|  |  |  |  |
| --- | --- | --- | --- |
| Logo, company name  Description automatically generated | | **DS 2022** | |
| Data Science Project | | | |
| **Team nr:** 12 | **Student 1 :** Eduardo Moreira Miranda | | **IST nr:** 95569 |
| **Student 2 :** Miguel Paraíso Alves | | **IST nr:** 95650 |
| **Student 3 :** Rodrigo De Melo Pinto | | **IST nr:** 95666 |

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

Classification

# 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.**

## Data Dimensionality

In both datasets the number of records is way bigger than the dimensionality making them low sparsity, so we can cover all the domain.

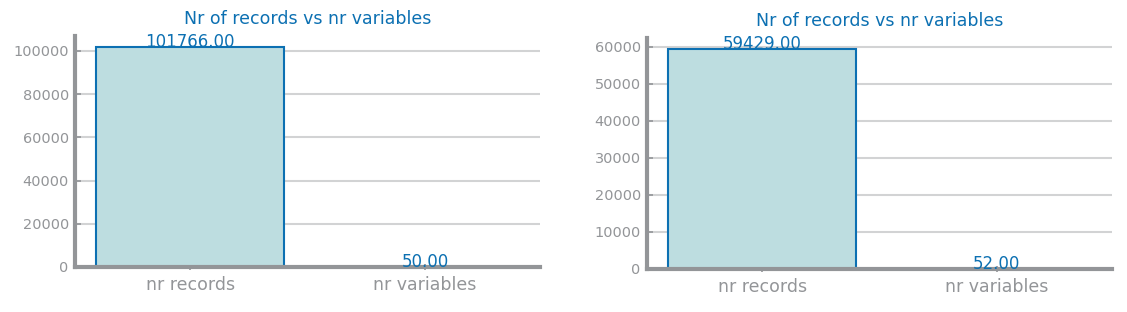


Figure 1 Nr Records x Nr variables for dataset 1 (left) and dataset 2 (right)



