TÉCNICO LISBOA

DS 2023

Data Science Project

Team nr: 13	**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

The 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

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.**

For the second dataset, the services domain one, we didn't do much processing prior to the study of the data for most forms of analysis. We simply noticed that the **age** variable had values with the character "_" which we removed for it to become a numeric variable as it is. There were several anomalies in the values for some of the variables. We decided to keep these values for the profiling.

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.**

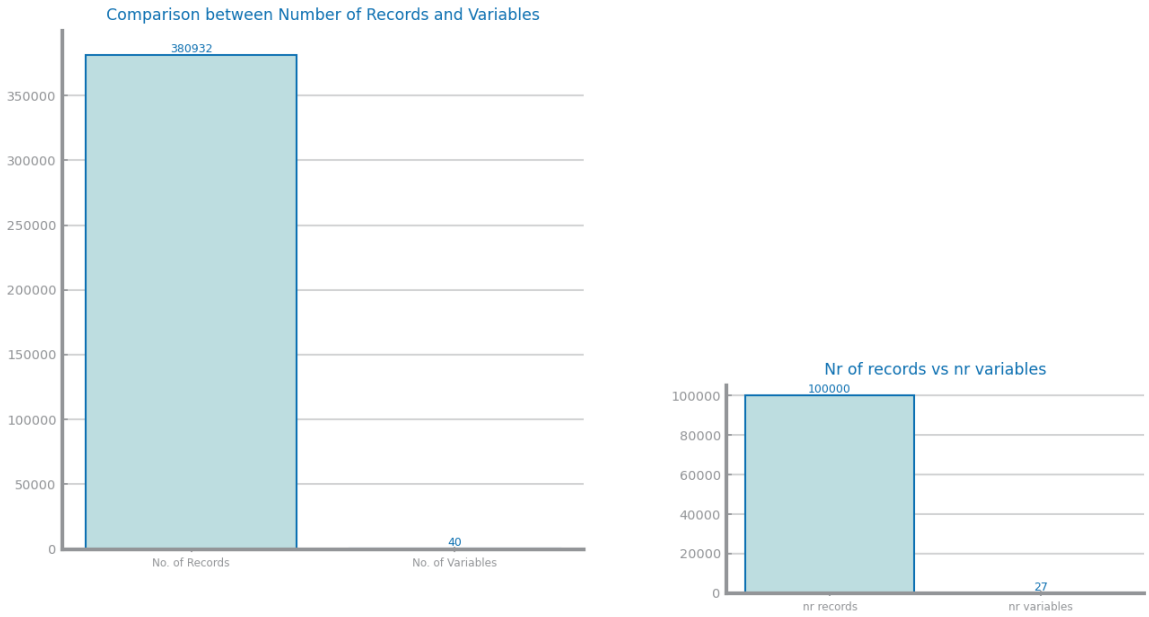


Figure 1: Nr Records x Nr variables for dataset 1 (left) and dataset 2 (right)

We can see that for both datasets there are much more records than variables, avoiding the curse of dimensionality.

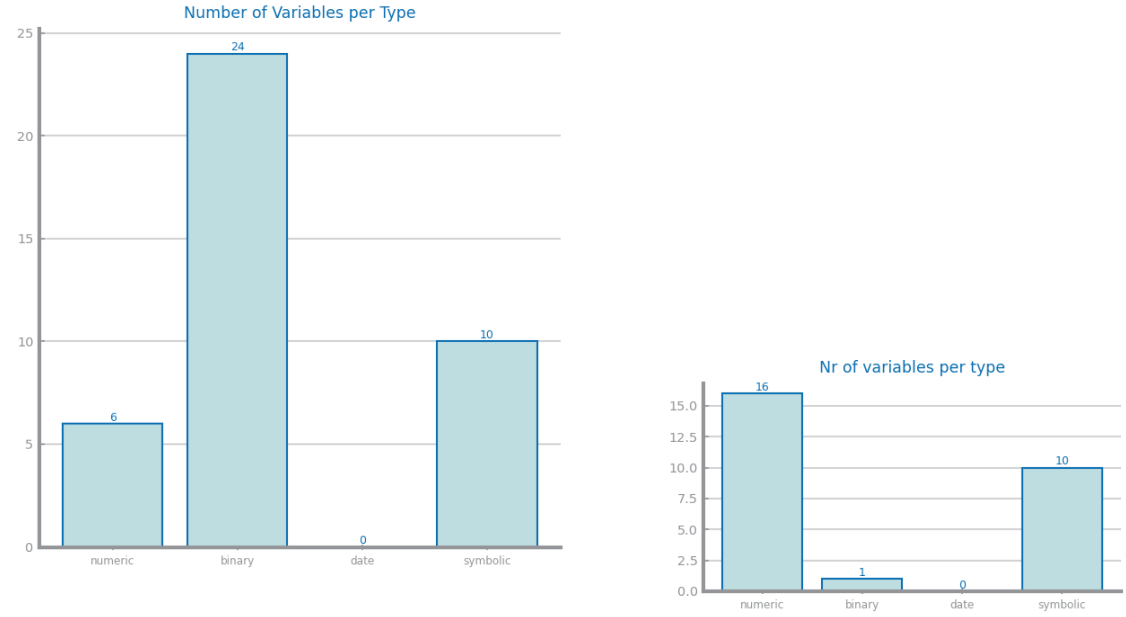


Figure 2: Nr variables per type for dataset 1 (left) and dataset 2 (right)

Neither dataset has *date* variables, meaning that theres is no time frame associated with them or the records. The dataset about health has predominantly symbolic variables, especially binary. This is expected as it is harder to quantify clinical observations about the state of a person. In contrast, the services dataset has mainly numerical variables. These

offer more precision and are more easily obtained in the financial context the data is in.

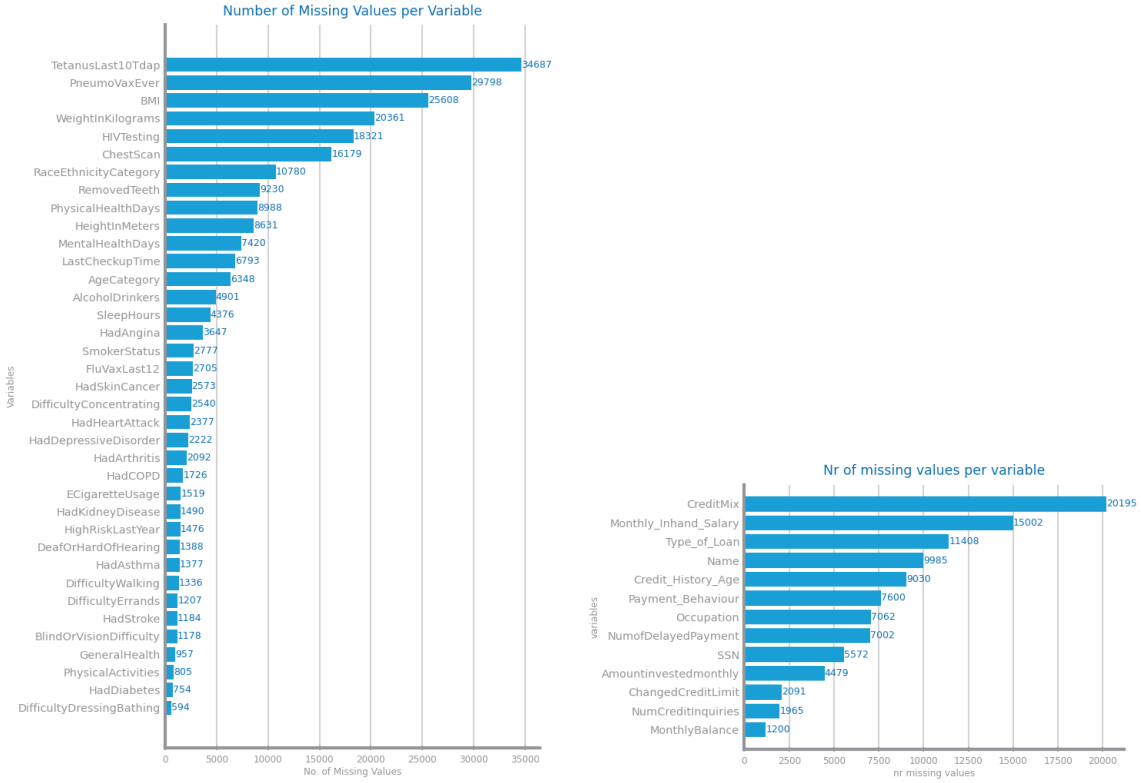


Figure 3: Nr missing values for dataset 1 (left) and dataset 2 (right)

For the health domain dataset, the number of records is 380.932. The variable with the most missing values has less than 10% of missing values. This indicates that some sort of imputation might be a viable option to address them. The second dataset has variables with a bigger ratio of missing values. However these ratios don't go over 20%, making it unlikely that it will be better to drop said variables.

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.**

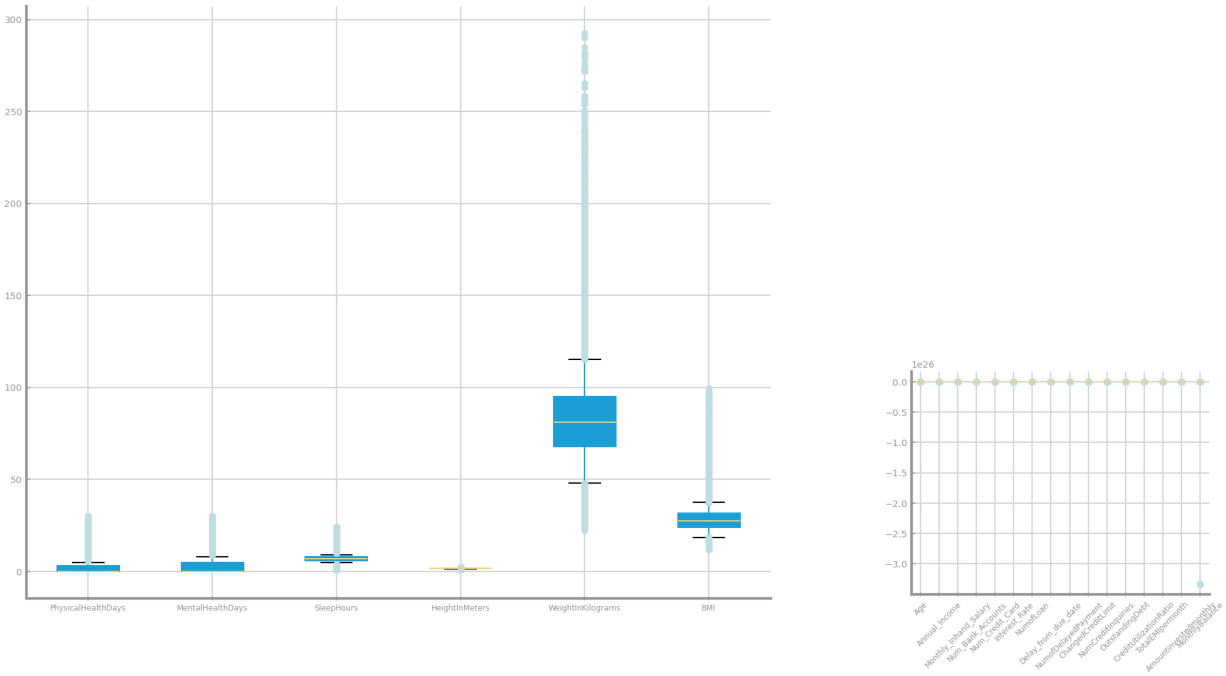


Figure 4: Global boxplots dataset 1 (left) and dataset 2 (right)

Global boxplots are rarely a good option to visualize data as they depend on the fact that all variables have the same scale. Given this, we can say, about the health dataset, that the variable with the biggest spread is *WeightInKilograms*. When examining the services dataset, the only thing we can tell is that the *MonthlyBalance* variable has a huge outlier. This completly distorts the overall representation of the data, rendering the figure nearly useless in terms of insights it could give us.

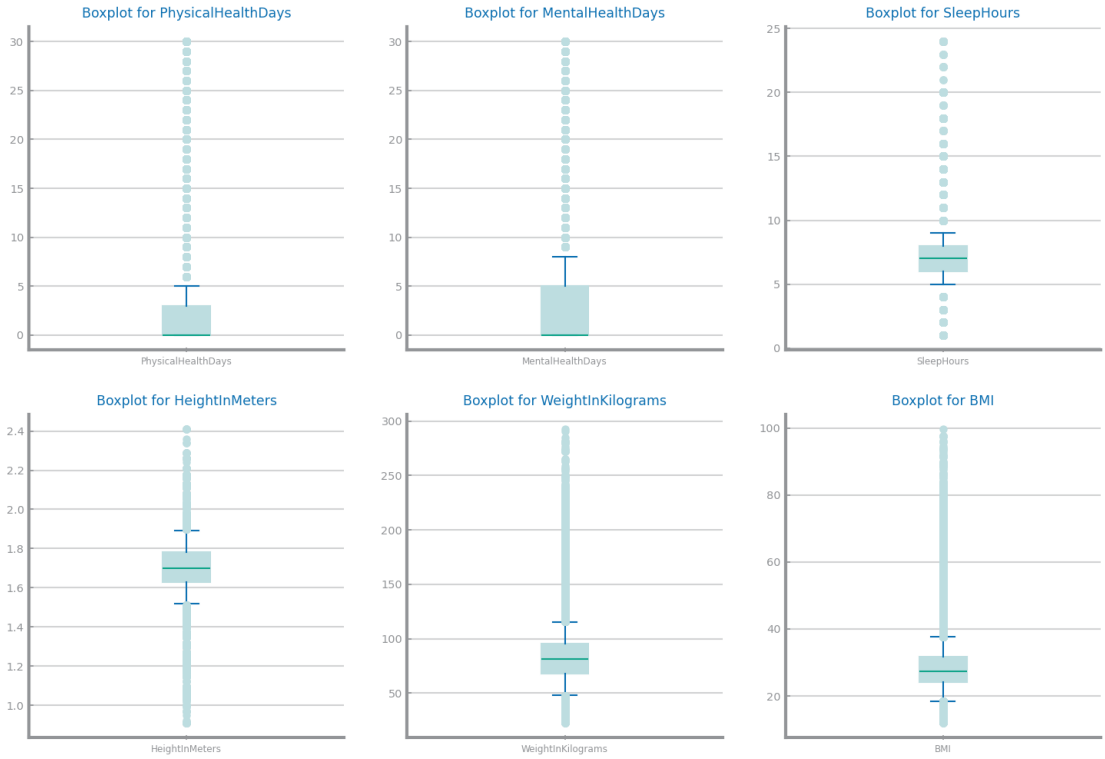


Figure 5: Single variables boxplots for dataset 1

From this we can conclude that all the numeric variables in this first dataset have quite high variability in their values.

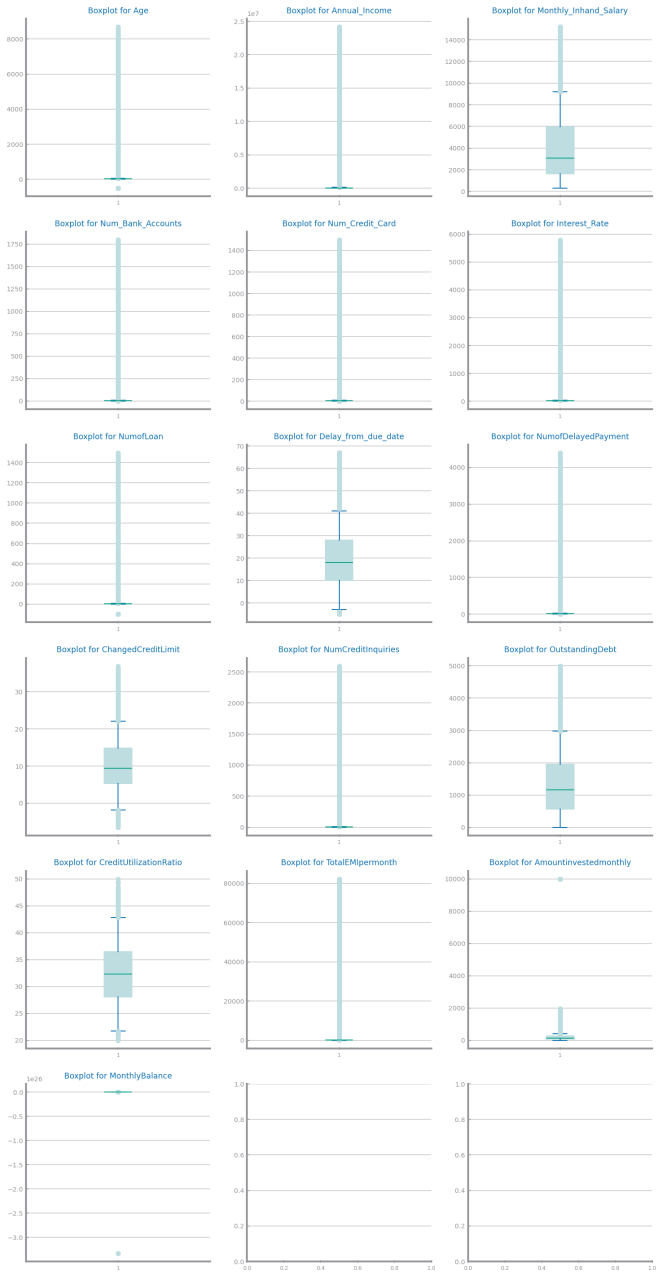


Figure 6: Single variables boxplots for dataset 2

Analysing these boxplots, we come to the conclusion that more than half the numeric variables in this dataset have massive variation. This means that a more aggressive approach during data preparation for the outlier values might be needed. Knowing this, it is important to remember that an outlier is a point that is far away from the regular ones. Given that for all but the *MonthlyBalance* variable all the points are continuous, some care must be taken in their treatment.

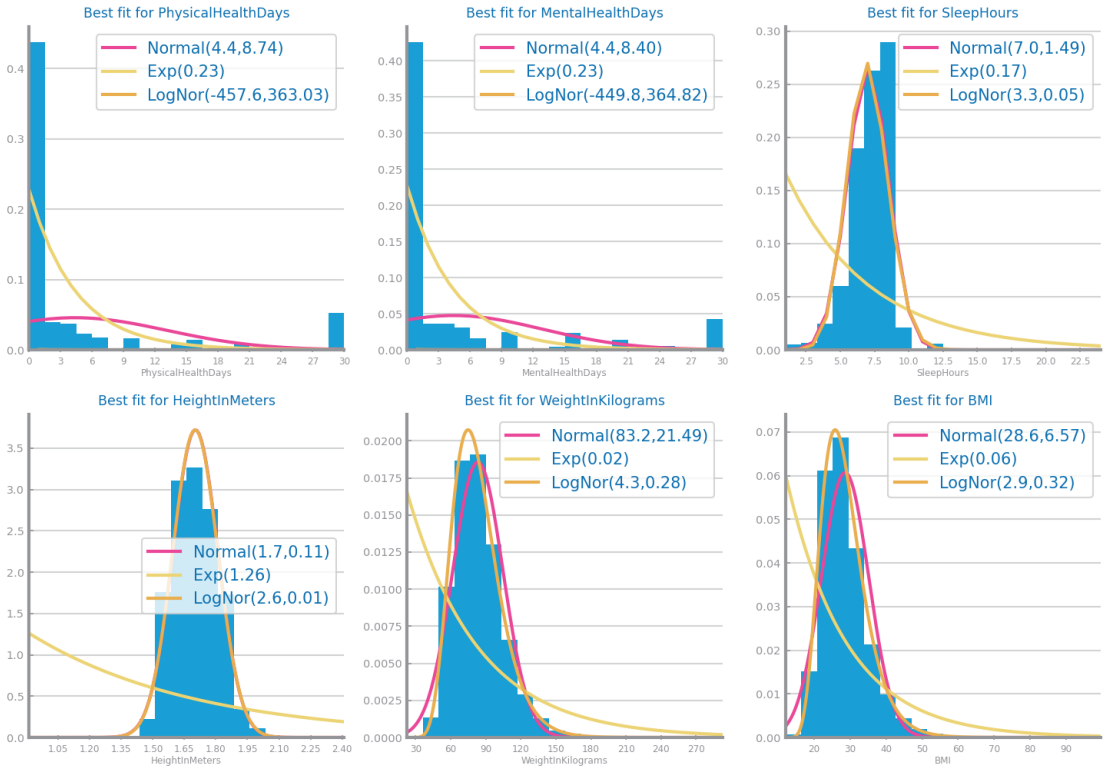


Figure 7: Histograms for dataset 1

From this visual representation we can extract that a majority of numerical variables follow a normal distribution quite well. The others have a shape closer to that of the exponential distribution. Maybe a normalization of the last would be advantageous. This would enable the application of models that rely on the data following normal distributions later on.

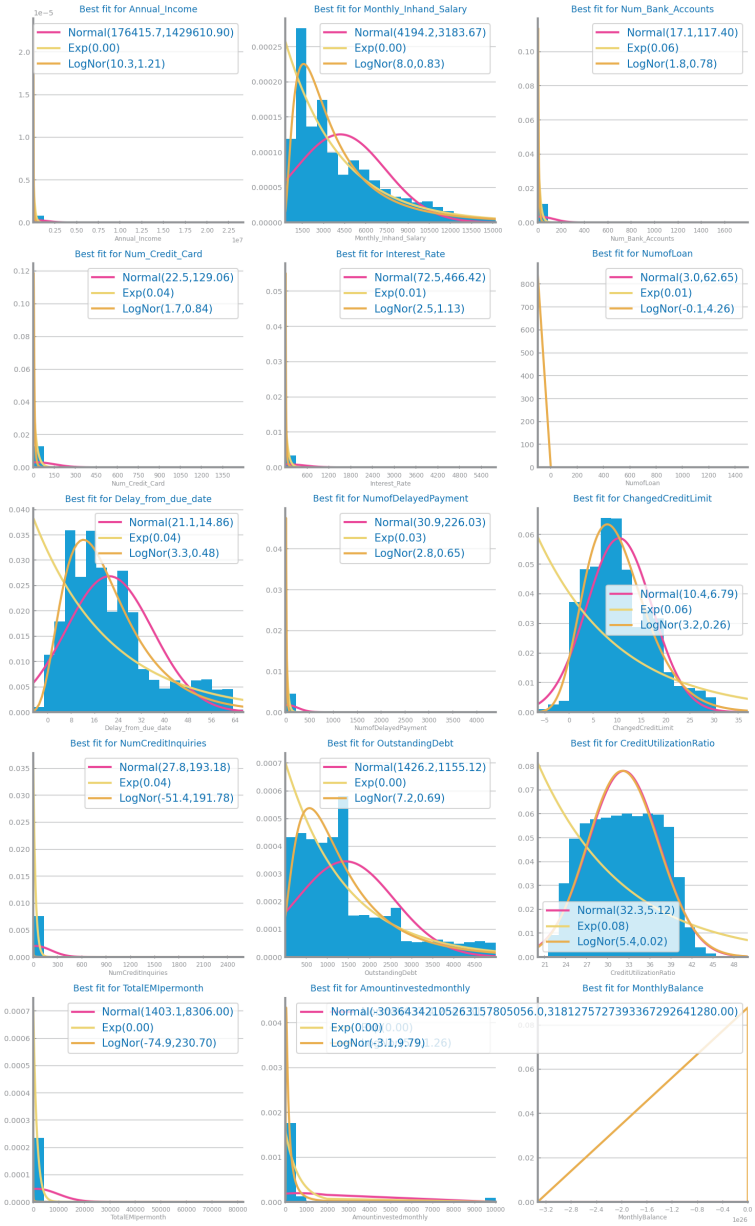


Figure 8: Histograms for dataset 2

The second dataset once again suffers from having very distant outliers. The histograms from which we can infer more information either follow a normal distribution or a lognormal one. The latter can easily be trasnformed to a normal one

if needed for some modelation.

