse presentation.	End of Week 12