TÉCNICO LISBOA

2024-2025

Data Science

Deadline – January 10th 2025 @ (23:59)

I. General description

A. Project Goal

The goal of the data science project is to help students **understand the impact of the different choices** made along the data science process (*KDD process*).

To achieve this goal, students are asked to address two distinct domains and two different tasks: classification and forecasting. In both situations, students shall train models from the available data by adequately selecting and preparing them, followed by assessing the learned models.

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

B. Delivery

Students must deliver a report describing the results obtained from exploring both datasets and tasks. The report should contain a technical description of the procedures performed on the data, the corresponding results, the decisions made, and possible justifications for those results.

You can imagine that 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 it has to follow the template, provide all the required charts, and not exceed 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 with 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 two distinct datasets, not considering any additional iteration. Data profiling, data preparation, modeling and evaluation steps have to be performed for each task.

There are two tasks to perform over the datasets: **classification** and **forecasting**.

In both 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.

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

A. Classification

The datasets for the classification task in this project were collected from the Kaggle platform and are available for **download** on **Fénix section Project**.

- **Security domain – NY_arrests**
  - classification **file** = class_ny_arrests.csv **target** = LAW_CAT_CD
  - description available on https://www.kaggle.com/datasets/mageentta/nypd-arrest-data
- **Economy domain – Financial Distress**
  - classification **file** = class_financial_distress.csv **target** = CLASS
  - description available on https://www.kaggle.com/datasets/shebrahimi/financial-distress

In both cases, the data available on Fénix was collected and processed from the original data, reducing the number of records and binary labeling the records. **You have to use the files available on Fénix.**

Data Profiling

For the first task, data should be characterized along the four perspectives: dimensionality, distribution, sparsity and granularity.

When in the presence of symbolic variables, and since `sklearn` is not able to deal with them correctly, students have to choose a **new encoding for those variables**, before proceeding with the correlation analysis.

Remember that data profiling is used to best understand the data and mostly to identify 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) to minimize the number of datasets to analyze.

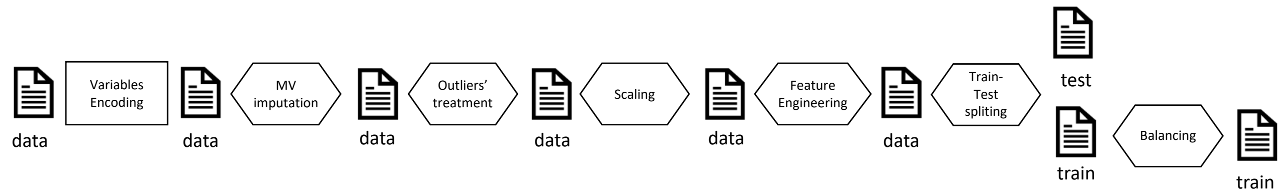

Figure 1 Data preparation methodology for the classification task

Variables encoding is the first step to apply, and it is only required in the presence of symbolic variables. This operation shall result directly from the *granularity* analysis performed in the data profiling step. Among the techniques available you find *transforming into numeric* and *dummification*. Different choices have usually to be made for each variable, however only a choice per variable shall be applied, without applying more than one alternative.

For the rest of the preparation steps, students have to apply at least two alternatives, evaluate each one's impact, and choose the most promising, according to the procedure illustrated in Figure 2.

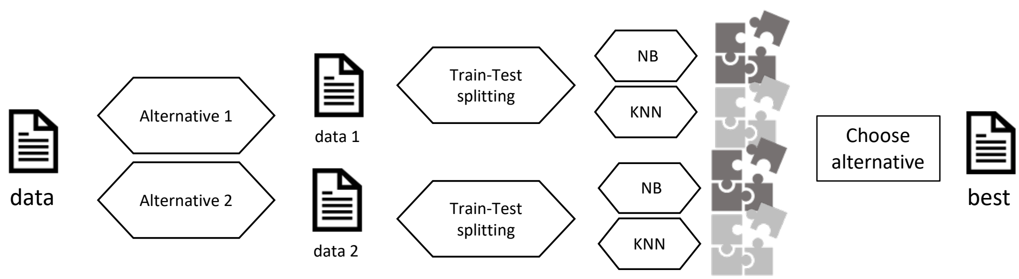

Figure 2 Decision process for each preparation step in the classification task

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.

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

We suggest using both Naïve Bayes and KNN to train these models due to their simplicity and the reduced number of parameters to tune. 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 to the next step.
- One of the alternatives led to the training of better models using both approaches, so we chose the dataset resulting from this transformation to proceed to the next step.
- Each learning technique has a different alternative supporting the improvement. Therefore, it is necessary to evaluate which model had the higher improvement 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 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. **However, that fact has to be mentioned in the report, and the decisions not to apply some preparation tasks** 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 must be applied only to the training dataset, and consequently, data partitioning must be done beforehand.
- When temporal data is present, data partition shall use older data to train and newer data to test, not use future data to classify past data. In case there are several rows concerning the same entity, some entities shall be used for training and others for testing. Otherwise, the partition shall be random.
- Feature selection may be applied before or after balancing. In either case, it may be studied as for the other preparation techniques, using only KNN and Naïve Bayes to assess its results.
- Feature generation will be done in the variable encoding process. Additional variable generation is optional.
- Feature Extraction is out of the scope of this project, but students may apply PCA optionally.