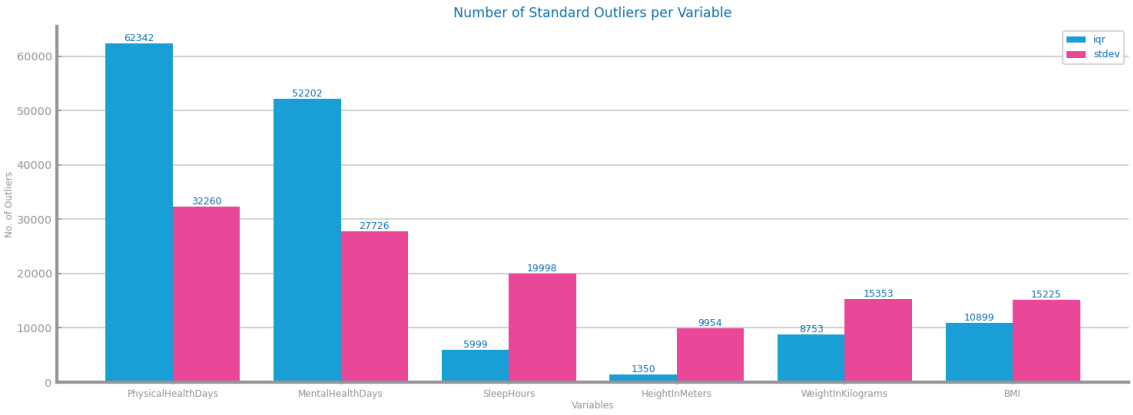


Figure 9: Outliers study dataset 1

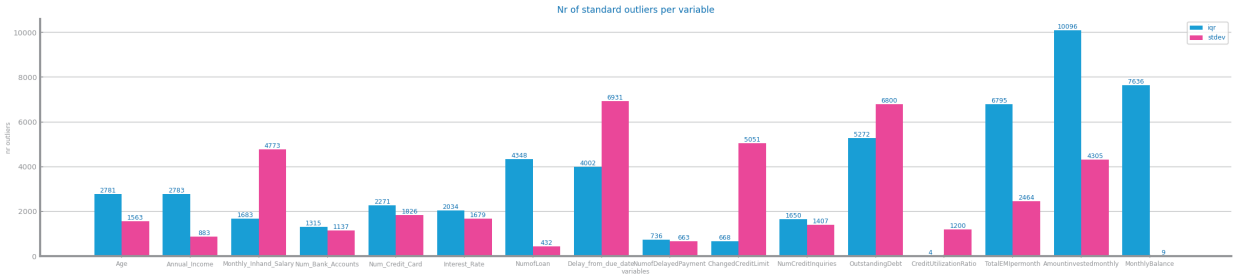


Figure 10: Outliers study dataset 2

For the outliers study per variable we used the IQR factor and the stantard deviation one. The IQR factor was of 1.5 and the stddev was of 2. IQR is more robust to extreme values, making it more desiarable for the study of the second dataset. The stddev method should only be used for variables that follow a normal distribution.

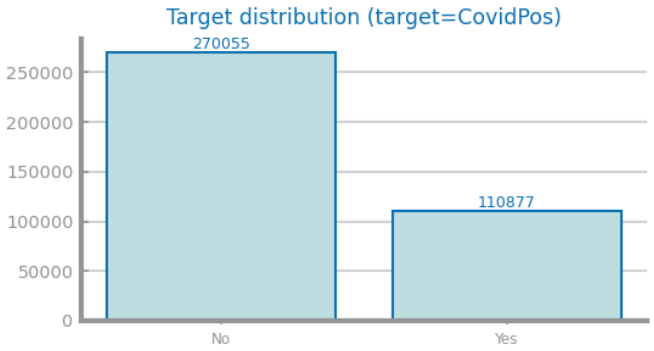


Figure 11: Class distribution for dataset 1

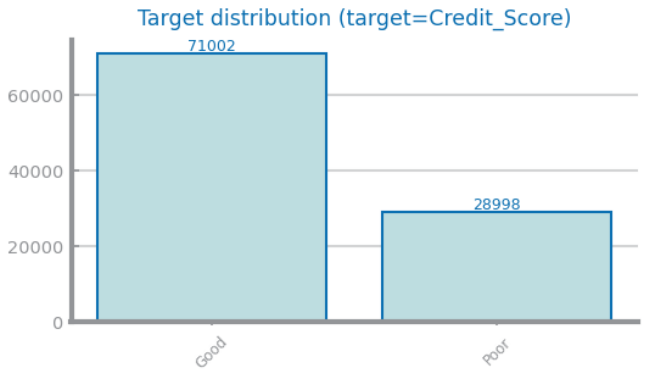


Figure 12: Class distribution for dataset 2

The class distribution is similar in both datasets, a 25%-75%. This means that probebly some balancing with be required during the data preparation phase.

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.**

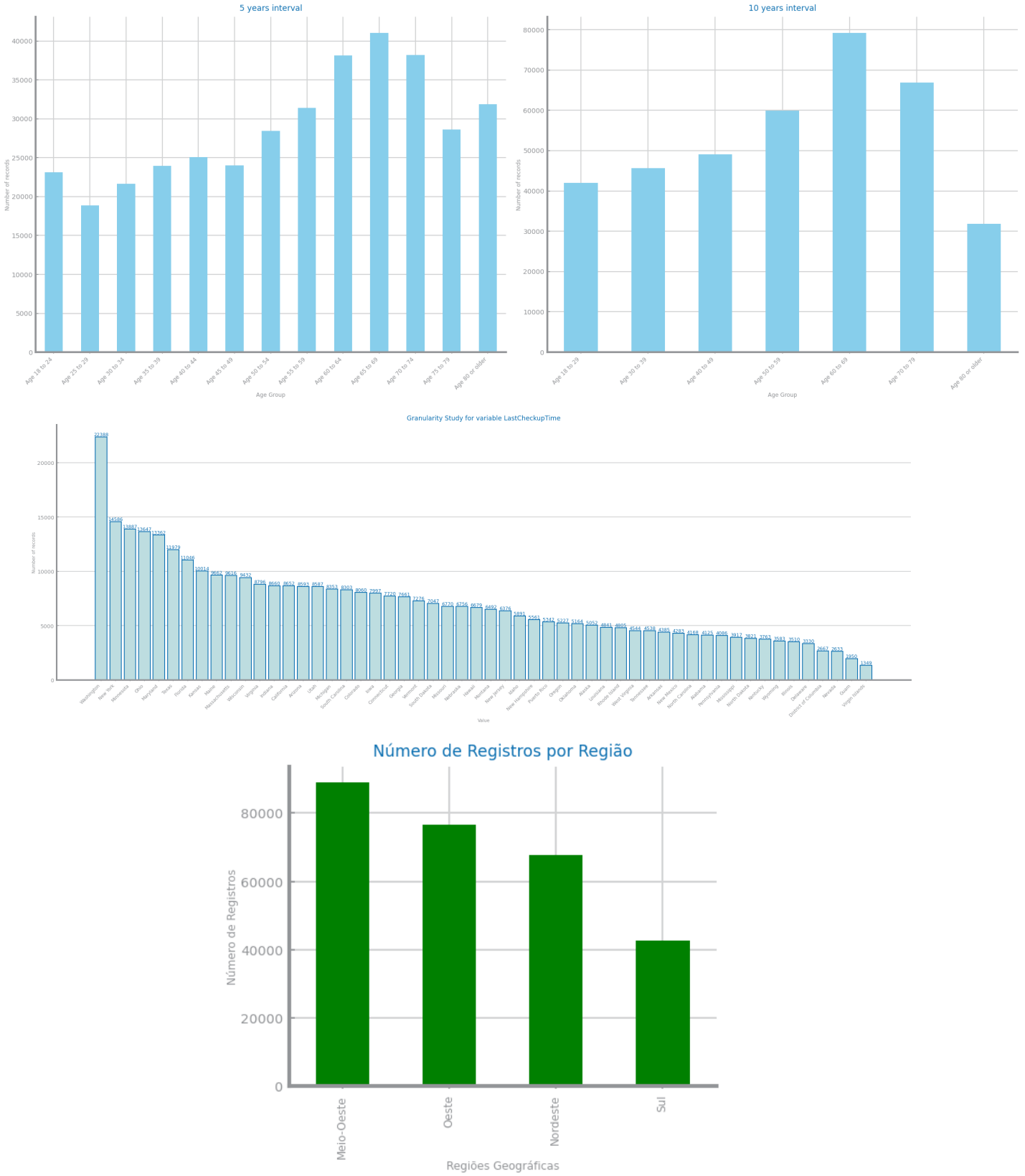


Figure 13: Granularity analysis for dataset 1

Granularity study for variables for which we thought to have interesting taxonomies to help us in the variable encoding process. The variables *LastCheckupTime*, *State* and *AgeCategory* have interesting taxonomies, allowing for reduction of variance without harming the distribution and information of these variables

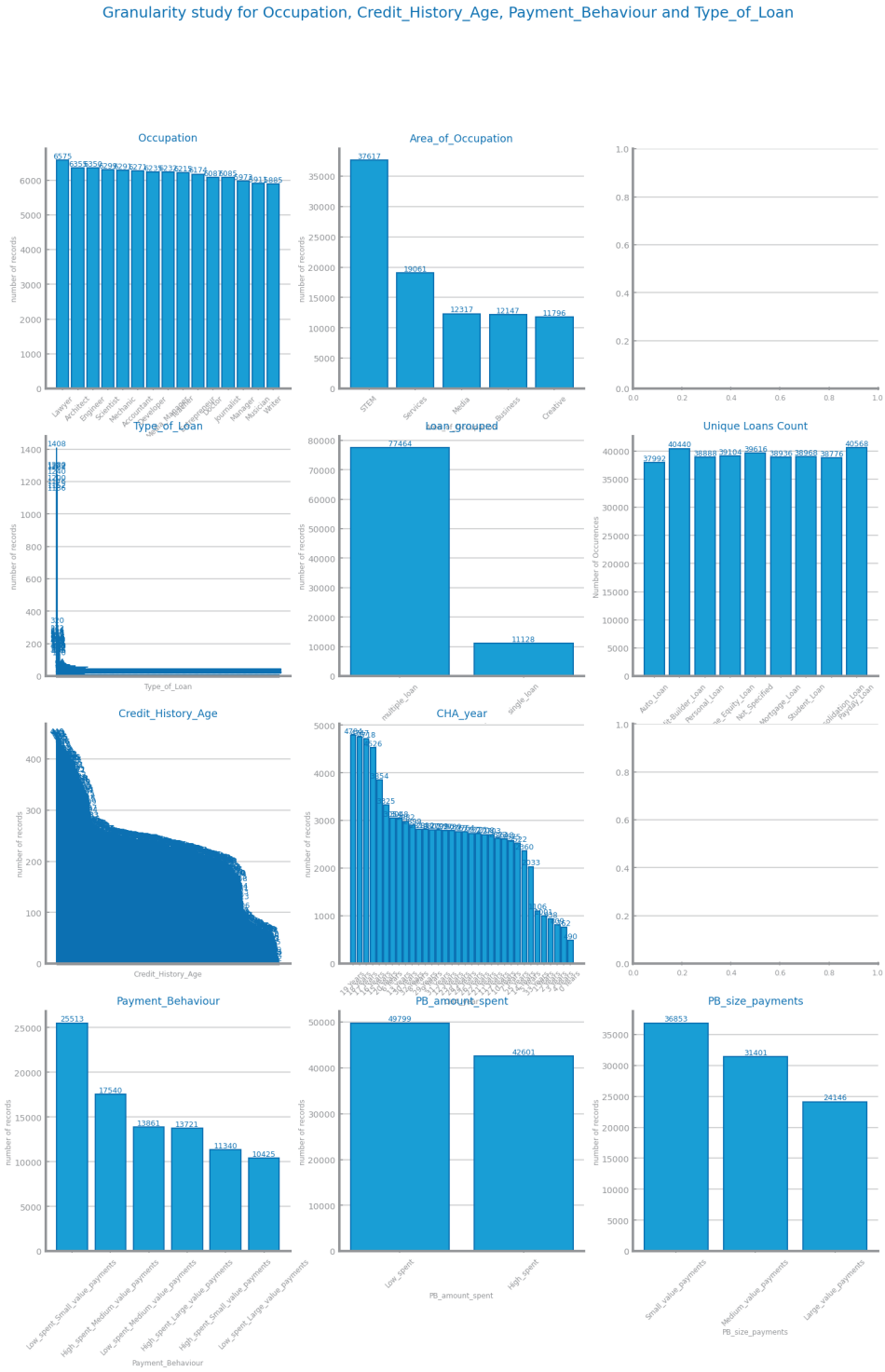


Figure 14: Granularity analysis for dataset 2

For dataset 2. Some of the variables clearly had too much differente values, this study allowed us to know if descending in the taxonomy hierachy would not signifincantly alter the distribution of these variables. Notable examples of this are the *Type_of_Loan* and *Credit_History_Age* variables.

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.**

For both datasets, we used variable encoding to transform the symbolic variables into numeric ones for the correlation analysis.

Figure 16: Sparsity analysis for dataset 2

Binary on binary is not interesting for sparsity evaluation. We can see which variables have more impact on the class and the correlation between them.

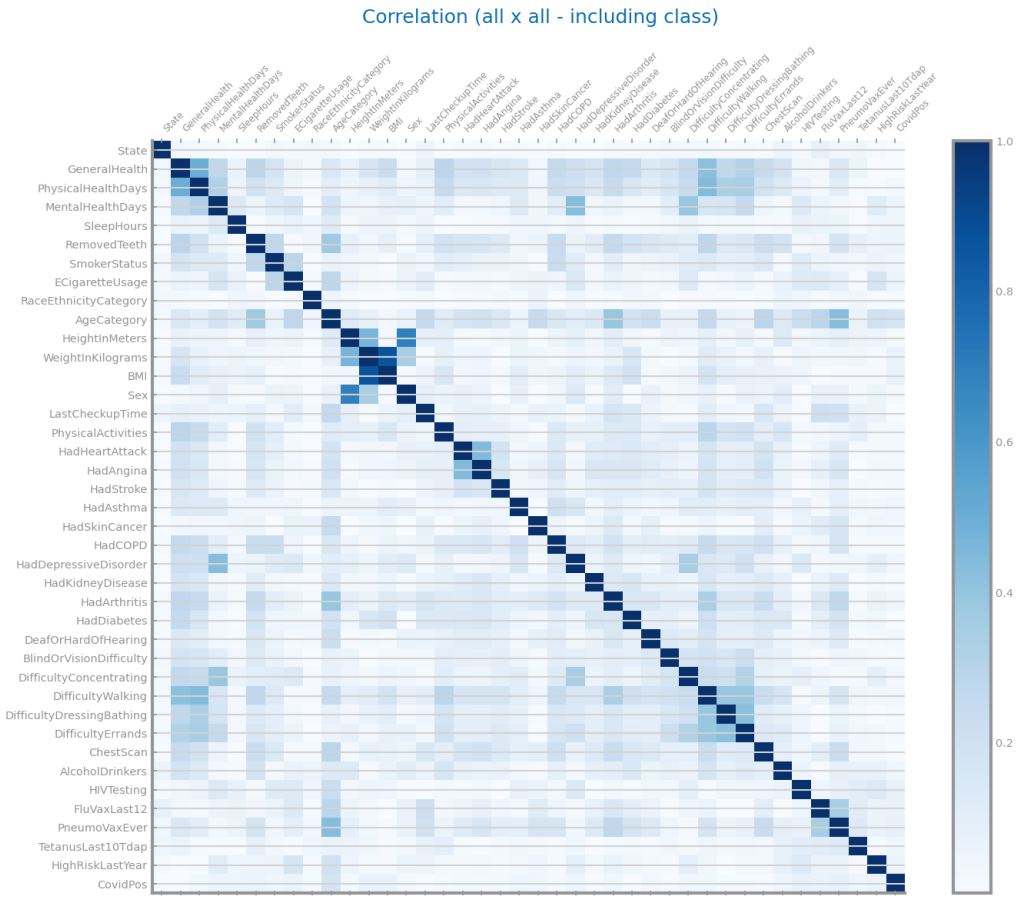


Figure 17: Correlation analysis for dataset 1

For both datasets, we used variable encoding to transform the symbolic variables into numeric ones for the correlation analysis. We can clearly observe that there are not many variables with high correlations, and the ones with higer ones aren't relevant - The gender of a person and their height, their BMI and their weight.

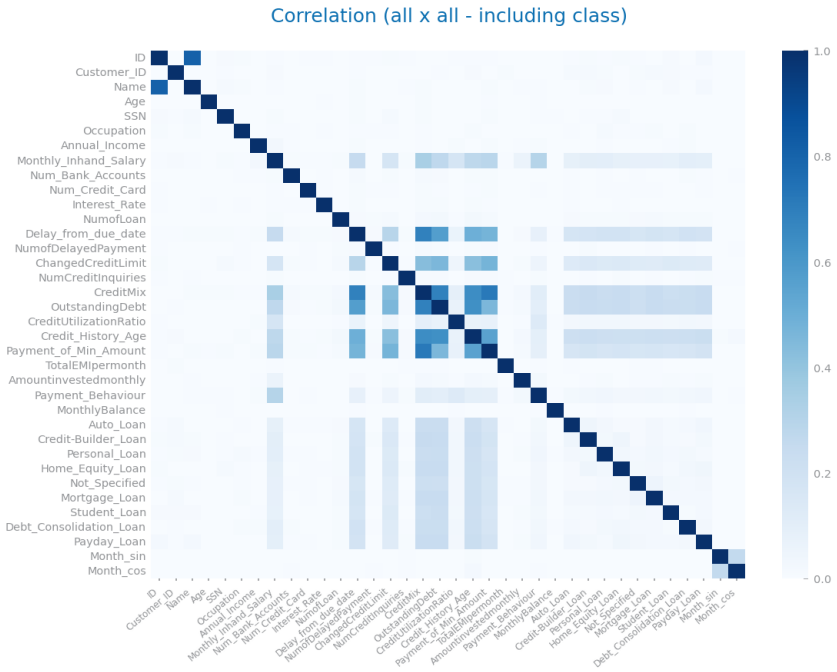


Figure 18: Correlation analysis for dataset 2

In the services dataset we can infer a bit more information, especially about the behaviour of people and what leads them to having a good or bad credit score. Despite this not all high correlations are relevant, per example, the *Name* and *ID* variables.

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.***

Dataset 1 - Health

In this step all symbolic variables were converted to numeric. All the binary got turned into zeros and ones. Ordinal variables preserved their order - *GeneralHealth*, *LastCheckupTime*, *RemovedTeeth*, *HadDiabetes*, *SmokerStatus*, *ECigaretteUsage*, *AgeCategory* and *TetanusLast10Tdap*. The *State* got its granularity reduced into regions according to the previously done study.

Dataset 2 - Services

In this step we converted all non numerical variables into numerical. For the *ID*, *Customer_ID*, *SSN* and *Name* were transformed into numbers, preserving their uniqueness. For variables that had order to their values we used ordinal

linear encoding, these were *Credit_Score*, *Credit_Mix*, *Payment_of_Min_Amount*, *Credit_History_Age* and *Payment_Behaviour* . For the *Month* we used cyclic encoding. We reduced the granularity of the *Occupation* variable, turning it more broad. For the *Type_of_Loan* variable we used dummification.

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.**

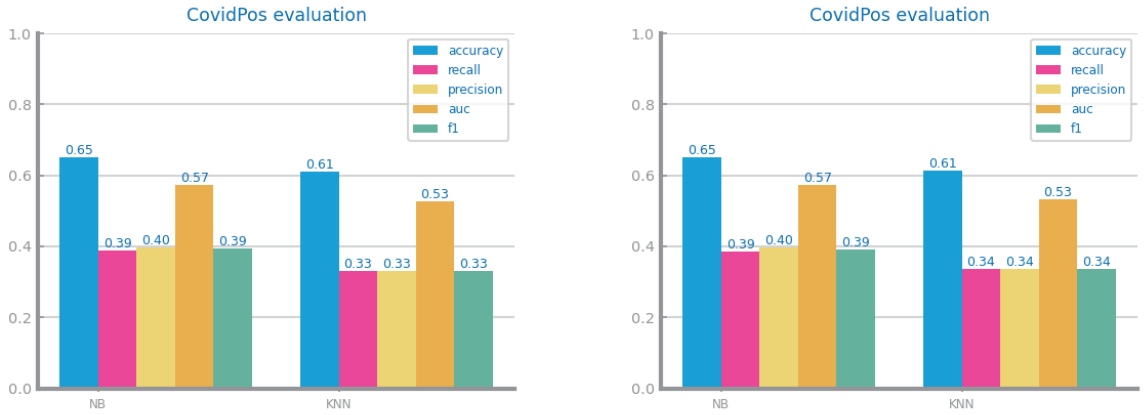


Figure 19: Missing values imputation results with different approaches for dataset 1. Frequence (left) and KNN (right) strategies

First drop all records with more than 80% missing values. Frequence strategy chosen to fill missing entries.

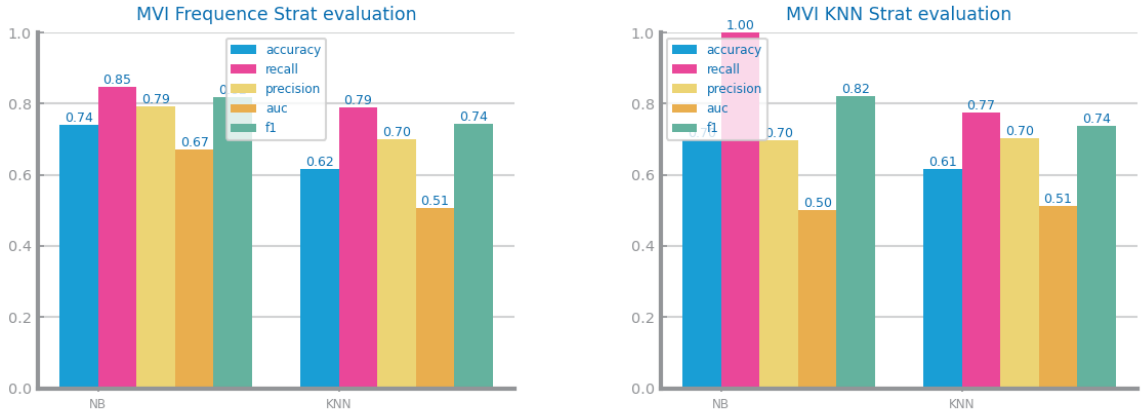


Figure 20: Missing values imputation results with different approaches for dataset 2

Firstly, we drop all the records with more than 90% of missing values, removing 11.482 entries. Then, we applied two different filling strategies, frequent and KNN, both had very similiar results after evaluation. However, the frequent

strategy came out on top and we chose it to fill the rest of the missing values.

