|  |  |  |  |
| --- | --- | --- | --- |
| Logo, company name  Description automatically generated | |  | |
| Data Science Engineering  Project Report | | | |
| **Team nr:** \_\_\_\_\_ | **Student 1 :** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | **IST nr:** \_\_\_\_\_\_\_ |
| **Student 2 :** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | **IST nr:** \_\_\_\_\_\_\_ |
| **Student 3 :** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | **IST nr:** \_\_\_\_\_\_\_ |

This document presents a template for the Data Science Project report. It specifies the mandatory format and suggests the structure to follow. All text with grey background shall be replaced with the analysis made over the datasets.

# Data Profiling

Dataset 1 Link:

Dataset 2 Link:

Dataset 3 Link:

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. **Shall not exceed 200 characters.**

## 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. **Shall not exceed 500 characters.**

Figure 1 Nr Records x Nr variables for dataset 1 (left) and dataset 2 (right)

Figure 2 Nr variables per type for dataset 1 (left) and dataset 2 (right)

Figure 3 Nr missing values for dataset 1 (left) and dataset 2 (right)

## Data Distribution

Shall contain all relevant information and charts respecting to the data distribution perspective, such as each variable distribution, type, domain and range. May be used to describe any useful observation about the data, and that was used in the current project. **Shall not exceed 500 characters**.

Figure 4 Global boxplots dataset 1 (left) and dataset 2 (right)

Figure 5 Single variable boxplots for dataset 1

Figure 6 Single variable boxplots s for dataset 2

Figure 7 Histograms for dataset 1 (with distributions is enough)

Figure 8 Histograms for dataset 2 (with distributions is enough)

Figure 9 Outliers study dataset 1

Figure 10 Outliers study for dataset 2

Figure 11 Class distribution for dataset 1

Figure 12 Class distribution for dataset 2

## 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. May present additional taxonomies if needed. **Shall not exceed 500 characters.**

Figure 13 Granularity analysis for dataset 1

Figure 14 Granularity analysis for dataset 2

## Data Sparsity

Shall contain all relevant information and charts respecting to the data sparsity perspective, such as domain coverage and correlation among variables. **Shall not exceed 500 characters.**

Figure 15 Sparsity analysis for dataset 1

Figure 16 Sparsity analysis for dataset 2

Figure 17 Correlation analysis for dataset 1

Figure 18 Correlation analysis for dataset 2

# Data Preparation

## Variables Encoding

Shall contain all relevant information respecting to the transformation of variables. 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. **Shall not exceed 500 characters for each dataset.**

## 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 modelling 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. **Shall not exceed 500 characters.**

Figure 19 Missing values imputation results with different approaches for dataset 1

Figure 20 Missing values imputation results with different approaches for dataset 2

## Outliers Treatment

Shall contain all relevant information and charts respecting to outliers’ imputation, such as the choices made and the impact of the different approaches on modelling 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. **Shall not exceed 500 characters**.

Figure 21 Outliers imputation results with different approaches for dataset 1

Figure 22 Outliers imputation results with different approaches for dataset 2

## 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 modelling 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. **Shall not exceed 200 characters.**

Figure 23 Scaling results with different approaches for dataset 1

Figure 24 Scaling results with different approaches for dataset 2

## 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 modelling 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. **Shall not exceed 500 characters**.

Figure 25 Balancing results with different approaches for dataset 1

Figure 26 Balancing results with different approaches for dataset 2

## Feature Selection

Shall contain all relevant information and charts respecting to feature selection based on filtering out **redundant** (based on correlation) and **relevant** (based on variation) variables. The different choices and their impact on the modelling 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 not exceed 500 characters.**

Figure 27 Feature selection of redundant variables results with different parameters for dataset 1

Figure 28 Feature selection of redundant variables results with different parameters for dataset 2

Figure 29 Feature selection of relevant variables results with different parameters for dataset 1 (variance study)

Figure 30 Feature selection of relevant variables results with different parameters for dataset 2 (variance study)

## Feature Generation

Shall contain all relevant information and charts respecting to feature generation. The different choices and their impact on the modelling results shall be presented and explained. Shall summarise all variables generated and the formula used to derive them (in a table). **Shall not exceed 200 characters.**

Figure 33 Feature generation results for dataset 1

Figure 34 Feature generation results for dataset 2

# Models’ Evaluation

Shall be used to point out any important decision taken during the training, including training strategy and evaluation measures used. **Shall not exceed 500 characters**

## 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.

**Shall not exceed 300 characters.**

Figure 35 Naïve Bayes alternatives comparison for dataset 1

Figure 36 Naïve Bayes alternative comparison for dataset 2

Figure 37 Naïve Bayes best model results for dataset 1 (left) and dataset 2 (right)

## KNN

Shall be used to present the results achieved through different similarity measures and KNN parameterisations. 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. **Shall not exceed 500 characters**

Figure 38 KNN different parameterisations comparison for dataset 1

Figure 39 KNN different parameterisations comparison for dataset 2

Figure 40 KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 41 KNN best model results for dataset 1 (left) and dataset 2 (right)

## Decision Trees

Shall be used to present the results achieved through different parameterisations 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. **Shall not exceed 500 characters**

Figure 42 Decision Trees different parameterisations comparison for dataset 1

Figure 43 Decision Trees different parameterisations comparison for dataset 2

Figure 44 Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 45 Decision trees best model results for dataset 1 (left) and dataset 2 (right)

Figure 46 Best tree for dataset 1

Figure 47 Best trees for dataset 2

## Random Forests

Shall be used to present the results achieved through different parameterisations 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. **Shall not exceed 500 characters**

Figure 48 Random Forests different parameterisations comparison for dataset 1

Figure 49 Random Forests different parameterisations comparison for dataset 2

Figure 50 Random Forests overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 51 Random Forests best model results for dataset 1 (left) and dataset 2 (right)

Figure 52 Random Forests variables importance for dataset 1 (left) and dataset 2 (right)

## Gradient Boosting

Shall be used to present the results achieved through different parameterisations 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. **Shall not exceed 500 characters**

Figure 53 Gradient boosting different parameterisations comparison for dataset 1

Figure 54 Gradient boosting different parameterisations comparison for dataset 2

Figure 55 Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 56 Gradient boosting best model results for dataset 1 (left) and dataset 2 (right)

Figure 57 Gradient boosting variables importance for dataset 1 (left) and dataset 2 (right)

## Multi-Layer Perceptrons

Shall be used to present the results achieved through different parameterisations 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. Shall be used to present the evaluation of the best model achieved. **Shall not exceed 500 characters**

Figure 58 MLP different parameterisations comparison for dataset 1

Figure 59 MLP different parameterisations comparison for dataset 2

Figure 60 Loss curves for dataset 1 (left) and dataset 2 (right)

Figure 60 MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 61 MLP best model results for dataset 1 (left) and dataset 2 (right)

# 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 shall 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 whether the models seem good enough for the problem at hand.

**Additional charts may be presented here. Shall not exceed 2000 characters.**