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

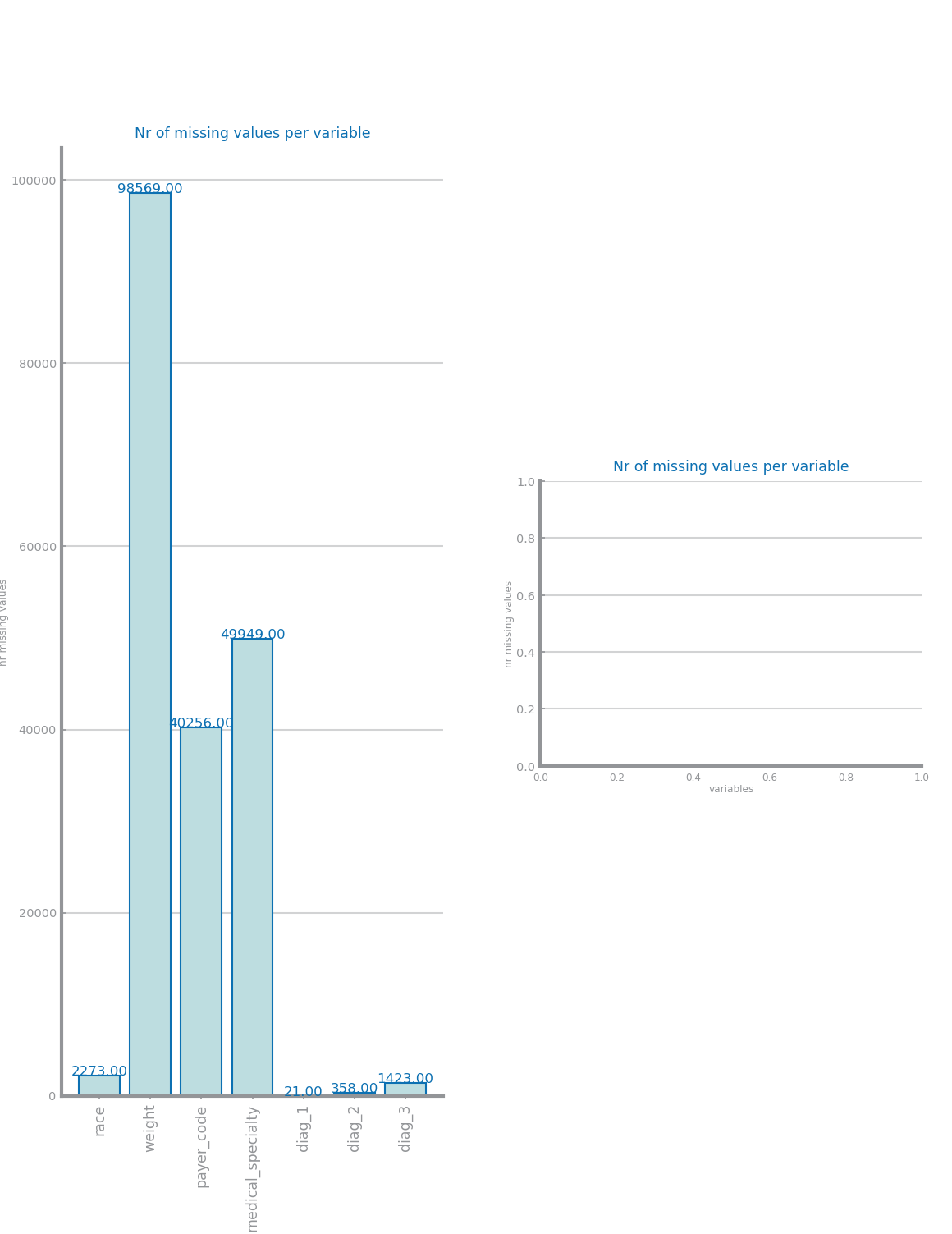


Figure 3 Nr missing values for dataset 1 (left) and dataset 2 (right)

## Data Distribution

In dataset 1 we can see that some variables only have one value in the entire dataset, the patient\_nbr and encounter\_id have a bigger value distribution than the other numeric variables and most numeric variables are discrete and do not follow a normal distribution. In the dataset 2 most numeric variables follow a normal distribution and some variables like SQ5 have a distribution where one value is a lot more common than the others.

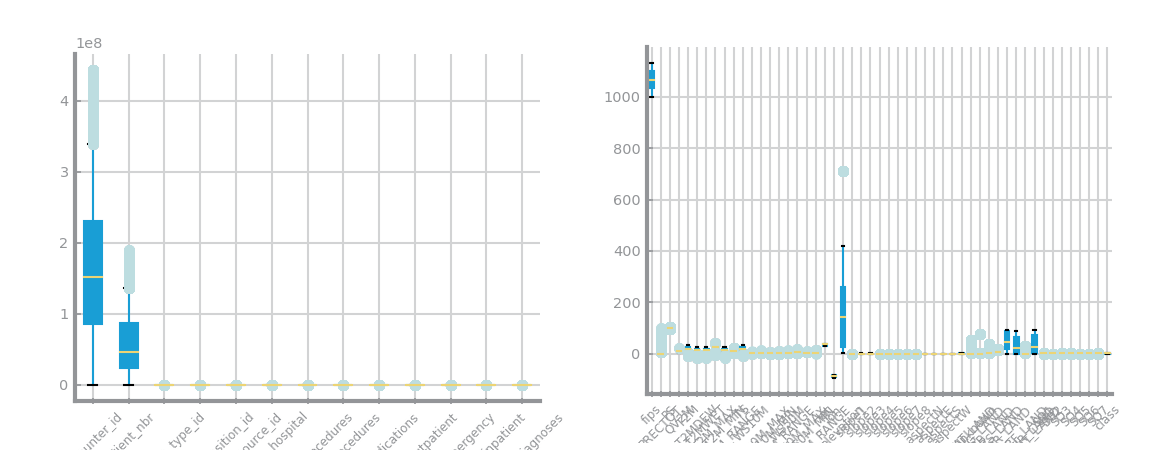


Figure 4 Global boxplots dataset 1 (left) and dataset 2 (right)



