TÉCNICO LISBOA

2023-2024

Programming for Data Science

Deadline – April 1st 2024 @ (23:59)

I. General description

A. Project Goal

The goal of the data science project is to help students **to understand the impact of the different choices** made along the data science process (*KDD process*), while training classification models, over numeric data.

B. Delivery

Students only have to deliver **a report** describing the results obtained over the exploration of both datasets and tasks. The report should contain a technical description of the procedures performed over the data, the corresponding results, the decisions taken and possible justifications for those results.

You can imagine, you are writing a report to be read by your supervisor, not your client, and so the description shall be technical and not from the domain point of view.

The report may be written in Portuguese or English, but has to follow the **template**, providing all the charts required and without exceeding the number of characters allowed per section. Exceeding text will not be considered. Additional charts are allowed and considered.

The report file shall be named **`report_X.pdf`** (replacing X by the team number) and has to be submitted through **Fénix** before the deadline stated on the first page.

Excellence

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.

II. WORK TO DEVELOP

The project consists of performing only **the first iteration of the KDD-process**, when training a set of models over a single dataset, not considering any additional iteration. In particular, data profiling, data preparation, modeling and evaluation steps have to be performed.

The goal is not only to describe the best models learnt, but to understand the impact of the available options on the produced models' performance.

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.

The dataset for the task was curated from Kaggle for studying the influential factors for life expectancy[1] and is available for **download** in **Fénix section Project**.

[1] https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Data Profiling

For the first task, data should be characterized along three perspectives: dimensionality, distribution and sparsity. Granularity is not required since there are only numeric variables.

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 following 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 shall be transformed solving the problems identified in the previous task.

For this purpose, students are asked to apply preparation techniques in a predefined order (shown in Figure 1), in a manner to minimize the number of datasets to analyze.

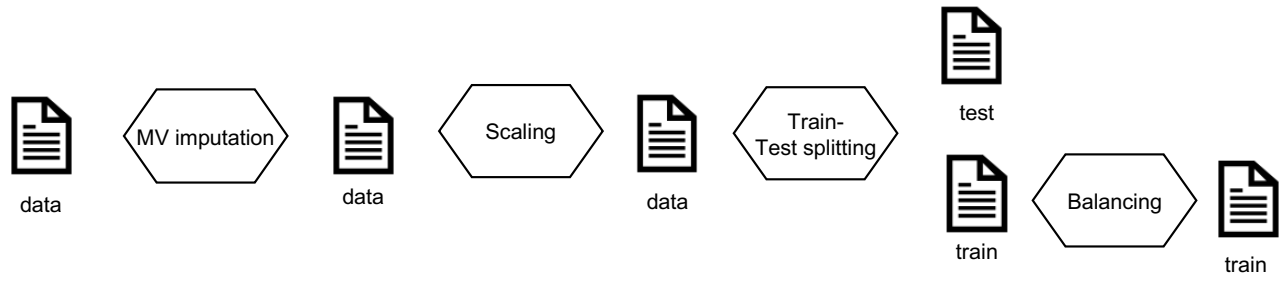

Figure 1 Data preparation methodology for the classification task

The proposal here is to process each alternative transformation and then assess the impact of the resulting datasets on training classification models through KNN and Naïve Bayes, by measuring their performance (see Figure 2).

In this manner, for each preparation step, students have to apply at least two different preparation techniques concerning a single preparation task, in order to obtain different prepared datasets. With each one of these datasets, you train both a KNN and a Naïve Bayes model. And then you compare the results obtained from the different datasets and identify the dataset that led to the best results and proceed with the chosen one to the next preparation step.

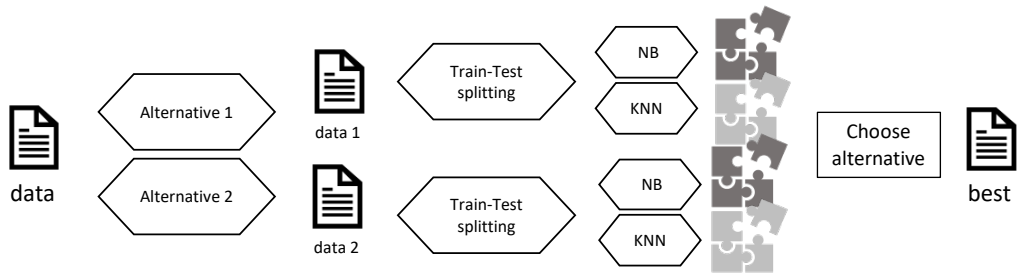

Figure 2 Decision process for each preparation step in the classification task

We suggest the use of both Naïve Bayes and KNN to train these models, due to their simplicity and the reduced number of parameters to tune. Besides that, the different nature of both approaches limits 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 alternative 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 if the data requires it. For example, if a dataset has no missing values, there is no need to perform missing values imputation. **Although, that fact has to be mentioned in the report,** and the decisions of not applying some preparation task have to be justified.

Some additional remarks:

- It is not possible to train models over missing values with `sklearn`; in this manner, the original dataset has to be replaced by one of the prepared ones to proceed to the next step.
- Scaling impact shall be only assessed through the use of KNN, theoretically it shouldn't change the results for Naïve Bayes.
- Balancing has to be applied only to the training dataset, and consequently data partition has to be done previously.

Modeling

During the modeling step, students are asked to train a set of classification models to learn the concept identified by the target variable for the dataset. In particular, students have to apply several machine learning methods and corresponding training algorithms, namely: ***Naïve Bayes***, ***kNN***, ***Decision Trees*** and ***Random Forests***.

Again, the goal is to study the impact of the different options available. This time the different parameterizations for each training algorithm.

The use of automatic optimizations offered on *autoML* frameworks are strongly discouraged, since they find the best parameters but do not give any intermediate results allowing for performing the impact analysis required.

The training data shall be the same for all the training methods, corresponding to the result of the preparation step – the dataset that led to the best performance of KNN and NB.

Evaluation

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 description of the best model found for each classification technique, and its performance.

- The study of overfitting when learning the best model.

Critical Analysis

After identifying the best models learnt with the different ML methods, a critical analysis shall be presented. In particular, students shall compare the best models for each method, concerning their content and performance. This analysis may incorporate an individual explanation for each model found, but mostly **a cross analysis** of the different results.

III. 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:

EVALUATION CRITERIA		
Data profiling	Dimensionality	2%
	Distribution	4%
	Sparsity	4%
Data preparation	Missing values imputation	5%
	Scaling	5%
	Balancing	5%
Modeling	Naïve Bayes	5%
	KNN	5%
	Decision Trees	7.5%
	Random Forests	7.5%
Evaluation	Results	5%
	Overfitting	7.5%
	Justifications	7.5%
Critical analysis		30%

Good Work ! ! !