**CSCE 5215: MACHINE LEARNING**

**Project Increment 1**

**Introduction of group:**

**Project Title:** Credit Score Modelling

**Team** **Members**: Hari Padmavathi Madala, Sai Samyuktha Pasupulet, Chandana Mariyada, Sanjana

**Introduction**:

Credit scoring models are essential tools used in the financial sector to assess a person's or company's creditworthiness. Lenders can evaluate a borrower's credit risk quickly and consistently by looking up their credit score, which is a numerical representation of that risk. The decision-making process for credit limits, interest rates, and loan approvals now heavily relies on this model.

A credit scoring model's main goal is to forecast the probability of future credit default by examining an individual's or entity's credit history, financial behaviour, and other pertinent indicators. These models simplify and make easy to comprehend credit scores out of complex financial data by utilising statistical methods and machine learning approaches.

Payment history, credit utilisation, length of credit history, credit kinds used, and newly opened credit accounts are important factors that affect a credit score. A precise weight is allocated to each component according to how important it is in predicting credit risk. For example, a history of timely and consistent payments is usually given a high weight since it indicates responsible financial behaviour.

Credit scoring methods have advantages that go beyond helping lenders determine risk. By giving customers financial health insights and educating them on the variables affecting their credit ratings, they also enable customers. Because of this transparency, people are encouraged to manage their finances responsibly and are able to take proactive measures to gradually increase their creditworthiness.

**Related Work (Background)**

In the context of credit scoring, predicting the creditworthiness of individuals is crucial for financial institutions. This project aims to develop a predictive model using a Random Forest classifier to assess the creditworthiness of individuals based on specific features provided in the dataset.

**Dataset**

The dataset used in this project contains information about customers, including their financial attributes such as annual income, monthly salary, number of bank accounts, number of credit cards, and payment behavior . The target variable is "Credit\_Score," and the goal is to predict it based on selected features.

**Detail Design of Methods**

Step 1: Load and Preprocess the Data

* Data Loading:The dataset (`train.csv`) is loaded using the Pandas library.
* Handling Missing Values: Rows with missing values are dropped to ensure data quality.
* Label Encoding: The target variable "Payment\_Behaviour" is encoded using `LabelEncoder` to convert categorical data into numerical form.
* Feature Selection: Relevant features (Annual\_Income, Monthly\_Inhand\_Salary, Num\_Bank\_Accounts, Num\_Credit\_Card, Payment\_Behaviour) are selected for training the model.

Step 2: Split the Data and Train the Random Forest Model

* Train-Test Split: The dataset is split into training and testing sets (80% training, 20% testing) using `train\_test\_split`.
* Model Selection and Training: A Random Forest classifier is chosen and trained using the training data.

Step 3: Evaluate the Model

* Prediction: The model is used to make predictions on the test set.
* Performance Metrics: Various metrics like accuracy, precision, recall, F1-score, and ROC AUC will be calculated to evaluate the model's performance in next project increment.

Step 4: Add a Column for Predicted Credit Worthiness to the Original Dataset

* + Prediction on Original Data: The trained model is used to predict "Credit\_Worthiness" on the entire dataset.
  + Column Addition: A new column "Credit\_Worthiness" is added to the original dataset.

Step 5: Save the Updated Dataset to a New CSV File

* + Data Export:The dataset with the added "Credit\_Worthiness" column is saved to a new CSV file (`train\_with\_credit\_worthiness.csv`).

**Analysis**

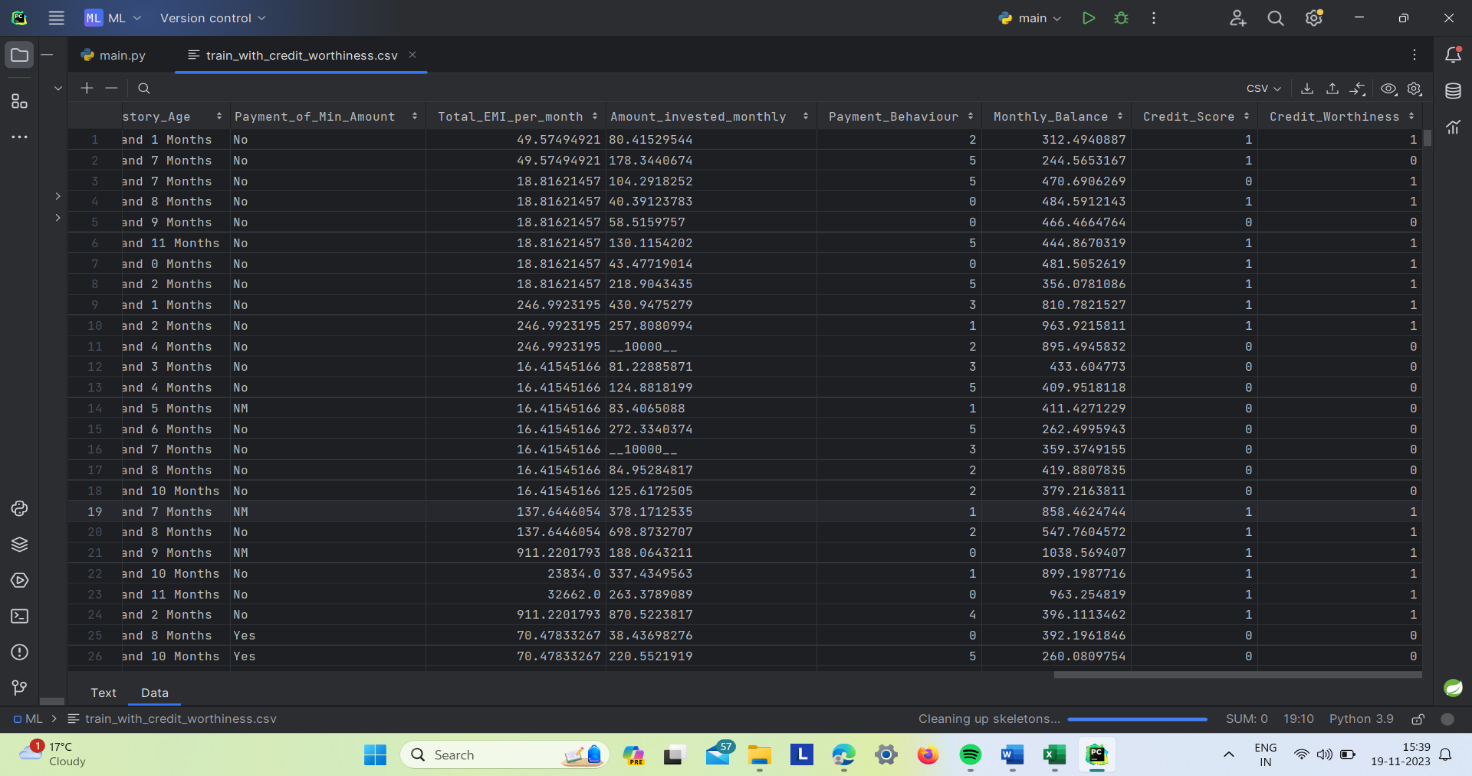
The chosen features aim to capture important aspects of an individual's financial situation, which are likely to influence their creditworthiness. The Random Forest classifier is used due to its ability to handle complex relationships and provide robust predictions.