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.**

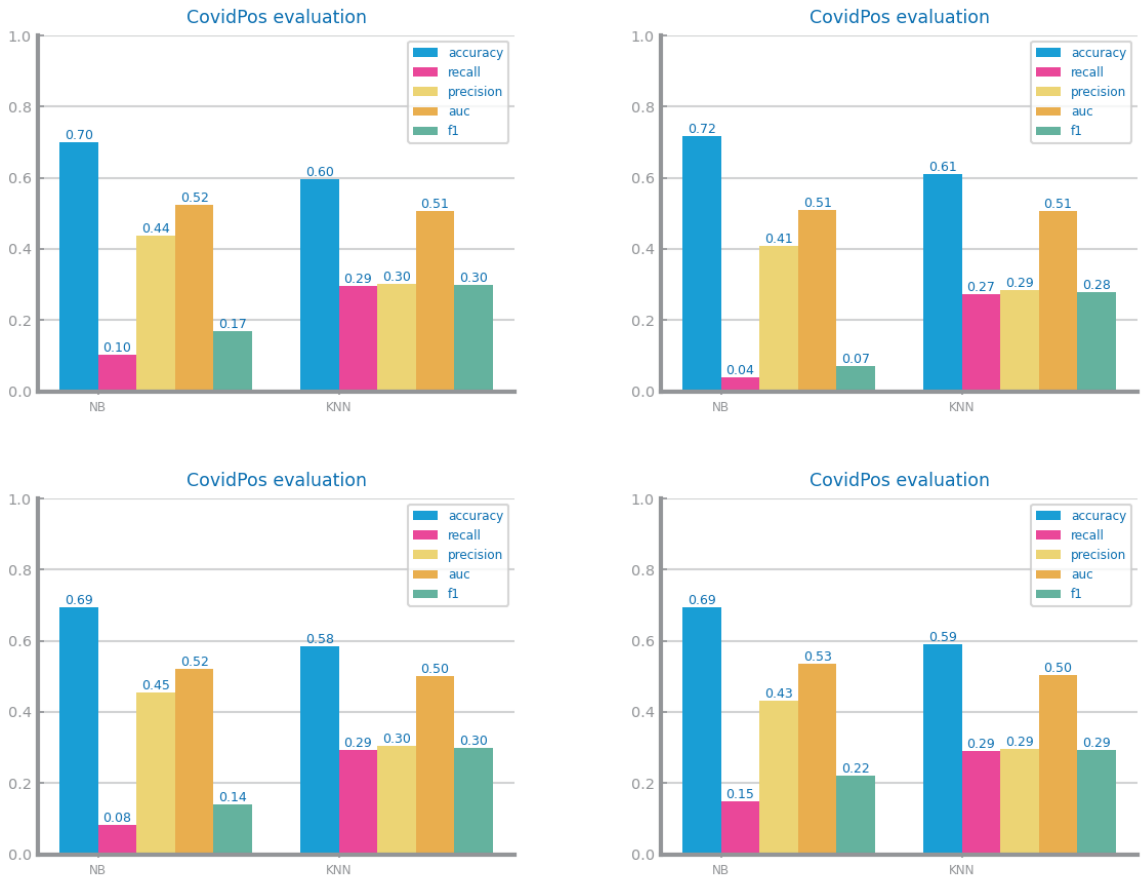


Figure 21: Outliers imputation results with different approaches for dataset 1. Rep_fixed_median (top left), rowDrop_NotStdBased (top right), rowDrop_StdBased (bottom left) and truncating_minmax (bottom right)

We applied four different strategies: replacing, truncating or dropping (either std based or iqr based) outliers. Best option was truncating minmax, looking at the recall metric.

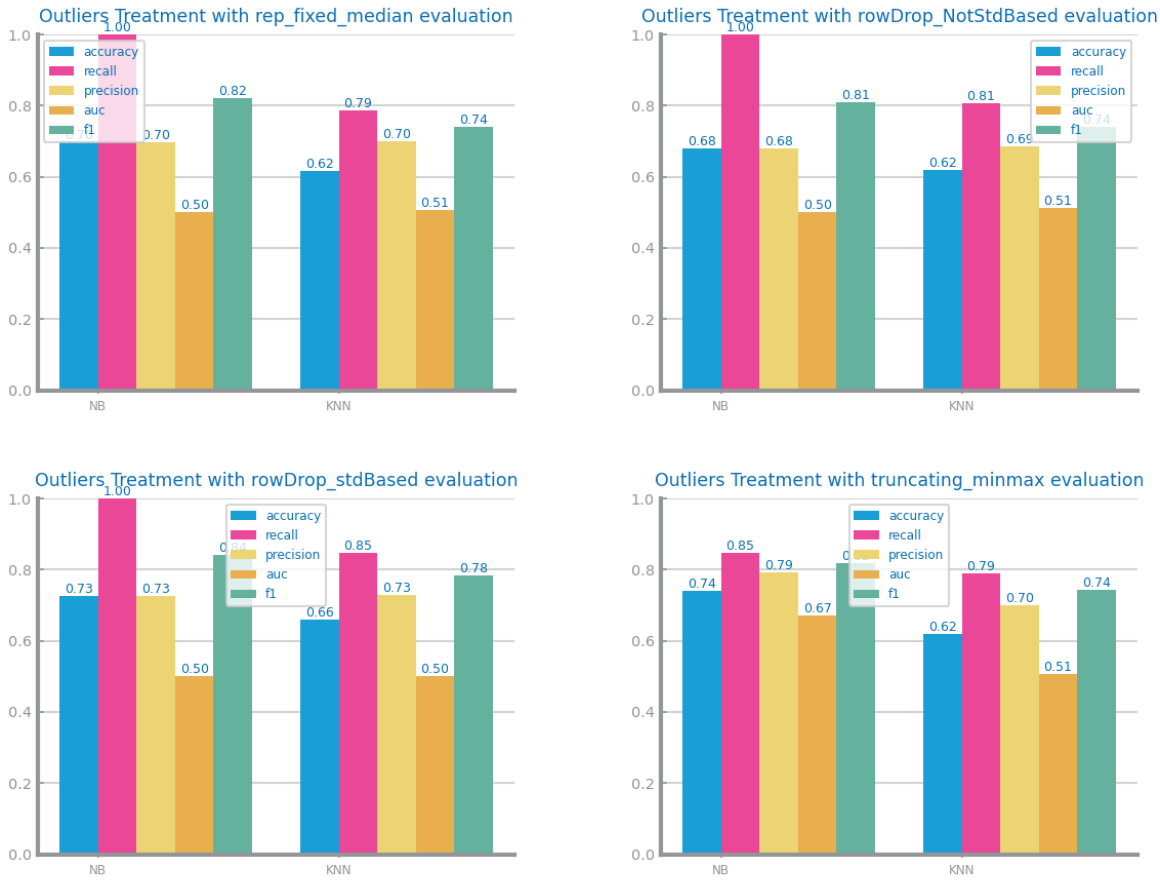


Figure 22: Outliers imputation results with different approaches for dataset 2

Same four strategies as dataset 1. They all attained similar results but the best evaluation values were the result of dropping std based (mainly looking at accuracy), therefore this was our choice moving forward.

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.**

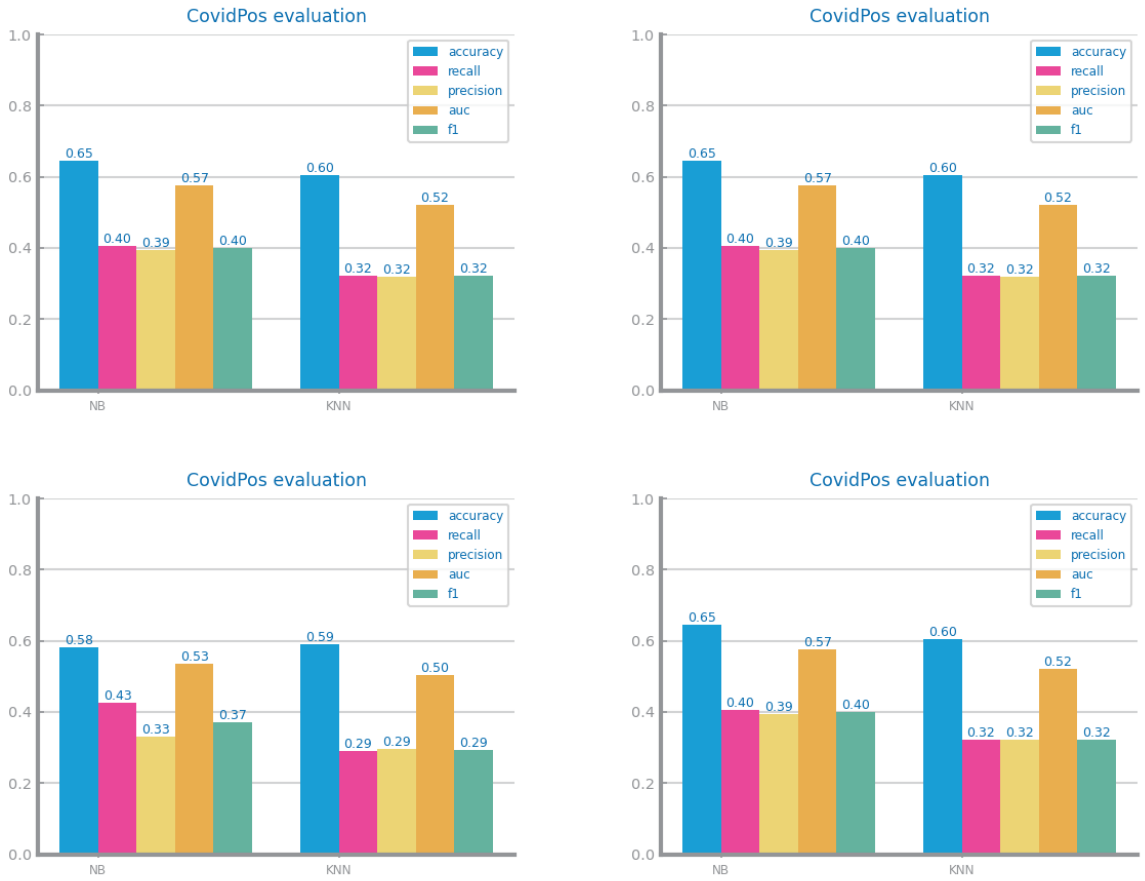


Figure 23: Scaling results with different approaches for dataset 1. MinMax[0,1] (top left), MinMax[0,10] (top right), Original (bottom left), Z-Score(bottom right)

Chosen strategy was MinMax from 0-1.

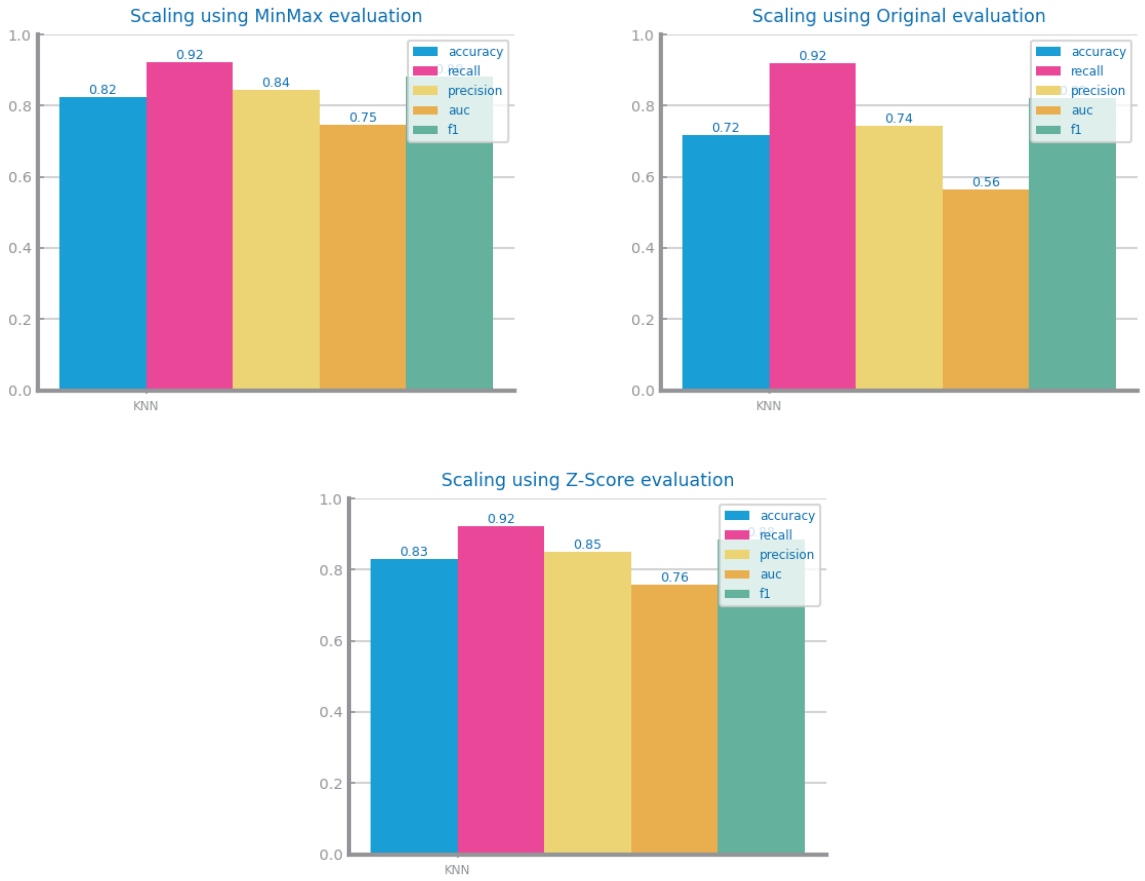


Figure 24: Scaling results with different approaches for dataset 2

We compared the same approaches as in dataset 1, given that both had better accuracys than the original but similar between each other we opted to scale using Z-Score as MinMax doesn't handle outliers very well.

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.**

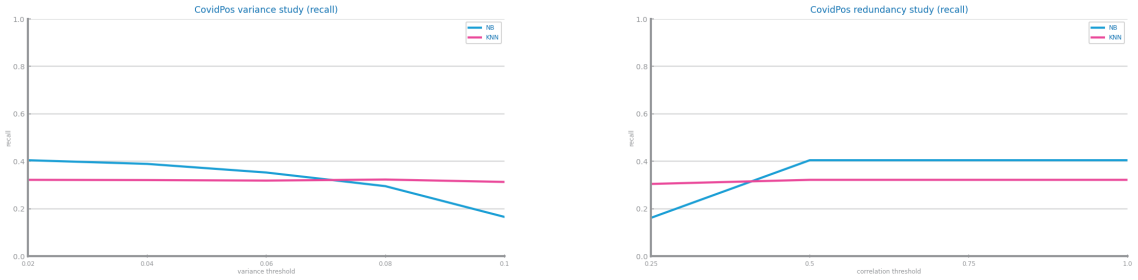


Figure 25: Feature selection of redundant variables results with different parameters for dataset 1

Recall was chosen as the metric to evaluate this dataset due to its domain. Drop all the variables with variance below 0.02 and all that have a correlation bigger than 0.5.

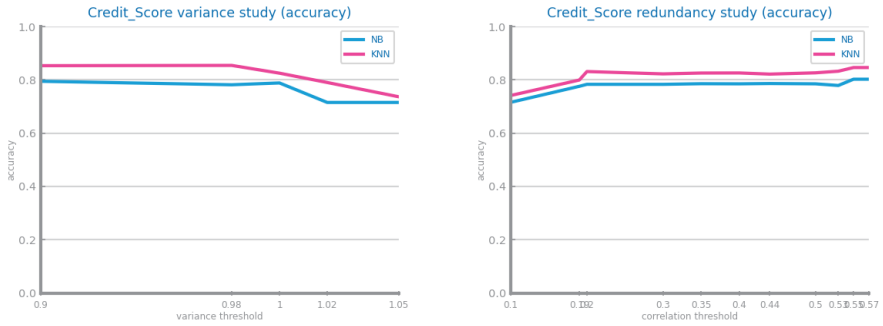


Figure 26: Feature selection of redundant variables results with different parameters for dataset 2

No variable is to be dropped due to low or high variance. (This is to be expected as we chose Z-Score scaling and all variables have the same variance of 1) In terms of redundacy, the best result is to drop no variables as well (correlation threshold of 0.57).

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.**

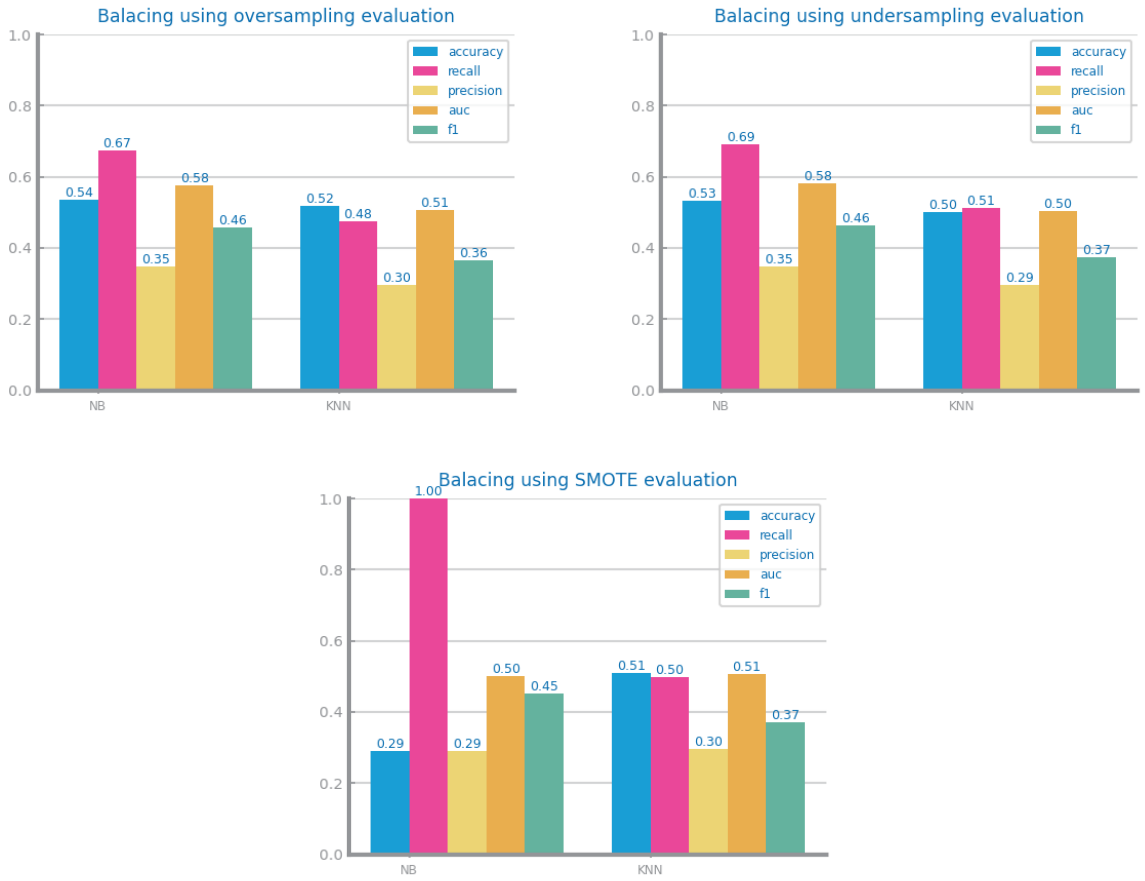


Figure 27: Balancing results with different approaches for dataset 1

For dataset 1, the SMOTE presented interesting results for recall, but the model was evidently flawed. So we chose to balance the data with undersampling, as the set contains a lot of noise (at least 16% of the records have an outlier value).

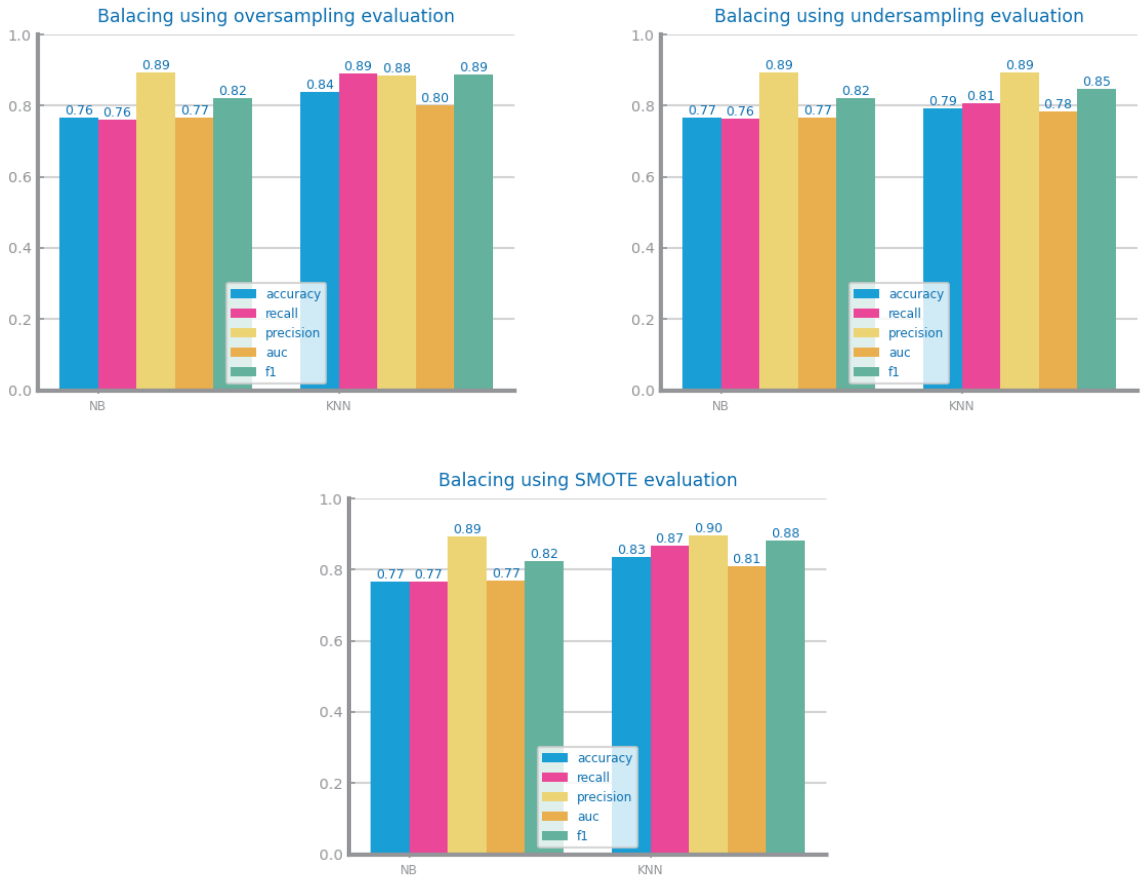


Figure 28: Balancing results with different approaches for dataset 2

We chose SMOTE balancing transformation.

3 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.**

For metric, we pursued recall for the dataset 1, since this means the system values more avoiding FN and maximizing TP. This is ideal because, as we are talking about health, it is always better to be safe than sorry.

For the dataset 2, we chose accuracy as the statistic, because, in this case, it makes sense to have system where we aim to get maximize TP and TN classifications, since a bad decision has a very high cost for the business.

In both datasets we used the hold-out strategy for training.

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.**

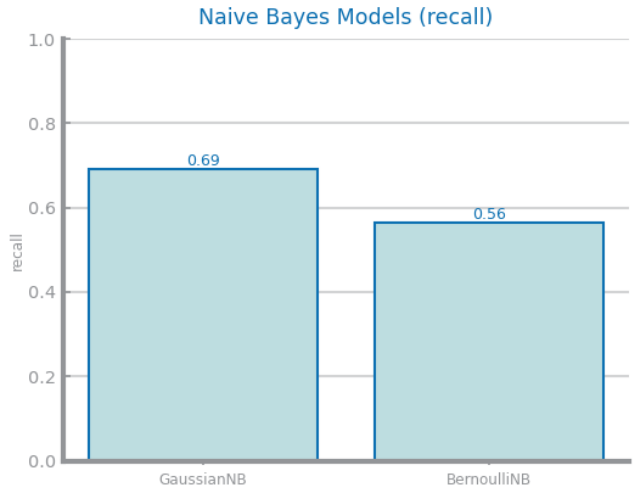


Figure 29: Naïve Bayes alternatives comparison for dataset 1

For the dataset 1, the Gaussian NB presents the better value, since the majority of numerical variables follow a normal distribution.

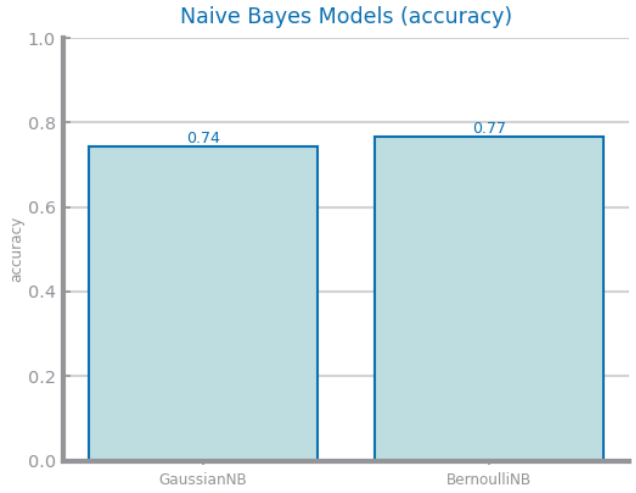


Figure 30: Naïve Bayes alternative comparison for dataset 2

For the dataset 2, the Bernoulli NB provides the best results, as the dataset is pretty sparse overall, thus making it harder to follow a normal distribution.

