Team 21: Risk Assessment and Loan Approval Prediction

1 Introduction

The `21.csv` **Synthetic Dataset for Risk Assessment and Loan Approval Modeling** is designed to facilitate predictive modeling for **financial risk assessment and loan approval decisions**. This dataset contains **20,000 records** covering personal, financial, and credit-related attributes, making it suitable for both **regression and classification tasks**. Financial institutions and lenders can use this dataset to develop machine learning models that help optimize credit risk assessment and improve the accuracy of loan approval predictions.

2 Dataset Description

The dataset consists of multiple features that provide insights into an applicant's financial stability and creditworthiness:

2.1 Demographic and Employment Information

- **Age** – Applicant's age.
- **Marital Status** – Single, Married, Divorced, etc.
- **Number of Dependents** – Number of individuals financially dependent on the applicant.
- **Education Level** – Highest level of education attained.
- **Employment Status** – Job situation (e.g., employed, self-employed, unemployed).
- **Experience** – Work experience in years.
- **Job Tenure** – Duration of the applicant's current job.

2.2 Financial and Credit Information

- **Annual Income** – Yearly earnings of the applicant.
- **Monthly Income** – Monthly earnings.
- **Total Assets** – Total value of owned assets.
- **Total Liabilities** – Total outstanding debts.
- **Net Worth** – Financial worth after liabilities.
- **Home Ownership Status** – Own, Rent, or Mortgage.
- **Savings Account Balance** – Amount in savings account.
- **Checking Account Balance** – Amount in checking account.
- **Debt-to-Income Ratio** – Ratio of total debt to income.
- **Total Debt-to-Income Ratio** – Overall debt burden.
- **Monthly Debt Payments** – Recurring debt obligations.
- **Utility Bills Payment History** – Payment record for utilities.

2.3 Credit History and Loan Information

- **Credit Score** – Creditworthiness score.
- **Credit Card Utilization Rate** – Percentage of available credit used.
- **Number of Open Credit Lines** – Count of active credit lines.
- **Number of Credit Inquiries** – Number of recent credit checks.
- **Bankruptcy History** – Record of past bankruptcies.
- **Length of Credit History** – Duration of credit history.
- **Previous Loan Defaults** – Instances of defaulting on past loans.
- **Payment History** – Record of timely or missed payments.

2.4 Loan Attributes

- **Loan Amount** – Requested loan size.
- **Loan Duration** – Repayment period.
- **Loan Purpose** – Reason for loan application.
- **Base Interest Rate** – Initial interest rate before adjustments.
- **Interest Rate** – Final applied interest rate.
- **Monthly Loan Payment** – Monthly installment for loan repayment.
- **Application Date** – Date of loan application.

2.5 Target Variables

- **Loan Approved (Binary Classification Task)** – Indicates whether the loan was approved (1) or denied (0).
- **Risk Score (Regression Task)** – A continuous risk score representing an applicant's likelihood of default or financial instability.

3 Tasks and Requirements

This dataset enables two primary machine learning tasks: classification and regression.

3.1 Loan Approval Classification (Supervised Learning)

- Develop a classification model to predict whether a loan will be approved.
- Train models.
- Evaluate performance using **accuracy, precision, recall, and F1-score**.
- Identify key financial and credit factors influencing loan approval.

3.2 Risk Score Prediction (Supervised Learning - Regression)

- Develop a regression model to predict an applicant's risk score.
- Apply models such as **Linear Regression, Decision Trees, Random Forest Regressor, and XGBoost**.
- Evaluate performance using **Mean Squared Error (MSE) and R-squared values**.
- Determine which financial indicators contribute most to risk assessment.

3.3 Clustering Analysis (Unsupervised Learning)

- Perform clustering to segment applicants based on financial stability and credit behavior.
- Apply clustering algorithms such as **K-Means, DBSCAN, and Hierarchical Clustering**.
- Use **Elbow Method and Silhouette Score** to determine the optimal number of clusters.
- Identify different risk categories among applicants.

3.4 Visualization and Reporting

- Generate histograms and scatter plots to visualize financial distribution trends.
- Create correlation heatmaps to analyze relationships between variables.

4 Submission Requirements

- A well-structured report detailing the methodology, results, and analysis in a given report format.
- Python code is used for implementation.
- A presentation summarizing key findings and recommendations in a given presentation format.