TÉCNICO LISBOA

Data Science 2021/2022

Deadline – January 28th 2021 @ (23:59)

PROJECT GOAL

Critical application of data science techniques to discover information in two distinct problems.

Students are asked to explore the datasets and, in accordance with their findings, adequately select and learn models for the available data, as well as assess and relate those models.

Additionally, students should be able to criticize the results achieved, hypothesize causes for the limited performance of the learned models, and identify opportunities to improve the mining process.

DATA

The datasets for analysis in this project are the following, and **they are available for download in Fénix section Project**.

- **NYC Motor Vehicle Collisions to Person**
  - classification **file** = NYC_collisions_tabular.csv **target** = PERSON_INJURY
  - non-supervised **file** = NYC_collisions_tabular.csv (remove the target)
  - forecasting **file** = NYC_collisions_timeseries.csv **target** = NR_COLLISIONS
  - description on https://www.kaggle.com/kukuroo3/nyc-motor-vehicle-collisions-to-person
- **Air Quality in China**
  - classification **file =** air_quality_tabular.csv **target** = ALARM
  - non-supervised **file** = **=** air_quality_tabular.csv (remove the target)
  - forecasting **file =** air_quality_timeseries.csv **target** = AQI_BEIJING
  - description on https://en.wikipedia.org/wiki/Air_quality_index#China

METHODOLOGY

The project should be developed according to one of the standard data science processes, for example CRISP-DM. Among those steps, only *data profiling*, *data preparation*, *modelling* and *evaluation* will be considered. Despite the iterative nature of those processes, we expect students to perform a single iteration.

Students may choose the mining tool to apply, between **python** (using *scikit-learn*), **R** and any other language. Other business intelligence platform may be used but discouraged, since they are not prepared to deliver the charts required.

Data Profiling

In terms *of data profiling,* data should be characterized along the four perspectives: dimensionality, distribution, sparsity and granularity.

Remember that data profiling is used as a mean to best understand the data and mostly for identifying the required transformations to apply to the original data, in the next step. These transformations aim to improve the performance of classification techniques, to be applied during the modeling phase.

In particular, students should perform a statistical analysis of the datasets in advance and summarize relevant implications in the report, such as the underlying distributions and hypothesize feature dependency.

Data Preparation

At this stage, data should be transformed in accordance with the properties of the original dataset, identified during the previous task. In this manner, the students are allowed to apply preprocessing techniques when needed or under a solid conjecture of its potential impact on learning.

The available and studied techniques to solve those problems should be discussed and applied. If there is no suspicion that one technique is more appropriate than another, both should be applied and the results obtained evaluated.

It is of particular importance the imputation *of missing values* and *dummification*, since the *sci-kit learn* methods' implementation do not deal neither with missing values nor with symbolic non-binary variables.

A third aspect to consider is the distribution of the class attribute, which when very unbalanced does not allow the correct validation of the results. In this situation it is mandatory to apply data *balancing* techniques.

The fourth and final aspect to mention is the *feature engineering* (*feature selection* in particular), whose impact is admittedly significant for some of the studied algorithms.

Additionally, *scaling* transformation and *outlier imputation* may also be performed.

In all cases, the application of each one of the preparation techniques should be assessed. This evaluation should be made by comparing the modeling results before and after the application of these techniques, verifying the impact of each one on the final results.

Suggested methodology

Choosing the best set of transformations to apply to the data, before moving on to the modeling phase, is not trivial. This choice implies the experimentation of each technique over the data, followed by the recognition of the improvements on the results obtained. However, the different preparation techniques, and their alternatives, do not have the same impact on all modelling approaches, and therefore their choice has to be careful.

With infinite time and resources, we could try out all the combinations among preparation and modeling techniques, and find the best models from them. (Using some optimizations this is the path chosen by *autoML* tools, strongly discouraged in this project).

However, there would be a combinatory explosion, which would disturb our understanding of the phenomena and the impact of each preparation techniques per se. It is therefore suggested, students follow a simplification of the process, evaluating the impact of each preparation technique separately.

