TÉCNICO LISBOA

DS 2023

-14em

1cm

Data Science Project

46em**Team nr:** Insert here	**Student 1:**	Gonçalo Gonçalves	**IST nr:** 99226
	Student 2:	José Cruz	**IST nr:** 99260
	Student 3:	Jorge Santos	**IST nr:** 99258
	Student 4:	Matilde Heitor	**IST nr:** 99284

hlcolorThe present 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. Put your charts in the *images* folder, and set the name of the file in the *includegraphics* command, after uncommenting it.

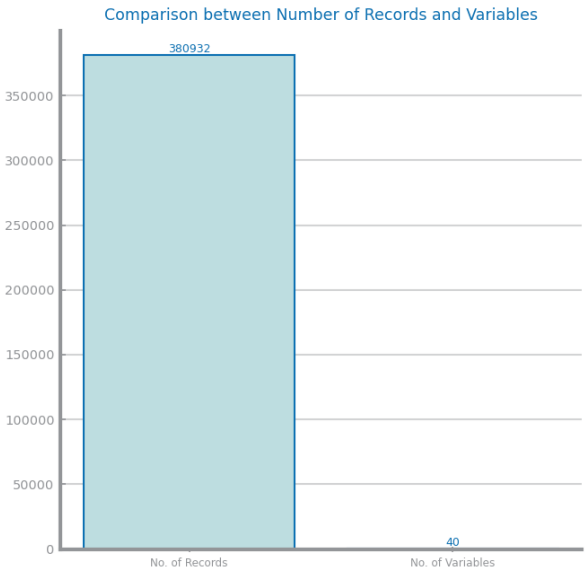
CLASSIFICATION

1 DATA PROFILING

hlcolorMay 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.**

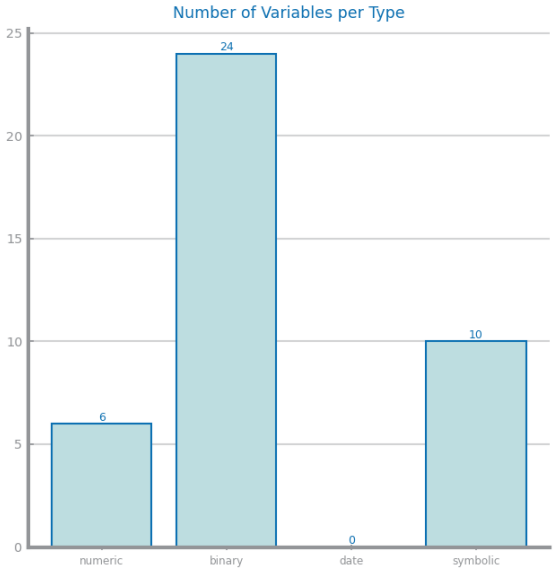
Data Dimensionality

hlcolorShall 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.**



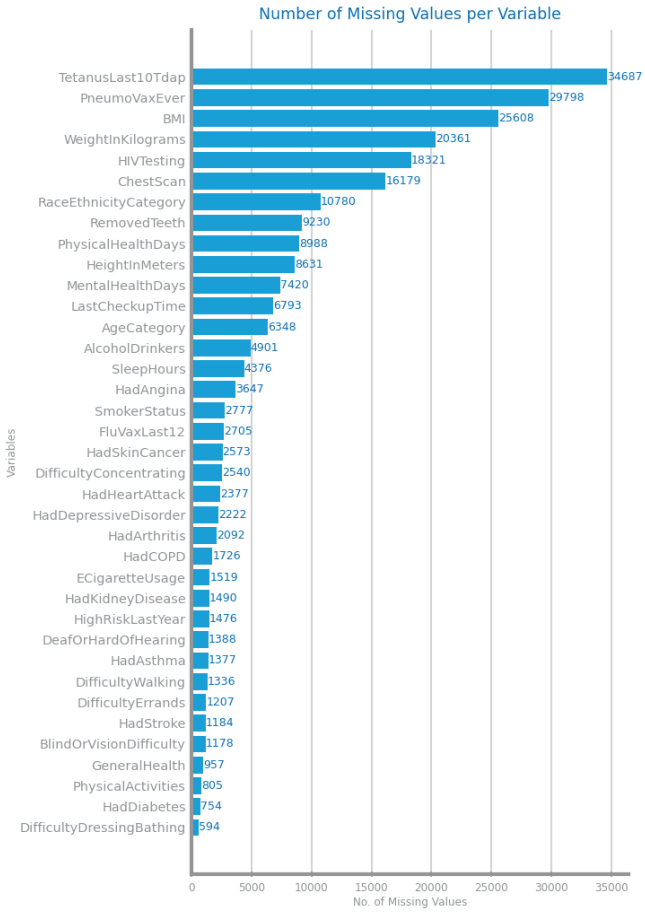
images/dataset2/data_profiling/Credit_Sc

