TÉCNICO LISBOA

2025-2026

Data Science

Deadline – January 9th 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 either Portuguese or English, but it must adhere to the template, include all required charts, and not exceed the character limit 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 must be submitted through **Fenix** 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 hypotheses 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 unacceptable, and there is always something to 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 iterations. Data profiling, data preparation, modeling, and evaluation steps must 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 also to understand the impact of the available options on the performance of the produced models.

Students may choose the mining tool to apply from Python (using Sci-Kit Learn), R, or 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 in the Fenix section, Project**.

- **Security domain – Traffic Accidents**
  - classification **file** = traffic_accidents.csv **target** = CRASH_TYPE
  - description available on https://www.kaggle.com/datasets/oktayrdeki/traffic-accidents
- **Economy domain – Flight Status Prediction**
  - classification **file** = combined_flights_2022.csv **target** = CANCELLED
  - description available on https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022

Data Profiling

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

When working with symbolic variables, and since `sklearn` is unable to handle them correctly, students must choose a **new encoding for those variables** before proceeding with the correlation analysis.

Remember that data profiling is used to gain a deeper understanding of the data and primarily to identify the necessary transformations to apply to the original data in the next step. These transformations aim to enhance the performance of classification techniques used during the modeling phase.

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 (as shown in Figure 1) to minimize the number of datasets to be analyzed.

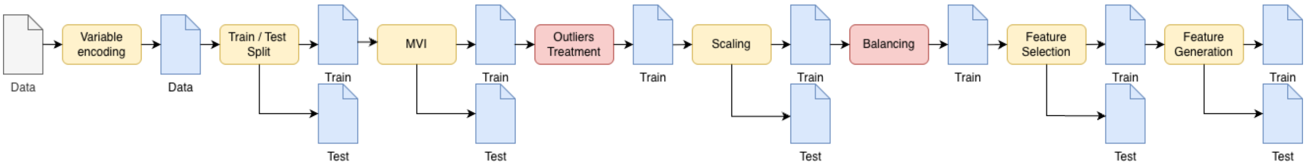

Figure 1 Data preparation methodology for the classification task

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 *dummy variables*. Different choices must be made for each variable; however, only one choice per variable shall be applied, without applying more than one alternative.

For the remaining preparation steps, students must apply at least two alternatives, evaluate the impact of each, and select the most promising option, as illustrated in Figure 2.

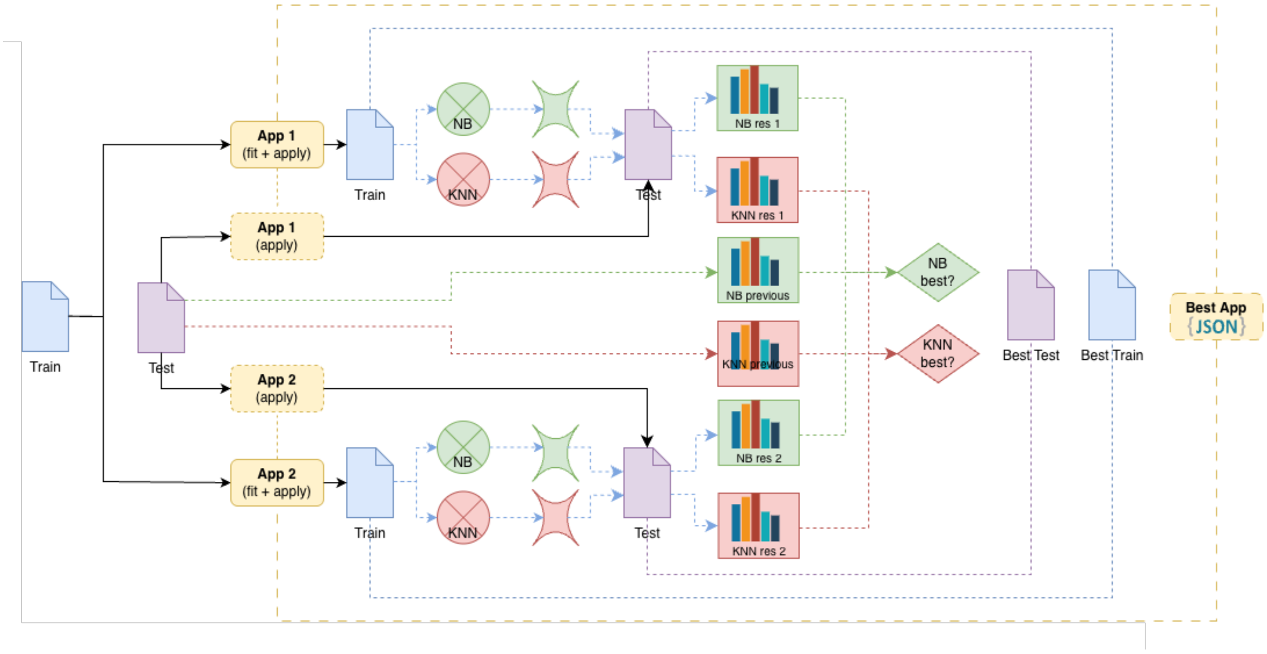

Figure 2 Decision process for each preparation step in the classification task

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 must apply at least two different preparation techniques to a single preparation task 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 distinct nature of both approaches limits the likelihood of selecting a technique best suited for a particular approach.

