Data Mining Project

A **project** consists in data analysis based on the use of data mining tools. The project has to be performed by a team of 2/3 students. It has to be performed by using Python. The guidelines require to address specific tasks and results must be reported in a unique paper. The total length of this paper must be **max 20 pages** of text including figures. The students must deliver both: paper and well commented Python notebooks.

Task 1 Data Understanding and Preparation (30 points):

Task 1.1: Data Understanding: Explore the dataset with the analytical tools studied and write a concise "data understanding" report describing data semantics, assessing data quality, the distribution of the variables and the pairwise correlations.

Task 1.2: Data Preparation: Improve the quality of your data and prepare it by extracting *new features* interesting for describing the customer profile and his purchasing behavior. These indicators have to be extracted for each customer. Indicators to be computed are:

- I: the total number of items purchased by a customer during the period of observation.
- Iu: the number of distinct items bought by a customer in the period of observation.
- Imax: the maximum number of items purchased by a customer during a shopping session
- E: the Shannon entropy on the purchasing behaviour of the customer

It is MANDATORY that each team defines **additional indicators** leading to the construction of a customer profile that can lead to an interesting analysis of customer segmentation.

Once, the set of indicators will be computed the team has to explore the new features for a statistical analysis (distributions, outliers, visualizations, correlations).

Subtasks of DU

- Data semantics
- Distribution of the variables and statistics
- Assessing data quality (missing values, outliers)
- Variables transformations & generation
- Pairwise correlations and eventual elimination of redundant variables

Task 2: Clustering analysis (30 POINTS - 32 with optional subtask)

Based on the customer's profile explore the dataset using various clustering techniques. Carefully describe your decisions for each algorithm and which are the advantages provided by the different approaches.

Subtasks

- Clustering Analysis by K-means:
    1. Identification of the best value of k
    2. Characterization of the obtained clusters by using both analysis of the k centroids and comparison of the distribution of variables within the clusters and that in the whole dataset
    3. Evaluation of the clustering results
- Analysis by hierarchical clustering
    1. Compare different clustering results got by using different version of the algorithm
    2. Show and discuss different dendrograms using different algorithms
- **Optional (2 points):** Explore the opportunity to use alternative clustering techniques in the library: https://github.com/annoviko/pyclustering/

Task 3: Predictive Analysis (30 POINTS)

Consider the problem of predicting for each customer a label that defines if (s)he is a **high-spending** customer, **medium-spending** customer or **low-spending** customer. The students need to:

1) define a customer profile that enables the above customer classification. Please, reason on the suitability of the customer profile, defined for the clustering analysis. In case this profile is not suitable for the above prediction problem you can also change the indicators.
2) compute the label for any customer. Note that, the class to be predicted must be nominal.
3) perform the predictive analysis comparing the performance of different models (at least Decision Tree and Random Forest) discussing the results and discussing the possible preprocessing that they applied to the data for managing possible problems identified that can make the prediction hard. Note that the evaluation should be performed on both training and test sets.

Task 4: XA (32 POINTS - Optional)

Explanation Analysis. Consider the non-interpretable models used in Task 3 (eg. SVM, ensemble methods, etc) and study the global explanation with SHAP and the local explanation with LIME and SHAP. Use the evaluation metrics presented during the XAI Laboratory (for using LORE, you can download the library at the link: https://github.com/rinziv/XAI_lib_HAI-net_Tutorial and follow the instructions contained in the notebook presented during the XAI Laboratory).

Rules for final delivery and Exam

Project Delivery.

Each group must deliver by email a zipped folder named **DM_GroupID.zip** and containing 4 folders and 1 pdf file:

1. a folder named **DM_GroupID_TASK1,** containing source code of data understanding
2. a folder named **DM_GroupID_TASK2,** containing source code of data clustering
3. a folder named **DM_GroupID_TASK3,** containing source code of classification
4. a folder named **DM_GroupID_TASK4,** containing source code of sequential pattern mining
5. a pdf file with maximum 20 pages including figures discussing the results of the 4 tasks. The name of this file must be: **DM_Report_GroupID.pdf.** The file must contain the list of authors (i.e., members of the group).

Remember that the final submission can contain updated versions of the work already delivered in the previous deadlines.

Exam

There are two possible options for the exam:

1. project presentation + questions on the whole program
2. project presentation + paper presentations (in the dates already fixed)

I prefer to have group presentations of the project. If this is impossible we can find a solution together.

Final Grade

The final grade of the exam is given by the weighted average of the project evaluation and oral/paper presentation evaluation. I will assign a weight of 70% to the project work and 30% to the oral/paper presentation. Consider that the project evaluation also includes the project presentation (my suggestion is using slides). Remember that any student must be able to answer any question on the project work.