TÉCNICO LISBOA

DS 2021

Data Science Project

Team nr:_____	**Student 1 :**_____	**IST nr:** _____
	Student 2 :_____	**IST nr:** _____
	Student 3 :_____	**IST nr:** _____

The present document presents a template for the Data Science Project report. It specifies the mandatory format and suggests the structure to follow. Intermediate headings (non-numbered ones) are not mandatory since they occupy to much space, but their usage is encouraged. The report **cannot exceeds 15 pages** (excluding the appendix).

All text with grey background shall be removed on the final report.

1 DATA PROFILING

May be used to describe any useful observation about the data, and that was used in the current project. An example is the use of any domain knowledge to process the data or evaluate the results.

Data Dimensionality

Shall contain all relevant information and charts respecting to the data dimensionality perspective, such as the number of records and number of dimensions, and their impact on the following analysis.

Data Granularity

Shall contain all relevant information and charts respecting to the data granularity perspective, such as the impact of different granularities considered for each variable. Redundant or non-fundamental charts can be appended in Appendix.

Data Distribution

Shall contain all relevant information and charts respecting to the data distribution perspective, such as each variable distribution, type, domain and range. Redundant or non-fundamental charts can be appended in Appendix.

Data Sparsity

Shall contain all relevant information and charts respecting to the data sparsity perspective, such as domain coverage and correlation among variables. Redundant or non-fundamental charts can be appended in Appendix.

2 DATA PREPARATION

May be used to summarize any useful observation derived from data profiling to better understand the choices followed ahead. Can contain the preparation of date variables, but not the prediction variable.

Missing Value Imputation

Shall contain all relevant information and charts respecting to missing values imputation, such as the choices made and the impact of the different approaches on modeling results. Shall also clearly reveal the approach selected to proceed with the processing.

If not applied explain the reason for that, based on data characteristics.

Dummification and other transformations

Shall contain all relevant information respecting to the transformation of variables, including *dummification*,. The list of variables under each one of the transformations shall be presented.

If not applied explain the reason for that, based on data characteristics.

Outliers Imputation

Shall contain all relevant information and charts respecting to outliers imputation, such as the choices made and the impact of the different approaches on modeling results. Shall also clearly reveal the approach selected to proceed with the processing.

If not applied explain the reason for that, based on data characteristics.

Scaling

Shall contain all relevant information and charts respecting to scaling transformation, such as the choices made and the impact of the different approaches on modeling results. Shall also clearly reveal the approach selected to proceed with the processing.

If not applied explain the reason for that, based on data characteristics.

Balancing

Shall contain all relevant information and charts respecting to balancing transformation, such as the choices made and the impact of the different approaches on modeling results. Shall also clearly reveal the approach selected to proceed with the processing.

If not applied explain the reason for that, based on data characteristics.

2.1 Feature Engineering

Feature Selection

Shall contain all relevant information and charts respecting to feature selection based on filtering out **redundant** variables. The different choices and their impact on the modeling results shall be presented and explained. Should also clearly reveal the approach selected to proceed with the processing.

All explanations shall be based on data characteristics.

Shall contain all relevant information and charts respecting to feature selection based on filtering out **irrelevant** variables. The different choices and their impact on the modeling results shall be presented and explained. Should also clearly reveal the approach selected to proceed with the processing.

All explanations shall be based on data characteristics.

Feature Extraction

Shall contain all relevant information and charts respecting to feature extraction, in particular PCA. The different choices and their impact on the modeling results shall be presented and explained.

Feature Generation

Shall contain all relevant information and charts respecting to feature generation. The different choices and their impact on the modeling results shall be presented and explained. Shall list all variables generated and the formula used to derive them.

3 CLASSIFICATION

Shall be used to summarize the preparation applied to the training dataset, used on the classification phase.

3.1 Naïve Bayes

Shall be used to present the results achieved with each one of Naïve Bayes implementations, comparing and proposing explanations for them. If any of the implementations is not used, a justification for it shall be presented.

Shall be used to present the evaluation of the best model achieved.

3.2 KNN

Shall be used to present the results achieved through different similarity measures and parametrizations of KNN. The results shall be compared and explanations for them shall be presented. The justification for the chosen similarity measures shall be presented.

Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it.

Shall be used to present the evaluation of the best model achieved.

3.3 Decision Trees

Shall be used to present the results achieved through different parametrizations for the train of decision trees. The results shall be compared and explanations for them shall be presented.

Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it.

Shall be used to present the evaluation of the best model achieved.

Shall be used to present the best tree achieved and its succinct description.

3.4 Random Forests

Shall be used to present the results achieved through different parametrizations for the train of random forests. The results shall be compared and explanations for them shall be presented.

Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it.

Shall be used to present the evaluation of the best model achieved.

May be used to present the most important variables in the model.

3.5 Gradient Boosting

Shall be used to present the results achieved through different parametrizations for the train of gradient boosting. The results shall be compared and explanations for them shall be presented.

Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it.

Shall be used to present the evaluation of the best model achieved.

May be used to present the most important variables in the model.

3.6 Multi-Layer Perceptrons

Shall be used to present the results achieved through different parametrizations for the train of MLPs. The results shall be compared and explanations for them shall be presented.

Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it. In particular by analysing the `loss_curve_` available at the end of each train.

Shall be used to present the evaluation of the best model achieved.

4 CLUSTERING

Shall be used to summarize the preparation applied to perform the clustering task.

Shall be used to present the results achieved through different algorithms. The results shall be compared and explanations for them shall be presented.

Shall be used to present the results achieved after applying PCA. The results shall be compared with the previous ones and explanations for them shall be presented

5 Association Rules

Shall be used to summarize the preparation applied to perform the discovery of association rules.

Shall be used to present the results achieved, providing explanations for them.

6 Time Series Analysis

Shall be used to summarize the preparation applied to the time series corresponding to the prediction variables.

Matrix Profile

Shall be used to present the results achieved through Matrix Profile, identifying the set of best motifs and anomalies.

Forecasting

Shall be used to present the results achieved through the different forecasting algorithms The results shall be compared and explanations for them shall be presented

7 Critical Analysis

Shall be used to present a summary of the results achieved with the different modeling techniques, and the impact of the different preparation tasks on their performance.

A cross-analysis of the different models may also be presented, identifying the most relevant variables common to all of them (when possible) and the relation among the patterns identified within the different classifiers.

A critical assessment of the best models shall be presented, clearly stating if the models seem to be good enough for the problem at hand.

APPENDIX (OPTIONAL)

To be used only for presenting less relevant charts for data profiling, and following the same order as before. No text will be considered, only the charts.