Modeling

During the modeling step, students are asked to train a set of classification models to learn the concepts identified by the target variable for both datasets. In particular, students have to apply several machine learning methods and corresponding training algorithms, namely Naïve Bayes, kNN, Decision Trees, Multi-Layer Perceptrons, Random Forests, and ***Gradient Boosting***.

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 is 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 Naïve Bayes.

Evaluation

The obtained models should be evaluated as usual through confidence measures and evaluation charts. A thorough comparison of their adequacy 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 learned with the different ML methods, a critical analysis will be presented. In particular, students will compare the best models for each method concerning their content and performance. This analysis may incorporate an individual explanation for each model found, but it will mostly be a cross-**analysis** of the different results.

B. Forecasting

The datasets for the forecasting task were collected from the same domains as the data used for classification, and again can be downloaded in **Fénix section Project**.

- **Security domain – NY_arrests**
  - classification **file** = forecast_ny_arrests.csv **target** = Manhattan
  - description available on https://www.kaggle.com/datasets/mageentta/nypd-arrest-data
- **Economy domain – Financial Distress**
  - classification **file** = forecast_gdp_europe.csv **target** = GDP
  - there is no description available

Note that in both cases, the data to be used was sampled from the original data and is ready to use, not requiring additional selection.

Data Profiling

In the forecasting context, profiling pays particular attention to the granularity analysis of the target variable and also to its distribution and stationarity.

Data Preparation

Like classification, data preparation shall follow a pre-defined sequence of operations to reduce the number of datasets to analyze.

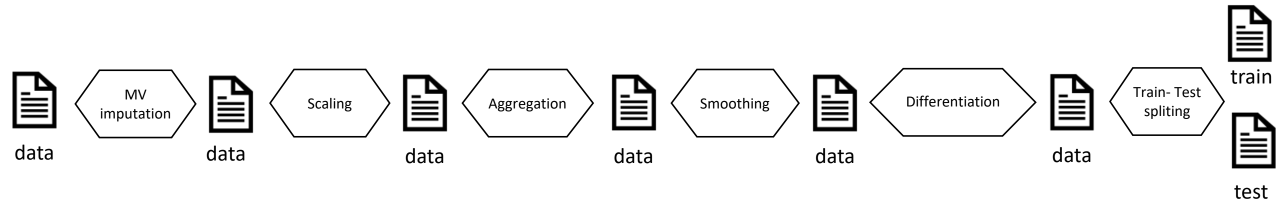

Figure 3 Data preparation methodology for the forecasting task

Now, after missing value imputation, a scaling transformation shall be applied, followed by the study of the best aggregation and smoothing operations. The last transformation to apply shall be the differentiation and any other transformation you think appropriate.

Aggregation shall consider two levels of aggregation, and differentiation shall be tested as the first and second derivatives.

Regarding data partition, remember that time series are temporal data, and so test data shall always be posterior to any train data. Note that aggregation and differentiation have to be applied to both train and test datasets since there is a change in the data space, but smoothing shall only be applied over the training data to help find a good model.

Remember that the Persistence model predicts the following value based on the last one known. So, we can consider two scenarios: the best – corresponding to the one-step horizon, and the rough one – when we use the last value of the training set to predict all the future values. In this manner, this model provides us two baselines for comparing all the other results.

The decision over which operation leads to better results shall be based on the Linear Regression model.

Modeling

The forecasting task has to explore the application of ***Simple Average**, **Persistence model**, **Rolling Mean**, **Exponential Smoothing**, **Linear Regression**, **ARIMA** and **LSTMs***, for training a single model for each domain. All but LSTMs and ARIMA only deal with univariate data, and students shall explore these last two in both situations.

Different parametrizations shall be applied over the same dataset. Again, the use of *autoML* tools is discouraged for the same reasons.

Evaluation

Like before, the obtained models should be evaluated as usual through confidence measures and evaluation charts, now in the forecasting context. A thorough comparison of the models' adequacy shall be presented, taking into consideration the adequacy of their behavior against the properties of each dataset and their observed performances.

For this purpose, the analysis of each forecasting technique shall be done at two different levels:

- The analysis of the impact of <u>the different parameters</u> on models' performance.
- The description of the <u>best model</u> found for each forecasting technique.

Critical Analysis

As before, the critical analysis shall <u>compare the different best models obtained</u>, explaining the achievements obtained through the different techniques.

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:

CLASSIFICATION	50%	**FORECASTING**	45%
Data profiling	5%	Data profiling	5%
Data preparation	10%	Data preparation	10%
Modeling and Evaluation		Modeling and Evaluation	
Naïve Bayes	1%	Simple Average and Persistence	1%
KNN	2%	Rolling mean	2%
Decision Trees	2%	Exponential smoothing	3%
Multi-layer perceptron	5%	Linear regression	3%
Random Forests	5%	ARIMA – one and multi var	7%
Gradient Boosting	5%	LSTMs – one and multi var	7%
Critical analysis	15%	Critical analysis	12%

Good Work ! ! !