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



images/dataset2/data_profiling/Credit_Sc

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

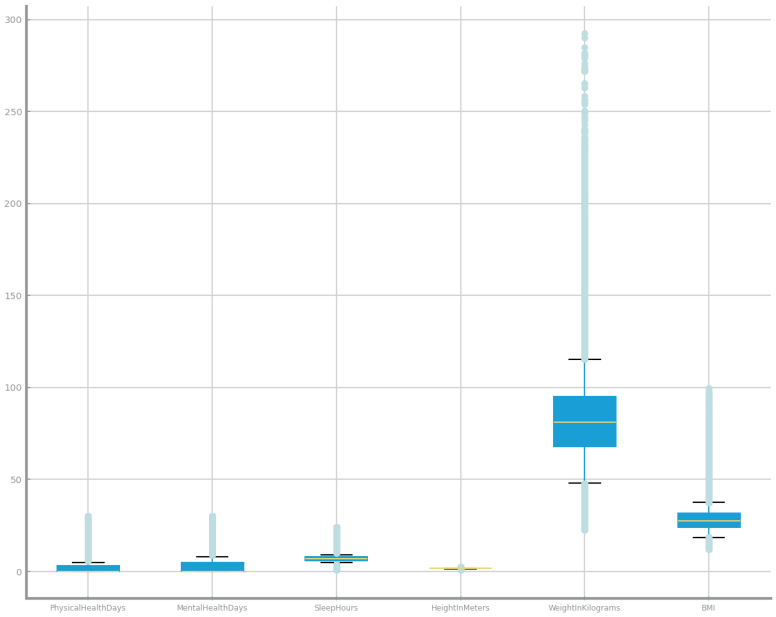


images/dataset2/data_profiling/Credit_Sc

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

Data Distribution

hlcolorShall 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.**



images/dataset2/data_profiling

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

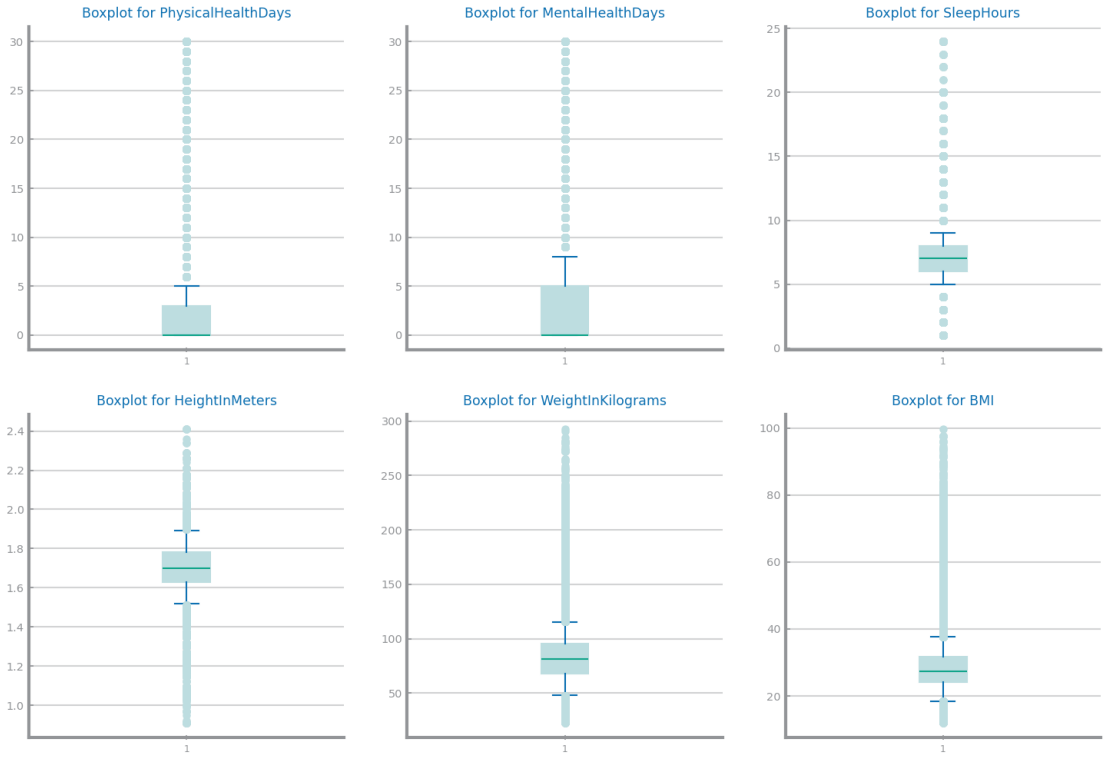


Figure 5: Single variables boxplots for dataset 1

images/dataset2/data_profiling/Credit_Score_single_boxplots.pn

Figure 6: Single variables boxplots for dataset 2

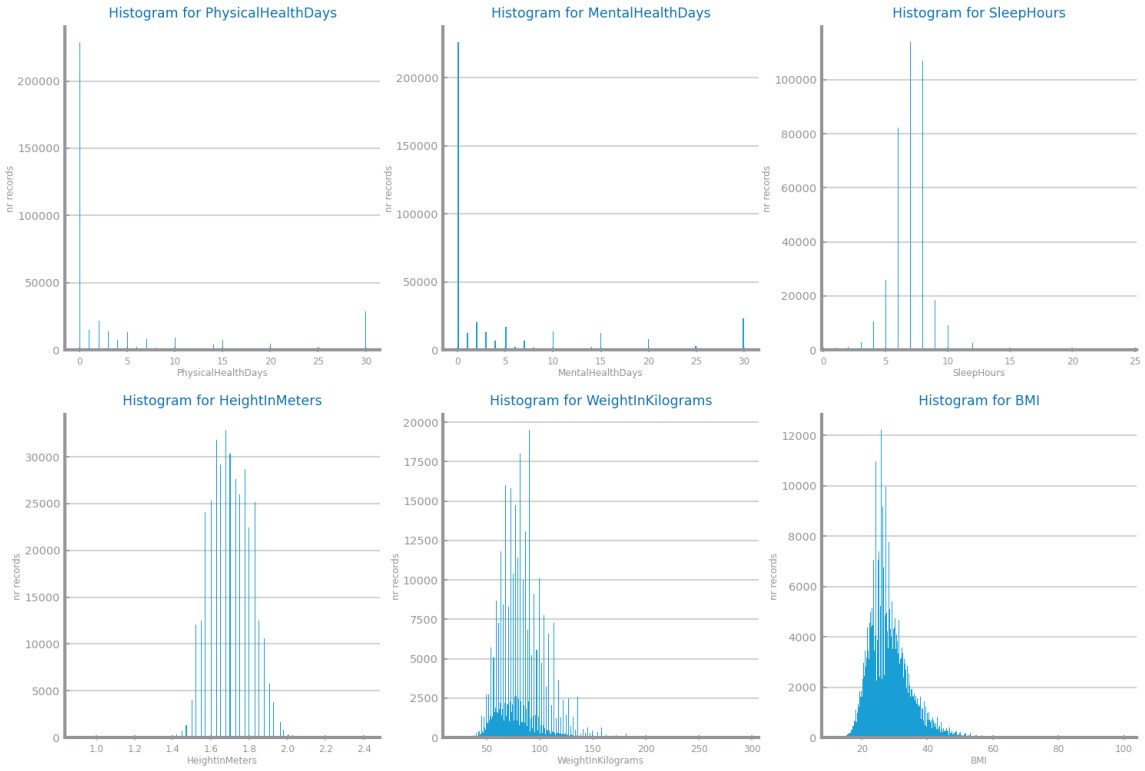


Figure 7: Histograms for dataset 1

images/dataset2/data_profiling/Credit_Score_single_histograms_

Figure 8: Histograms for dataset 2

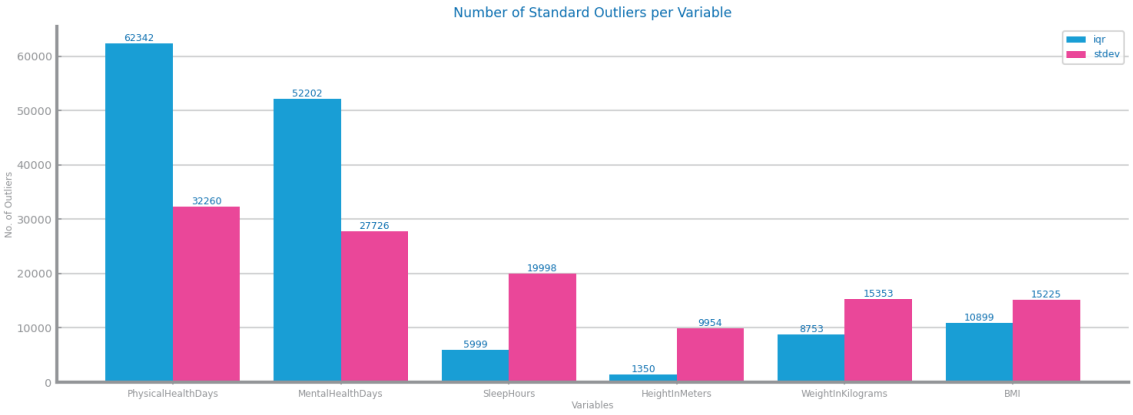


Figure 9: Outliers study dataset 1

Figure 10: Outliers study dataset 2

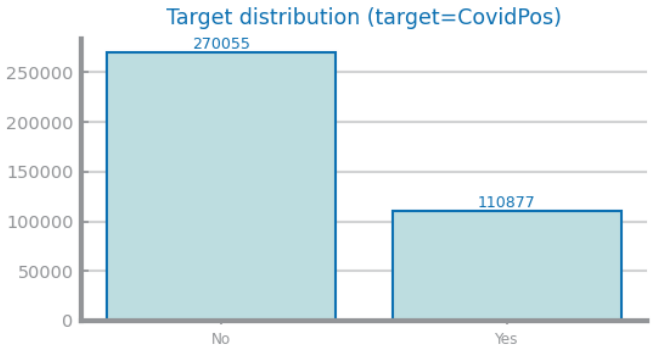


Figure 11: Class distribution for dataset 1

Figure 12: Class distribution for dataset 2

Data Granularity

hlcolorShall 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

hlcolorShall 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

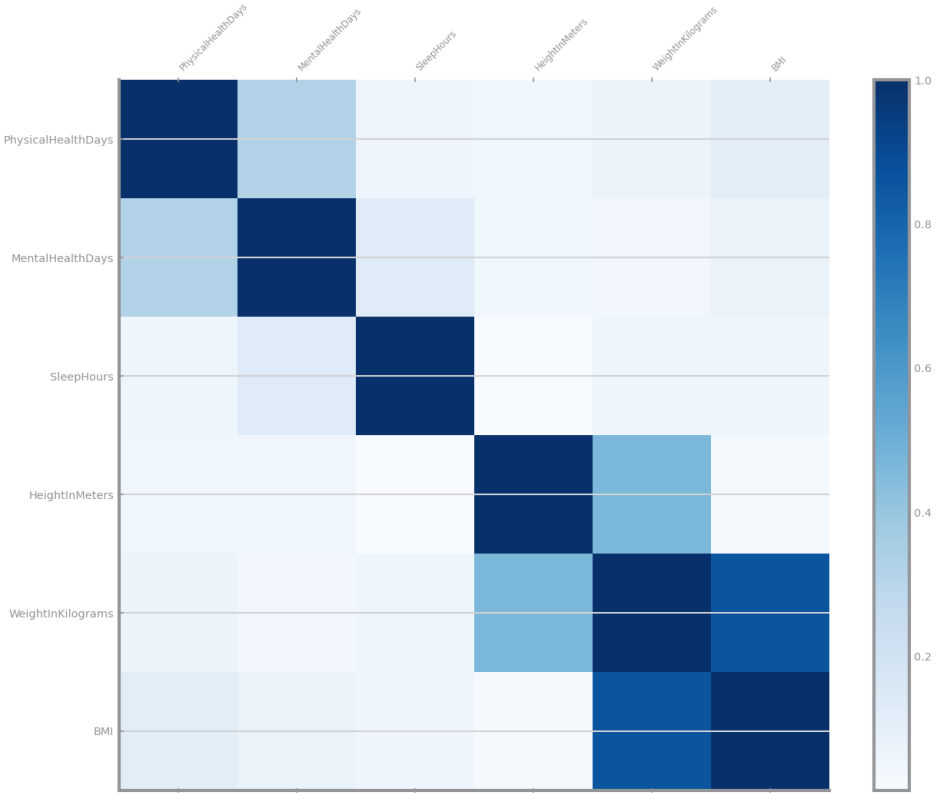


Figure 17: Correlation analysis for dataset 1

Figure 18: Correlation analysis for dataset 2

2 DATA PREPARATION

Variables Encoding

hlcolorShall 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

hlcolorShall 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.**

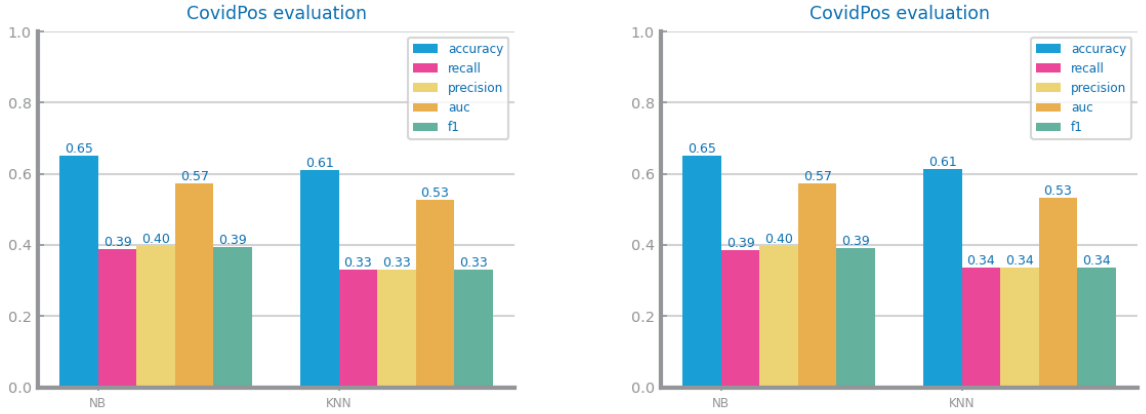


Figure 19: Missing values imputation results with different approaches for dataset 1

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

Outliers Treatment

hlcolorShall 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.**

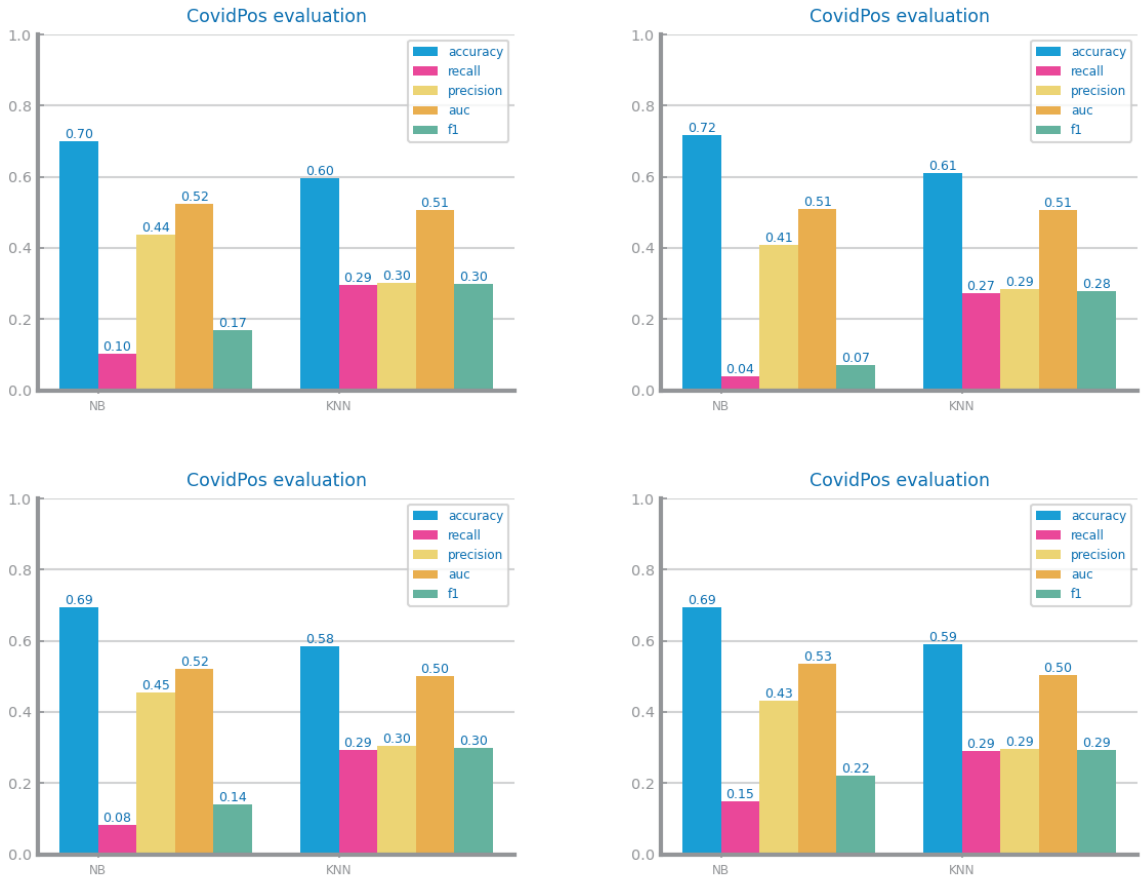


Figure 21: Outliers imputation results with different approaches for dataset 1

Figure 22: Outliers imputation results with different approaches for dataset 2

Scaling

hlcolorShall 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.**

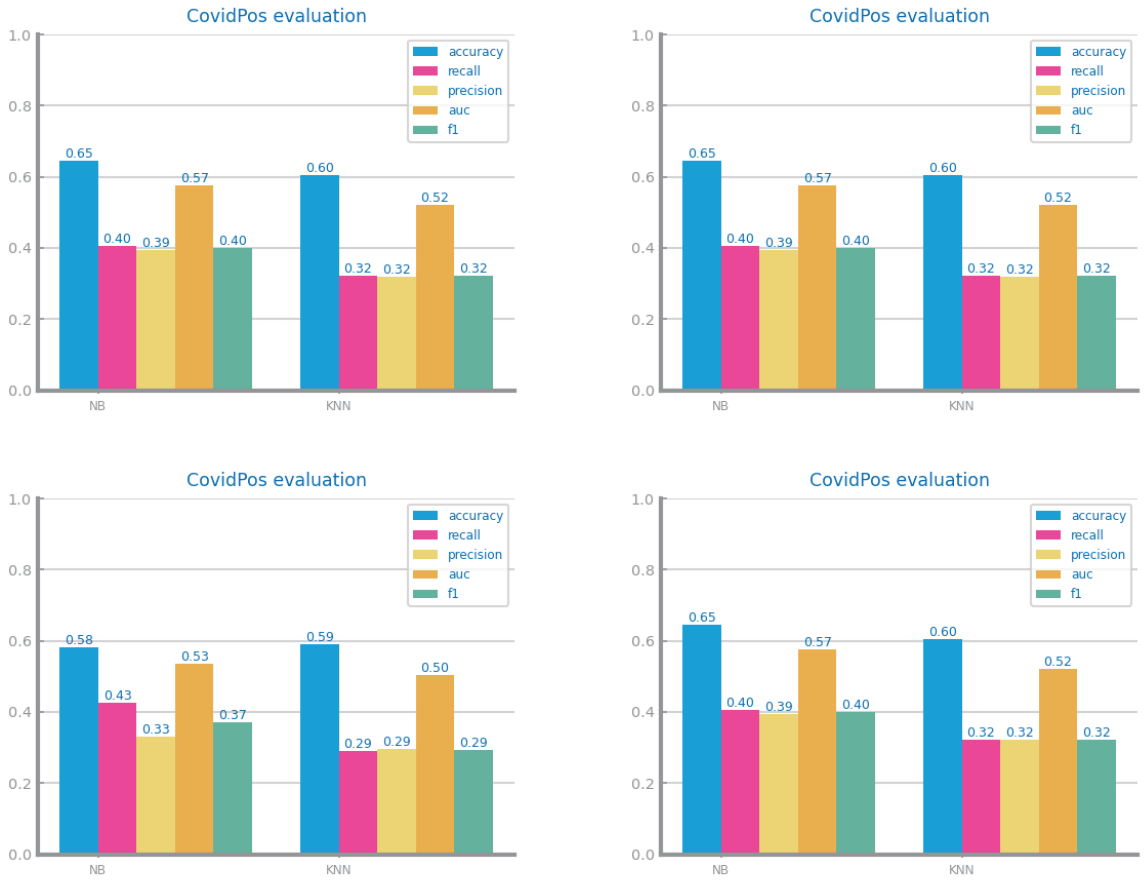


Figure 23: Scaling results with different approaches for dataset 1

Figure 24: Scaling results with different approaches for dataset 2

Balancing

hlcolorShall 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.**

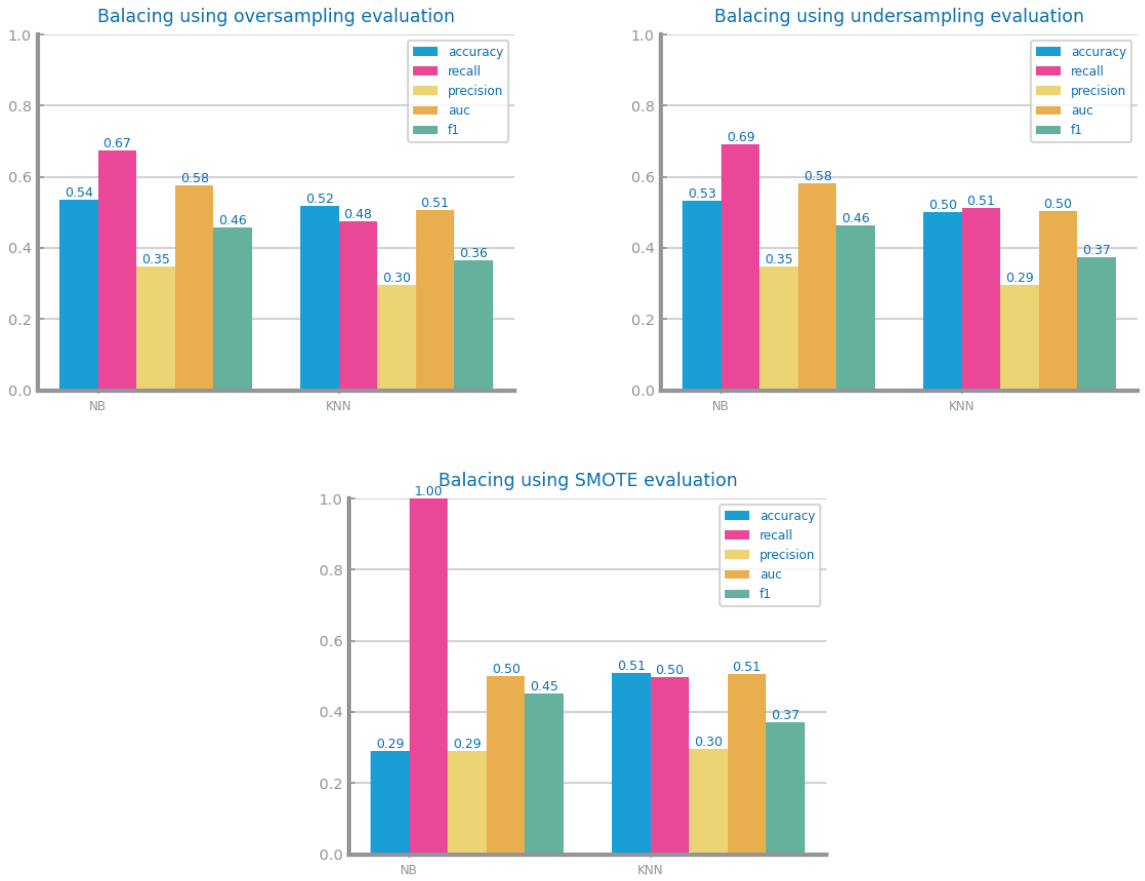


Figure 25: Balancing results with different approaches for dataset 1

Figure 26: Balancing results with different approaches for dataset 2

Feature Selection

hlcolorShall 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.**

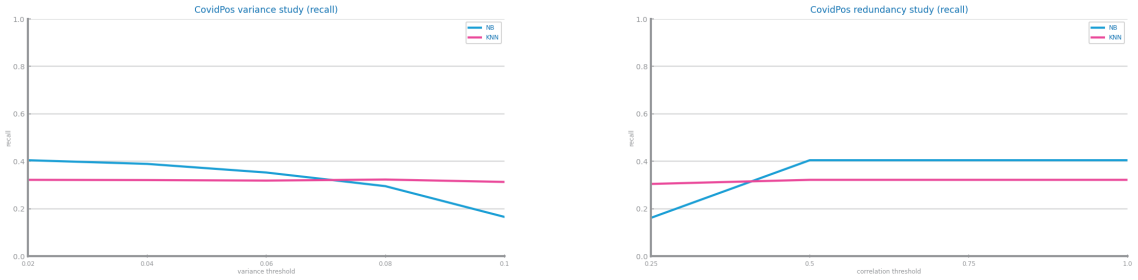


Figure 27: Feature selection of redundant variables results with different parameters for dataset 1

