Course Project Description: Predicting Bank Loan Approval

Overview

In this data analytics project, students will engage in a comprehensive analysis of a bank loan dataset. The primary goal is to develop predictive models that can accurately determine whether a loan will be approved based on various applicant features. This project is designed to provide hands-on experience with real-world data preparation, exploration, and predictive modeling techniques in the context of financial decision-making.

The dataset contains the following fields:

Loan_ID: A unique identifier for the loan (numeric or alphanumeric).

Applicant_ID: A unique identifier for the applicant (numeric or alphanumeric).

Gender: Male, Female, Other.

Married: Yes, No.

Dependents: Number of dependents (0, 1, 2, 3).

Education: Graduate, Not Graduate.

Self_Employed: Yes, No.

ApplicantIncome: Income of the applicant (numeric).

CoapplicantIncome: Income of the co-applicant (numeric).

Has_CreditCardd: Yes (applicant has an active credit card), No (applicant has no active credit cards.

LoanAmount: The loan amount in thousands (numeric).

Loan_Amount_Term: Term of loan in months (numeric).

Owns_Car: Yes, No.

Owns_House: 1 (owns house), 0 (does not own house).

Property_Area: Urban, Semiurban, Rural.

Loan_Status: Y (Loan approved), N (Loan not approved).

Objectives:

A) Preprocessing

1- Delete unnecessary columns (identifiers, low variance, too many missing values, duplicated).
2- Clean the ApplicantIncome and CoapplicantIncome from outliers.
3- Impute (using kNN or median, or mode any missing data for columns that are at least 50% complete (aka columns that do not have 40% or more missing values). If a column has 50% or more missing data, please delete it.
4- Write a section that answers the following questions:
    a. How many empty values did each column contain?
    b. How many outliers did ApplicantIncome contain?
    c. How many outliers did CoapplicantIncome contain?
    d. Which columns did you delete, and why?

Deliverable: Please dump (write to a new file) the dataset **after you finished cleaning** it and submit it to the dropbox. Also, include the answers for the questions above in the report

B) Visualization

- Figure 1: Provide a visualization that describes the Married columns for all accepted loans
- Figure 2: Provide a visualization that describes the Married columns for all rejected loans
- Figure 3: Provide a line chart (or scatter plot) for the ApplicantIncome and CoapplicantIncome
- Figure 4: Provide a Box plot for each of the following columns:
    - ApplicantIncome
    - CoapplicantIncome
    - LoanAmount

Deliverable: provide the visualization plots in the project report (in MS Word or Google Docs).

C) Descriptive Analytics

Calculate and include in the report the 5-number-summary for the following:

- ApplicantIncome
- CoapplicantIncome
- LoanAmount

D) Predictive Analytics

Build two machine learning models to predict the Loan_Status column. Then, please write a summary of your training testing split, accuracy, which model is better and why and include it in your project report.

Deliverables:

Please submit the **code** you used for all tasks for this project

Please submit a separate **report** (MS Word, Google doc) that includes:

- A summary of the data cleaning activities you performed on the data as described in the preprocessing section.
- The clean dataset in CSV format after finishing the preprocessing section.
- The required visualization plots.
- Five-number summary for the columns described in the descriptive analytics section (a screenshot or a table that states the column and provide the 5-number-summary for each column.
- A section that describes the used machine learning models as described the predictive analytics section.