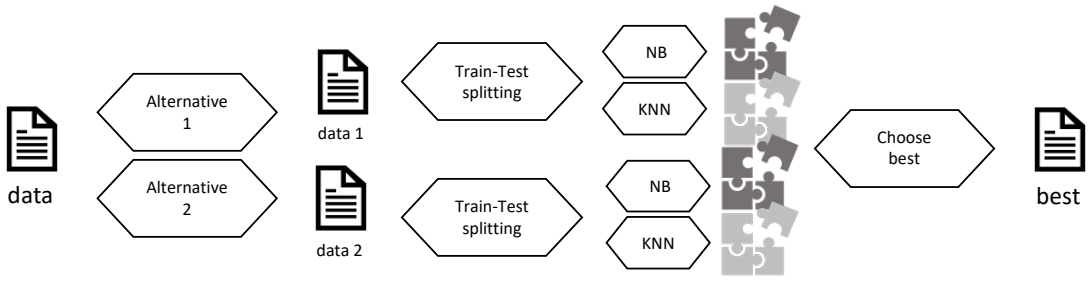
The following figure illustrates the proposed process.

The application of preparation techniques should follow the order illustrated on the figure, carefully chosen to avoid the duplication of efforts. Note `sklearn` doesn't allow for training models over data with *missing values*, and so it has to be the first one to be applied. In the end, data balancing has to be only applied over the training data, in a manner that the test data set doesn't suffer any transformation on the original data distribution. In this manner, the train-test splitting has to be done before data balancing.

In the majority of situations, each preparation technique has diverse alternatives, but each one may have different impact on the modeling techniques results.

The proposal here is to process each transformation and then assess the impact of each alternative. Such impact has to be measured over the models trained over the datasets resulting from those alternative transformations.

The figure below illustrates the application of a generic transformation with two alternatives. And it works as follows.



First, we apply the different two alternatives to a single dataset, generating two new datasets. Then, we apply a train-test split and train new models from each alternative dataset. We suggest the use of both Naïve Bayes and KNN to train these models, resulting in four different models. Our choice is supported by the simplicity of both techniques and the reduced number of parameters to tune. Besides that, the different nature of both approaches limit the chances of choosing a technique best suited for a particular approach.

After training the different models, we chose the preparation technique that presents the best improvement when compared with the previous dataset. In this manner, after the training we may face 4 possibilities:

- none of the alternatives preparation techniques applied improve the results: so, we should **keep the previous dataset** and proceed for the next step;
- one of the alternatives lead to the training of better models using both approaches: so, we **chose the dataset** resulting from this transformation to proceed for the next step;
- the alternative supporting the improvement is different for each learning technique: so, it is necessary to evaluate in which of the models the improvement was higher, and choose the technique responsible for that increase;
- the improvements are residual: so, it is our choice to continue with the previous dataset, or to follow with the technique that theoretically should present higher improvements.

Remember that you should only consider applying the technique, i.e. using one of the resulting files, if in fact there is an improvement in the performance of the models when compared to the performance in the original dataset.

There are two exceptions to this rule:

- the imputation of *missing values*, since sklearn does not allow the application of training algorithms in data with missing *values;*
- and the separation in train and test, in order to warrant an unbiased and independent evaluation of models' performance.

Another important aspect, is that each technique only applies to solve a specific situation. It makes no sense to impute *missing values,* in datasets without *missing values*, for example.

A word about **feature engineering**. Given the different impact on the different algorithms studied, it shall be measured for each of them individually, and therefore shall be studied as one of the key factors for models performance.

Modeling

The project includes three components: classification, unsupervised learning and forecasting. Unsupervised exploration must be done through clustering and association rule mining.

In all situations, the goal is not only to describe the best models learned, but to understand the impact of the available options on the produced models performance.