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 Extraction (optional)

hlcolorShall contain all relevant information and charts respecting to feature extraction, in particular PCA. The different choices and their impact on the modelling results shall be presented and explained. **Shall not exceed 200 characters.**

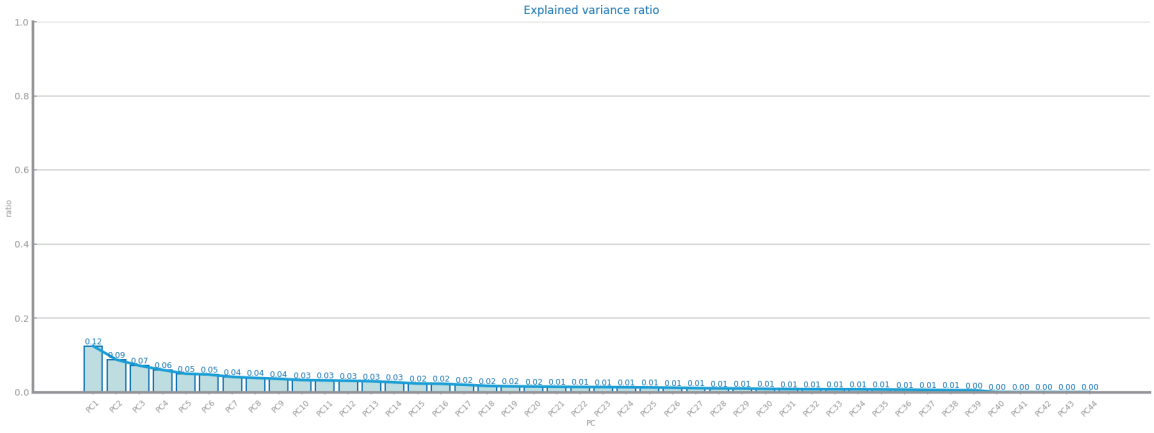


Figure 31: Principal components analysis and feature extraction results for dataset 1

Figure 32: Principal components analysis and feature extraction results for dataset 2

Additional Feature Generation (optional)

hlcolorShall 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

3 MODELS' EVALUATION

hlcolorShall 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

hlcolorShall 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.**

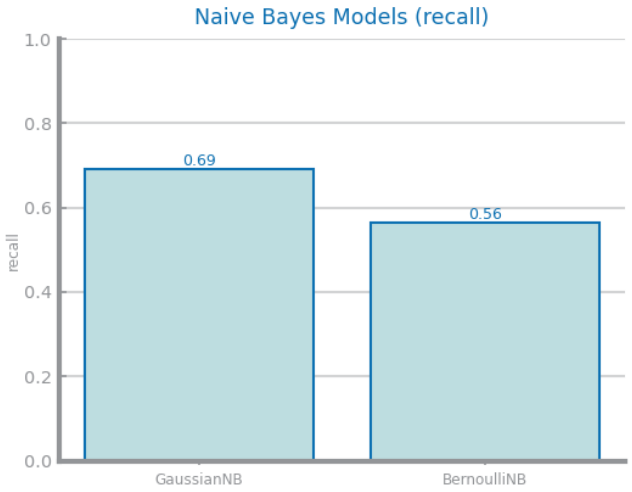


Figure 35: Naïve Bayes alternatives comparison for dataset 1

Figure 36: Naïve Bayes alternative comparison for dataset 2

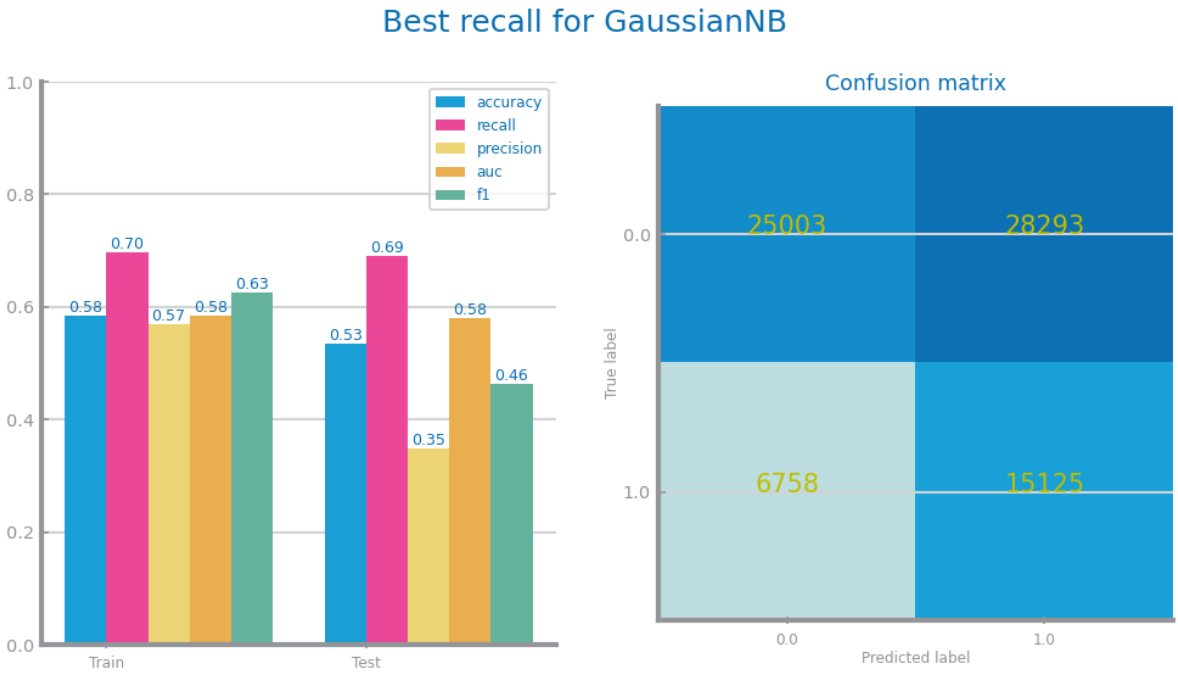


Figure 37: Naïve Bayes best model results for dataset 1 (left) and dataset 2 (right)

KNN

hlcolorShall 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.**

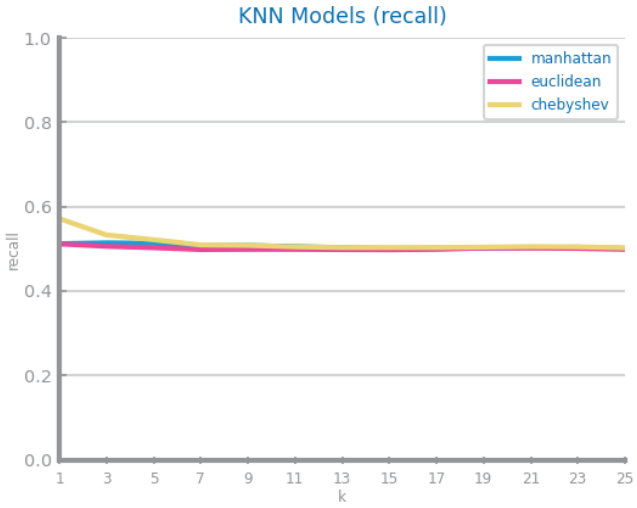


Figure 38: KNN different parameterisations comparison for dataset 1

Figure 39: KNN different parameterisations comparison for dataset 2

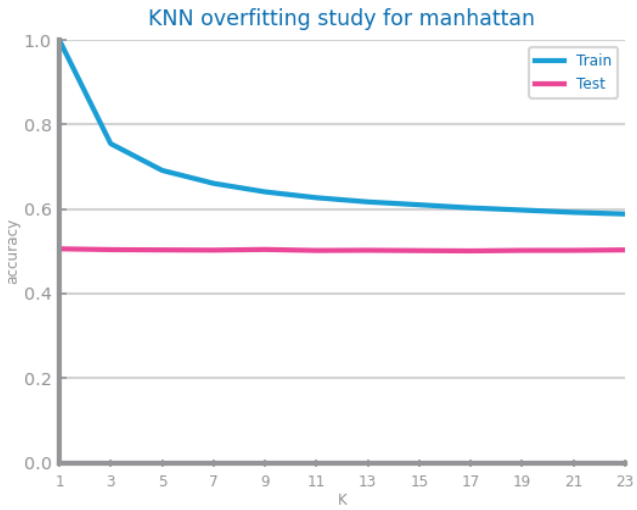


Figure 40: KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

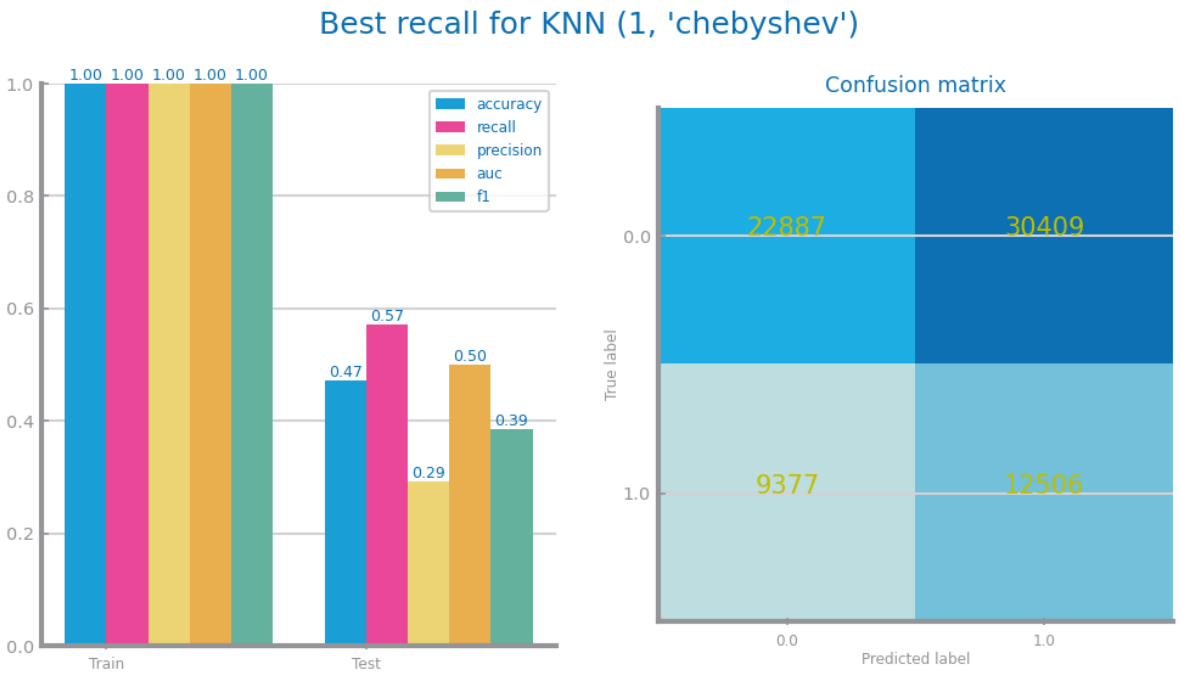


Figure 41: KNN best model results for dataset 1 (left) and dataset 2 (right)