Figure 5 Single variable boxplots for dataset 1

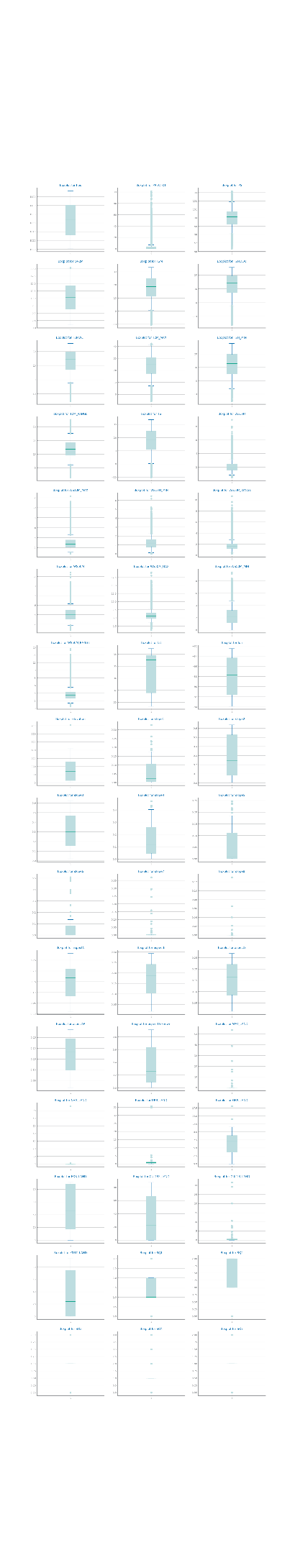


Figure 6 Single variable boxplots s for dataset 2

