Project instructions
Data Mining I

The objective of the Data Mining I project is to test yourself on a real knowledge discovery process using the methods and algorithms learned during the course. This is, by design, an **independent** activity, where you are expected to identify relevant questions that can be answered using Data Mining methods, autonomously reason about the encountered problems and identify appropriate solutions, and it is a task based on *real data*: nothing has been simplified to force some educational concepts to emerge, as done instead in the preparatory assignments.

These are the basic instructions:

GROUPS
- The project is performed by the groups formed on the student portal.

EXAMINATION
- The project is on the G/U grading scale (pass/not_pass).
- To pass the project, you must:
    1. Choose one Data Mining **question/problem** regarding the data you have chosen. (That is, you must recognize that what you are asking for requires the application of at least one of the Data Mining algorithms studied in this course, and cannot be performed e.g. just using SQL or basic descriptive statistics.)
    2. Identify the Data Mining **algorithm(s)** that is (are) appropriate to answer your question. Notice that we require both a question/problem (1), expressed without any references to how you are going to address it – for example, "Can we identify any discrimination between men and women based on census data?" – and the technical way in which you are going to answer (2) – e.g., setting up a classifier based on salaries, education, etc.
    3. Preprocess the data appropriately, to make it ready for the selected Data Mining algorithms.
    4. Execute the chosen Data Mining algorithms and present the obtained results in a written report. Given that you work on real data, it may happen that you cannot identify any patterns in the data.
    5. Interpret your results and be able to convince your examiner that your results can be trusted (whether you have found patterns or you are claiming that there are no patterns in the data).
- You must submit a short report on the student portal before the deadline indicated online. The report must be essential (we give no page limit, but we expect typical reports not to exceed 3/4 pages), but provide all the information necessary to verify that the relevant items in the checklist at the end of this document have been achieved. It will help if you organize the report based on the 5 items above.

DATA & TOOLS

You are free to choose any tools to perform your analysis. A recommended combination is MySQL or sqlite (if you need a lot of initial preprocessing power and you want to exploit your knowledge of SQL) and RapidMiner, or only RapidMiner if you do not need database methods, but you are free to choose other tools if you prefer, e.g., R, Weka, Orange, etc. You do not need any approval for this.

For your analysis, you can use a dataset of your choice and interest (that must be approved by your tutor) or one of the following three datasets (that do not need approval):

- A traditional dataset containing data from the 1990 U.S. Census, that can be downloaded from the following link: https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)
- A dataset obtained by monitoring an Online Social Network, with user posts, likes and a following/follower network. Please, notice that you *do not have to* use all the tables, but you can (and should) choose those suiting your analysis questions: for example, you can focus on the "entries" or "comments" tables. These can be obtained from the following link: https://drive.google.com/folderview?id=0B_D5tuT1vDQtckFGWkk1aTh5VlE&usp=sharing
- Data from the Global Health Observatory Data Repository. Please, notice that the data are spread through several tables, which will need to be integrated following the formulation of your data mining questions. The data are available from the following link: http://apps.who.int/gho/data/?theme=home

Independently of the chosen dataset, consider that **you will probably need to spend most of the project time understanding, retrieving and pre-processing the data**.

SUPPORT

This project tests your ability to independently design and execute a knowledge discovery process on real data – that is, not "academically" prepared to be simple to understand or where existing patterns have been "made ready for discovery".

Therefore, you should work independently on this project.

However, you can have multiple meetings with your tutor (one will be assigned to you) to check that you are on the right track and get feedback. You also need your tutor to check and approve your dataset if you use one that is not in the list above. **You are encouraged to use your own datasets**, but do not have to.

CHECK LIST

□ The problem is clearly stated at the beginning.

□ The problem requires a Data Mining approach.

□ The chosen Data Mining approach is appropriate.

□ The chosen data representation (table, set, graph, …) is appropriate.

□ [for tables] Each attribute has been given the correct type (nominal, …).

□ The chosen algorithm is appropriate (motivate the choice wrt the features of the data).

□ [for text] the list of applied pre-processing operations (stemming, …) has been motivated.

□ [for tables and text] an appropriate proximity function (Jaccard, Manhattan, …) has been used, if required by the algorithm.

□ [for tables] Scaling issues have been addressed.

□ [for tables] Correlation issues have been addressed.

□ [for tables] Dimensionality has been reduced, if necessary.

□ If relevant, a good test dataset has been generated.

□ If necessary, overfitting has been addressed.

□ If necessary, class imbalance has been addressed.

□ If necessary, noise has been reduced.

□ The results are supported by sufficient data (large-enough leaf size, high-enough support, …).

□ The results have been interpreted, and related to the original problem (Was the problem solved? How can the results be used? Any new hypotheses have been generated?)