Decision Trees

hlcolorShall 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.**

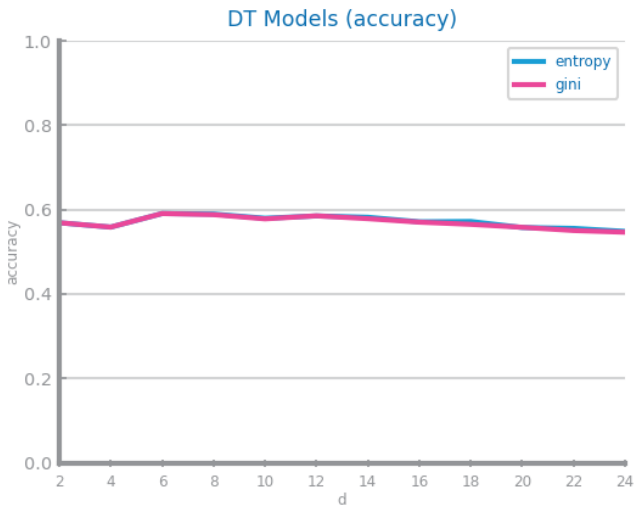


Figure 42: Decision Trees different parameterisations comparison for dataset 1

Figure 43: Decision Trees different parameterisations comparison for dataset 2

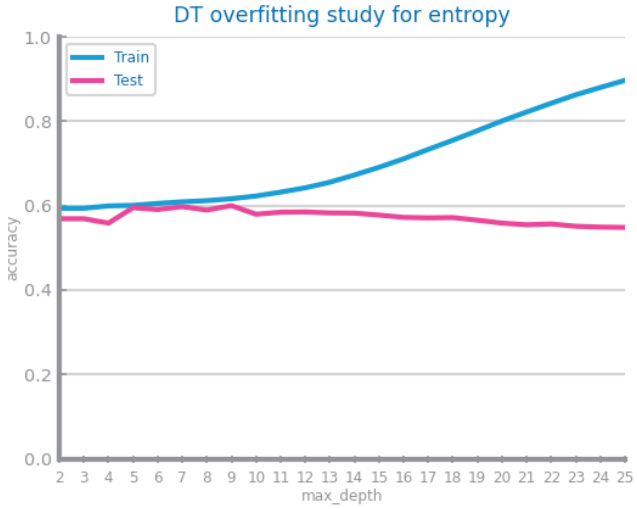


Figure 44: Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

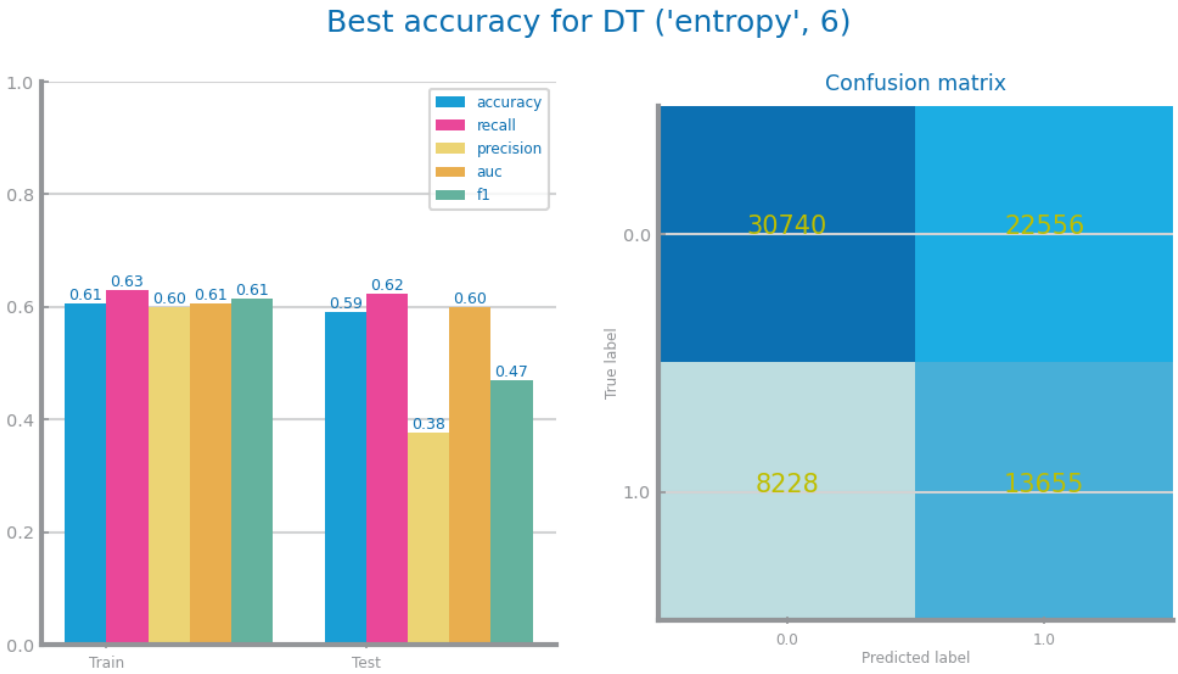


Figure 45: Decision trees best model results for dataset 1 (left) and dataset 2 (right)

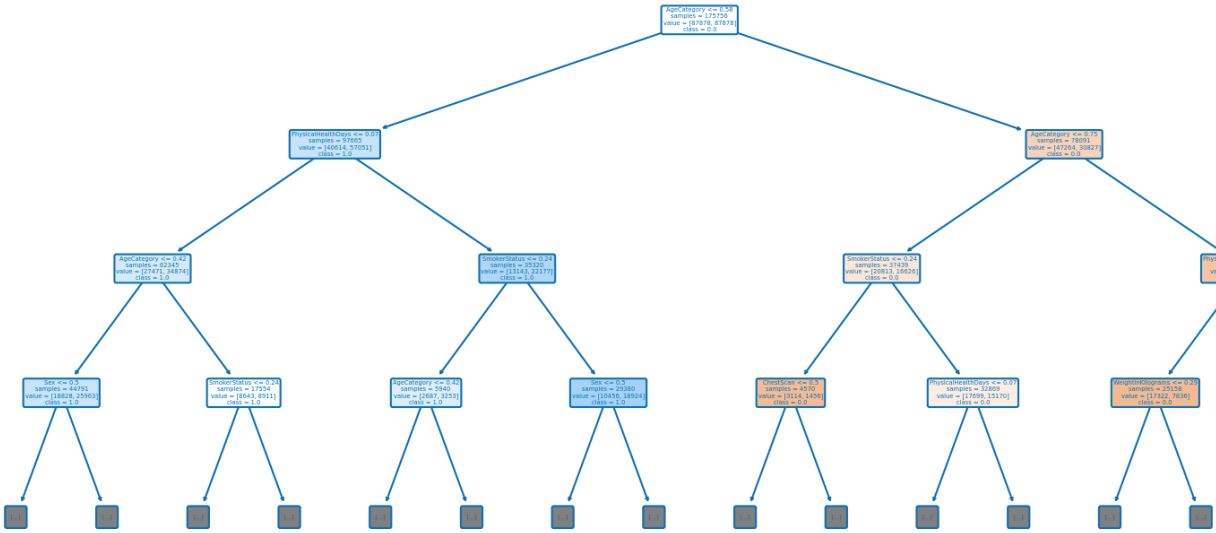
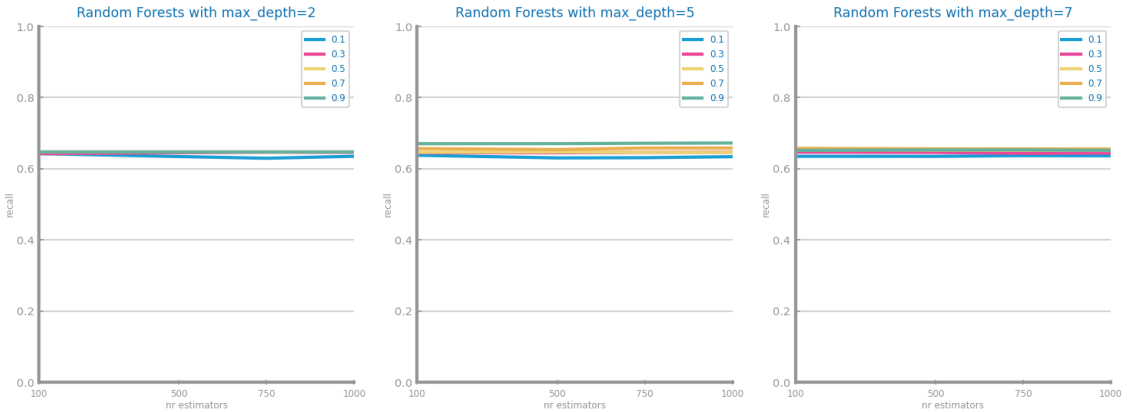


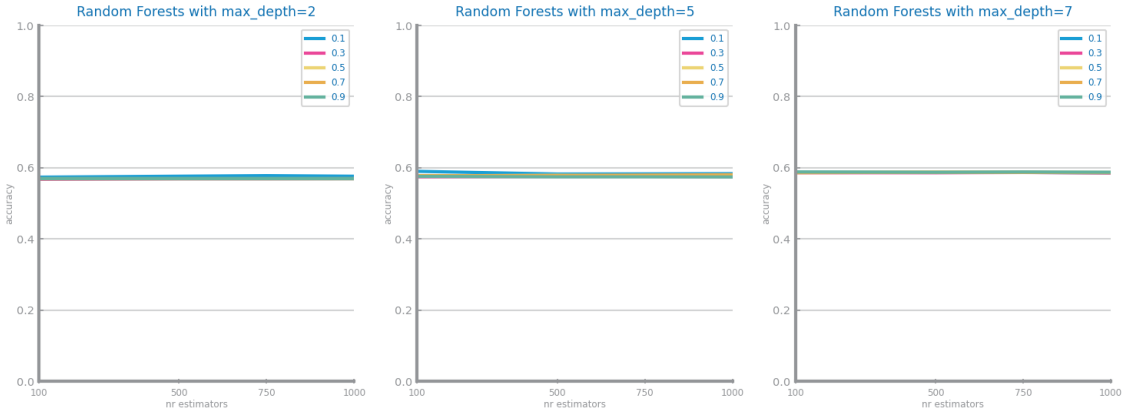
Figure 46: Best tree for dataset 1

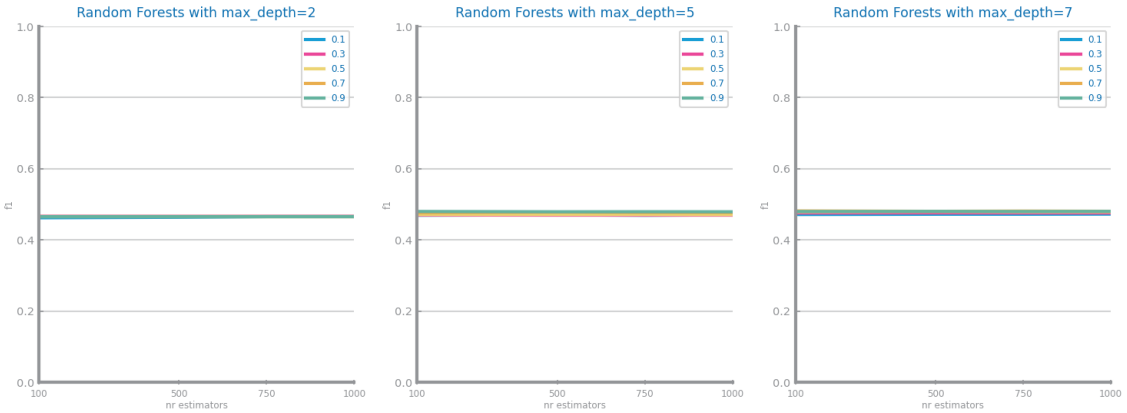
Figure 47: Best tree for dataset 2

Random Forests

hlcolorShall 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.**


Random Forests with max_depth=2
Random Forests with max_depth=5
Random Forests with max_depth=7
recall
nr estimators
0.1
0.3
0.5
0.7
0.9


Random Forests with max_depth=2
Random Forests with max_depth=5
Random Forests with max_depth=7
accuracy
nr estimators
0.1
0.3
0.5
0.7
0.9


Random Forests with max_depth=2
Random Forests with max_depth=5
Random Forests with max_depth=7
f1
nr estimators
0.1
0.3
0.5
0.7
0.9

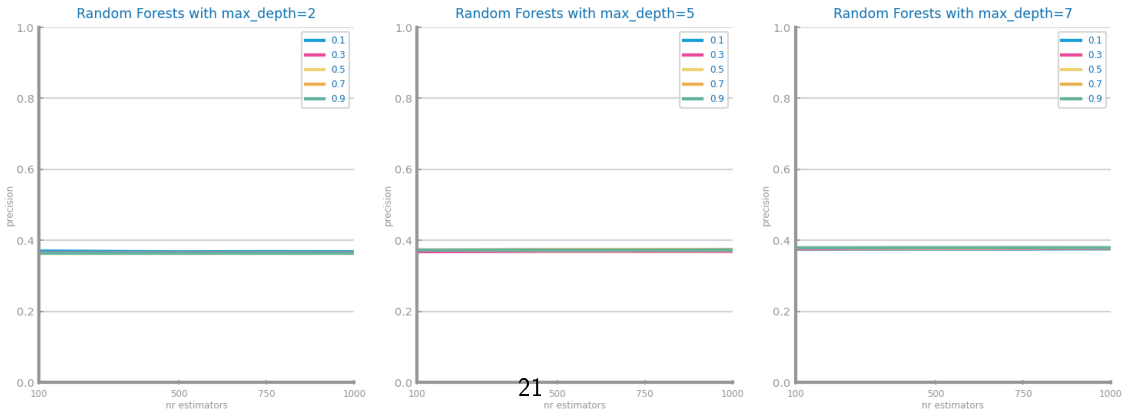

Random Forests with max_depth=2
Random Forests with max_depth=5
Random Forests with max_depth=7
precision
nr estimators
0.1
0.3
0.5
0.7
0.9

Figure 49: Random Forests different parameterisations comparison for dataset 2

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

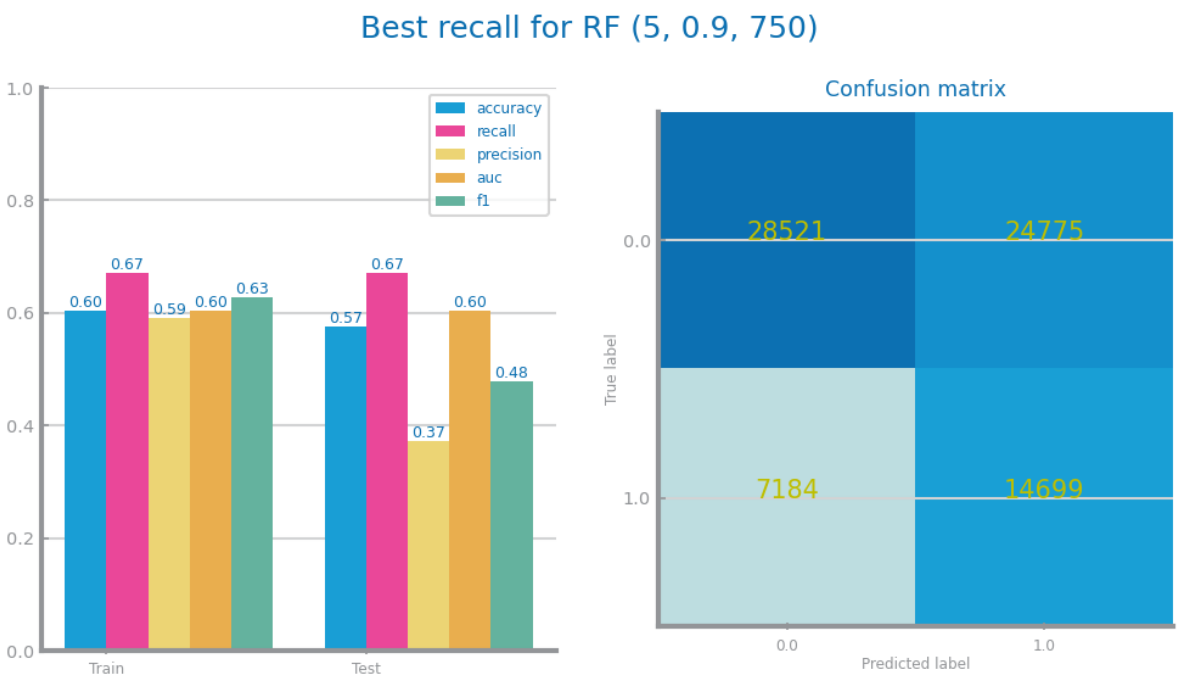


Figure 51: Random Forests best model results for dataset 1 (left) and dataset 2 (right)