After training the different models, we selected the preparation technique that yields the best improvement compared to the previous dataset. In this manner, after the training, we may face 4 possibilities:

- None of the alternative preparation techniques applied improves 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 value imputation. **However, this fact must be mentioned in the report, and the decisions not to apply certain preparation tasks** must be justified.

Some additional remarks:

- It is not possible to train models on datasets with missing values using `sklearn`; therefore, the original dataset must be replaced by one of the prepared ones to proceed to the next step.
- Scaling impact shall be only assessed using KNN. Theoretically, it shouldn't change the results for Naïve Bayes.
- When temporal data is present, the data partition should use older data to train and newer data to test, rather than using future data to classify past data. In cases where there are multiple rows concerning the same entity, some entities should be used for training, while others should be used for testing. Otherwise, the partition shall be random.

- Feature selection may be applied before or after balancing. In either case, it can be studied using the same methods as the other preparation techniques, with only KNN and Naïve Bayes used to assess its results.
- Feature generation will be done in the variable encoding process. Additional variable generation is optional.
- Feature Extraction is outside the scope of this project, but students may optionally apply PCA, although it will not be scored.

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. Students must apply several machine learning methods and corresponding training algorithms, including Naïve Bayes, k-Nearest Neighbors (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, as they find the best parameters but do not provide any intermediate results, thereby preventing the impact analysis required.

The training data shall be the same for all yield training methods, corresponding to the result of the preparation step – the dataset that led to the best performance for 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 in relation to 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:

- Analysis of the impact of different parameters on the models' performance.

- The description of the best model found for each classification technique, along with 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. 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; however, it will primarily 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 can also be downloaded in the **Fenix section of the Project**.

- **Security domain – Traffic Prediction**
  - classification **file** = TrafficTwoMonth.csv **target** = TOTAL
  - description available on https://www.kaggle.com/datasets/hasibullahaman/traffic-prediction-dataset
- **Economy domain – Global Economic Indicators**
  - classification **file** = economic_indicators_dataset_2010_2023.csv **target** = INFLATION RATE (USA) https://www.kaggle.com/datasets/heidarmirhajisadati/global-economic-indicators-dataset-2010-2023

Data Profiling

In the forecasting context, profiling pays particular attention to the granularity analysis of the target variable and 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. Now, after imputing missing values, a scaling transformation will be applied, followed by an examination of the best aggregation and differentiation operations. The final transformation to be applied will be smoothing, along with any other transformations you deem appropriate.

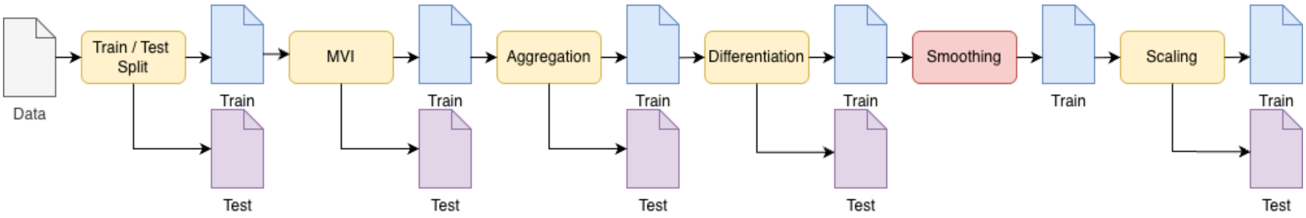

Figure 3 Data preparation methodology for the forecasting task

Aggregation will be considered at two levels, and differentiation will be tested using the first and second derivatives.

Regarding data partition, remember that time series are temporal data; therefore, test data should always follow any training data. Note that aggregation and differentiation must be applied to both the training and test datasets, as there is a change in the data space. However, smoothing should only be applied to the training data to help find an effective model.

The decision about which operation yields better results will be based on both the Persistence and Linear Regression models.

Modeling

The forecasting task involves exploring the use of ***Exponential Smoothing**, **Multi-Layer Perceptron**, **ARIMA,** and **LSTMs*** to train a single model for each domain. All except LSTMs and ARIMA only handle univariate data, and students should explore these last two in both scenarios.

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

Evaluation

As before, the obtained models should be evaluated using confidence measures and evaluation charts, now within the forecasting context. A thorough comparison of the models' adequacy will be presented, considering how well their behavior aligns with the properties of each dataset and their observed performances.

For this purpose, each forecasting technique will be analyzed at two different levels:

- The analysis of how different parameters affect models' performance.

- Description of the top model for each forecasting method.

Critical Analysis

As before, the critical analysis will compare the best models obtained, explaining the achievements achieved through different techniques.

C. Deployment

The guidelines for deploying the best classification models for one of the datasets will be published as soon as possible.

III. Evaluation Criteria

Below are the evaluation criteria broken down by topic. If students choose to deploy the classification models for one of the problems, forecasting will only be applied to the first dataset, which counts for 20% of the total project score.

Classification	**60%**	**Forecasting**	**40%**
Data profiling	5%	Data profiling	5%
Data preparation	10%	Data preparation	10%
Modeling and Evaluation		Modeling and Evaluation	
Naïve Bayes	3%	Exponential smoothing	2%
Logistic Regression	4%	Multi-layer perceptron	3
KNN	4%	ARIMA – one and multi var	5%
Decision Trees	4%	LSTMs – one and multi var	5%
Multi-layer perceptron	5%	Critical analysis	10%
Random Forests	5%	**Deployment**	**20%**
Gradient Boosting	5%	Preparation pipeline	7%
Critical analysis	15%	Each model	1%
		Web-based system	6%

Good Work!