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 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/

Delivery of the first draft of the report with Task 1.1, Task 1.2 and Task 2: 5 November

Note: The final report delivered within the end of December can also improve the already delivered tasks.

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 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 set.

Delivery of the Task 3: 2 December