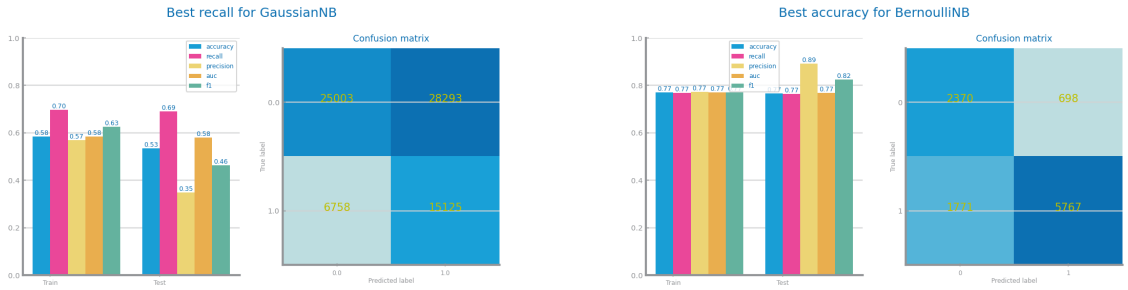


Figure 31: Naïve Bayes best model results for dataset 1 (left) and dataset 2 (right)

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.**

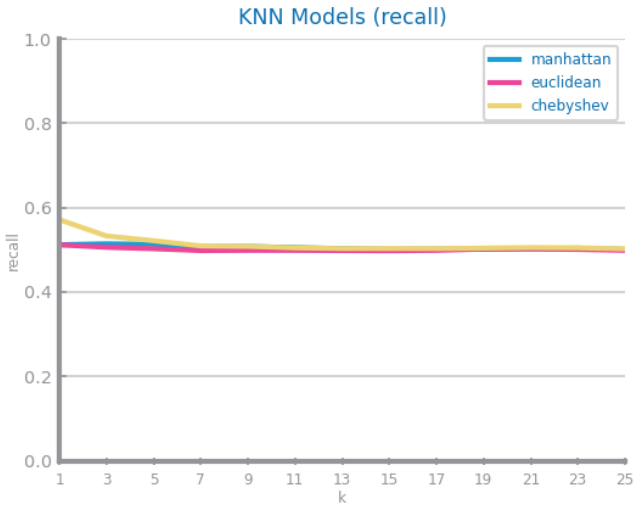


Figure 32: KNN different parameterisations comparison for dataset 1

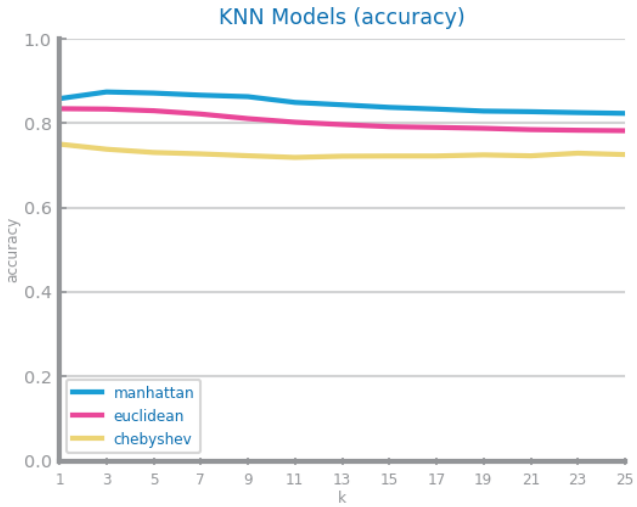


Figure 33: KNN different parameterisations comparison for dataset 2

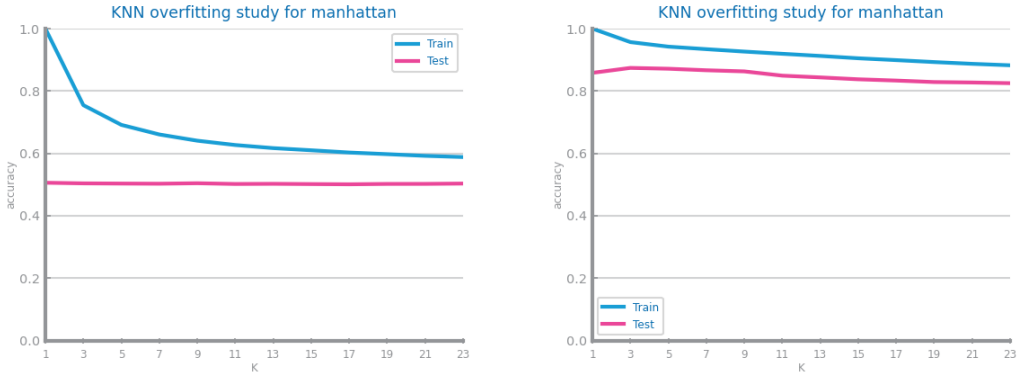


Figure 34: KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

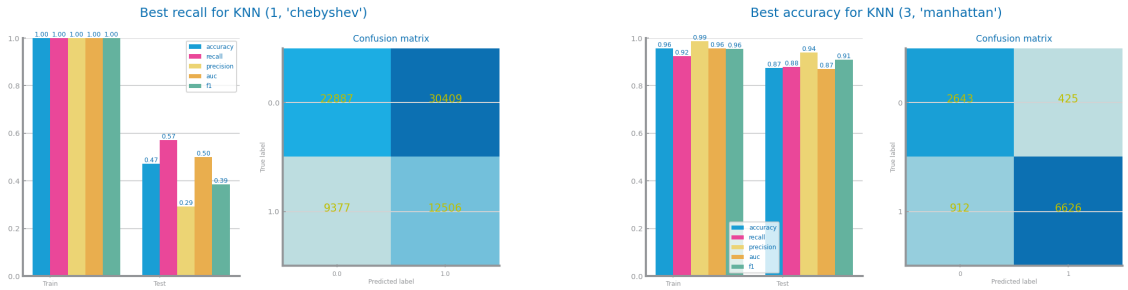


Figure 35: KNN best model results for dataset 1 (left) and dataset 2 (right)

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.**

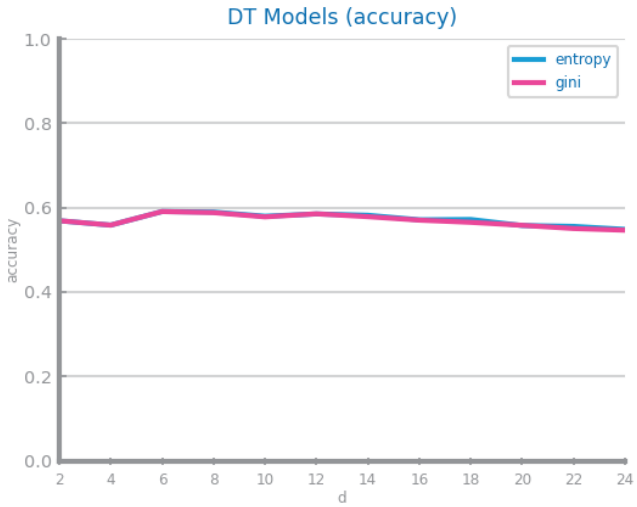


Figure 36: Decision Trees different parameterisations comparison for dataset 1

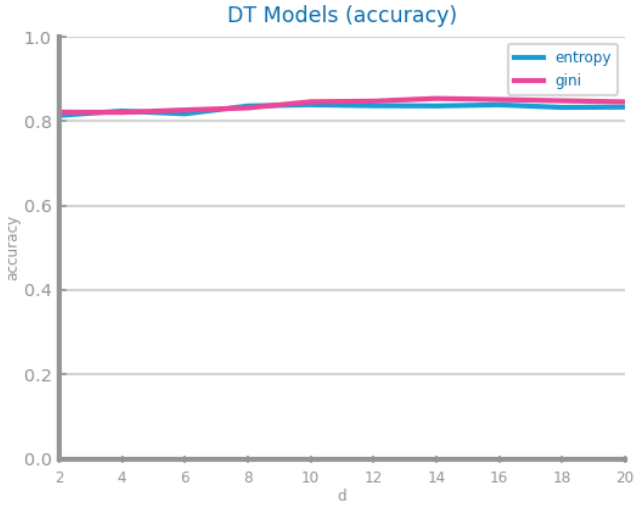


Figure 37: Decision Trees different parameterisations comparison for dataset 2

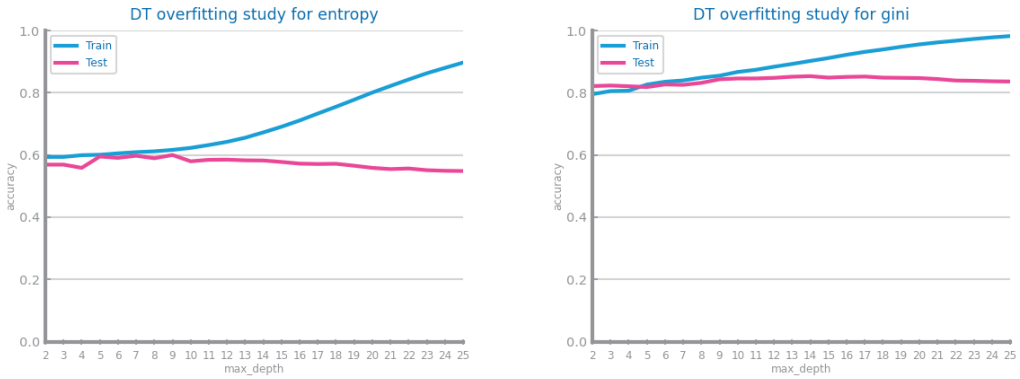


Figure 38: Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

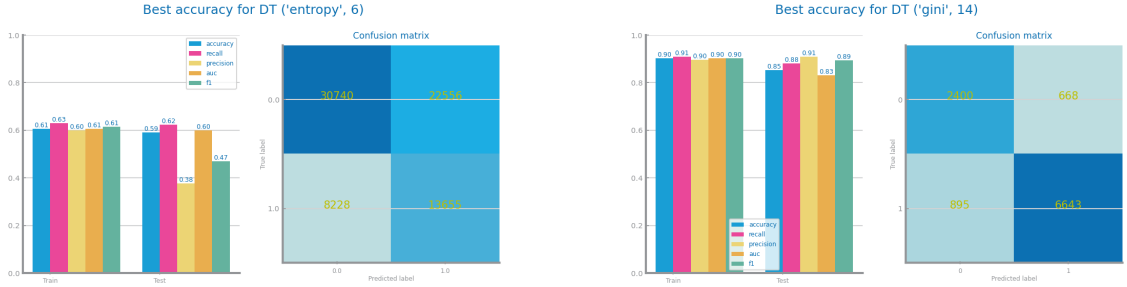


Figure 39: Decision trees best model results for dataset 1 (left) and dataset 2 (right)

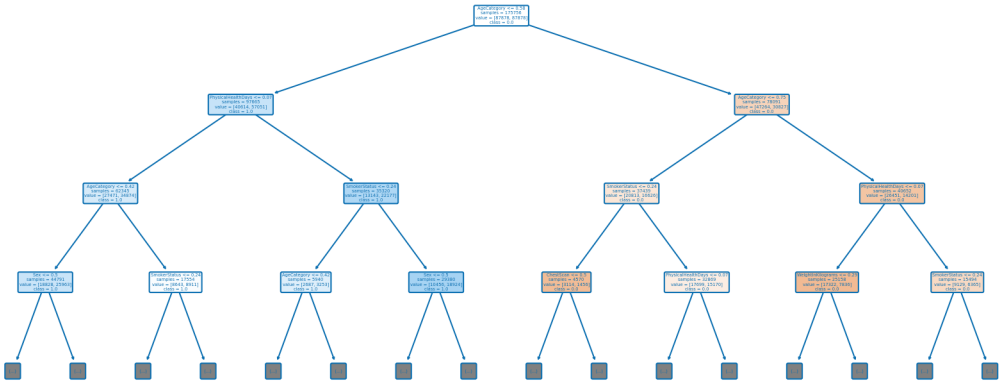


Figure 40: Best tree for dataset 1

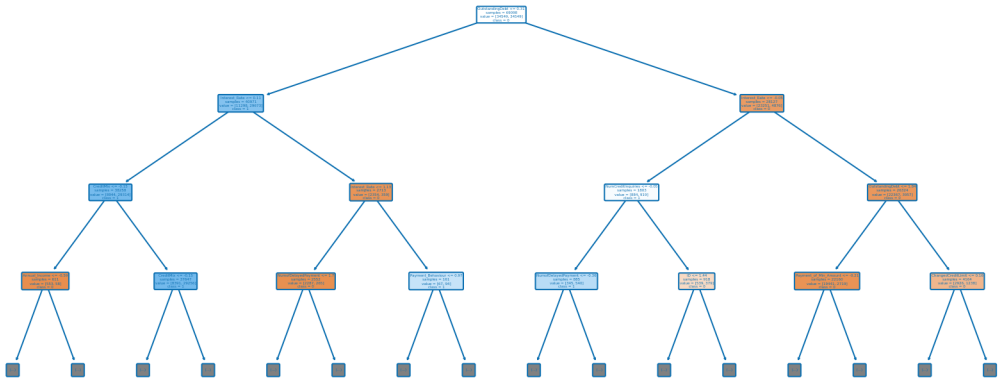
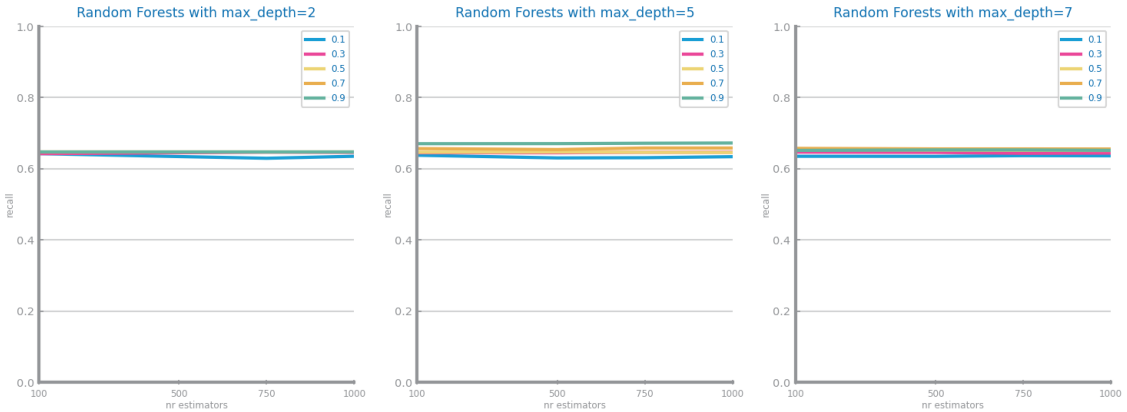
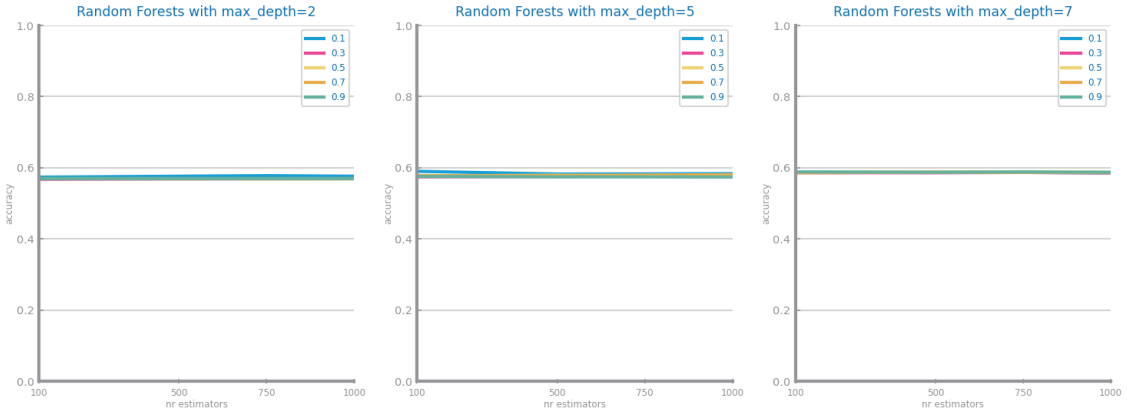


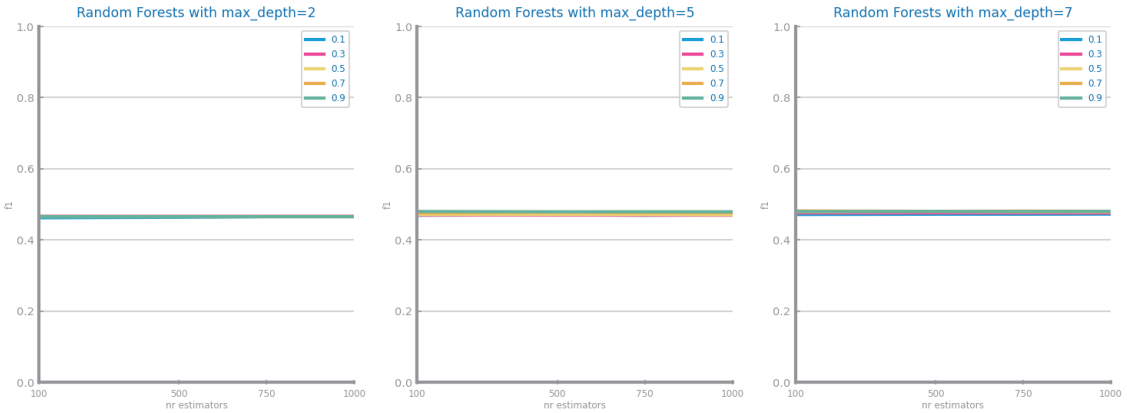
Figure 41: Best tree for dataset 2

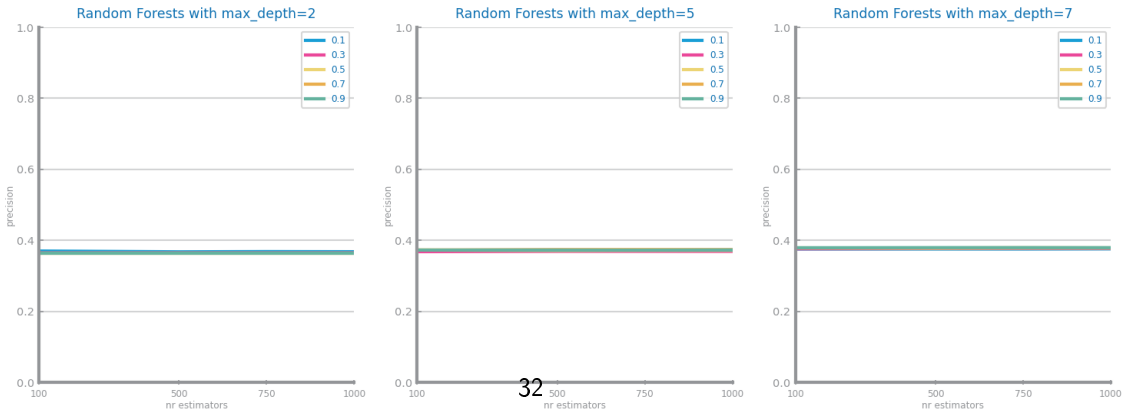
Random Forests

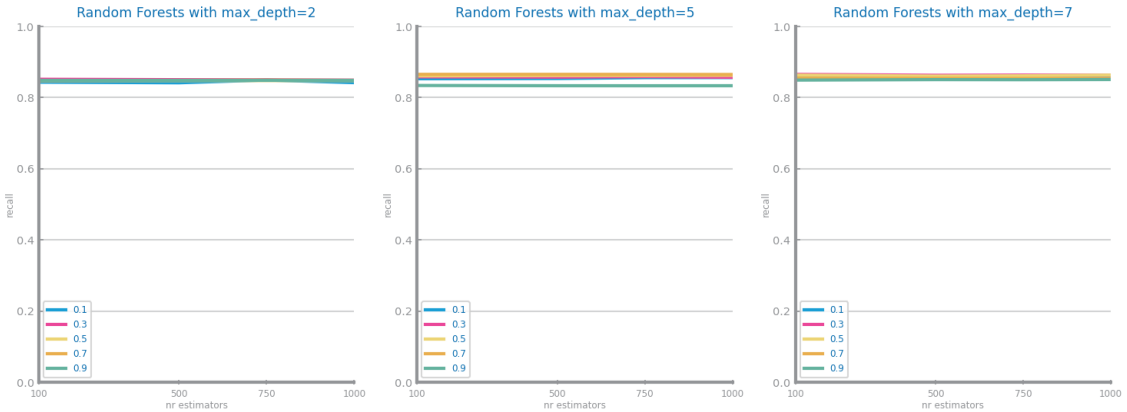
Shall be used to present the results achieved through different parameterisations for the train of random forests. The results shall be compared and explanations for them shall be presented. Shall be used to address the overfitting phenomenon, studying the conditions under which models face it. Shall 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.**

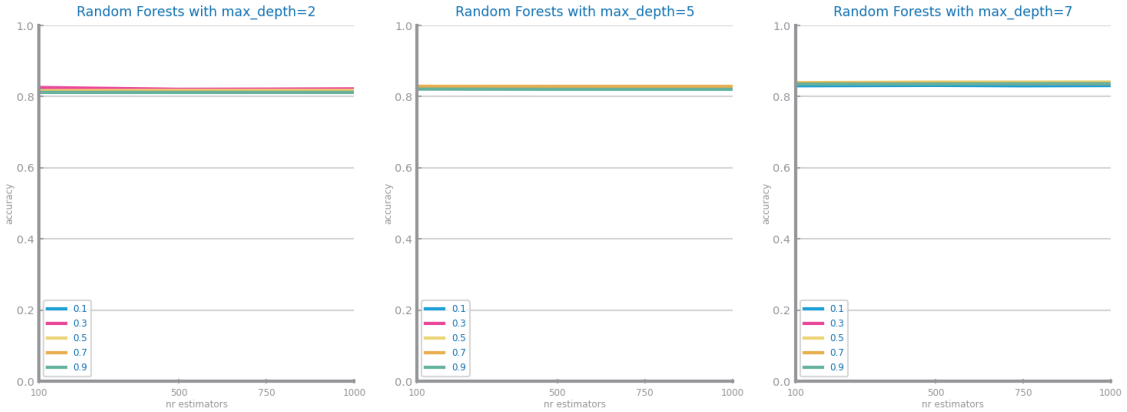

Random Forests with max_depth=2
Random Forests with max_depth=5
Random Forests with max_depth=7
recall
nr estimators
0.1
0.3
0.5
0.7
0.9

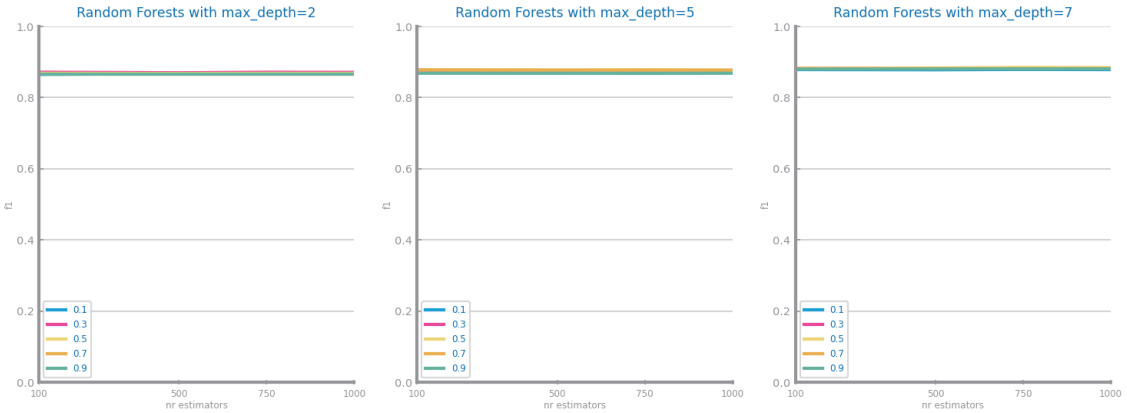

Random Forests with max_depth=2
Random Forests with max_depth=5
Random Forests with max_depth=7
accuracy
nr estimators
0.1
0.3
0.5
0.7
0.9