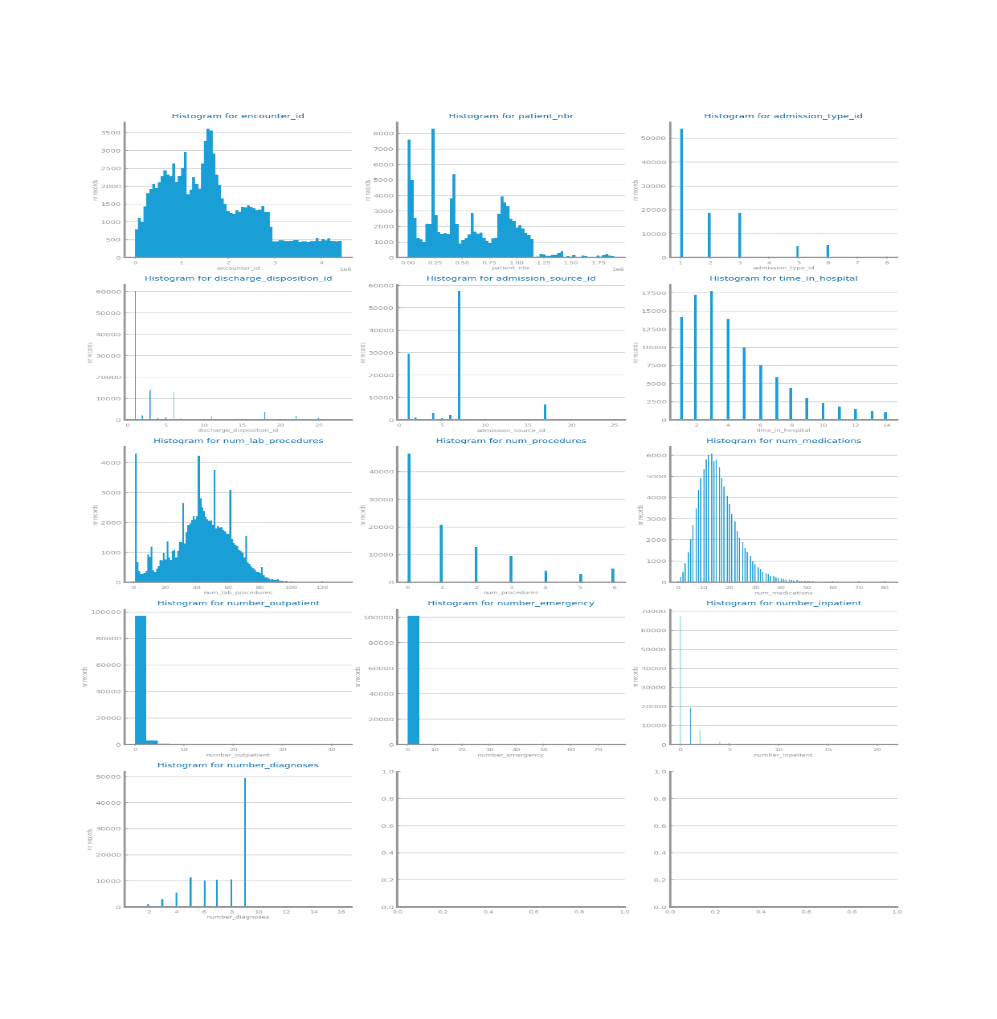
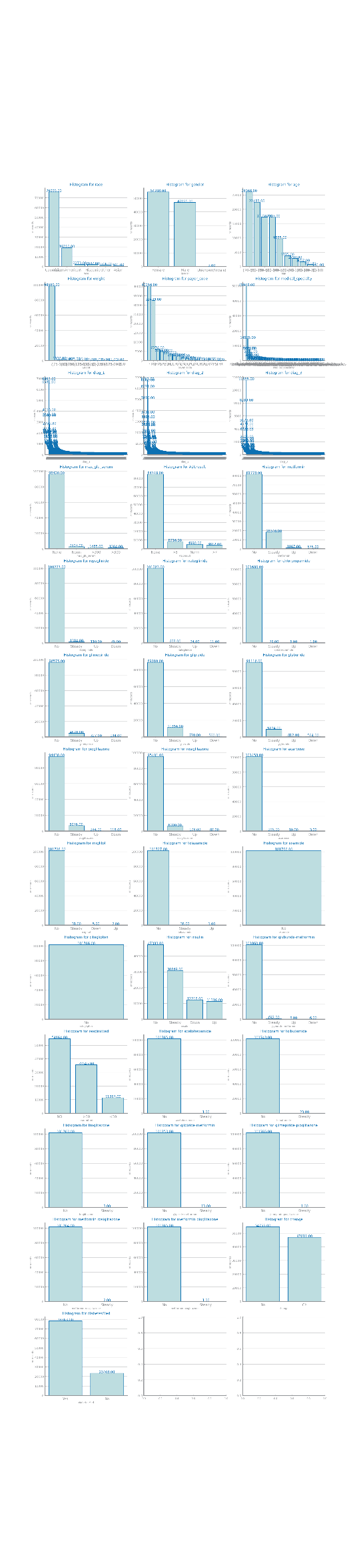
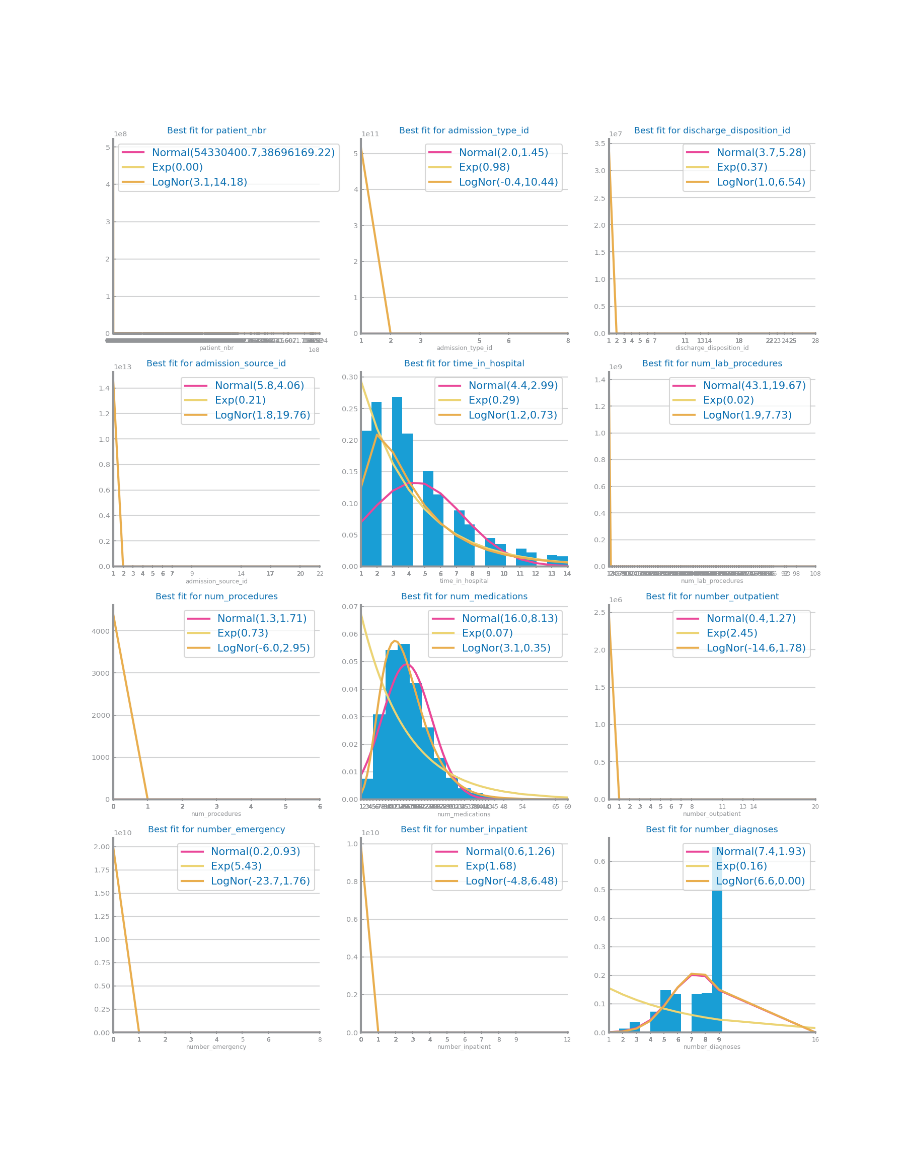