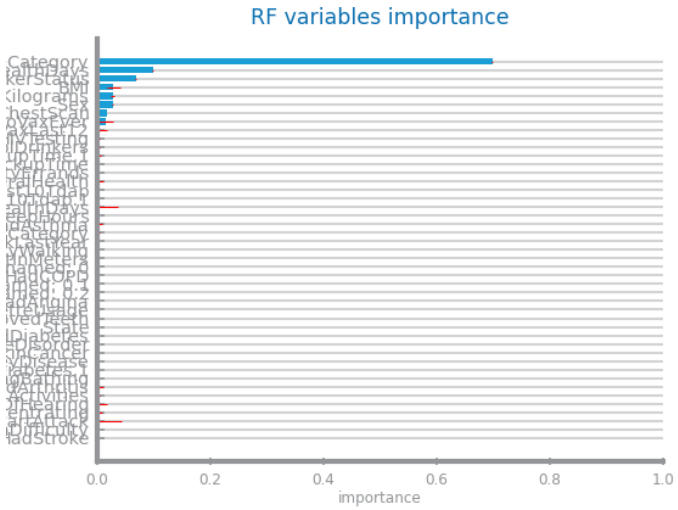
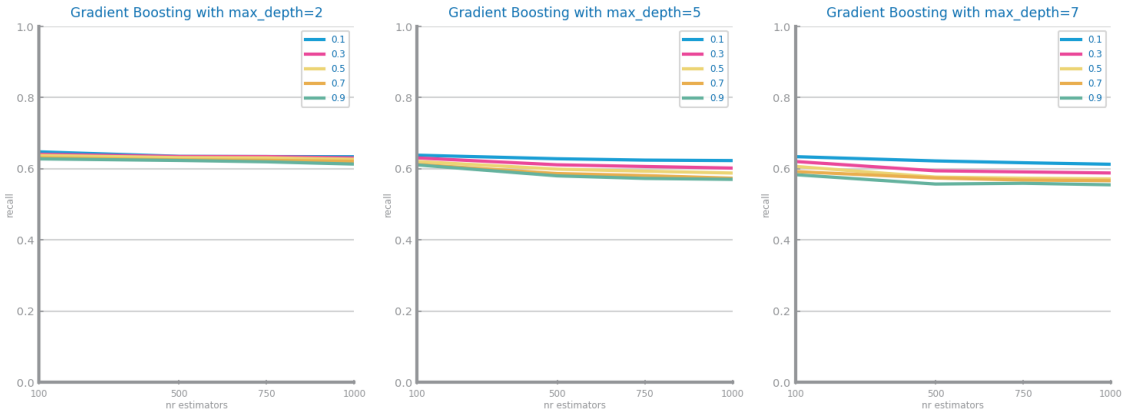
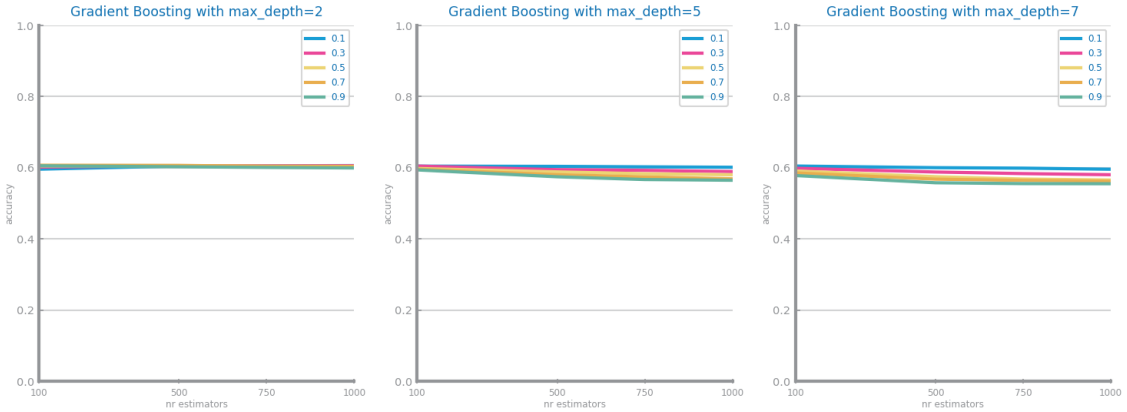


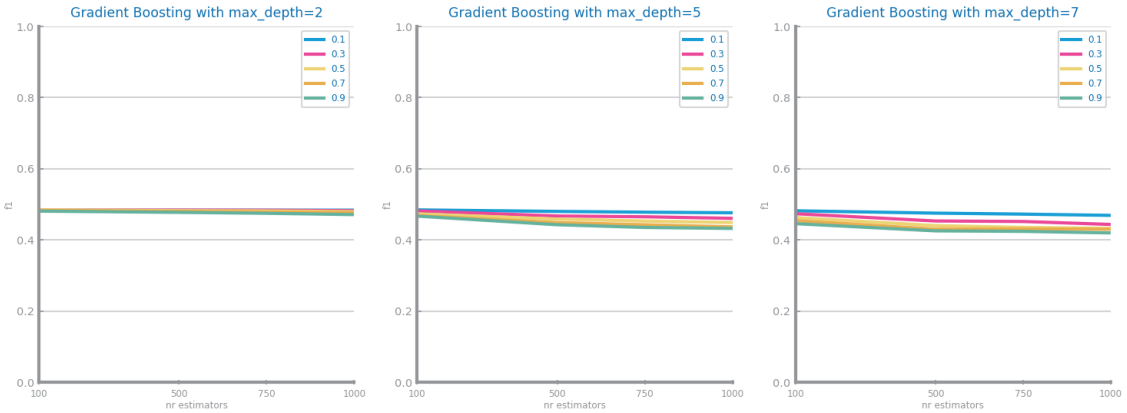
Figure 52: Random Forests variables importance for dataset 1 (left) and dataset 2 (right)

Gradient Boosting

hlcolorShall 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.**


Gradient Boosting with max_depth=2
Gradient Boosting with max_depth=5
Gradient Boosting with max_depth=7
recall
nr estimators
0.1
0.3
0.5
0.7
0.9


Gradient Boosting with max_depth=2
Gradient Boosting with max_depth=5
Gradient Boosting with max_depth=7
accuracy
nr estimators
0.1
0.3
0.5
0.7
0.9


Gradient Boosting with max_depth=2
Gradient Boosting with max_depth=5
Gradient Boosting with max_depth=7
f1
nr estimators
0.1
0.3
0.5
0.7
0.9

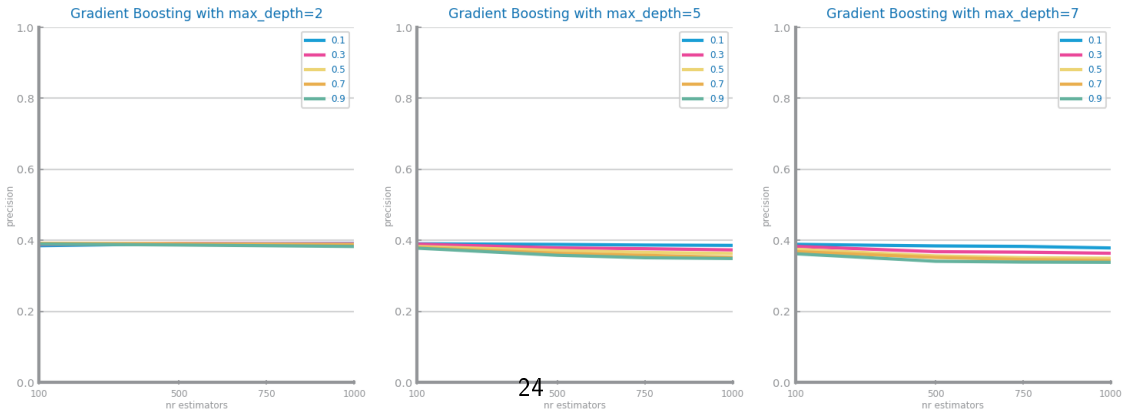

Gradient Boosting with max_depth=2
Gradient Boosting with max_depth=5
Gradient Boosting with max_depth=7
precision
nr estimators
0.1
0.3
0.5
0.7
0.9

Figure 54: Gradient boosting different parameterisations comparison for dataset 2

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

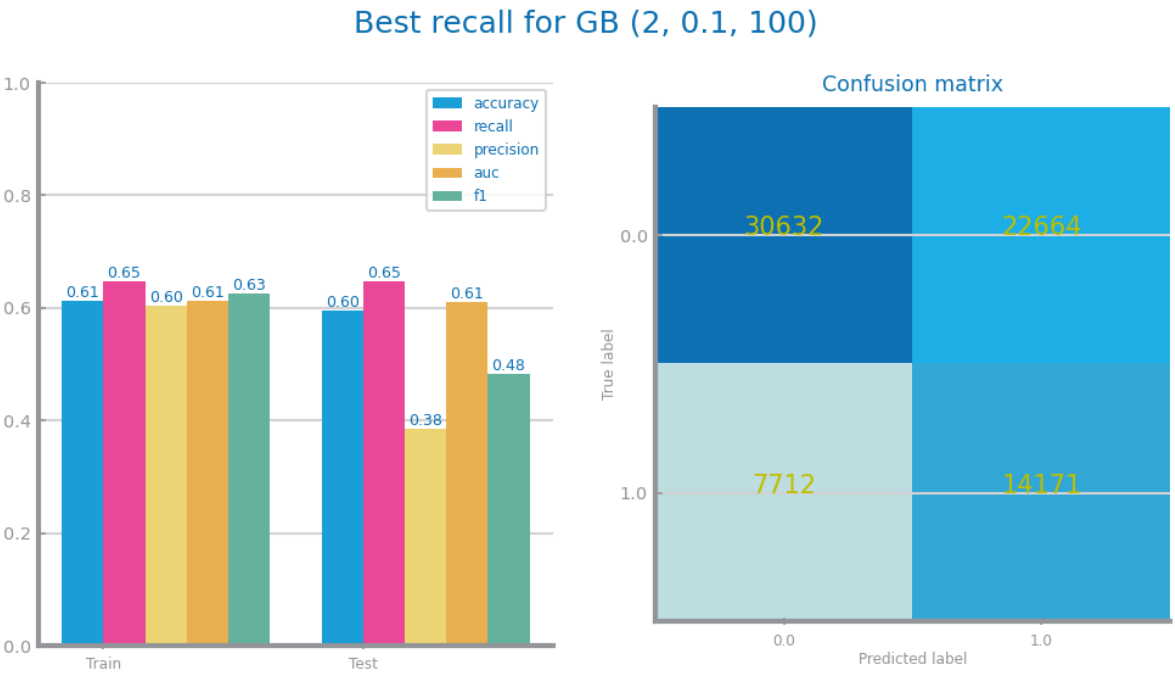


Figure 56: Gradient boosting best model results for dataset 1 (left) and dataset 2 (right)

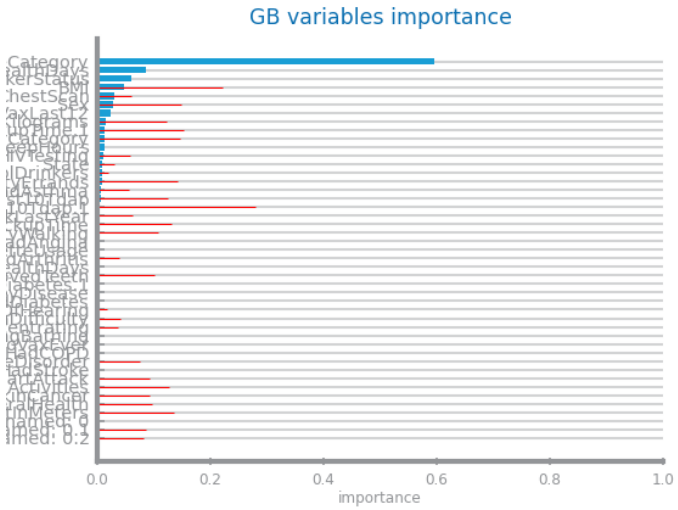


Figure 57: Gradient boosting variables importance for dataset 1 (left) and dataset 2 (right)

Multi-Layer Perceptrons

hlcolorShall 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.**

MLP with constant
MLP with invscaling
MLP with adaptive
recall
accuracy
f1
precision
nr iterations
0.5
0.05
0.005
0.0005