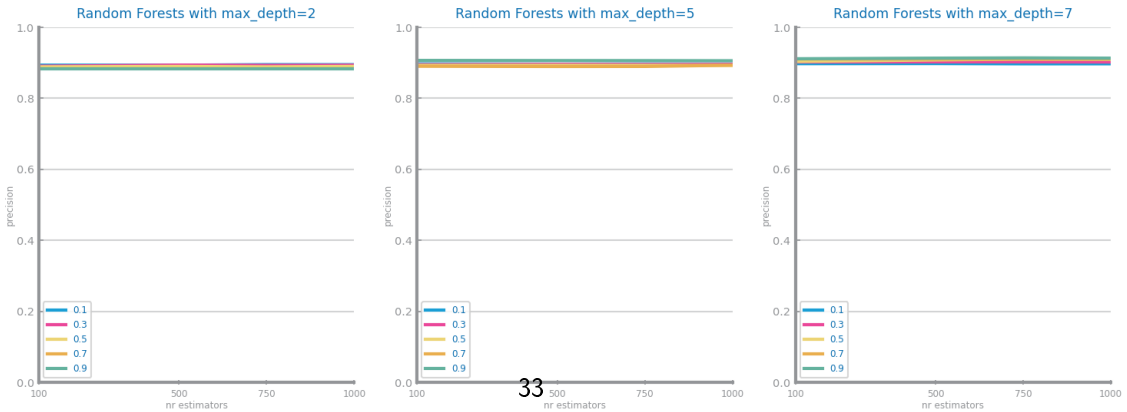

Random Forests with max_depth=2
Random Forests with max_depth=5
Random Forests with max_depth=7
f1
nr estimators
0.1
0.3
0.5
0.7
0.9


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


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

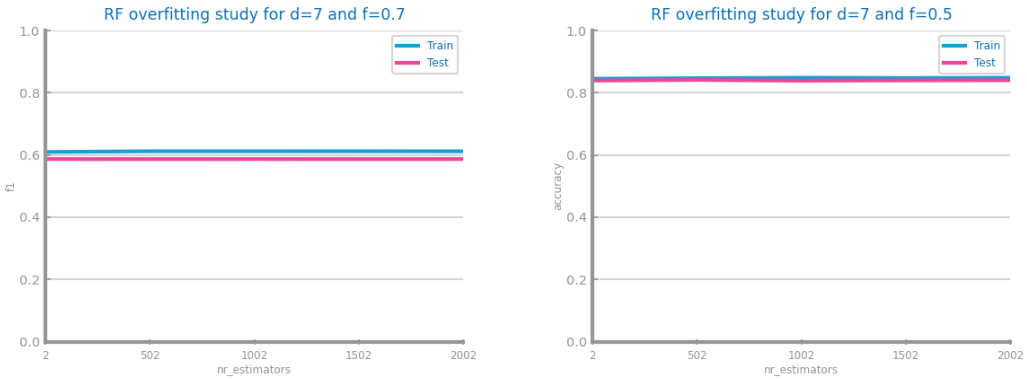


Figure 44: Random Forests overfitting analysis for dataset 1 (left) and dataset 2 (right)

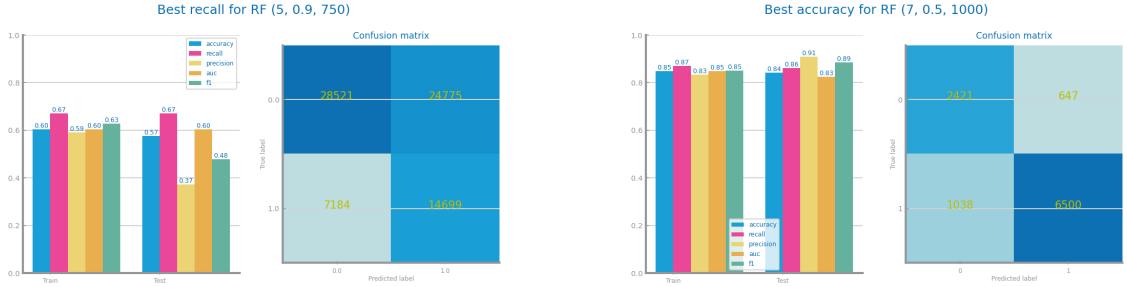


Figure 45: Random Forests best model results for dataset 1 (left) and dataset 2 (right)

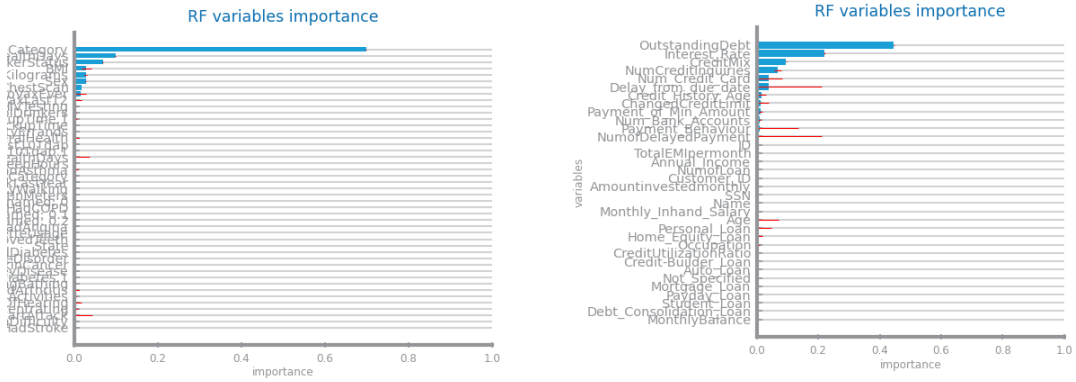
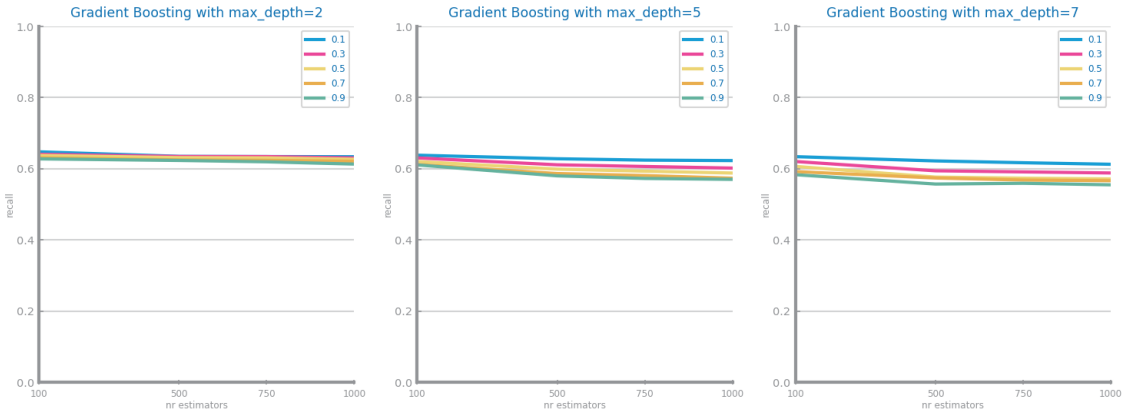
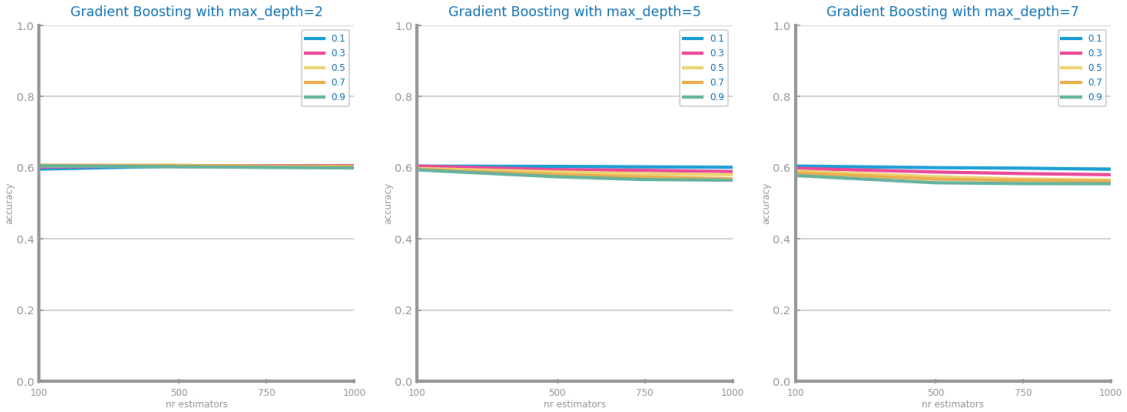


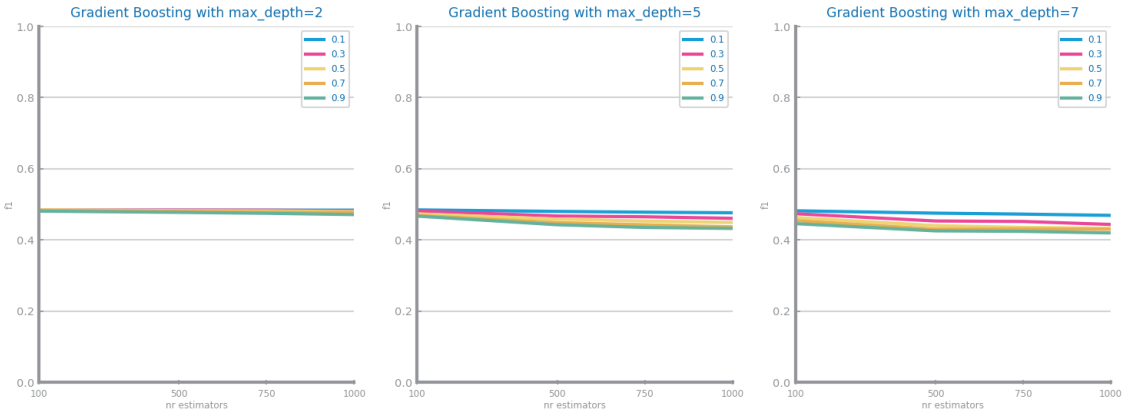
Figure 46: Random Forests variables importance for dataset 1 (left) and dataset 2 (right)

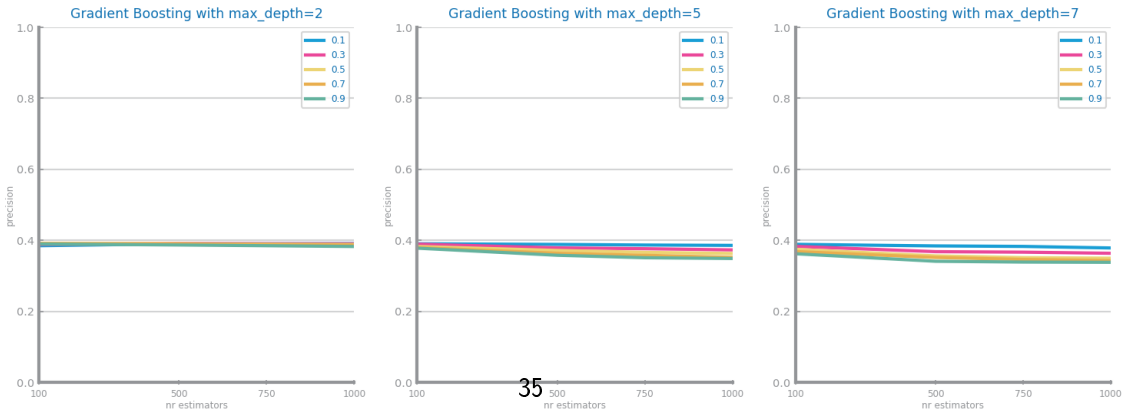
Gradient Boosting

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.**

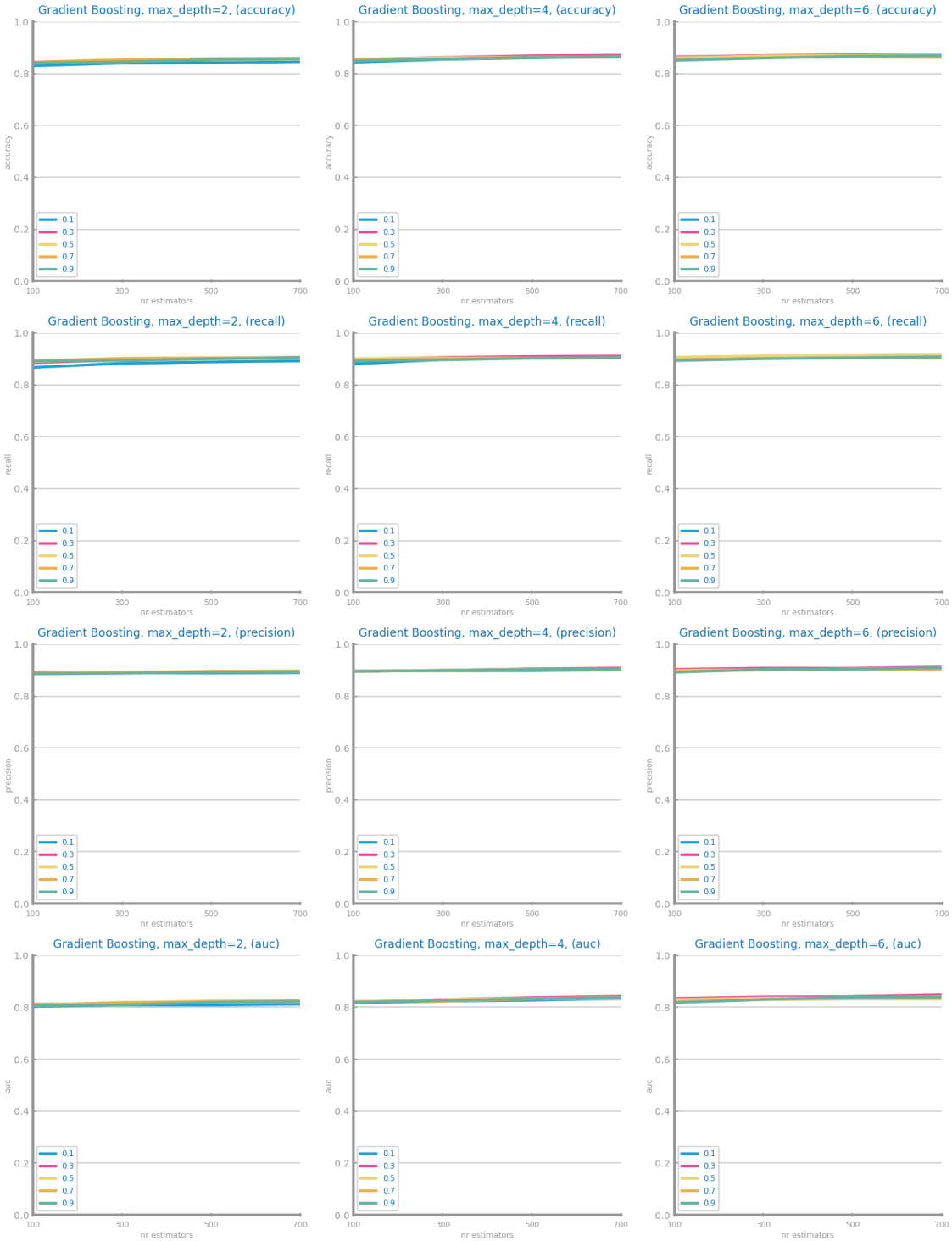

Gradient Boosting with max_depth=2
Gradient Boosting with max_depth=5
Gradient Boosting with max_depth=7
recall
nr estimators
0.1
0.3
0.5
0.7
0.9


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

Gradient Boosting study for different parameters



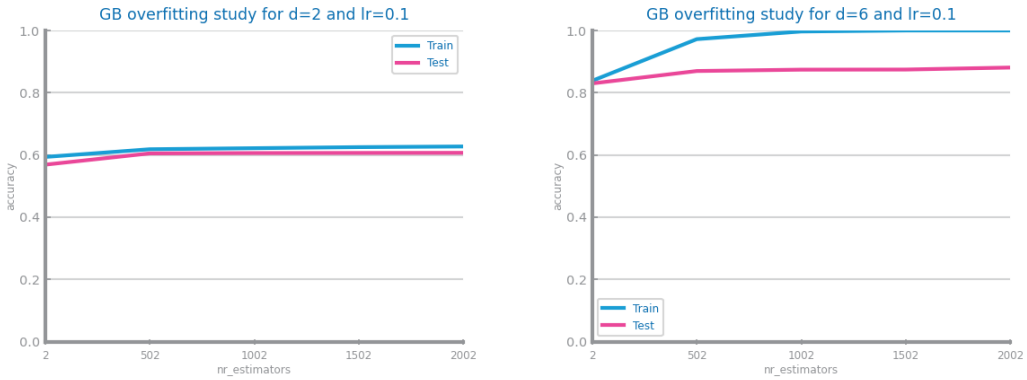


Figure 49: Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

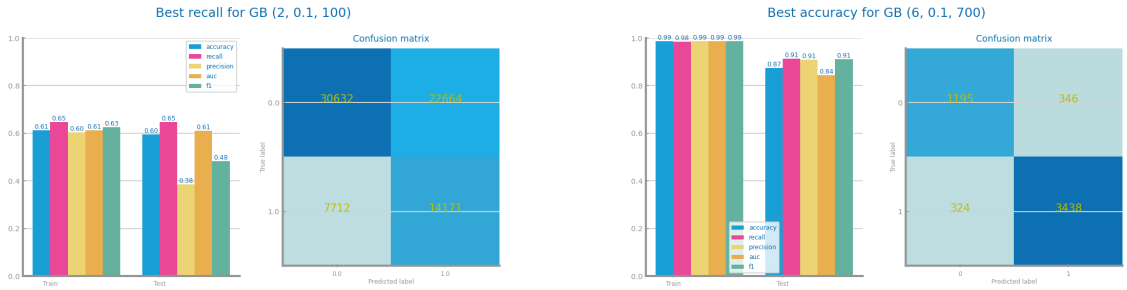


Figure 50: Gradient boosting best model results for dataset 1 (left) and dataset 2 (right)

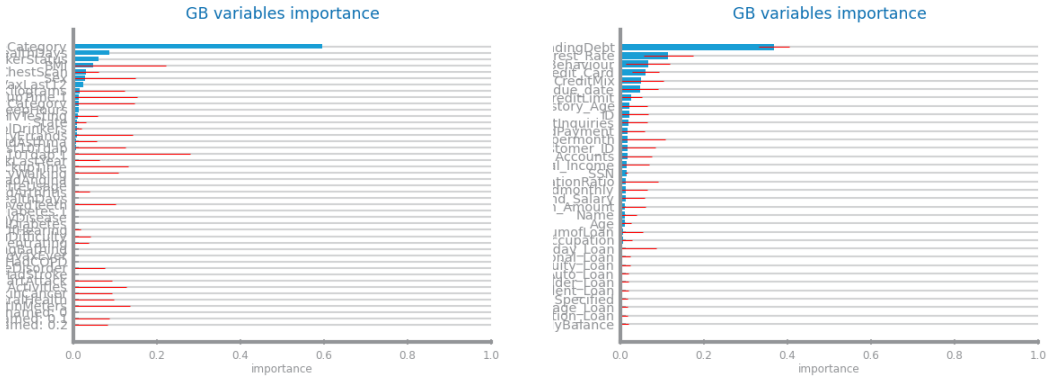


Figure 51: Gradient boosting variables importance for dataset 1 (left) and dataset 2 (right)

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.**


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


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

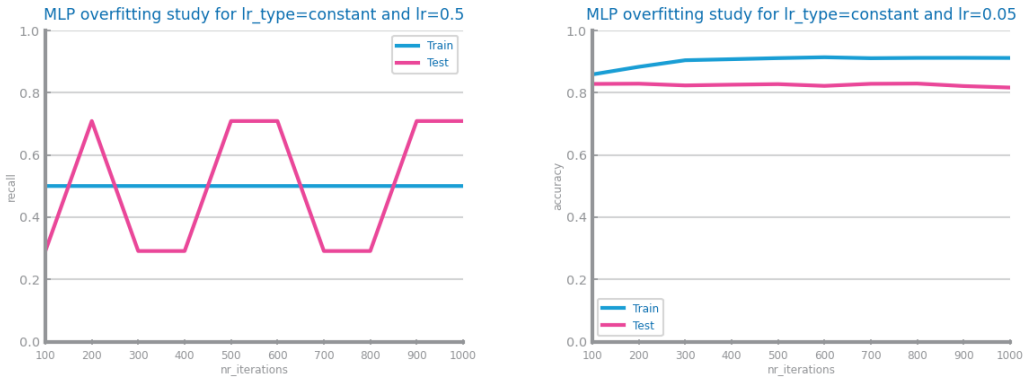


Figure 54: MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

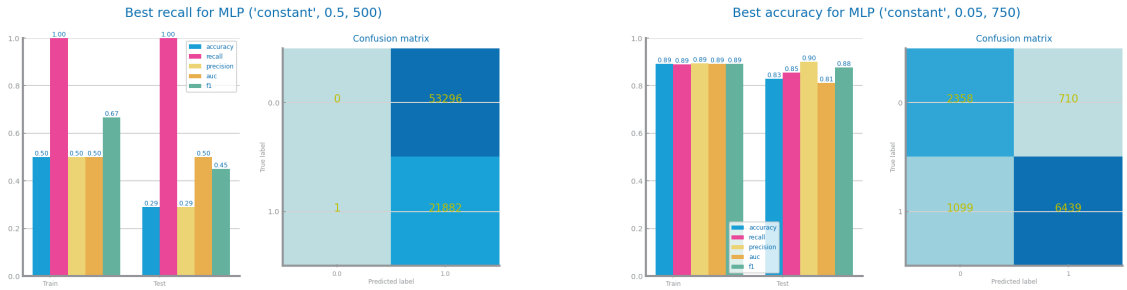


Figure 55: MLP best model results for dataset 1 (left) and dataset 2 (right)

4 CRITICAL ANALYSIS

Shall 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

We used the "sum" function as "agg_fun" for both datasets. We studied the granularity at three different levels, for dataset 1, weekly (atomic), monthly and quarterly with an upwards trend but no seasonality or cyclical behaviour. For dataset 2 by 15 minutes (atomic), hourly and daily, with no visible trend, daily seasonality on each morning and evening

and also weekly cyclicality with a busier day a week (usually Mondays except the first spike) both corresponding to heavier traffic flows. **Shall not exceed 500 characters.**