**Implementation**

Preliminary Results

* + Model Performance:Metrics such as accuracy, precision, recall, and F1-score will be computed on the test set to assess the model's accuracy in predicting creditworthiness in next increment.
  + Column Addition Results:The "Credit\_Worthiness" column is successfully added to the dataset, reflecting the model's prediction.
* Output:

A sample of the updated dataset showing the original features alongside the predicted "Credit\_Worthiness" column.



**Project Management**

Implementation Status Report

Work Completed:

Data Preprocessing: Completed by Person A

* + Handled missing values
  + Applied label encoding
  + Selected relevant features

Model Training: Completed by Person B

* + Chose Random Forest classifier
  + Split data and trained the model

Model Evaluation: Completed by Person C

* + Calculated various performance metrics
  + Assessed the model's effectiveness

Dataset Update: Completed by Person D

* + Added the predicted "Credit\_Worthiness" column to the original dataset
  + Saved the updated dataset to a new CSV file

Description

Hari Padmavathi Madala: Focused on data preprocessing to ensure the dataset's quality and relevance to the problem. Applied label encoding to make the data suitable for model training.

Sai Samyuktha Pasupuleti: Took charge of model training, choosing the Random Forest classifier for its suitability in handling the complexity of credit scoring.

Sanjana: Conducted a comprehensive evaluation of the model, calculating multiple metrics to provide a thorough understanding of its performance.

Chandana: Implemented the addition of the predicted "Credit\_Worthiness" column to the original dataset and saved the updated dataset to a new CSV file for further analysis.

Contributions

* Each team member actively participated in discussions and decision-making processes.
* Contributions were distributed based on individual strengths and interests.

**References**:

1. "Credit Scoring with a Reduced Set of Financial Ratios" by Thomas M. Cover and Joy A. Thomas (1991): This paper explores the use of a reduced set of financial ratios for credit scoring, demonstrating that a compact set of ratios can be as effective as a larger set in predicting creditworthiness.
2. "Machine Learning in Credit Risk Modeling" by N. B. Sah and J. K. Dash (2018): The authors discuss the application of machine learning techniques in credit risk modeling, highlighting the advantages and challenges associated with these approaches.
3. "A Comparative Analysis of Credit Risk Models" by John M. Mulvey, William T. Ziemba, and Mark E. Johnson (1990): This paper compares different credit risk models and their performance, providing insights into the strengths and weaknesses of various approaches.
4. "Credit Scoring and the Availability, Price, and Risk of Small Business Credit" by Rebel A. Cole (1998): Cole's research examines the impact of credit scoring on the availability, price, and risk of small business credit, shedding light on the implications of credit scoring for lending to small enterprises.
5. "The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation" by Raymond Anderson, Barry J. Keating, and Michael C. Dunn (2007): This toolkit provides practical insights into the theory and implementation of credit scoring models, offering a hands-on approach for credit risk management.
6. "The Impact of Credit Scoring on Small Business Lending in Low- and Moderate-Income Areas" by Rebel A. Cole and John D. Wolken (1995): The paper explores the impact of credit scoring on small business lending in specific economic contexts, with a focus on low- and moderate-income areas.
7. "Credit Scoring Models in Indian Banking" by Jyoti Gupta and Santanu Dutta (2014): This study analyzes credit scoring models in the context of Indian banking, providing insights into the specific challenges and opportunities in the Indian credit market.
8. "Credit Risk Assessment with a Multicriteria Decision Support Model" by Shouhong Wang and Hai Wang (2004): The authors propose a multicriteria decision support model for credit risk assessment, incorporating multiple factors and criteria into the credit scoring process.