Figure 59: MLP different parameterisations comparison for dataset 2

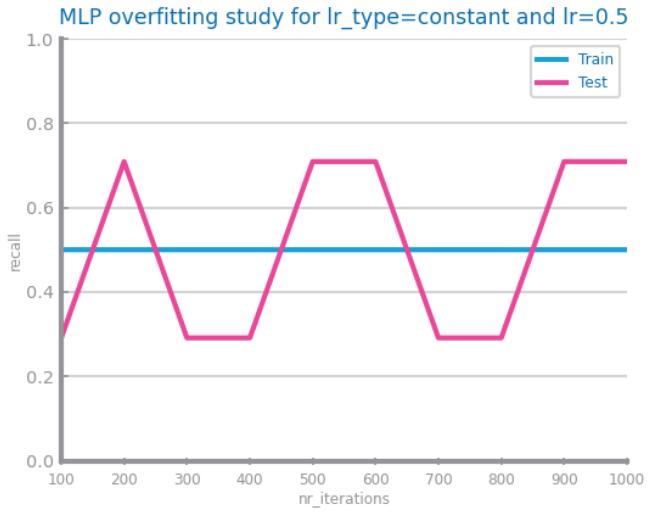


Figure 60: MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

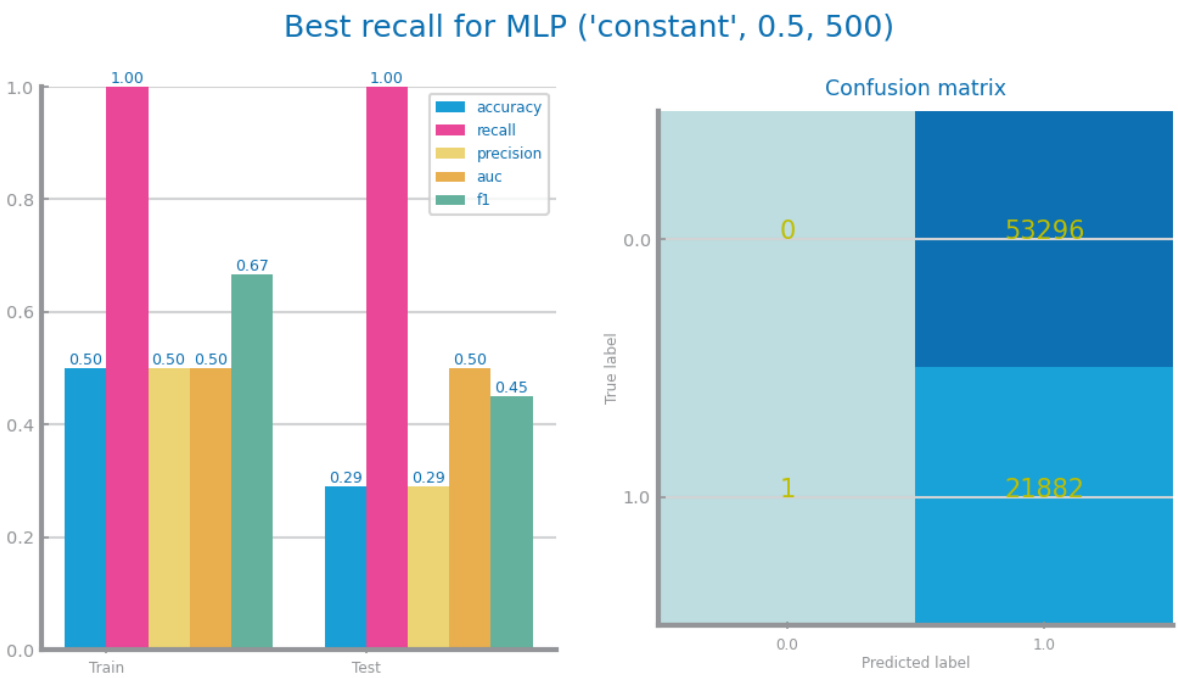


Figure 61: MLP best model results for dataset 1 (left) and dataset 2 (right)

4 CRITICAL ANALYSIS

hlcolorShall be used to present a summary of the results achieved with the different modelling 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. **Additional charts may be presented here. Shall not exceed 2000 characters.**

TIME SERIES ANALYSIS

5 DATA PROFILING

Data Dimensionality and Granularity

hlcolorMay be used to identify the most atomic granularity and two other different granularities to consider. **Shall not exceed 500 characters.**

Figure 62: Time series 1 at the most granular detail

Figure 63: Time series 1 at the second chosen granularity

Figure 64: Time series 1 at the third chosen granularity

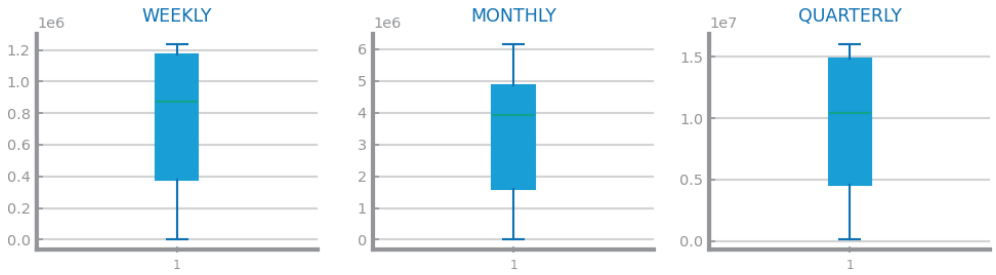
Figure 65: Time series 2 at the most granular detail

Figure 66: Time series 2 at the second chosen granularity

Figure 67: Time series 2 at the third chosen granularity

Data Distribution

hlcolorShall be used to perform the data analysis at those three different granularities, concerning the series distribution. **Shall not exceed 500 characters.**



	WEEKLY	MONTHLY	QUARTERLY
count	1.990000e+02	1.990000e+02	4.600000e+01
mean	7.743876e+05	7.743876e+05	3.350068e+06
std	4.356312e+05	4.356312e+05	1.937462e+06
min	0.000000e+00	0.000000e+00	1.000000e+00
25%	3.775975e+05	3.775975e+05	1.602803e+06
50%	8.691990e+05	8.691990e+05	3.919796e+06
75%	1.172049e+06	1.172049e+06	4.874868e+06
max	1.238650e+06	1.238650e+06	6.172614e+06

Name: deaths, dtype: float64

Figure 68: Boxplot(s) for time series 1

Figure 69: Boxplot(s) for time series 2

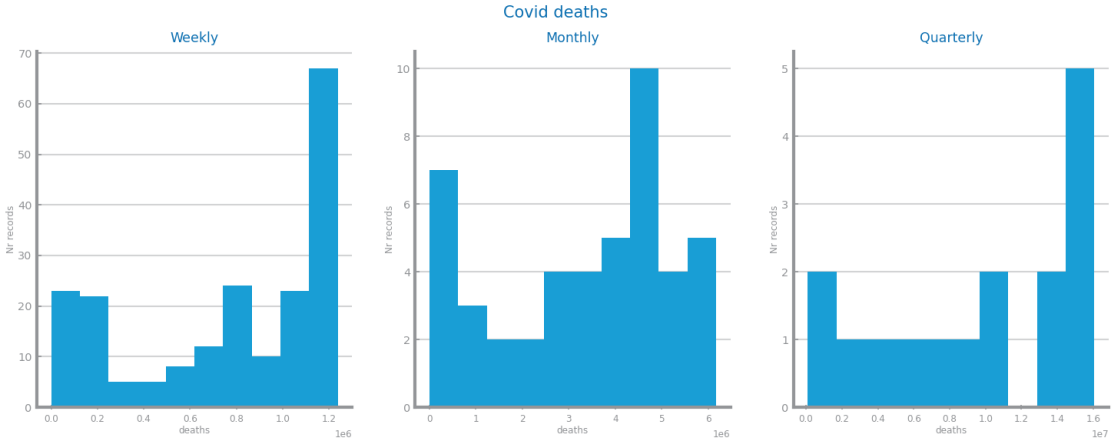


Figure 70: Histogram(s) for time series 1

Figure 71: Histogram(s) for time series 2

Data Stationarity

hlcolorShall be used to perform the data analysis at those three different granularities, concerning the series stationarity. **Shall not exceed 300 characters.**

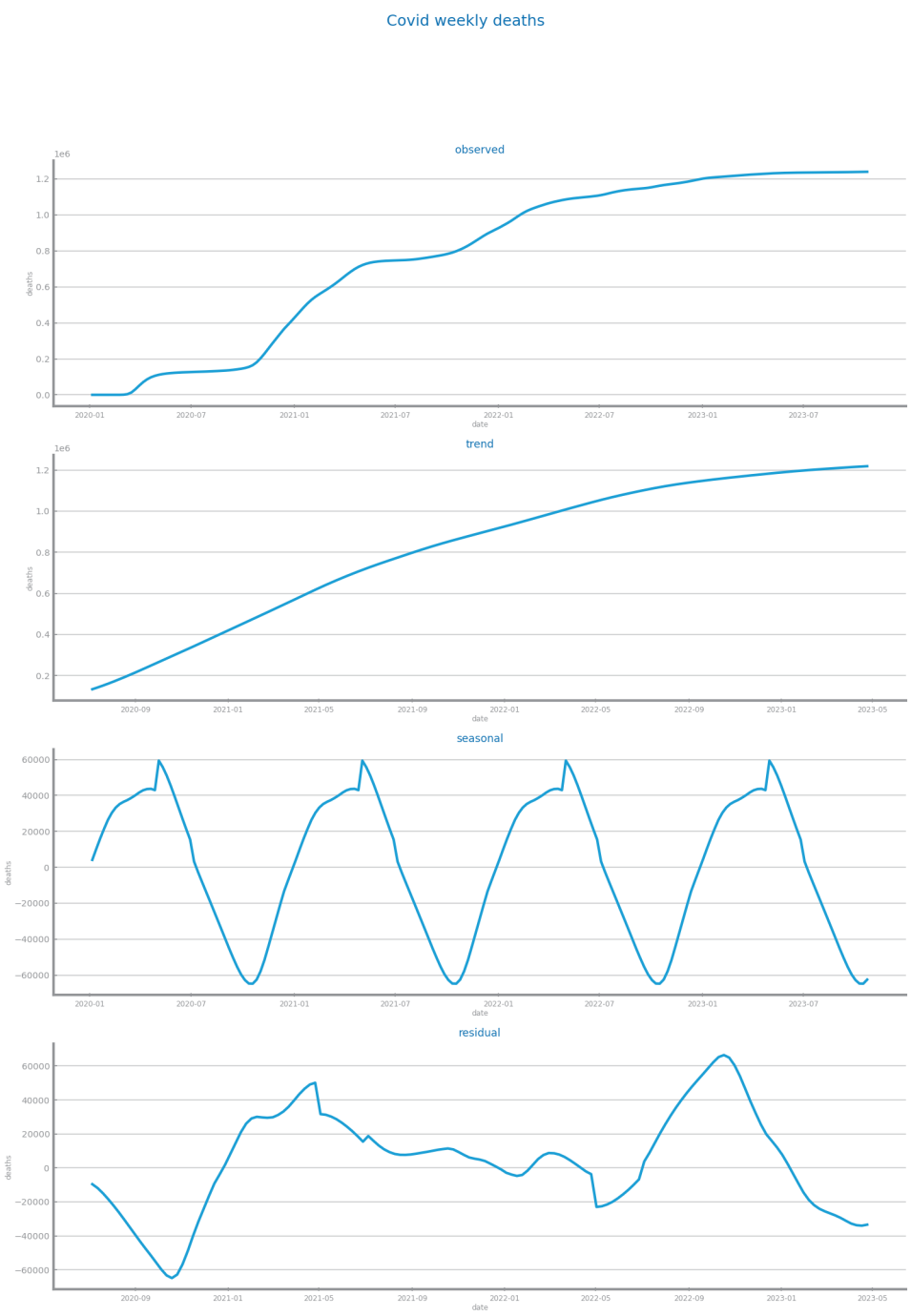


Figure 72: Components study for time series 1

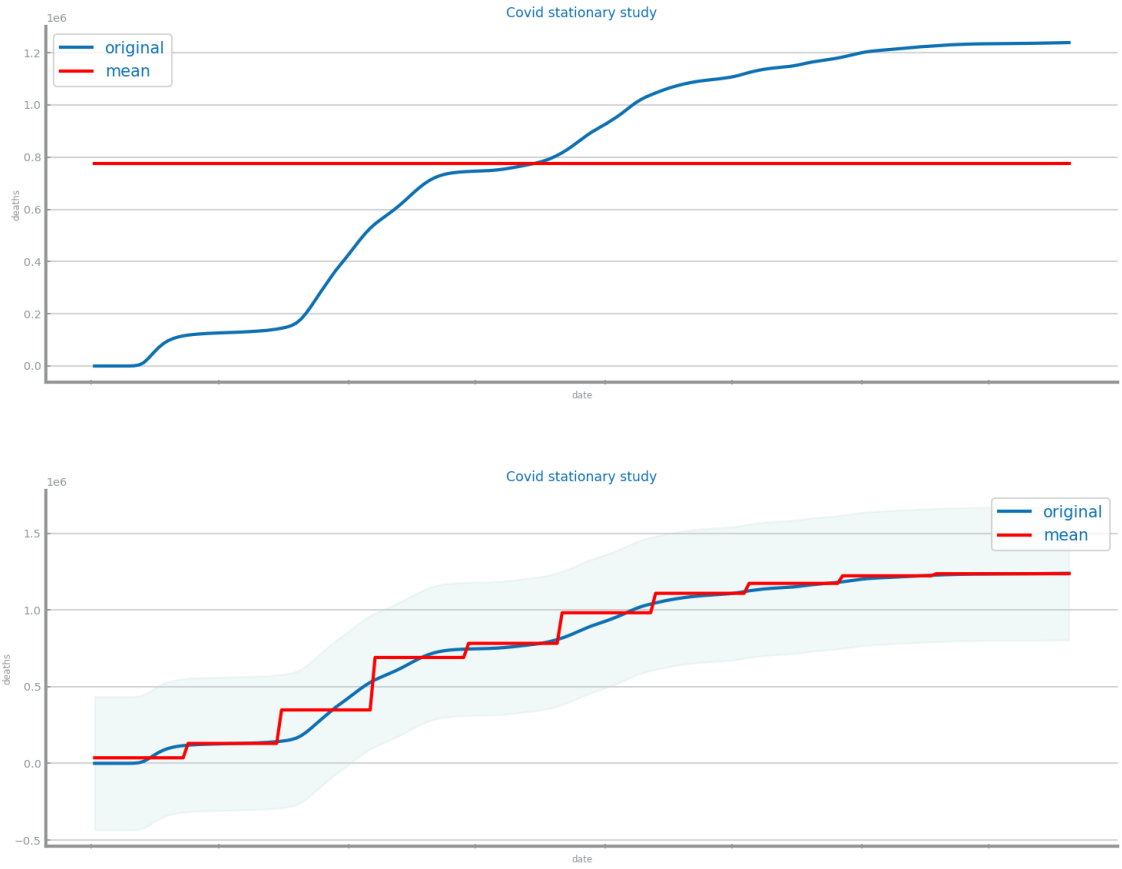


Figure 73: Stationarity study for time series 1

Figure 74: Components study for time series 2

Figure 75: Stationarity study for time series 2

6 DATA TRANSFORMATION

Aggregation

hlcolorShall describe the results of applying three different aggregations over both datasets, and identifying the granularity chosen to proceed. **Shall not exceed 300 characters.**