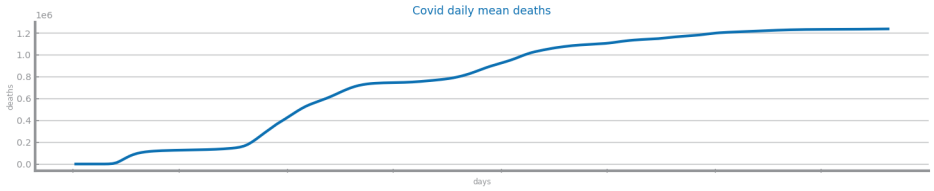


Figure 56: Time series 1 at the most granular detail

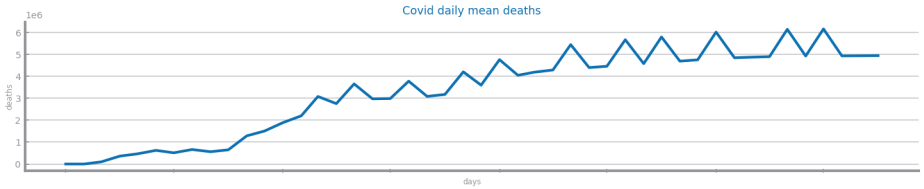


Figure 57: Time series 1 at the second chosen granularity

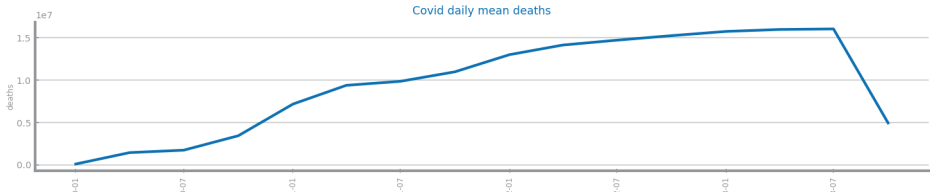


Figure 58: Time series 1 at the third chosen granularity

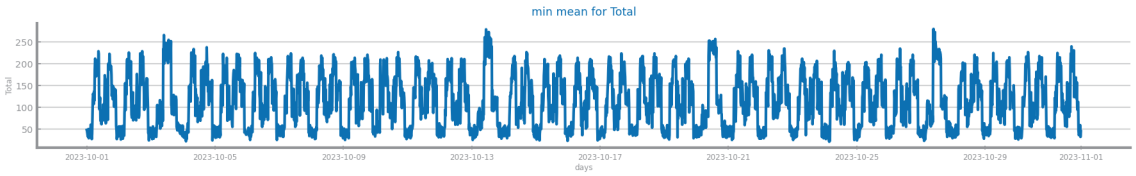


Figure 59: Time series 2 at the most granular detail

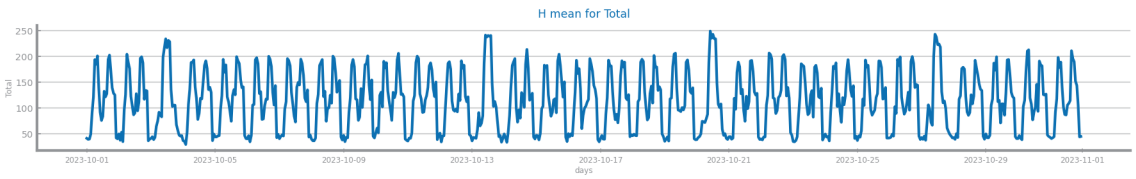


Figure 60: Time series 2 at the second chosen granularity

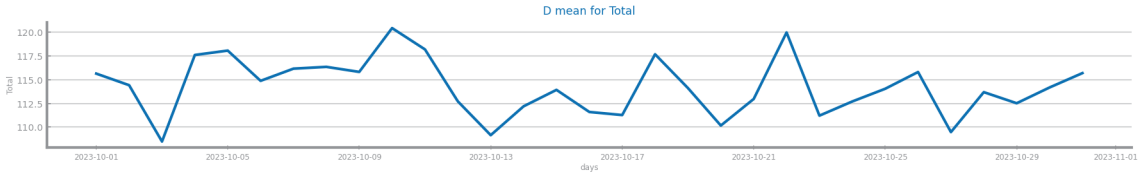
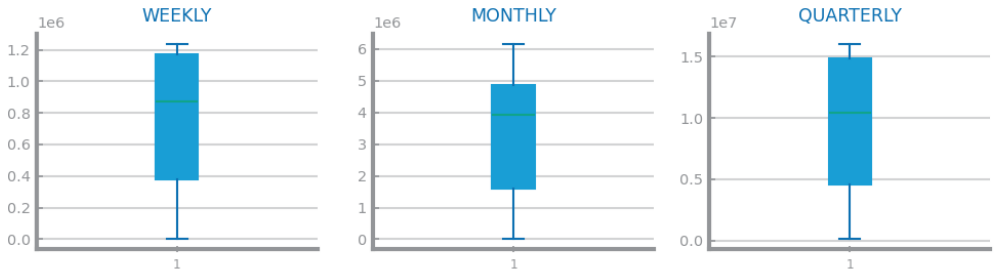


Figure 61: Time series 2 at the third chosen granularity

Data Distribution

Shall 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 (WEEKLY) | Name: deaths, dtype: float64 (MONTHLY) | Name: deaths, dtype: float64 (QUARTERLY)

Figure 62: Boxplot(s) for time series 1

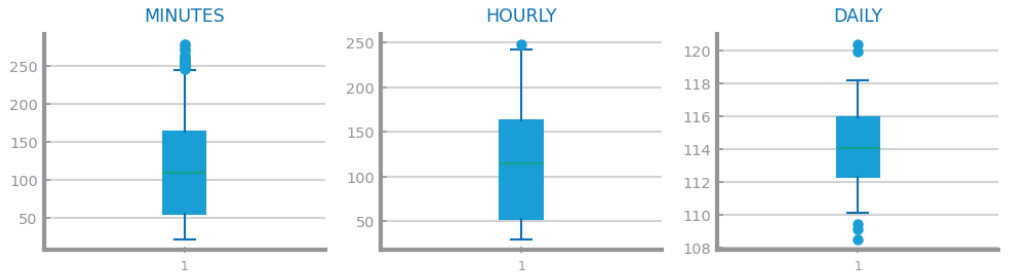


Figure 63: Boxplot(s) for time series 2

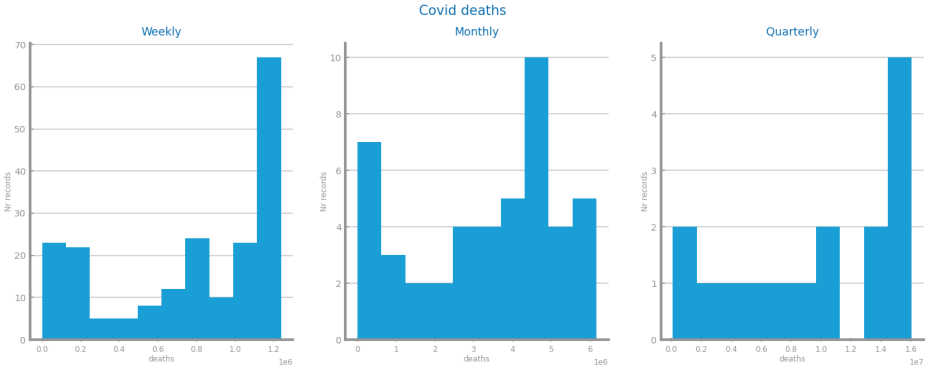


Figure 64: Histogram(s) for time series 1

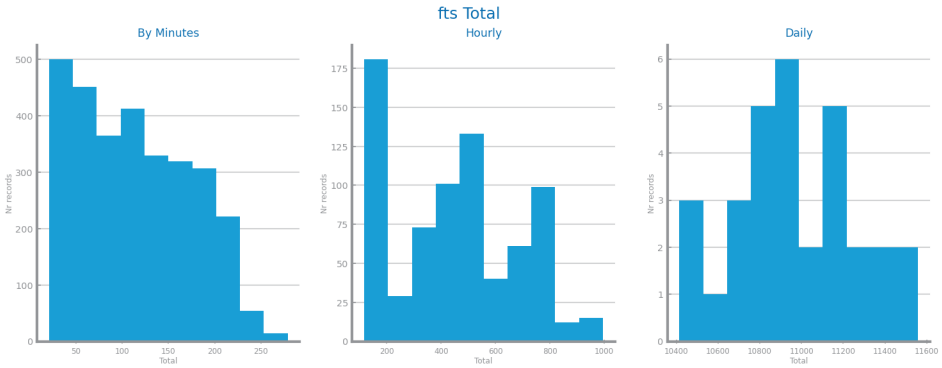


Figure 65: Histogram(s) for time series 2

Data Stationarity

For the 1st dataset we obtained a p-value of 0.223. Looking at the graphs, we can see that there is a trend in the first 2

and that there is evidence of a seasonal trend in the 3rd. For the 2nd dataset we obtained a p-value of 0. and that there is no trend but there is some seasonality. **Shall not exceed 300 characters.**

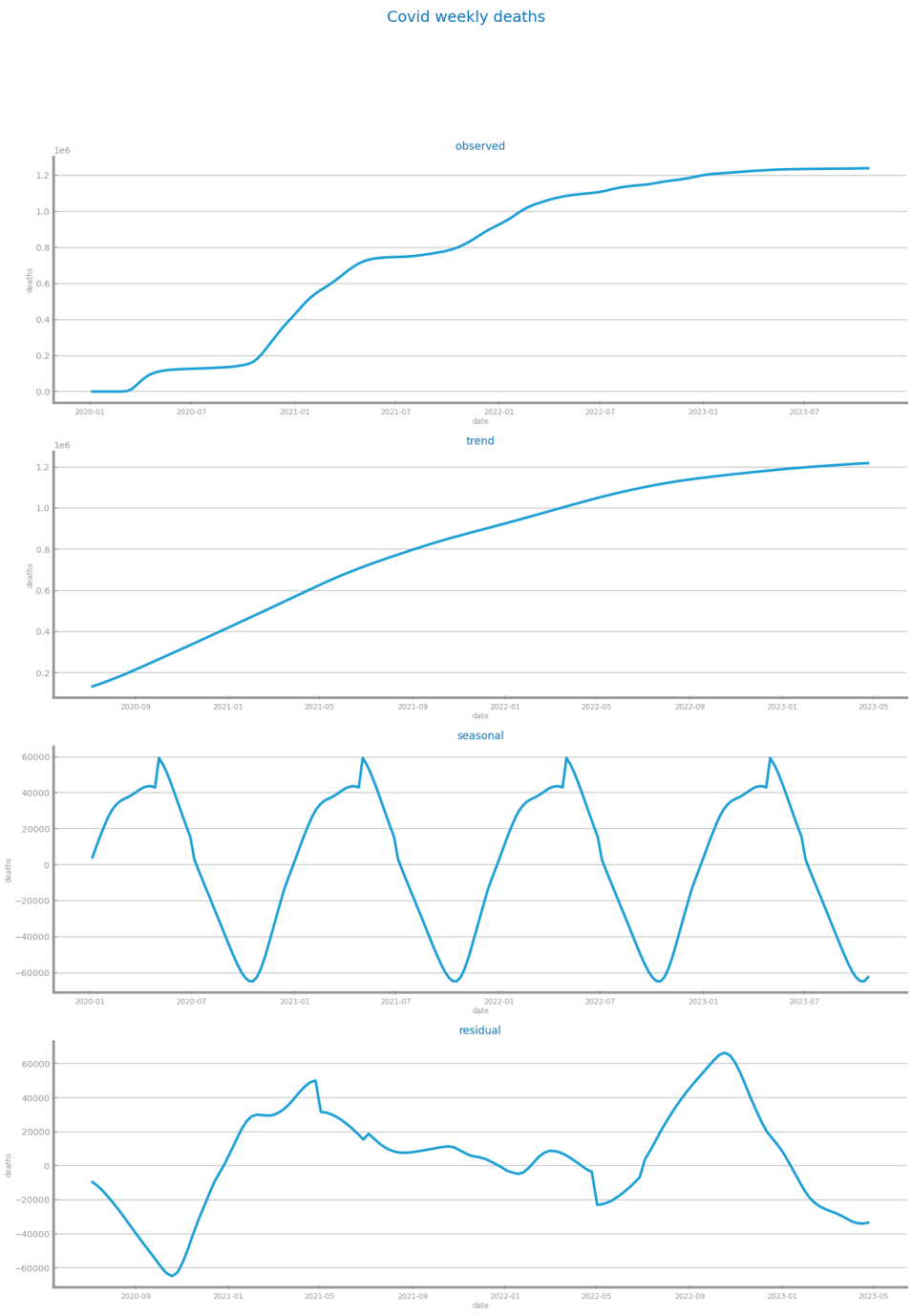


Figure 66: Components study for time series 1

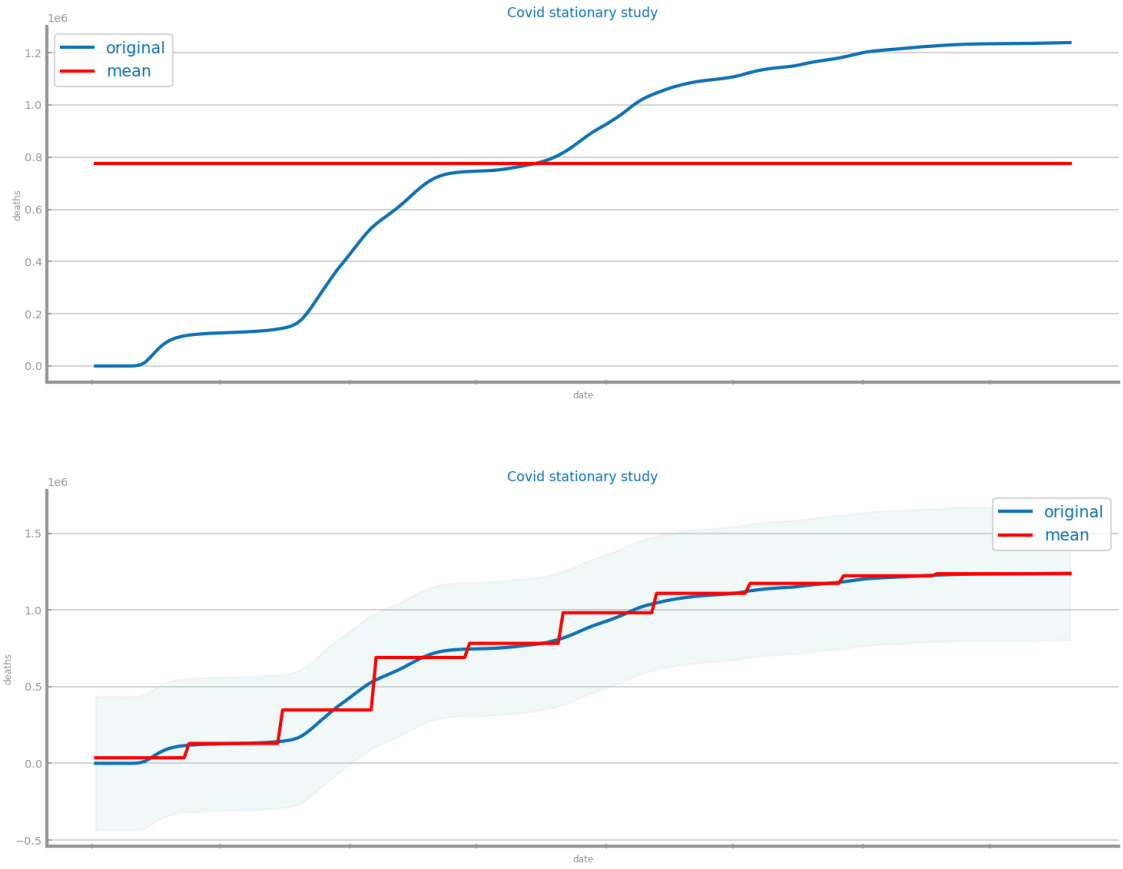


Figure 67: Stationarity study for time series 1

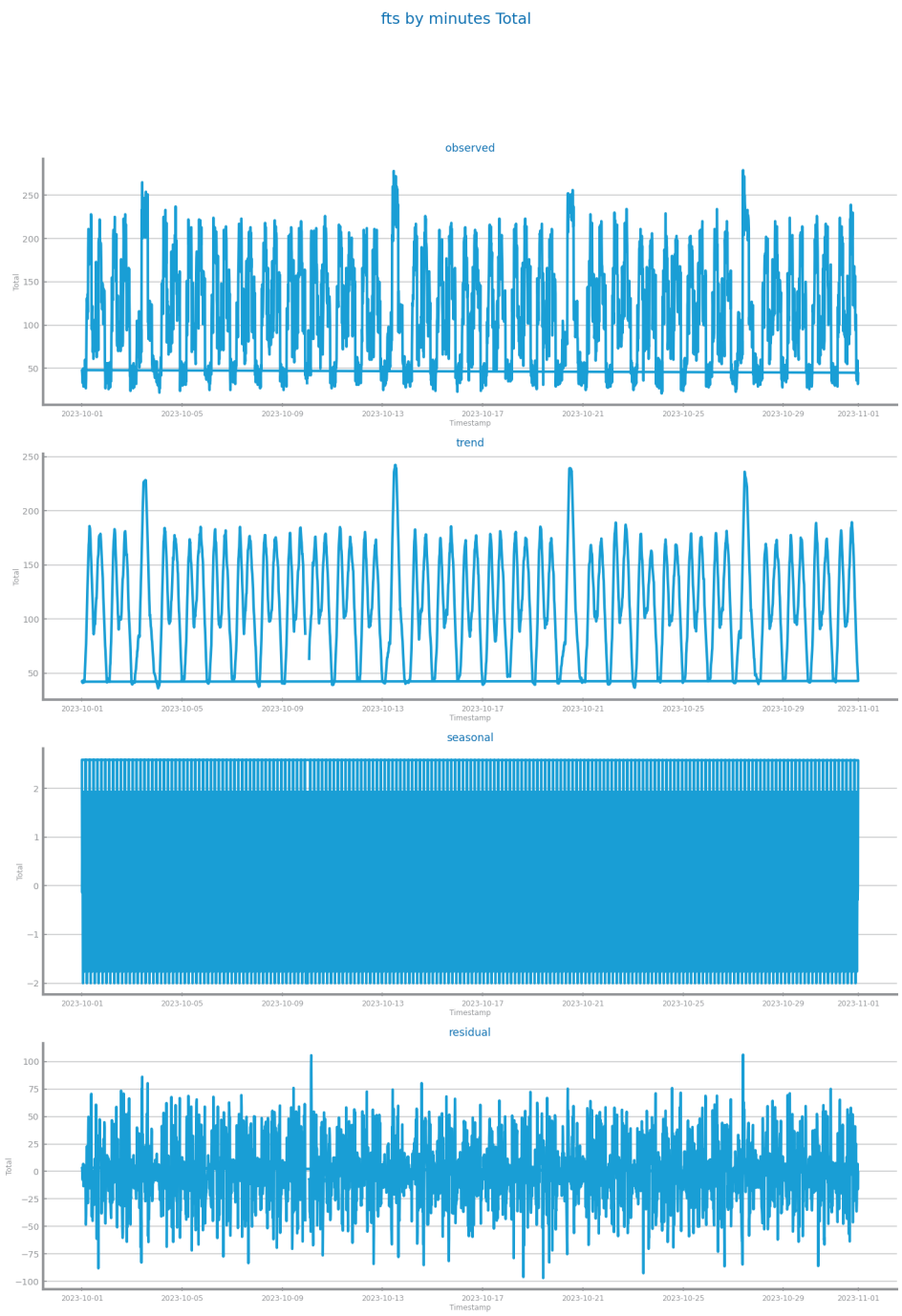


Figure 68: Components study for time series 2

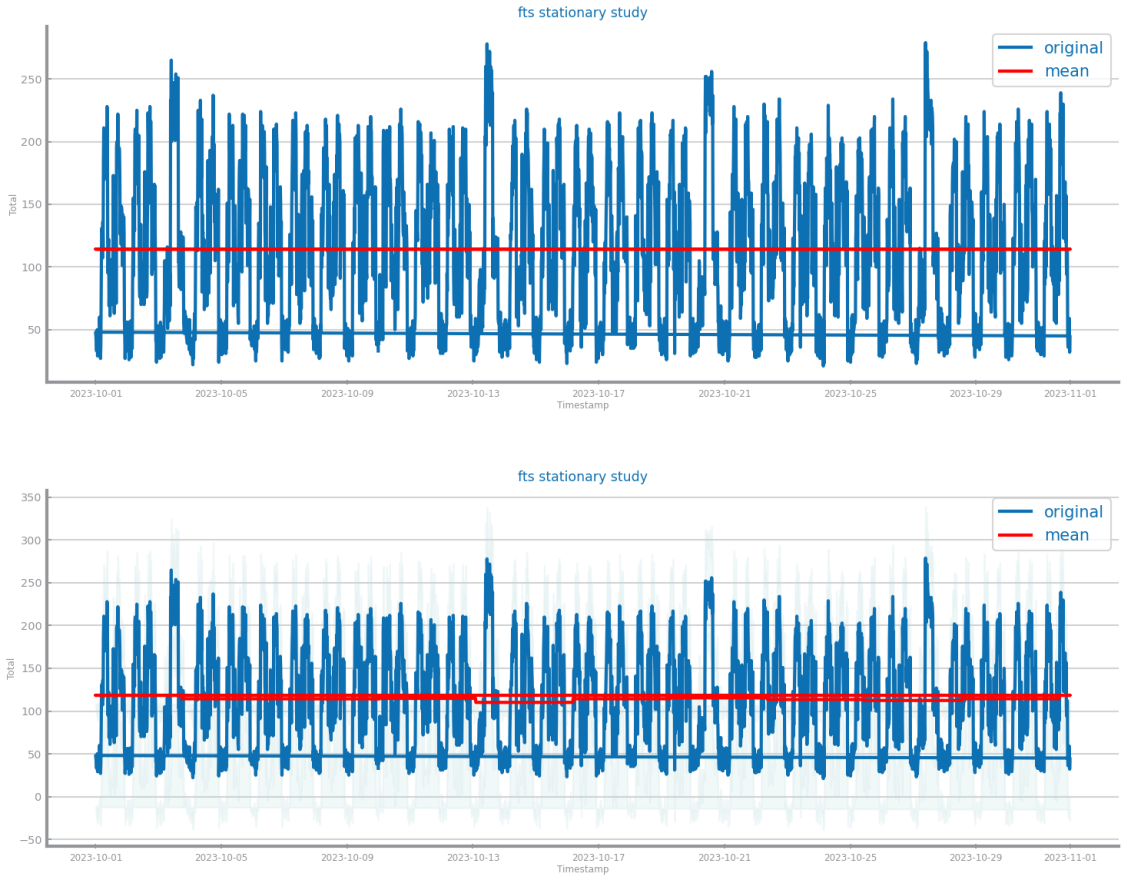


Figure 69: Stationarity study for time series 2

6 DATA TRANSFORMATION

Aggregation

To study the best aggregation, we applied the model at 3 different levels studied in Data Profiling. In dataset 1, chose the weekly aggregation and for dataset 2, hourly aggregation as they obtained lower values for the different errors, simplifying the model without losing information or context. **Shall not exceed 300 characters.**