Figure 7 Histograms for dataset 1

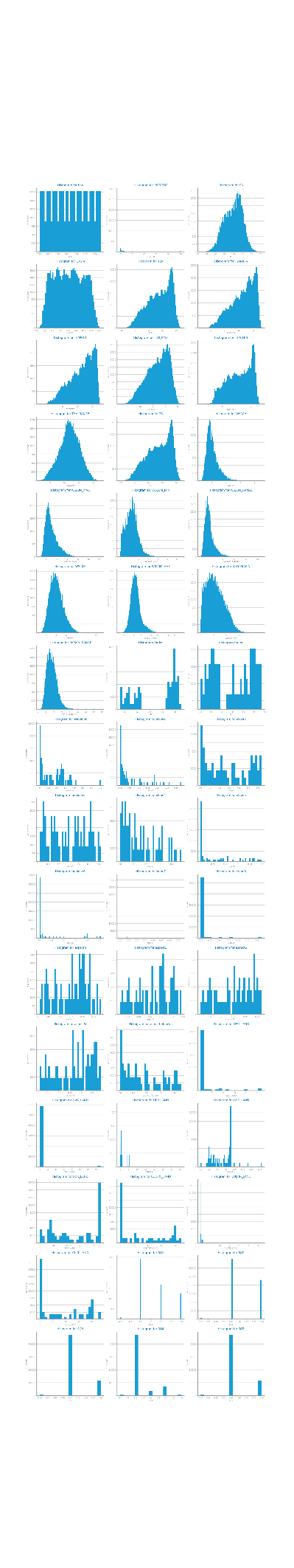
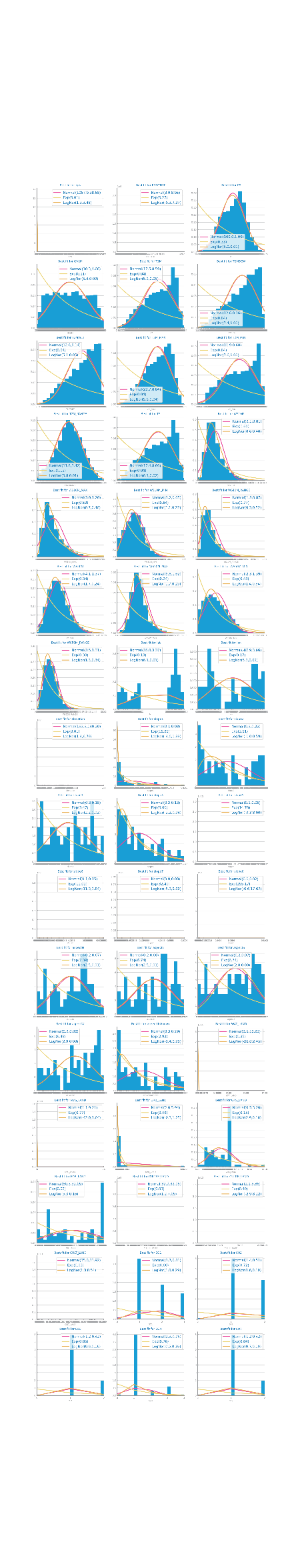
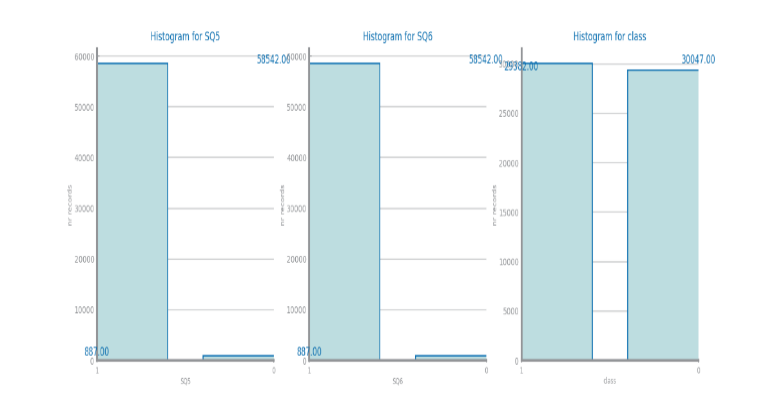


