JM0150-M-6 - Data Mining Course

2nd Group Assignment

This assignment aims to assess students' understanding and application on advanced concepts and techniques in Data Mining domain and the design and implementation of an end-to-end pipeline with optimized models.

Instructions:

- This is the 2nd Group Assignment (5-6 students per group).
- It is graded with 20points in total.
- The code of your assignment consists of four parts that are also assessed respectively as shown below.
- Along with your **iPython notebook** you have to also submit a **technical report** (max 6 pages and in PDF format), covering the topics that are indicated in the respective template that can be found here. Please upload both files in a zip or rar file by the deadline
- The **domains and datasets** that you can select to implement your Group Assignment are as per the below table.

Domain of Interest	Dataset
Financial	Marketing Campaign
e-Commerce	E-Commerce Data
Banking	Bank Marketing

- Please feel free to select the dataset and domain of your preference.
- Clear explanations and well-documented code are expected.
- You should pay high attention in the interpretation and justification of your results and code. **Results without any explanation and justification will not be graded.**

You are requested to:

a) apply clustering and association rules mining techniques towards identifying relationships and patterns. Applying such techniques on the pre-processing step can improve the predictive capabilities of your models.

b) train, evaluate, and compare three different predictive models. Finetune and optimize their performance and detail any trade-off between accuracy, use of resources, training time etc..

The overall structure of your **Group Assignment** should include the below parts (in high-level). Of course, please feel free to proceed with your own implementations and structuring of your code. In parenthesis you can find the overall assessment process and the allocated per part points.

Part 1: Motivation and Problem Statement (2 points)

Define a real-world problem that can be solved based on the dataset and domain that you have selected to proceed.

- 1 point: Provide a clear explanation of the problem, key objectives, and how data mining techniques (clustering, association rule mining, and prediction) will be beneficial.
- 1 point: Identify and explain the types of data mining tasks that need to be applied (e.g., cleaning, clustering, association rule mining, and prediction) and are relevant to the problem. Provide an initial pipeline and architecture of your approach.

Part 2: Advanced Exploratory Data Analysis (EDA) (8 points)

After designing and elaborating on the proposed architecture, you are requested to proceed with the implementation of an advanced and extensive EDA by also applying techniques from clustering and association rule mining domains. In this step you are also requested to proceed with any needed data processing and transformation task that you consider necessary.

- 1 point: Discuss potential challenges, limitations, and propose solutions related to data quality, computational complexity, and data mining techniques, and discuss approaches of how to tackle and address them through your approach. Perform the identified data cleaning (deal with outliers etc.) tasks. Summarize the dataset using descriptive statistics and create visualizations (histograms, boxplots etc.) to describe the distribution of key variables. Provide adequate explanations on the insights that can be derived from these initial steps of the EDA.
- 1 point: Apply advanced feature engineering techniques (e.g., scaling, normalization etc.), and deal with categorical data (e.g., one-hot encoding). Perform more advanced visualizations (pair plots, correlation heatmaps) and provide insights and explanations based on these visualizations. Detail the univariate and multivariate analysis that you perform and justify its specific case.

- 2 points: Apply two different clustering algorithms (e.g., DBSCAN, k-means, hierarchical) and evaluate clusters using metrics such as silhouette score or within-cluster variance. Provide explanations on the optimal number of clusters to be selected and detail the use of the Elbow method. Provide enhanced visualizations of the clusters using 2D or 3D scatter plots (if applicable), color-coding the data points based on their assigned clusters. For a deeper analysis, create heatmaps or pair plots to examine how each feature contributes to cluster formation and whether there are distinguishable patterns within and across clusters.
- 1 points: Explore how association rules could be extracted from the clustered data. Perform association rule mining (e.g., Apriori or FP-Growth) and generate rules. Evaluate the rules using metrics like confidence and lift, and discuss their potential value for the business problem.
- 1 point: Interpret the insights derived from clustering and association rules in the context of the problem and discuss potential business or practical applications. Explain how these results align with business objectives and guide the next steps in predictive modeling.
- 1 point: Propose any improvements for the clustering algorithm or association rule mining process, such as trying different parameter configurations or adjusting the minimum support and confidence thresholds.
- 1 point: Build respective data processing pipelines for each one of the models implemented in the next step by using the scikit-learn library.

Part 3: Predictive Modeling (8 points)

Build three different predictive models to solve the problem. You are free to use any algorithm (e.g., any type of Naïve Bayes, ANN, SVM, Ensemble Learning models, Random Forest, etc.) of your preference, however the final selection of the models should be justified from **technical, business, and scientific** perspective, by also providing references to related research works and State-of-the-Art techniques related to your problem. You are also requested to perform cross-validation and model fine-tuning for improved performance of your models, by also justifying your selections and final parameters usage.

- 2 points: Split the data into training and testing sets. Train three simple models (e.g., XGBoost, ANN, Logistic Regression, Random

Forest etc.) on the dataset. Provide an in-depth evaluation using multiple different metrics (accuracy, precision, recall, F1-score, confusion matrix, ROC curve, MSE, MAE etc.) depending on your selected predictive case.

- 2 points: Finetuning and continuous validating the performance of your models by applying cross-validation techniques to evaluate the model's performance more robustly. Compare the performance between the three different models and their versions that you will finally select to apply in your scenario by focusing on the cross-validation and finetuning steps.
- 2 points: Interpret the outcomes of the models, perform error analysis, and assess limitations. Identify ways to improve performance, balancing training time, accuracy, and variance. Justify any trade-offs made in the fine-tuning process. Provide a well-justified final selection between these three models. Discuss and propose ways to improve models' performance and provide a detailed explanation of the process.
- 2 points: Conclude on the final insights and outcomes of your overall implementation and discuss how this approach is in alignment with the initially set objectives and goals. Highlight the impact and added-value in terms of technical and business perspective.

Part 4: Report (2 points)

You are requested to submit a technical report along with your iPython notebook. The report should be submitted in **PDF format** and follow the instructions and structure of the template as indicated in the introduction. The delivery of the report will be evaluated with a score of up to 2 points, based on its clarity, structure, level of justification, and overall efficiency in demonstrating and detailing your approach, implementations, and results.