FINAL PROJECT MILESTONE

1. Dataset

For our group, we have chosen the “German Credit Card” dataset that is essential for studying credit risk assessment and predictive modeling in the financial sector.

When a bank receives a loan application, it needs to determine whether to proceed with approving the loan. This decision is made based on the application’s profile. The primary aim of the “German Credit Data” is to reduce the risk of loans to applicants while maximizing the chances of profiting from good loans. Loan managers evaluate an applicant’s demographic and socio-economic profiles before making a decision on their loan application.

The task requires exploring the data and constructing a predictive model (classification) to offer guidance to a bank manager in making loan approval decisions. The bank’ decision carries 2 types of risks:

- Customers categorized as having a good credit risk are likely to repay the loan. If their loan applications are not approved, the bank faces a loss of potential business opportunities.
- On the other hand, customers classified as having a bad credit risk are unlikely to repay the loan. Approving their loan applications results in a financial loss for the bank.

The “German Credit Data” dataset consists of information on 20 variables and classifies whether each of the 1,000 loan applications is deemed to be a **Good** or **Bad** credit risk. It contains various a range of attributes that provide insights into the creditworthiness of customers. These may include:

- Age (years)
- Sex (male and female)
- Job status (unemployed, employed, self-employed, etc)
- Housing type (own, rent, or for free)
- Saving/Stocks (high, little, low, no saving)
- Duration of credit (months)
- Purpose of credit (car, education, home, business)
- Other financial indicators

2. Project Approach (Using Python-pandas and sklearn)

Data Preparation, EDA:

- ***Attribute Types***: Utilize Python's pandas’ library to discern attribute types (nominal, ordinal, quantitative) for comprehending the data structure.

- ***Missing Values:*** Python's pandas to check missing values.
- ***Descriptive Statics:*** Computing descriptive statistics, including maximum, minimum, mean, and standard deviation, offers valuable insights into the distribution and attributes of the data. (Using Python or R)
- ***Distribution Analysis:*** Understanding of the distribution of attributes associated with financial indicators and demographics can yield valuable insights into the traits of credit applicants and their probability of default. (Python's matplotlib or seaborn libraries)
- ***Correlation Analysis***: Analyzing correlations between attributes help identify relationships and dependencies that may influence the credit approval decisions.

Predictive Modeling (Classification): Using the Decision tree and Naive Bayes predictive modeling, we are going to use the following steps for the dataset we choose.

Data Preprocessing:

- Load the dataset into my programming environment (Python, R, etc.).
- Preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features if necessary.
- Split the dataset into training and testing sets.

Model Training:

- Train a decision tree classifier and a Naive Bayes classifier using the training data.In Python.

Model Evaluation: Evaluate the performance of the models using the testing set.

Prediction: Once we are satisfied with the models' performance, we will use them to make predictions on the dataset.

Conclusion and Recommendation:

- Summarize from data preparation and predictive modeling sections
- Provide recommendations