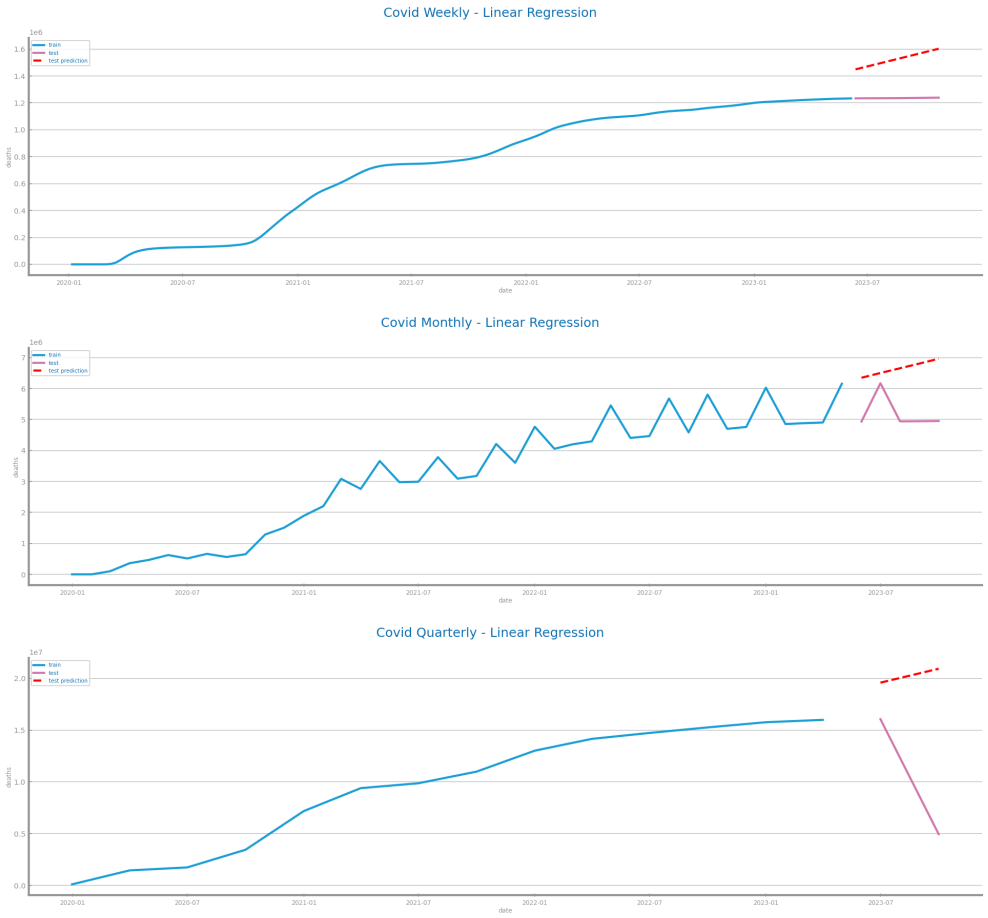


Figure 76: Forecasting plots after different aggregations on time series 1

Figure 77: Forecasting results after different aggregations on time series 1

Figure 78: Forecasting plots after different aggregations on time series 2

Figure 79: Forecasting results after different aggregations on time series 2

Smoothing

hlcolorShall describe the results of applying smoothing transformations over both datasets, and identifying the best result to proceed. **Shall not exceed 300 characters.**

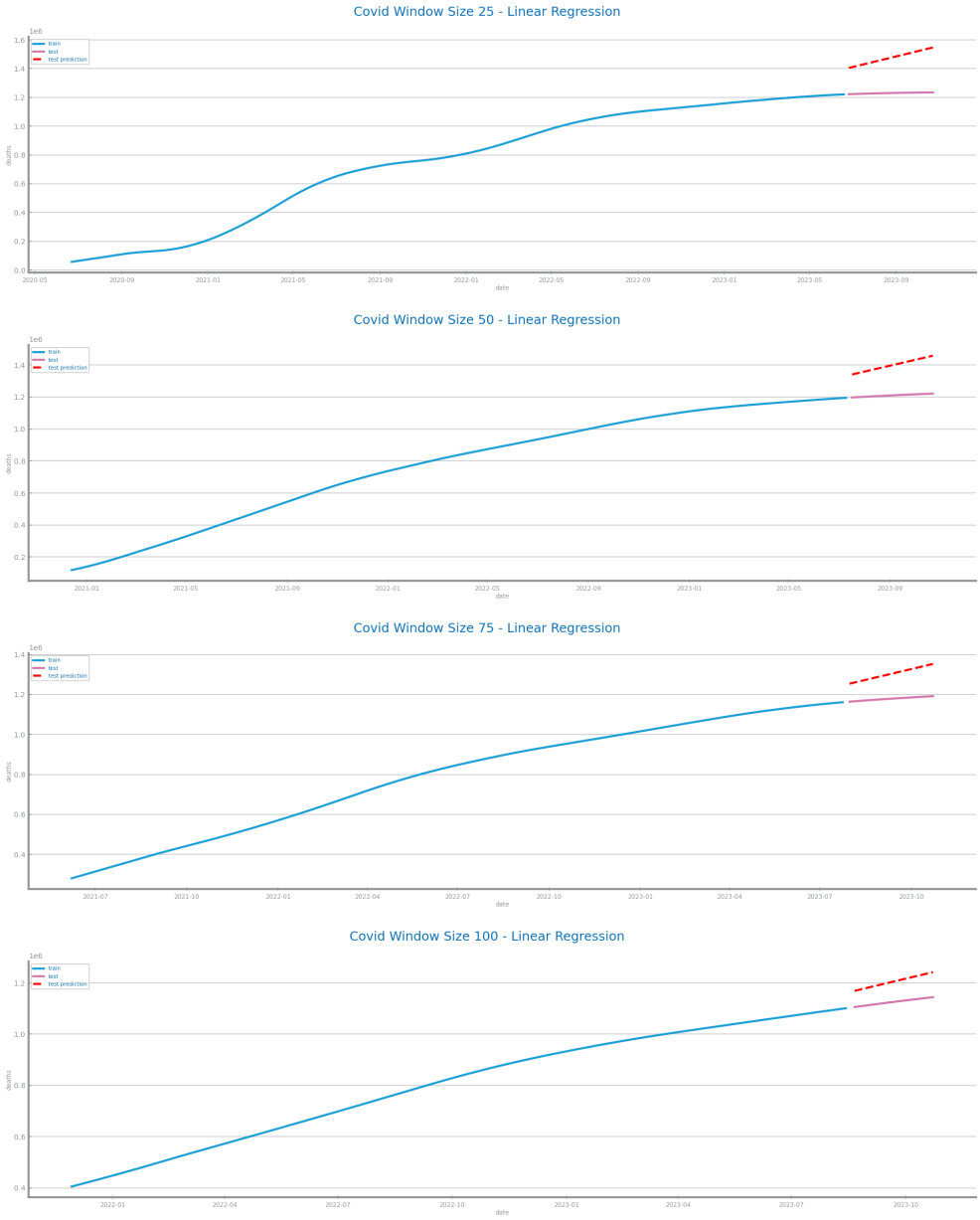


Figure 80: Forecasting plots after different smoothing parameterisations on time series 1

Figure 81: Forecasting results after different smoothing parameterisations on time series 1

Figure 82: Forecasting plots after different smoothing parameterisations on time series 2

Figure 83: Forecasting results after different smoothing parameterisations on time series 2

Differentiation

hlcolorShall describe the results of applying two consecutive differentiation of both datasets, and identifying the best result to proceed. **Shall not exceed 300 characters.**

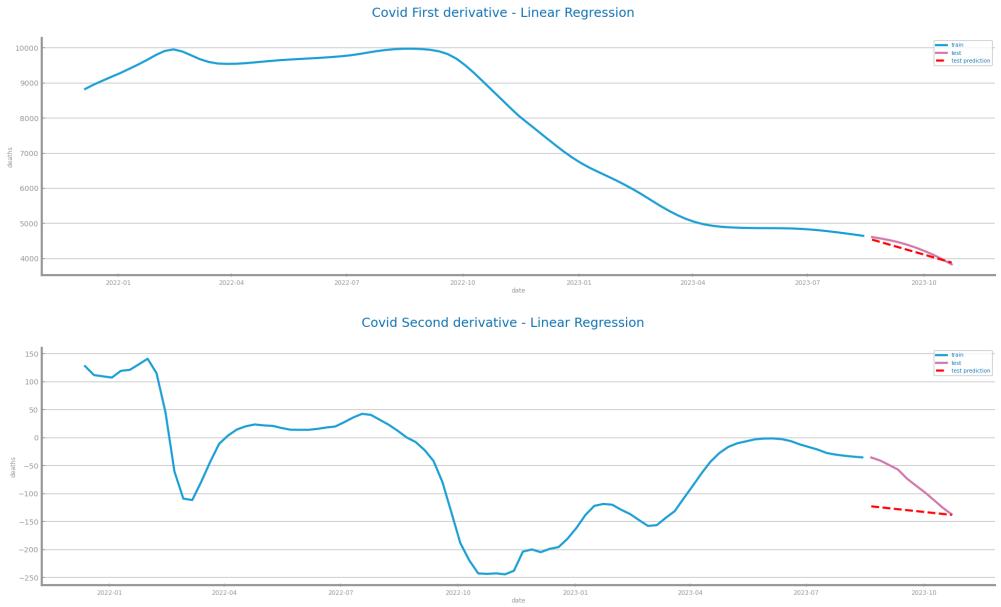


Figure 84: Forecasting plots after first and second differentiation of time series 1

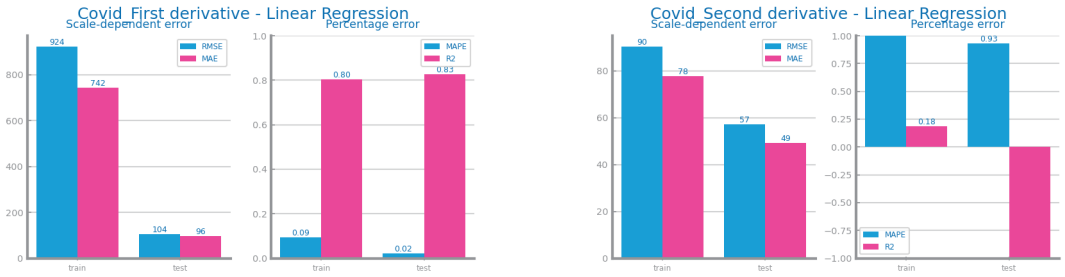


Figure 85: Forecasting results after first and second differentiation of time series 1

Figure 86: Forecasting plots after first and second differentiation of time series 2

Figure 87: Forecasting results after first and second differentiation of time series 2

Other transformations (optional)

hlcolorShall describe the results of applying other transformations over both datasets, and identifying the best result to proceed. **Shall not exceed 500 characters.**

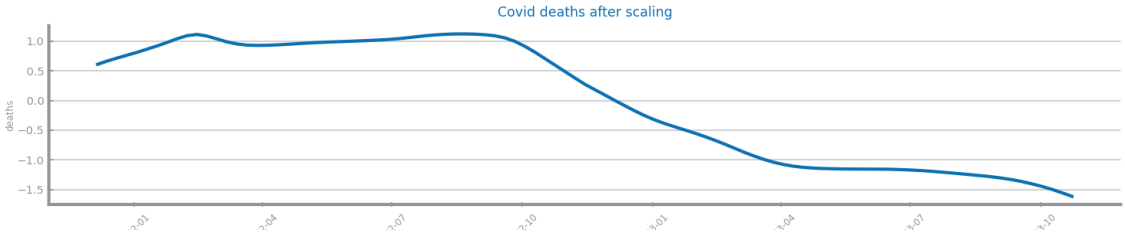


Figure 88: Forecasting plots after applying scaling over time series 1

Figure 89: Forecasting results after applying other transformations over time series 1

Figure 90: Forecasting plots after applying other transformations over time series 2

Figure 91: FForecasting results after applying other transformations over time series 2

7 MODELS' EVALUATION

hlcolorShall be used to summarise the transformations done over the original time series. **Shall not exceed 500 characters.**

Simple Average Model

hlcolorShall be used to present the results achieved through the simple average model. **Shall not exceed 200 characters.**

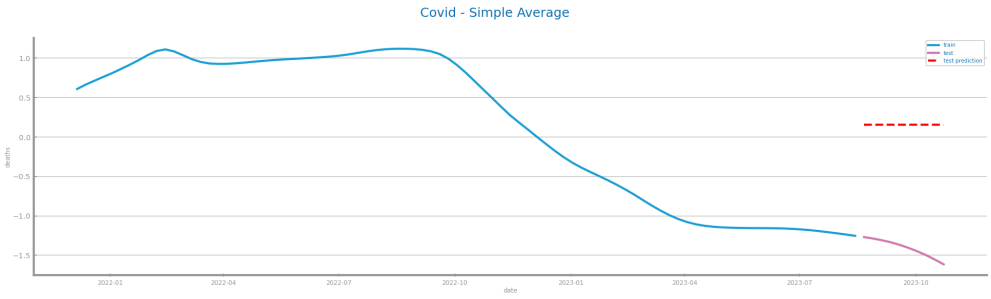


Figure 92: Forecasting plots obtained with Simple Average model over time series 1

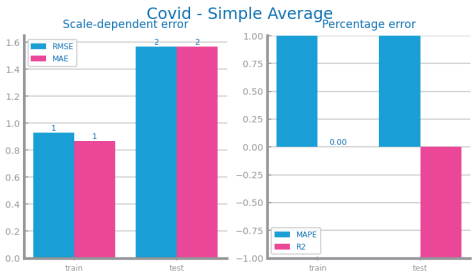


Figure 93: Forecasting results obtained with Simple Average model over time series 1

Figure 94: Forecasting plots obtained with Simple Average model over time series 2

Figure 95: Forecasting results obtained with Simple Average model over time series 2

