JM0150-M-6 - Data Mining Course

Individual Assignment

This assignment aims to assess students' individually on the understanding and application on advanced concepts from the Data Mining domain and their critical thinking, integration of different approaches and implementation of an end-to-end pipeline following the techniques and approaches discussed in the context of the course.

Instructions:

- The Individual Assignment is graded with 40points in total.
- The code of your assignment should be consisted of three different parts that are also assessed respectively as shown below.
- Along with your iPython notebook you have to also submit a **scientific report** (max 6 pages), 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.
- The **dataset** that you are requested to leverage under the scopes of this **Individual Assignment** is the CDC-NHANES dataset that incorporates information with regards to the health, demographic, and nutritional variables. This dataset contains six (6) different files, thus students are requested to proceed with merging the appropriate files for the implementation of a comprehensive analysis from their side.
- 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.**

The overall structure of your **Individual 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.

You are requested to:

a) establish a real-world scenario that can be extracted out of the given files of this dataset. Your final target is to create a classification model based on the available data by merging two or more files of the given dataset by analyzing the presence of conditions based on

health metrics. You are free to select your own target outcome for the classification, such as hypertension status or diabetes presence etc. You are strongly advised to go also beyond a binary classification problem and provide different levels of risk based on probabilities;

b) apply clustering and association rules mining techniques towards identifying relationships and patterns within the different files and health determinants. The latter will improve the interpretation of the results and the relationships/interlinking between different diseases, nutrition habits, demographics etc. Moreover, applying such techniques on the pre-processing step can improve the predictive capabilities of your models.
c) 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..

Part 1: Business Understanding and Problem Statement (6 points)

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

- 2 points: Provide a clear explanation of the problem, outline key objectives, and describe the potential benefits of solving the problem using data mining.
- 2 points: Provide a concrete background research and motivation based on related research work and surveys to better define your problem.
- 2 points: 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: Data Preparation and Exploratory Data Analysis (EDA) (12 points)

After designing and elaborating on the research questions and the proposed methodology to be followed, 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 to handle missing values (e.g., by imputation or removal), to discretize continuous variables where

applicable, especially for association rules, and to normalize or standardize variables for clustering and for the later step of classification.

- 3 points: Discuss the potential challenges and limitations that might be encountered within your introduced architecture and related to the data mining techniques (e.g., data quality, computational complexity), 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 and to better support and explain the merging approaches that you decided to follow. Provide adequate explanations on the insights that can be derived from these initial steps of the EDA.
- 2 points: 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 Elbow method, silhouette score, and within-cluster, between-cluster variance. 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.
- 3 points: Explore how association rules can be utilized in the context of your project. Identify significant association rules within the data to explore relationships between health and demographic variables. Select and interpret at least 3 association rules, focusing on support, confidence, and lift values. Try to select categorical variables to answer questions such as "*Is there any association between physical activity level and hypertension?*", "*Are there specific dietary habits associated with increased cholesterol levels?*" etc..
- 2 points: 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.

Part 3: Classification Modeling and Deployment (12 points)

Build four different predictive models to solve the problem. You are free to use any algorithm (e.g., any type of Logistic Regression, ANN, SVM, Ensemble Learning models, etc.) of your preference, however the final selection of the models should be in alignment with the below tasks and 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. It is highly advised to also link your selections with insights derived from the previous step. 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.

- 4 points: Develop and compare classification models to predict a specific health outcome and classify the patients based on this. Split the data into training and testing sets. Train two simple models (e.g., SVM, Logistic Regression etc.) on the dataset. Provide an in-depth evaluation using multiple different metrics (accuracy, precision, recall, F1-score, confusion matrix, ROC curve etc.) depending on your selected predictive case. Finetune 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 two different models and their versions that you will finally select to apply in your scenario by focusing on the cross-validation and finetuning steps
- 4 points: Implement an ensemble technique to improve the classification accuracy. Apply an ensemble method (e.g., Random Forest or XGBoost) to classify the same target variable used in the classification task. Compare the performance with the individual classifiers from the previous task. Discuss if the ensemble model improves upon the individual models, and provide reasons for the improvement or lack thereof.
- 4 points: Build a simple neural network to predict the chosen health outcome. Experiment with architecture (layers, neurons) and activation functions. Compare the neural network's performance with the previous classifiers, and discuss its strengths or weaknesses.

Note: Always interpret the outcomes of the models and justify any trade-offs made in the fine-tuning process. Provide a well-justified final selection between the implemented models.

Scientific Report (10 points)

The delivery of the scientific report presentation will be evaluated with a score of up to 10 points, based on its clarity, structure, and overall effectiveness in presenting the followed approach and implementations. You are requested to 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. The overall scaling for the assessment of this Scientific report is as per below:

- Conformance to the format of the IEEE template (2 points)
- Establishment of concrete and well-defined research questions (1 point)
- Number of references and related works discussed and analyzed (1 point)
- Justification of introduced research methodology and description of the overall workflow to be followed (2 points)
- Extensive discussion and interpretation of the results from scientific point of view (2 points)
- Discussion on the impact and possible implications of this research work (1 points)
- Discussion on the limitations and future work that you envision (1 point)