Figure 8 Histograms for dataset 2

## Data Granularity

Some variables like encounter\_id on dataset 1 benefit from a higher granularity because we can see more detail however most variables have few possible values so they do not benefit from higher granularity.

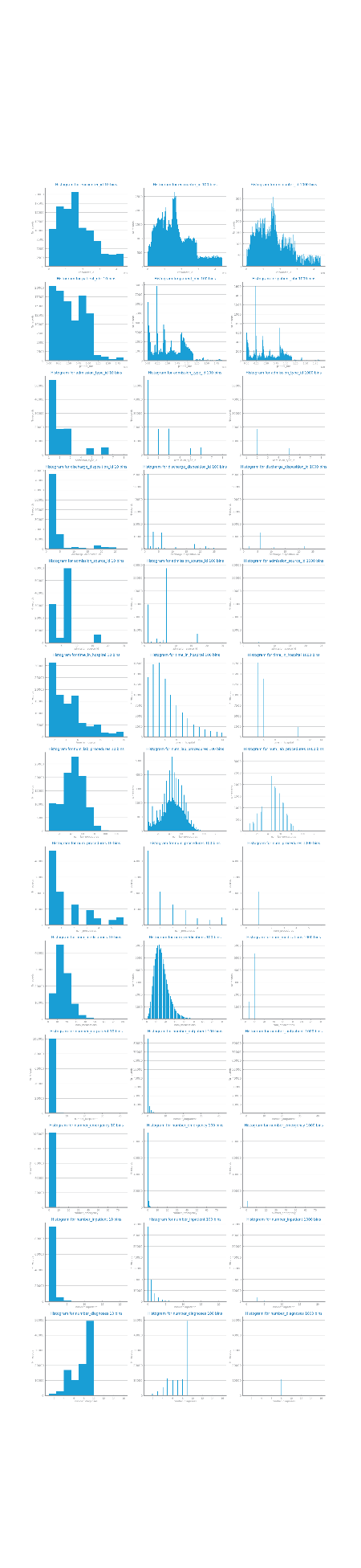


Figure 9 Granularity analysis for dataset 1

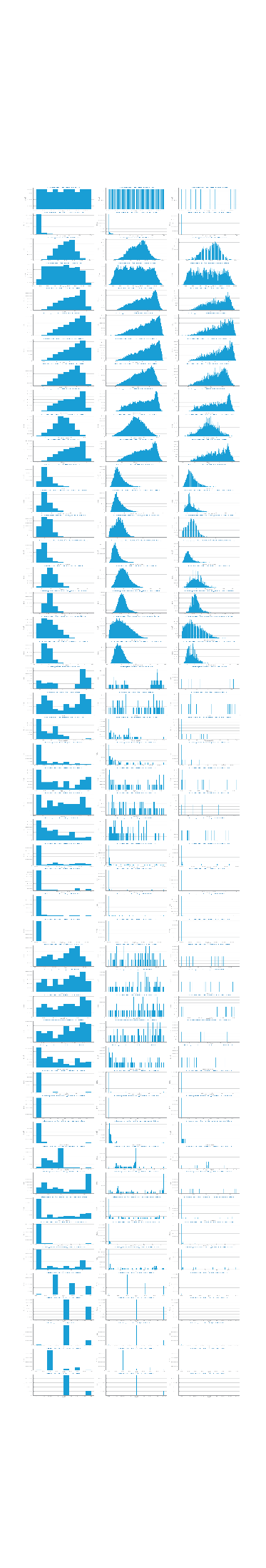