Persistence Model

hlcolorShall be used to present the results achieved through the persistence model. **Shall not exceed 500 characters.**

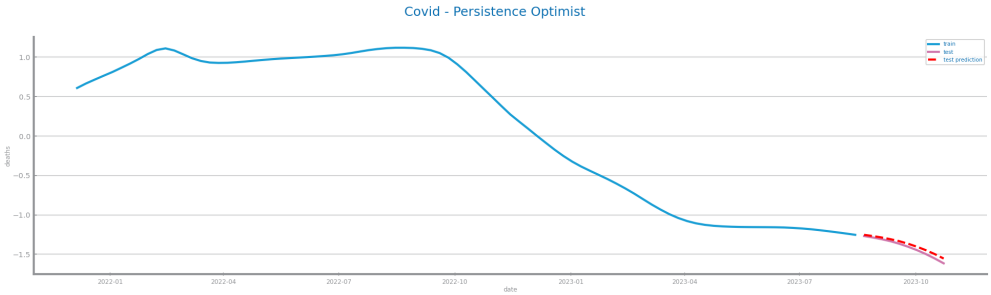


Figure 96: Forecasting plots obtained with Persistence model (long term) over time series 1

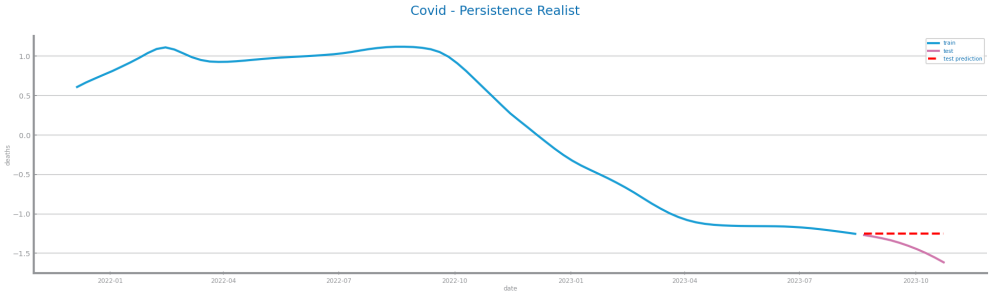


Figure 97: Forecasting plots obtained with Persistence model (one-set-behind) over time series 1

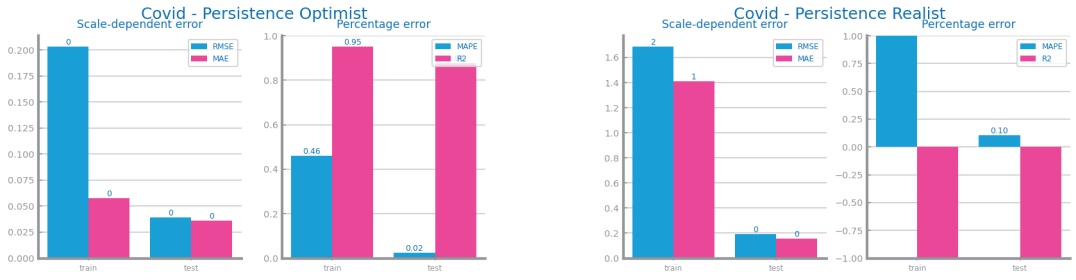


Figure 98: Forecasting results obtained with Persistence model in both situations over time series 1

Figure 99: Forecasting plots obtained with Persistence model (long term) over time series 2

Figure 100: Forecasting plots obtained with Persistence model (one-set-behind) over time series 2

Figure 101: Forecasting results obtained with Persistence model in both situations over time series 2

Rolling Mean Model

hlcolorShall be used to present the results achieved through the rolling mean forecasting algorithms. **Shall not exceed 500 characters.**

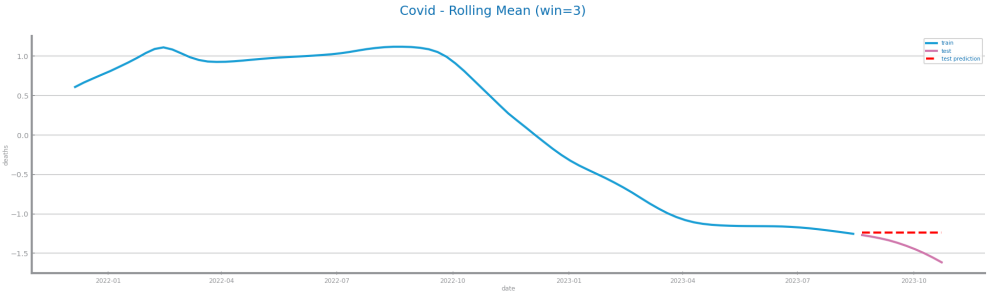


Figure 102: Forecasting study over different parameterisations of the rolling mean algorithm over time series 1

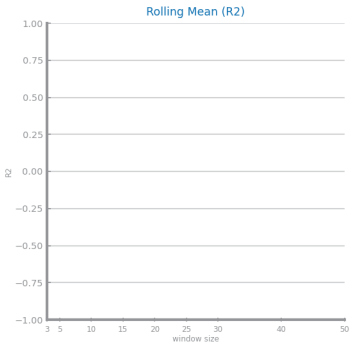


Figure 103: Forecasting plots obtained with the best parameterisation of rolling mean algorithm, over time series 1

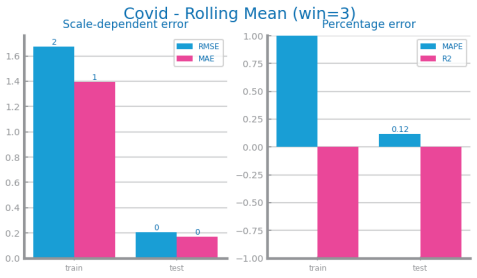


Figure 104: Forecasting results obtained with the best parameterisation of rolling mean algorithm, over time series 1

Figure 105: Forecasting study over different parameterisations of the rolling mean algorithm over time series 2

Figure 106: Forecasting plots obtained with the best parameterisation of rolling mean algorithm, over time series 2

Figure 107: Forecasting results obtained with the best parameterisation of rolling mean algorithm, over time series 2

ARIMA Model

hlcolorShall be used to present the results achieved through the ARIMA forecasting algorithms. **Shall not exceed 500 characters.**

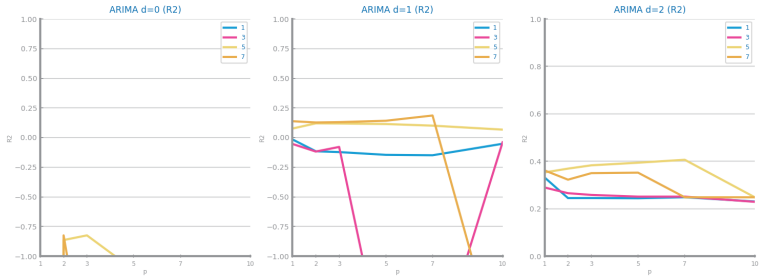


Figure 108: Forecasting study over different parameterisations of the ARIMA algorithm over time series 1

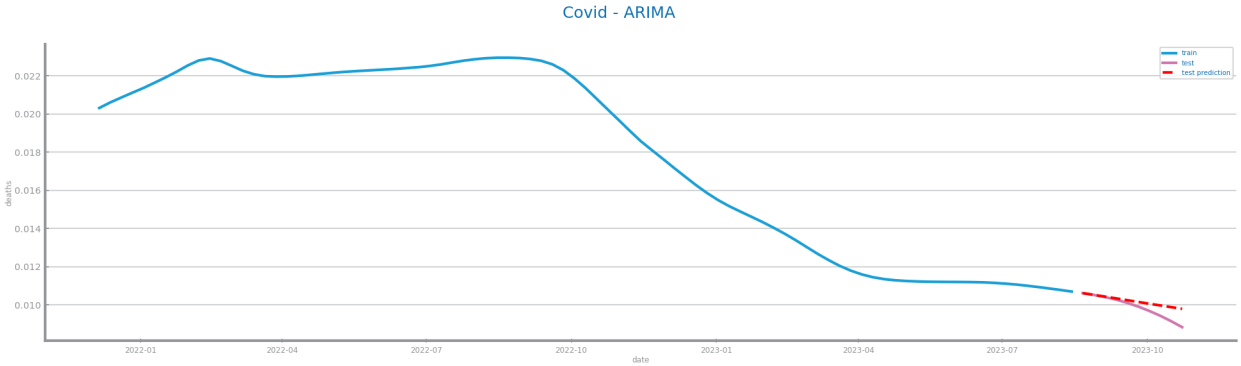


Figure 109: Forecasting plots obtained with the best parameterisation of ARIMA algorithm, over time series 1

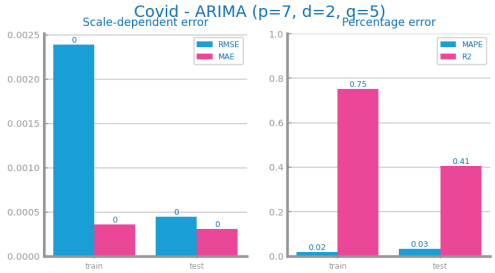


Figure 110: Forecasting results obtained with the best parameterisation of ARIMA algorithm, over time series 1

Figure 111: Forecasting study over different parameterisations of the ARIMA algorithm with external variables over time series 1

Figure 112: Forecasting plots obtained with the best parameterisation of ARIMA algorithm with external variables over time series 1

Figure 113: Forecasting results obtained with the best parameterisation of ARIMA algorithm with external variables over time series 1

Figure 114: Forecasting study over different parameterisations of the ARIMA algorithm over time series 2

Figure 115: Forecasting plots obtained with the best parameterisation of ARIMA algorithm, over time series 2

Figure 116: Forecasting results obtained with the best parameterisation of ARIMA algorithm, over time series 2

Figure 117: Forecasting study over different parameterisations of the ARIMA algorithm with external variables over time series 2

Figure 118: Forecasting plots obtained with the best parameterisation of ARIMA algorithm with external variables over time series 2

Figure 119: Forecasting results obtained with the best parameterisation of ARIMA algorithm with external variables over time series 2

LSTMs Model

hlcolorShall be used to present the results achieved through LSTMs. **Shall not exceed 500 characters.**

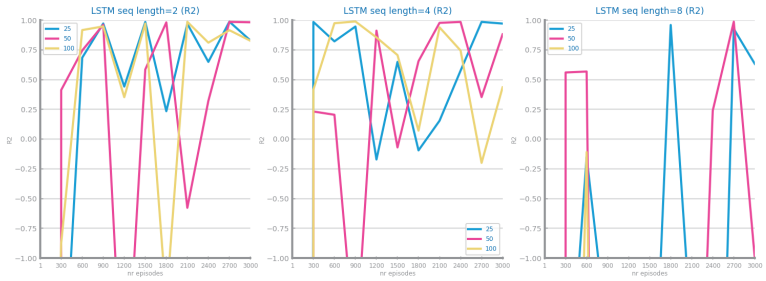


Figure 120: Forecasting study over different parameterisations of LSTMs over time series 1

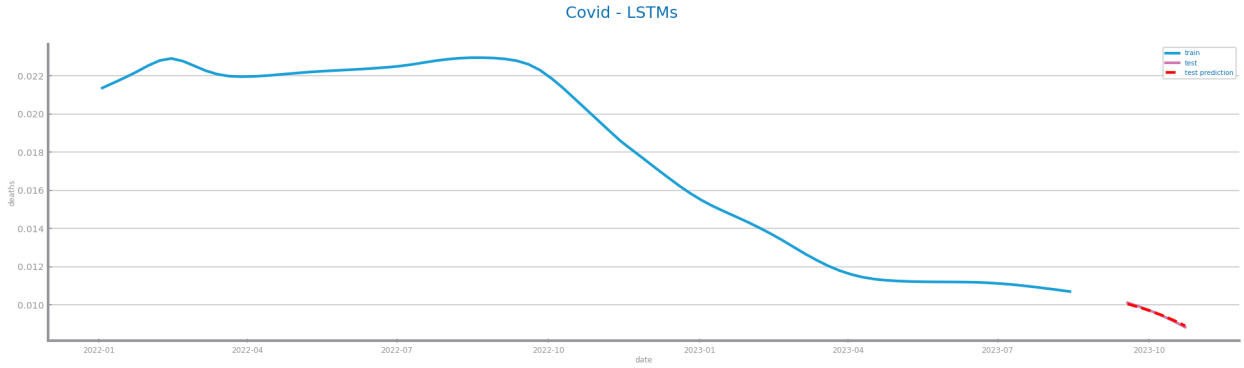


Figure 121: Forecasting plots obtained with the best parameterisation of LSTMs, over time series 1

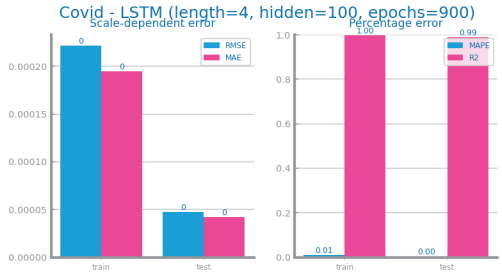


Figure 122: Forecasting results obtained with the best parameterisation of LSTMs, over time series 1

Figure 123: Forecasting study over different parameterisations of LSTMs with multiple variables over time series 1

Figure 124: Forecasting plots obtained with the best parameterisation of LSTMs with multiple variables over time series 1

Figure 125: Forecasting results obtained with the best parameterisation of LSTMs with multiple variables over time series 1

Figure 126: Forecasting study over different parameterisations of the LSTMs over time series 2

Figure 127: Forecasting plots obtained with the best parameterisation of LSTMs, over time series 2

Figure 128: Forecasting results obtained with the best parameterisation of LSTMs, over time series 2

Figure 129: Forecasting study over different parameterisations of LSTMs with multiple variables over time series 2

Figure 130: Forecasting plots obtained with the best parameterisation of LSTMs with multiple variables over time series 2

Figure 131: Forecasting results obtained with the best parameterisation of LSTMs with multiple variables over time series 2

8 CRITICAL ANALYSIS

hlcolorShall be used to present a summary of the results achieved with the different forecasting techniques, and the impact of the different preparation tasks on their performance. 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. Additional charts may be presented here. **Shall not exceed 2000 characters.**