Knowledge Discovery Process Project

The objective is to develop a complete Knowledge Discovery Process.

Deadlines:

- First Delivery: November 13th: Domain and Preprocessing stage (Points 1 and 2 bellow)
- Main delivery: Presentation live on December 18th, 15h00 (report sent before date of presentation)

Project:

The stages of the project are:

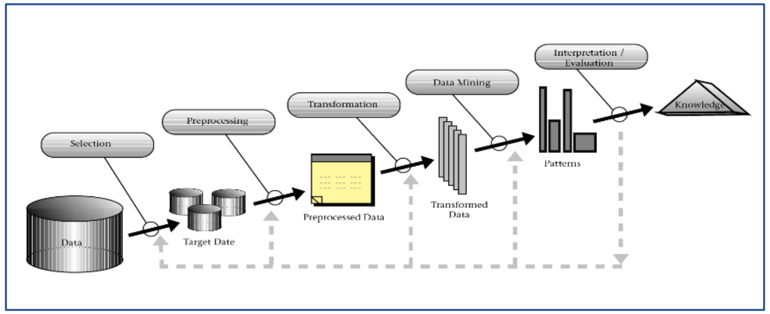
1. Students will choose a **domain** to which data they have access, are interested in, etc. For EIT Students it must include Time Series Data)
2. **Analyze the domain**, describe ithe data set, analyze data characteristics, and perform all the preprocessing tasks needed (cleaning, trasnforming, coding, missing values, irrelevant attributes, etc.)
3. Stablish the **objectives** to be achieved through the Data Mining Project. They will consider the different tasks that would be carried out in each stage of the Knowledge Discovery process according to the specific needs of the domain and the goals you want to achieve.
4. By using a Knowledge Discovery software tool (WEKA, Python or other tool/programming language), **Data Mining** algorithms will be applied to the data of each domain. In addition, the student will analyze the limitations of the algorithms available in the tool and possible improvements.
5. A **evaluation** plan will be made to assess the results obtained and the plan will be executed.

The Data Mining Project will be presented in class. Each group will have 10-12 minutes for the oral presentation The presentation will be made using a Power point presentation (or similar) by all the group members The presentation will address the key points of the project (domain description, KDD objectives, data mining techniques, results and evaluation) .

Please summarize key concepts and ideas in the presentation as the time is very limited.

The report will have a similar structure, but you can provide more details, results, etc. as there is no hard limit for the dimensions of the report. But in any case, we consider that 10-15 pages are more than enough to describe all stages and results of the project. Also, for the final delivery, upload the presentation and the code.

Course PROJECT



- Domain / Input Data set (data set with time series for EIT students) / Goal
  - Select the application domain
  - Domain understanding, data collection, data understanding. Take into account the gender perspective when makes sense (science, health and medicine, environment,….)
  - Goal definition. Match this goal with a data mining task (classification, clustering, TS forecasting,…). You can afford more than one task if you want
- Preprocess your Data
  - Inspect them (visual inspection, distributions, statistics,…). Remove useless features (unique values, …)
  - Clean data (duplicates, missing values, noise, outliers,…). If so, reduce dimensionality (features selection,…)
  - Prepare them to be appropriate as input for the data mining algorithm(s) you want to apply (data types, transformations,…)
  - Take into account all what you need to do (splitting, maybe oversampling/undersampling, etc.)
- Apply Data Mining algorithm(s)
  - Select parameters if needed, and apply on the data/features that you have prepared. Should apply different techniques (recommend also to create a benchmark, i.e., a very simple -or even random- model and compare your results against it)
- Results Validation
  - Use the appropiate metrics to validate the results (depending on the domain, the problem, the data type…)
  - Cross validation generally recomended (adapted to your needs, if so: stratified (imbalance data), “nested” (hyperpameters validation), grouped data,…)
  - Analyze how to improve (varying parameters, input data,…) and apply data mining again until the results are good enough (compare your results with the previous iterations, the benchmark,…)
- Results interpretation
  - Explain your results in an understable way (don't fall into the trap of misunderstanding statistics)

Remember:
Iterative and
interactive