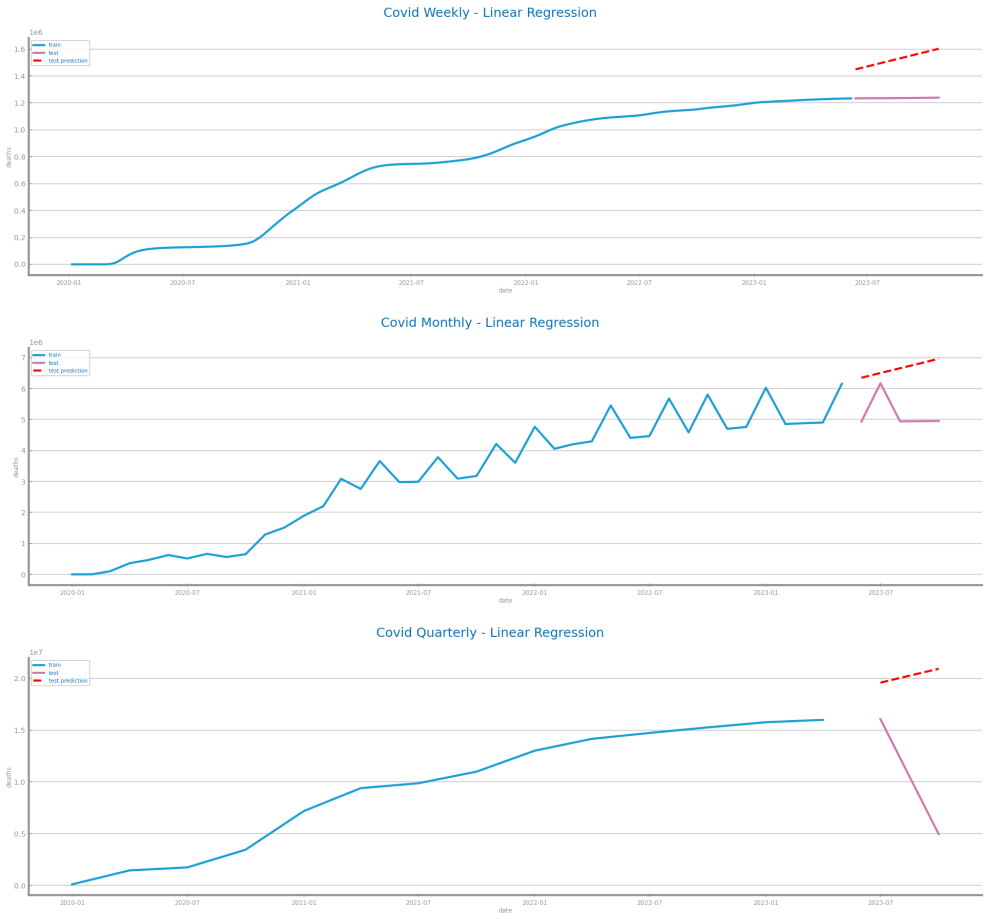


Figure 70: Forecasting plots after different aggregations on time series 1

Figure 71: Forecasting results after different aggregations on time series 1

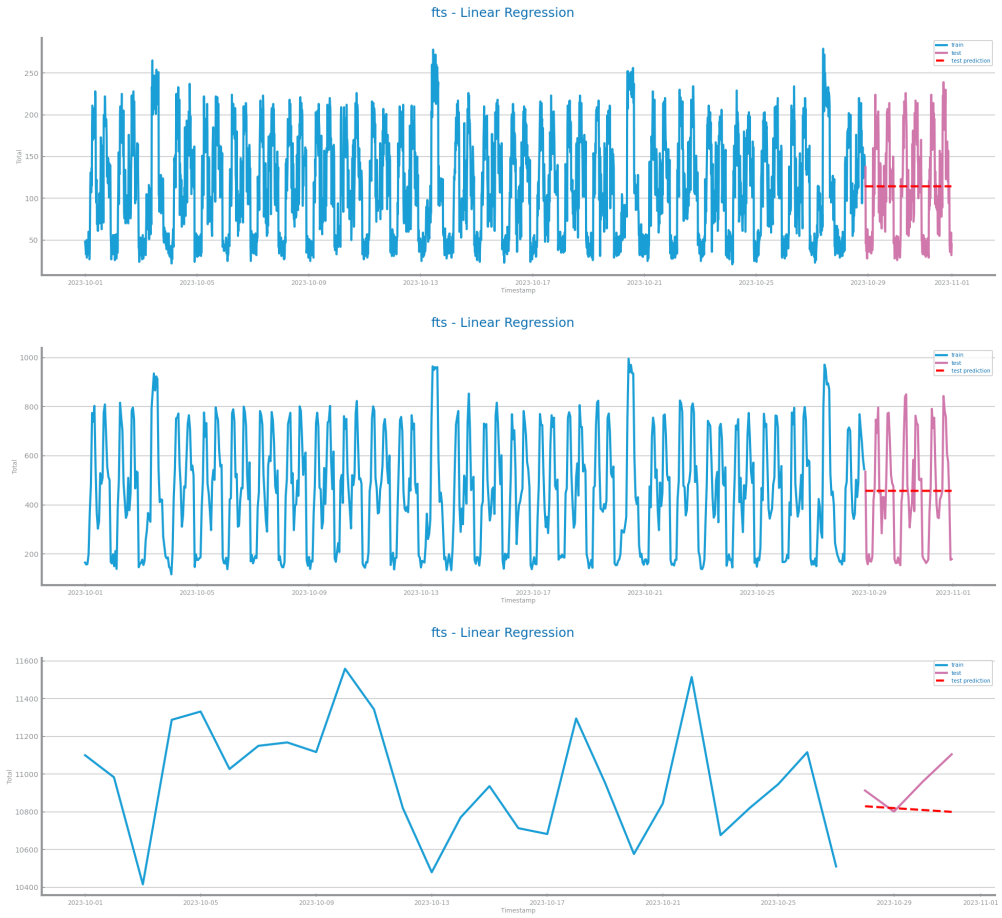


Figure 72: Forecasting plots after different aggregations on time series 2

Figure 73: Forecasting results after different aggregations on time series 2

Smoothing

To study the best Window Size, we applied the model to 4 different values (25, 50, 75 and 100). In this case we chose 100 for both datasets since it was where we obtained the lowest values for the different errors. **Shall not exceed 300 characters.**

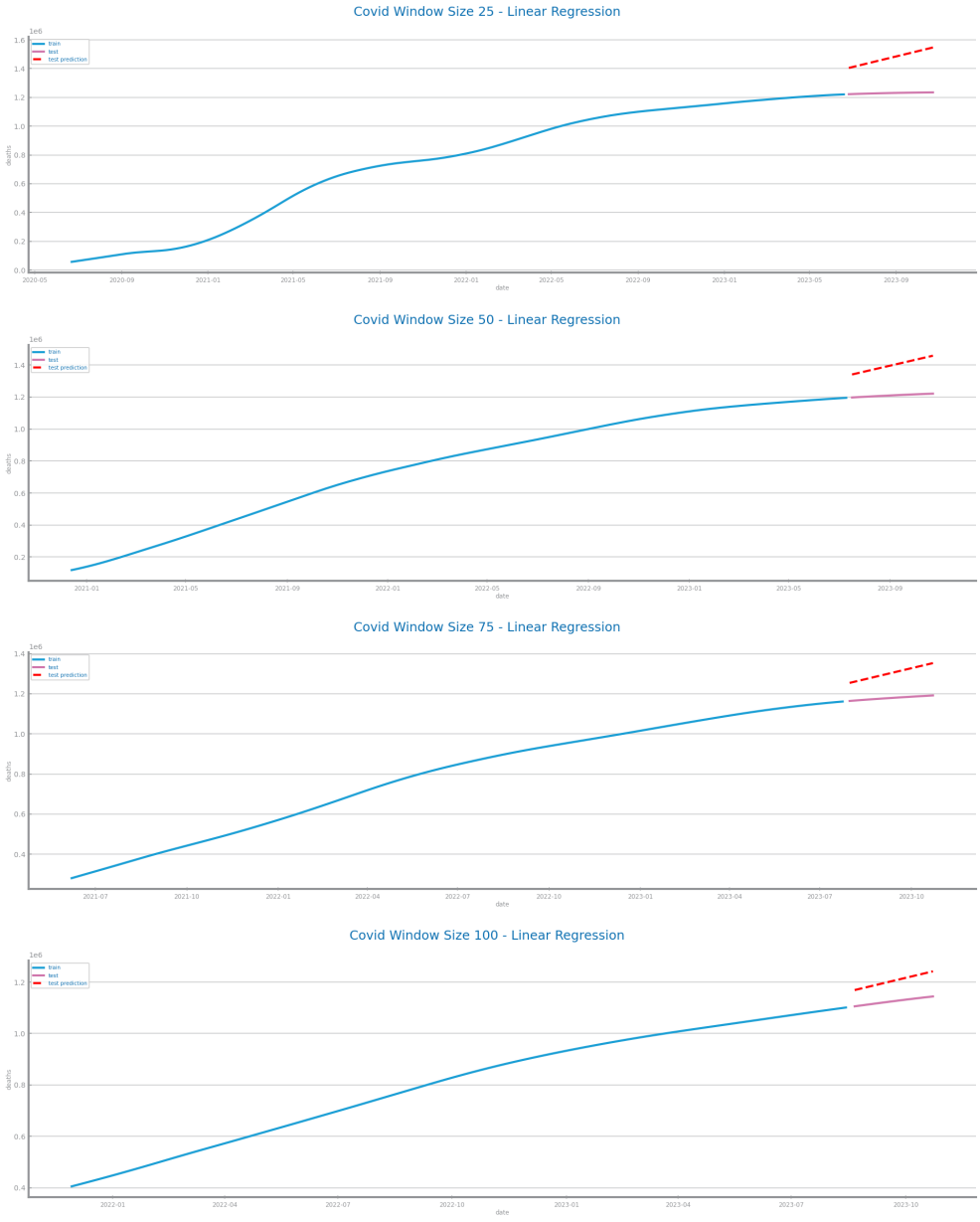


Figure 74: Forecasting plots after different smoothing parameterisations on time series 1

Figure 75: Forecasting results after different smoothing parameterisations on time series 1

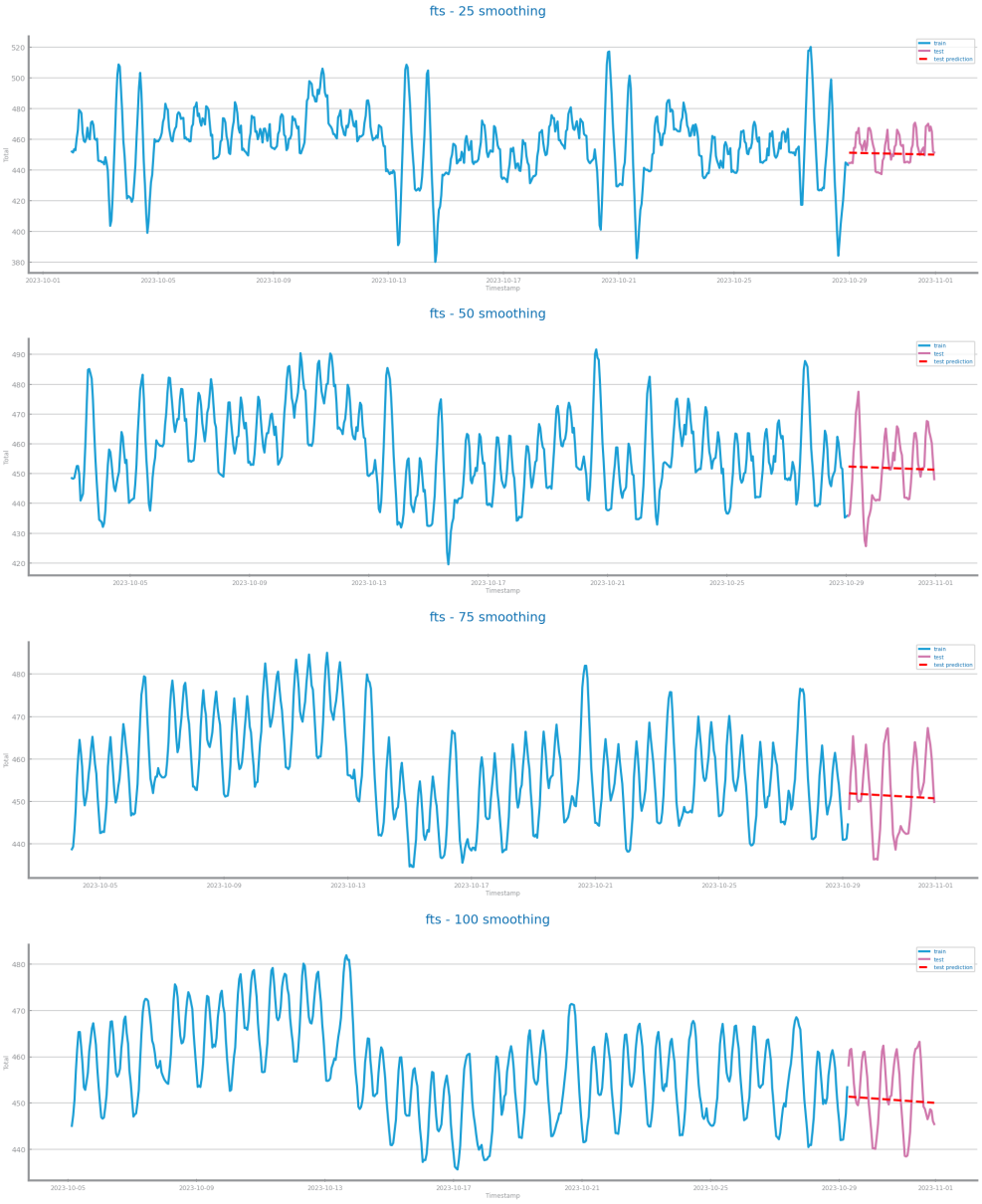


Figure 76: Forecasting plots after different smoothing parameterisations on time series 2

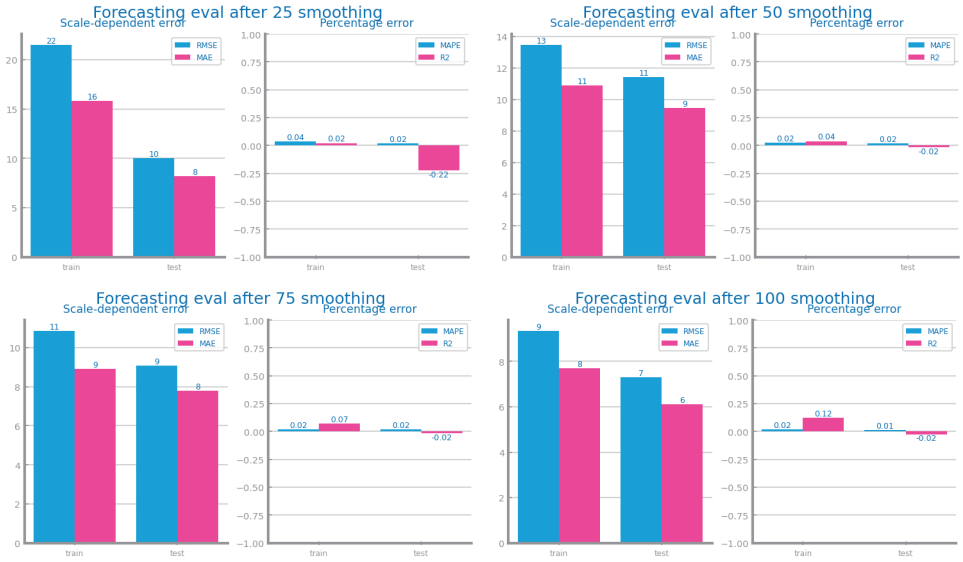


Figure 77: Forecasting results after different smoothing parameterisations on time series 2

Differentiation

To study differentiation, we applied the first two derivatives, favoring the first one for both datasets. In dataset 1, it helped remove quadratic trends and minimize the errors and for dataset 2, both derivatives removed seasonality, but the second added complexity, making it harder to predict. **Shall not exceed 300 characters.**

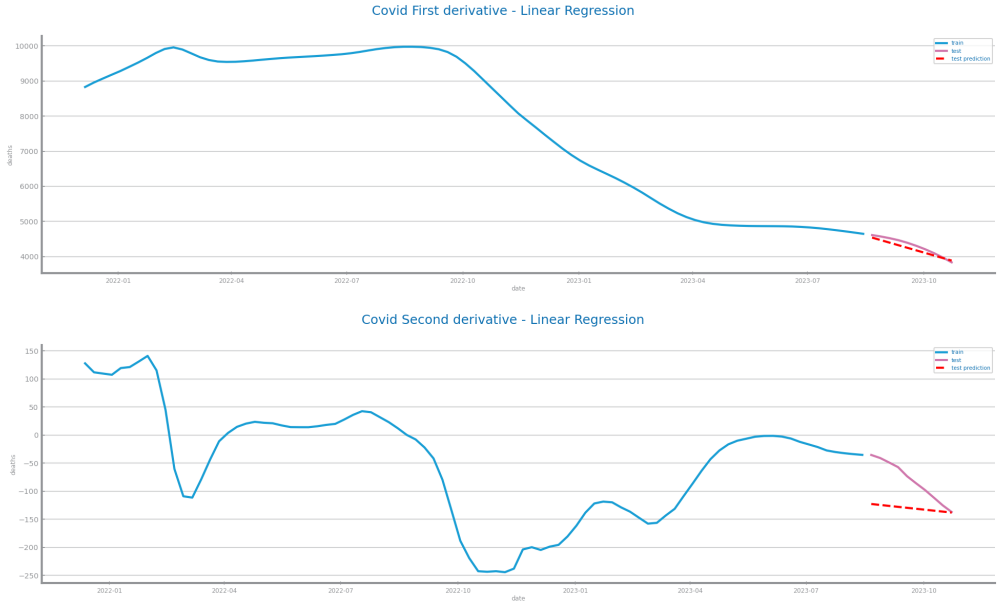


Figure 78: Forecasting plots after first and second differentiation of time series 1

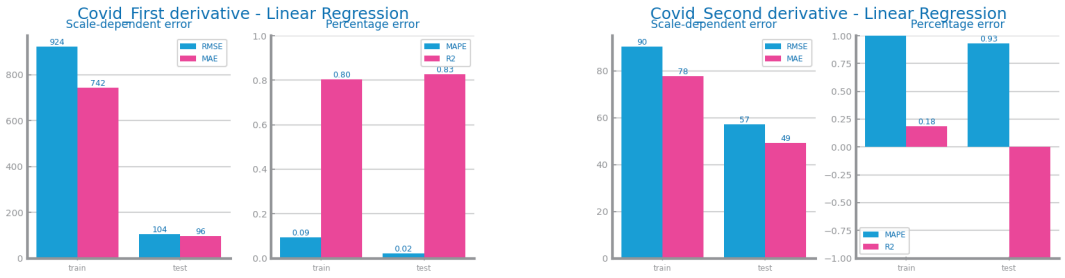


Figure 79: Forecasting results after first and second differentiation of time series 1

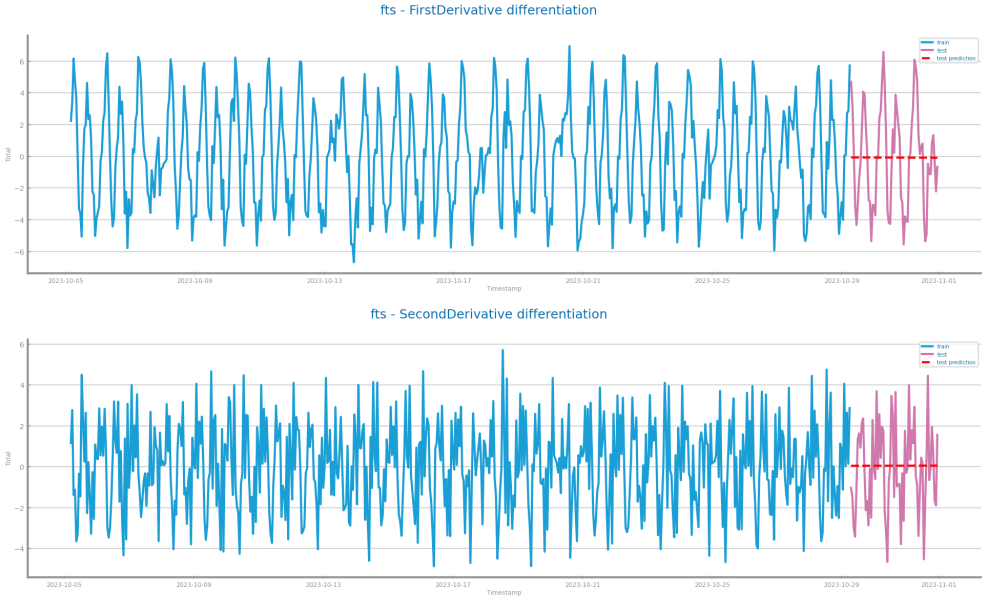


Figure 80: Forecasting plots after first and second differentiation of time series 2

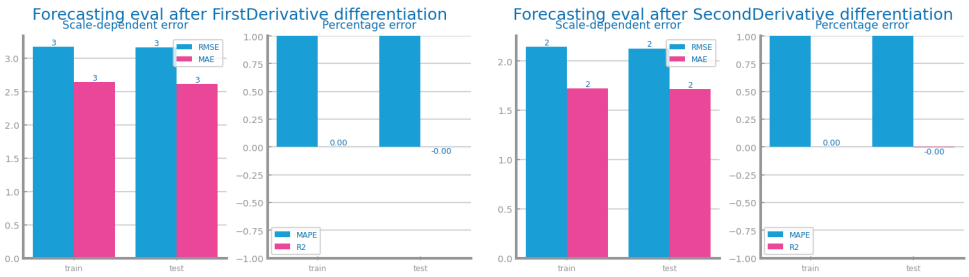


Figure 81: Forecasting results after first and second differentiation of time series 2

Other transformations (optional)

Finally, we applied scaling in both datasets in order to have best values to use in the models' evaluation specifically in the LSTM model. **Shall not exceed 500 characters.**

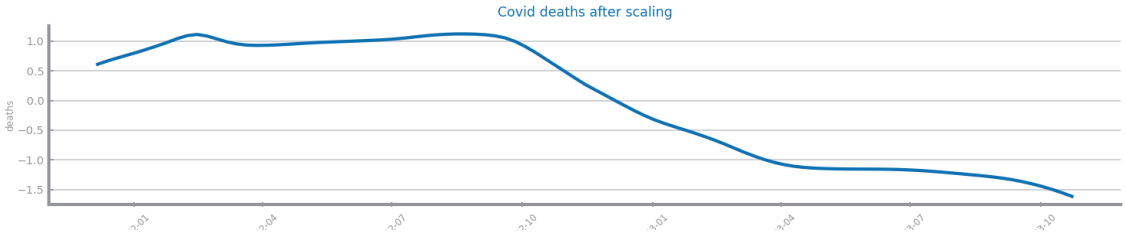


Figure 82: Forecasting plots after applying scaling over time series 1

Figure 83: Forecasting results after applying other transformations over time series 1

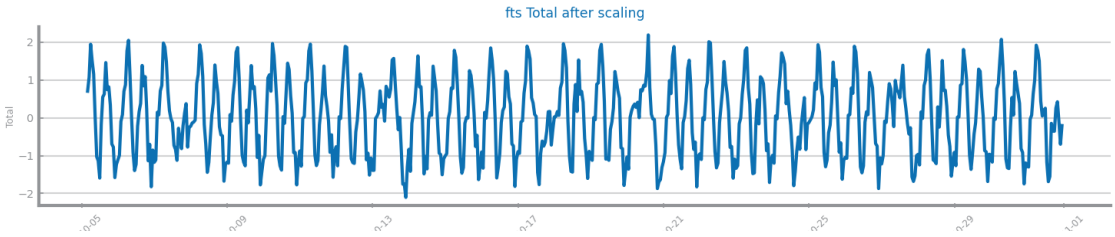


Figure 84: Forecasting plots after applying other transformations over time series 2

Figure 85: FForecasting results after applying other transformations over time series 2

7 MODELS' EVALUATION

For the first dataset we used the weekly aggregation with window size=100, the first derivative and then applied scaling. For dataset 2 we selected minutely aggregation, window size=100, first derivative and scaling. Results were surprisingly positive for dataset 1 when applying the linear regression compared to dataset 2 as the last is closer to the shape of a cosine function instead of linear. For the aggregation study, higher levels were not selected due to high loss of information.

Simple Average Model

Although this metric doesn't approximate any of the datasets correctly, the error and R2 values for dataset 2 seem better because there are some contact points between the real and predicted values.

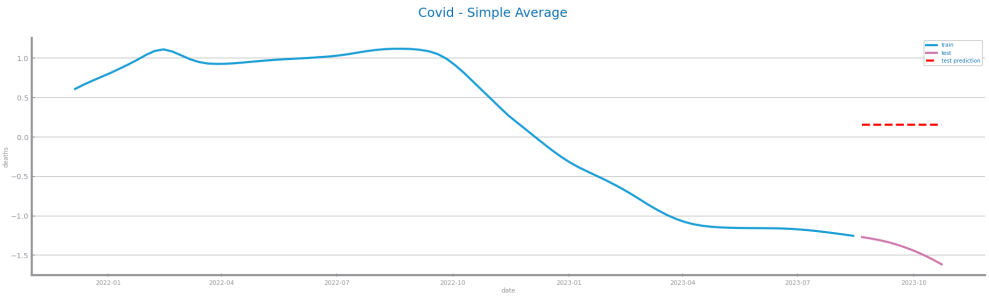


Figure 86: Forecasting plots obtained with Simple Average model over time series 1

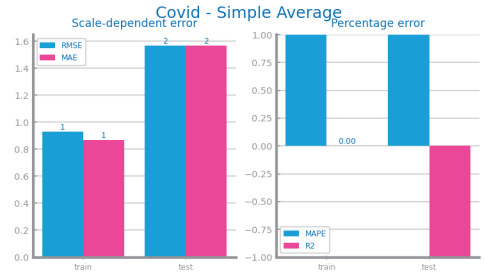


Figure 87: Forecasting results obtained with Simple Average model over time series 1

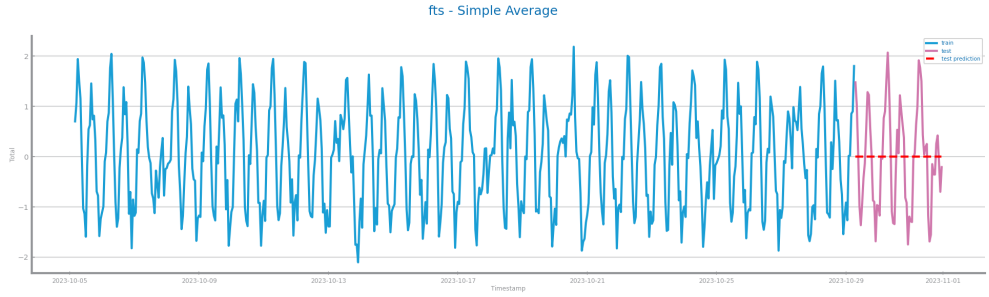


Figure 88: Forecasting plots obtained with Simple Average model over time series 2

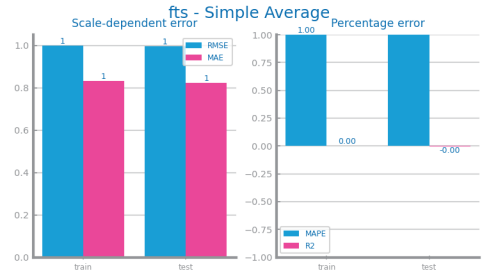


Figure 89: Forecasting results obtained with Simple Average model over time series 2

Persistence Model

The persistence model analysis displays bad results for the realist model and very good results for the optimistic. However, the optimistic model isn't capable to make long-term predictions, it can only accurately predict on a short term space whereas the realist model approximates for long distance. For these reasons, both are bad models for the datasets.