Classification

Prediction variables **cannot be used.** The supervised exploration must be done via the application of *Naïve Bayes, kNN, Decision Trees, Random Forests*, *Gradient Boosting* and *Multi-Layer Perceptrons*. For this purpose, the use of class variables is mandatory. Evaluation of the obtained models should be done as usual, through confidence measures and evaluation charts. A thorough comparison of the adequacy of the models

should be presented taking into consideration the adequacy of their behavior against the properties of each dataset and their observed performance.

For this purpose the analysis of each classification technique should be done at three different levels:

- the analysis of the impact of the different parameters on models performance;
- the analysis of feature selection on models performance;
- the description of the best model found for each classification technique;
- the comparison of different best models, explaining the different achievements with the different techniques.

In order to train models, don't forget to split the data into train and test datasets. The choice of which approach to use (*hold-out* or *cross-validation*) depends on the amount of data available. Using multiple runs allows for determining the confidence intervals as we studied.

Clustering

Class and prediction variables **cannot be used** to cluster the data, only to plot and evaluate the results, if desired. However, cohesion and separability have to be evaluated.

Feature extraction (in particular PCA) and feature selection (in particular variance and correlation based) shall be tried.

Again the different choices shall be analyzed and the results compared.

Association Rules

Class variables **may be** used for mining association rules, in order to establish some relation with the results from other learning tasks. However, students shall not put any constraints to the rules to be found. In this manner, the rules may have any variable as a consequent.

The evaluation of the rules discovered shall include the analysis of the number and quality of the rules discovered, in particular the support, confidence and lift.

Forecasting

The **only variables** to use are the **date** and the **target variable** identified as the target for forecasting, identified in the first page . Beside developing the forecast models with regression and LSTMs, data preparation shall be applied and described. Matrix Profile should be applied to find some motifs and anomalies.

REPORT

The report file should be named **`report_X.pdf`** (replacing X by the team number) and submitted through Fénix. It should follow the template, without changing the margins and fonts used, and should have at most **15 pages**. Each additional page with analysis won't be considered, but an appendix for data profiling charts is allowed.

The report may be written in Portuguese or English. It should describe the majority of experiments made over the data, from their profile to the discovered models. Beside the placed choices, preparation performed, applied parameterizations and found results for each dataset, their interpretation and critical analysis are mandatory. Additionally, it should include a comparison of the results achieved in both problems, and the relation among the information discovered through the different techniques.

Delivery

The project has to be delivered through Fenix system, after enrolling the team. Only one report per team has to be submitted.

The submission deadline is the one stated in the first page.

EXCELLENCE

A project that applies the suggested data mining techniques over the given datasets and provides a clear and *sound analysis of the collected results is not necessarily an excelling project*.

Excelling projects have three major characteristics.

First, they show an acute understanding of the data characteristics and their impact on the discovery, formulating hypothesis to explain differences in performance.

Second, robust assessments go beyond simple performance indicators, studying different and adequate parameters, and deriving trends from the experiments.

Third, poor results are not acceptable, and there is always something that we can learn from the data.

Plagiarism

Plagiarism is an act of fraud. We will apply state-of-the-art software to detect plagiarism. Students involved in projects with evidence of plagiarism will be reported to the IST pedagogical council in accordance with IST regulations.

EVALUATION CRITERIA

The project will be evaluated as a *whole*. Nevertheless, we provide below a decomposition of the total project score for the purpose of guidance and prioritization:

1. **Data profiling** (**10%**)
2. **Data preparation** (**10%**)
3. **Classification**
    a. Naïve Bayes (**2%**)
    b. KNN(**3%**)
    c. Decision Trees (**5%**)
    d. Random Forests (**5%**)
    e. GradientBoosting (**5%**)
    f. Multi-layer perceptron (**5%**)
4. **Unsupervised**
    a. Association Rules (**5%**)
    b. Clustering (**10%**)
5. **Forecasting**
    a. Time series preparation (**5%**)
    b. Matrix profile (**5%**)
    c. Regression (**5%**)
    d. LSTMs (**5%**)
6. **Evaluation and critical analysis (20%)**

Good Work ! ! !