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 vendor profile and his behavior. These indicators have to be extracted for each vendor.

The **additional indicators** should be useful for an interesting analysis of vendor 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 vendor'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 density-based clustering**:**
    1. Study of the clustering parameters
    2. Characterization and interpretation of the obtained clusters
- 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
- Final evaluation of the best clustering approach and comparison of the clustering obtained
- **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 profile a label that defines if it is a **big-seller** vendor, **small-seller** vendor. The students need to:

1) define a vendor profile that enables the above vendor classification. Please, reason on the suitability of the vendor 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 vendor. Note that, the class to be predicted must be nominal.
3) perform the predictive analysis comparing the performance of different models 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: Frequent Pattern mining and Association Rule Mining (30 POINTS)

Consider the problem of mining frequent common patterns of products bought together and extract the interesting association rules. Explore different parameters.

Optional (2 points): Study and compare the association rules of different geographical regions. One possibility is to consider the top regions or countries and compare the results obtained by mining patterns in each one.

Rules for final delivery and Exam

Project Delivery. The project has to be delivered at least 5 days before the oral exam.

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 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).

Exam

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.