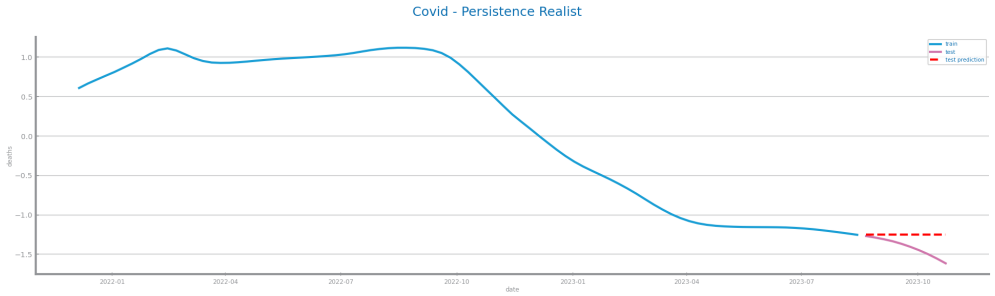


Figure 90: Forecasting plots obtained with Persistence model (long term) over time series 1

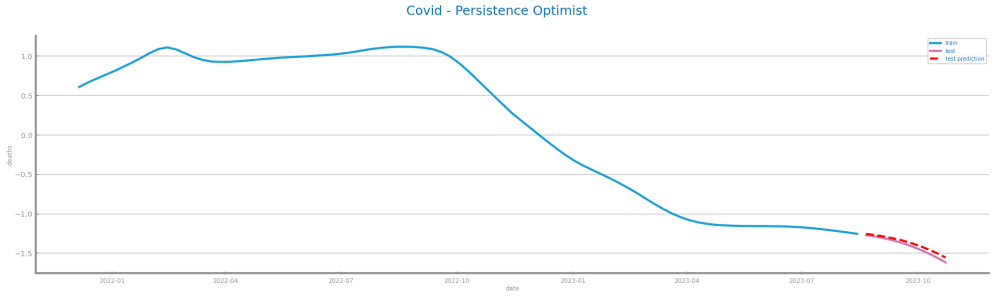


Figure 91: Forecasting plots obtained with Persistence model (one-set-behind) over time series 1

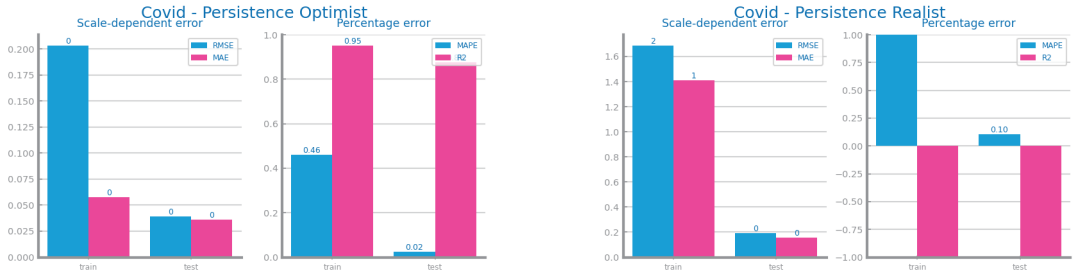


Figure 92: Forecasting results obtained with Persistence model in both situations over time series 1

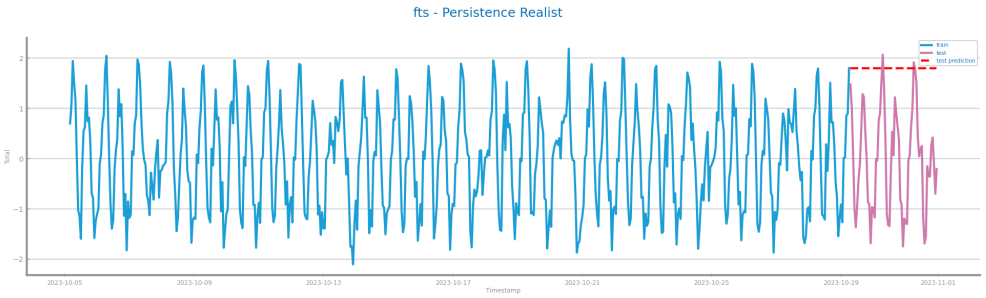


Figure 93: Forecasting plots obtained with Persistence model (long term) over time series 2

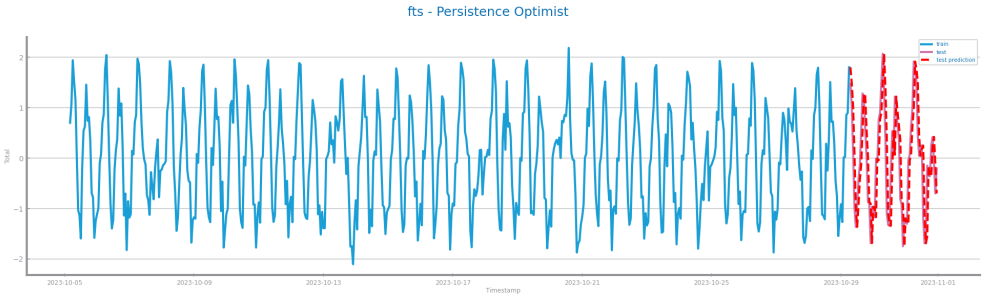


Figure 94: Forecasting plots obtained with Persistence model (one-set-behind) over time series 2

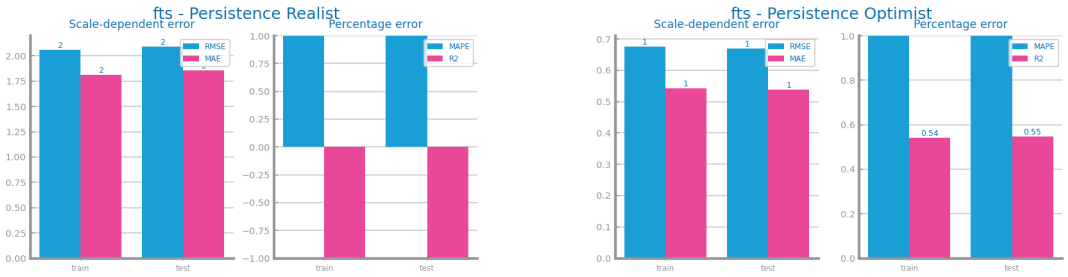


Figure 95: Forecasting results obtained with Persistence model in both situations over time series 2

Rolling Mean Model

Although this metric doesn't approximate any of the datasets correctly, the first dataset obtains better results for the MAE, RMSE and MAPE as it predicts values closer to the real ones but doesn't predict any correct value while dataset 2 has a better R2 because it fluctuates between somewhat symmetric high and low values obtaining a horizontal line between them so there are some contact points but the line remains very distant from the minimum and maximum points.

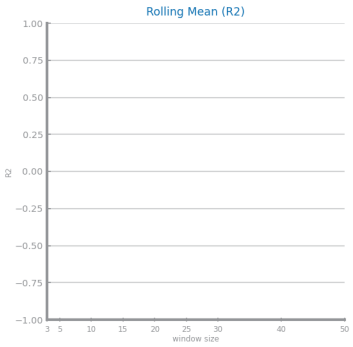


Figure 96: Forecasting study over different parameterisations of the rolling mean algorithm over time series 1

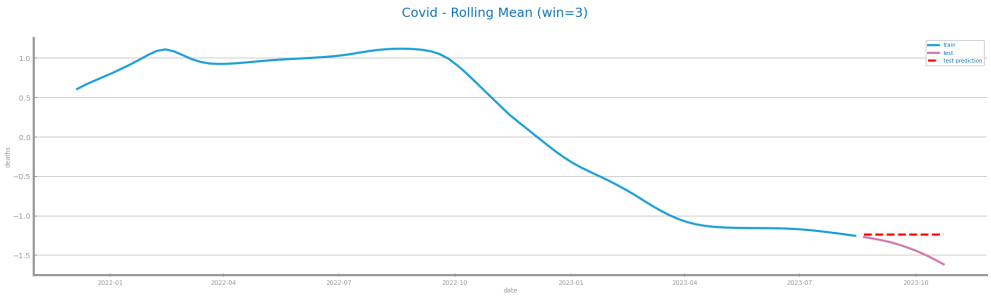


Figure 97: Forecasting plots obtained with the best parameterisation of rolling mean algorithm, over time series 1

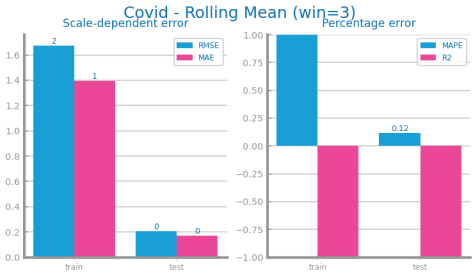


Figure 98: Forecasting results obtained with the best parameterisation of rolling mean algorithm, over time series 1

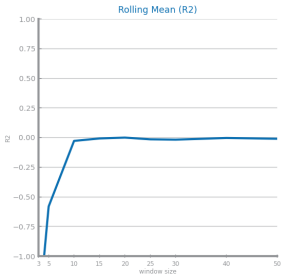


Figure 99: Forecasting study over different parameterisations of the rolling mean algorithm over time series 2

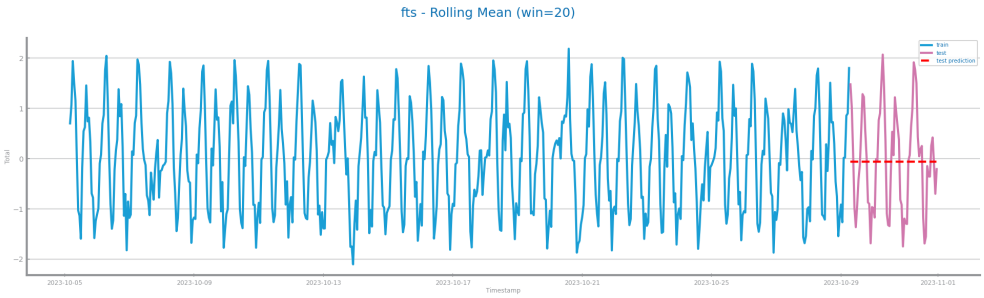


Figure 100: Forecasting plots obtained with the best parameterisation of rolling mean algorithm, over time series 2

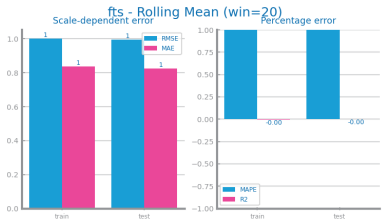


Figure 101: Forecasting results obtained with the best parameterisation of rolling mean algorithm, over time series 2

ARIMA Model

Dataset 1, with a trend-style pattern, benefits from parameters (p, d, q) set to (7, 2, 5). This configuration allows the model to capture and accommodate the complexities associated with trend-based data. On the other hand, Dataset 2, exhibiting a cosine-like shape, attains superior performance with parameters at (3, 0, 5). This parameter choice enables the model to capture the cyclical and periodic components in the dataset, showcasing the adaptability of ARIMA to diverse time series patterns.

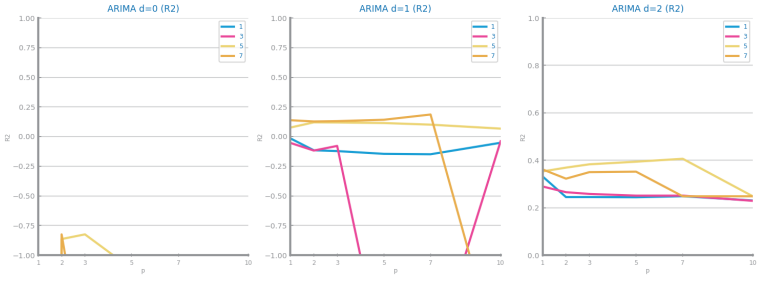


Figure 102: Forecasting study over different parameterisations of the ARIMA algorithm over time series 1

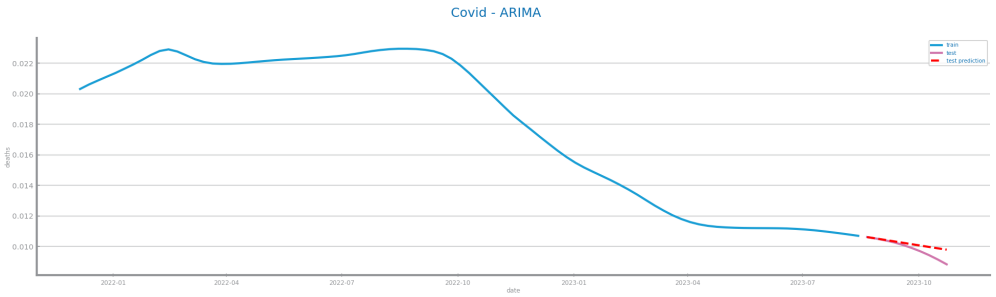


Figure 103: Forecasting plots obtained with the best parameterisation of ARIMA algorithm, over time series 1

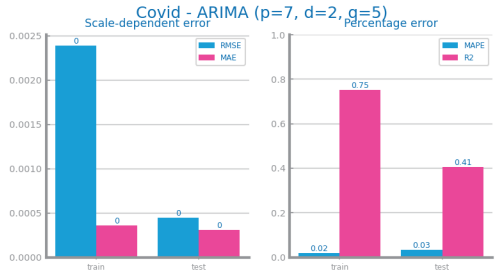


Figure 104: Forecasting results obtained with the best parameterisation of ARIMA algorithm, over time series 1

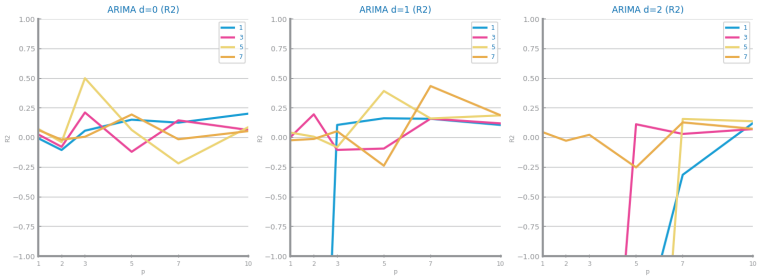


Figure 105: Forecasting study over different parameterisations of the ARIMA algorithm over time series 2

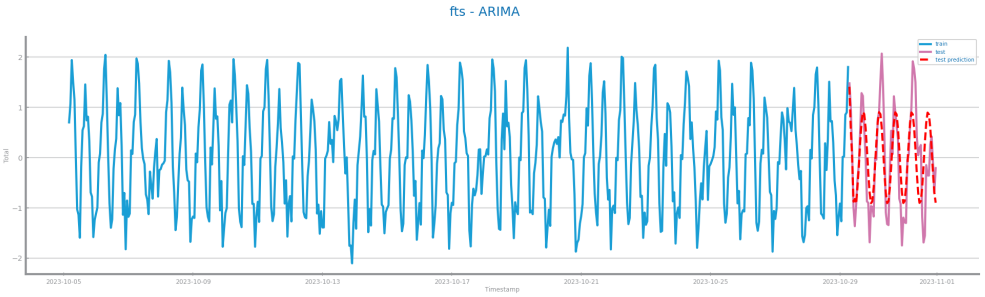


Figure 106: Forecasting plots obtained with the best parameterisation of ARIMA algorithm, over time series 2

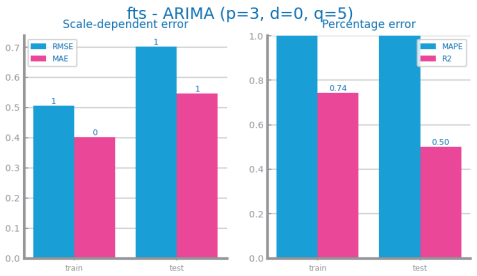


Figure 107: Forecasting results obtained with the best parameterisation of ARIMA algorithm, over time series 2

LSTMs Model

For Dataset 1, having a trend, the LSTM excels with parameters length=4, hidden=100 and nr_episodes=900, effectively capturing trend-oriented patterns. This adaptability extends to Dataset 2, featuring a cosine-like trend, where the same parameter configuration yields optimal results. As expected the LSTM's model achieves the best forecasting results for both datasets compared to the previous models.

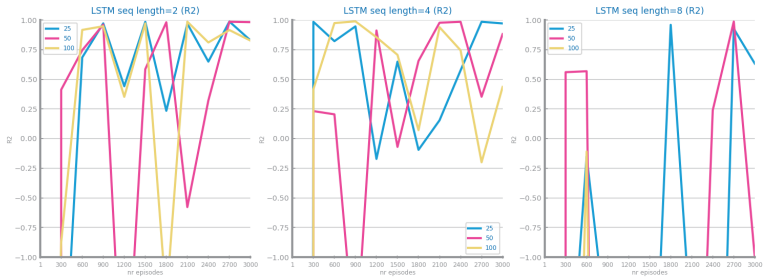


Figure 108: Forecasting study over different parameterisations of LSTMs over time series 1

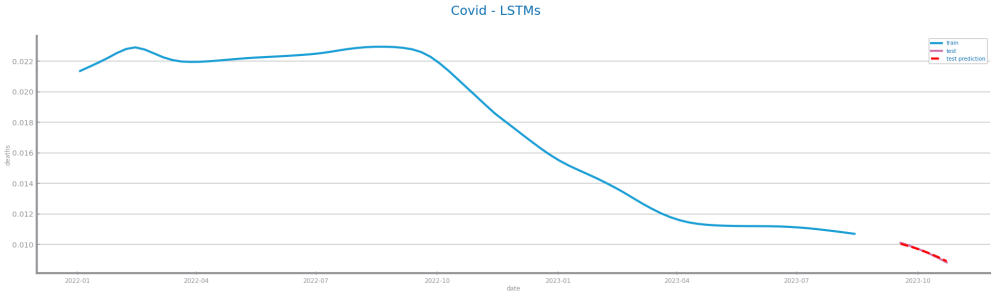


Figure 109: Forecasting plots obtained with the best parameterisation of LSTMs, over time series 1

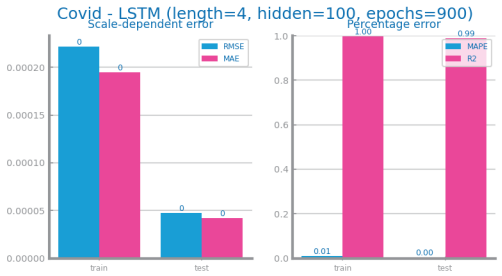


Figure 110: Forecasting results obtained with the best parameterisation of LSTMs, over time series 1

Figure 111: Forecasting study over different parameterisations of LSTMs with multiple variables over time series 1

Figure 112: Forecasting plots obtained with the best parameterisation of LSTMs with multiple variables over time series 1

Figure 113: Forecasting results obtained with the best parameterisation of LSTMs with multiple variables over time series 1

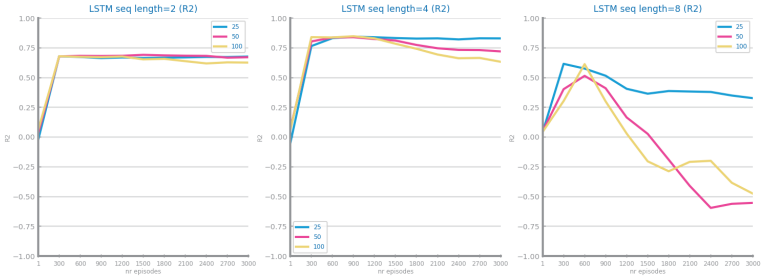


Figure 114: Forecasting study over different parameterisations of the LSTMs over time series 2

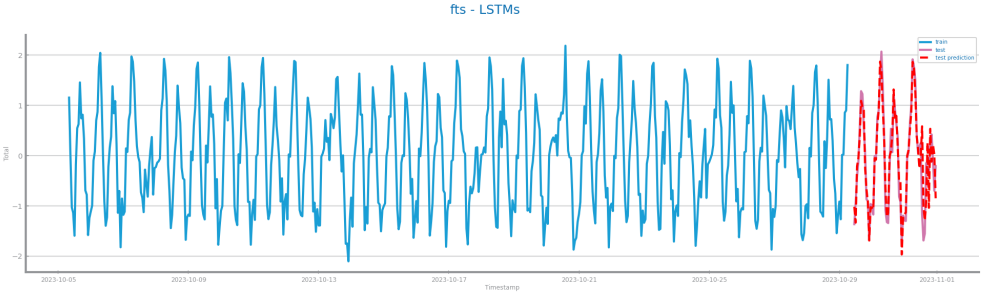


Figure 115: Forecasting plots obtained with the best parameterisation of LSTMs, over time series 2

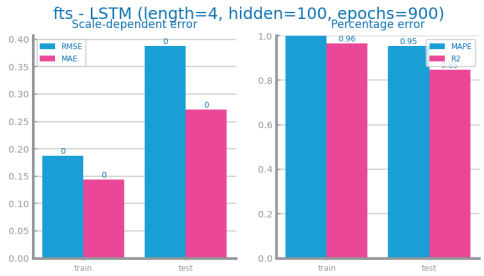


Figure 116: Forecasting results obtained with the best parameterisation of LSTMs, over time series 2

Figure 117: Forecasting study over different parameterisations of LSTMs with multiple variables over time series 2

Figure 118: Forecasting plots obtained with the best parameterisation of LSTMs with multiple variables over time series 2

Figure 119: Forecasting results obtained with the best parameterisation of LSTMs with multiple variables over time series 2

8 CRITICAL ANALYSIS

In a critical assessment, both ARIMA and LSTMs outperformed simpler models like simple average or rolling mean, emphasizing their adaptability and effectiveness. LSTMs, in particular, demonstrated superior performance across diverse datasets, positioning them as a favorable choice for both. The decision between these strands hinges on dataset specifics and the desired trade-off between interpretability and complexity. Dataset 1 reveals a non-stationary nature, characterized by an initial upward trend with it slowing down. Although on the smoothing phase this shape is flattened, this decline in Covid related death rates can be seen on the differentiation analysis, by employing the second derivative and removing the quadratic trend, recalling the beginning upwards activity and settling down later. In order to have best suited values for the LSTM model, we applied scaling. Notably, although we predicted it to be a good fit for this model, ARIMA is not a suitable fit for this series as it's prediction is too long-term based, not dealing with the abrupt decline. For dataset 2, also non-stationary, having applied both smoothing, differentiation and scaling to regularise the value spikes and revealing cyclical and seasonal behaviors of this series, but with no evident trend. As expected but differing from the first dataset, the ARIMA model provides a good fit, encapsulating the cyclicality of this series on a good level. Despite not being as accurate as LSTM, if the model complexity from the latest proves a bottleneck for the problem, ARIMA is a good substitute. The univariate nature of the time series hinders accurate prediction, leading to suboptimal model performance. For this reason and as the simple average model, persistence realist and optimist and rolling mean can only deal with linear series or short term predictions which are not the case for either datasets, none of these models seem good enough to solve these problems. **Shall not exceed 2000 characters.**