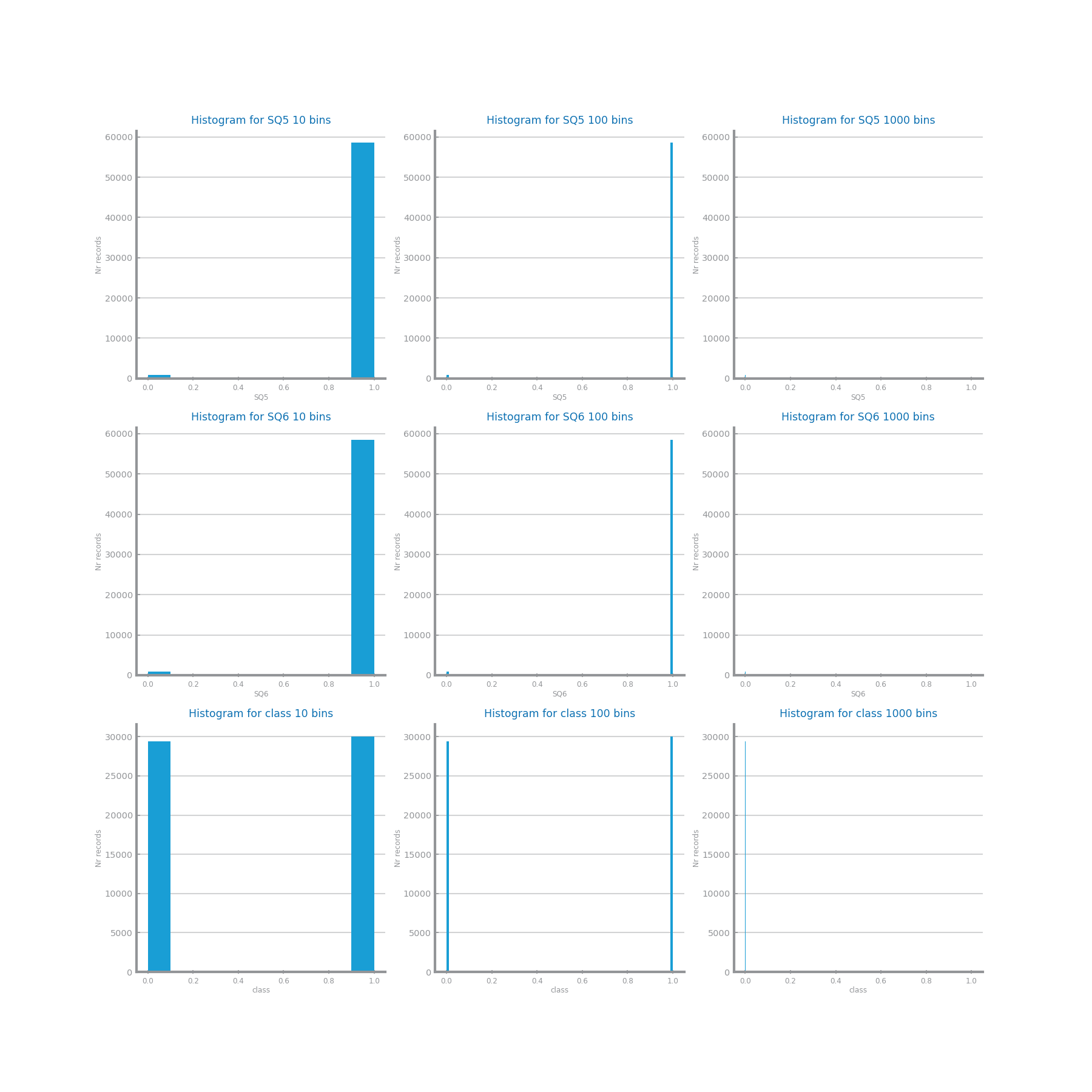


Figure 10 Granularity analysis for dataset 2

## Data Sparsity

In both dataset we have low sparsity because most values of the domain are covered. In our correlation analysis we can see that in dataset 2 there are a lot more correlations than in dataset 1. For example in dataset 2 the T2M variables are strongly correlated.

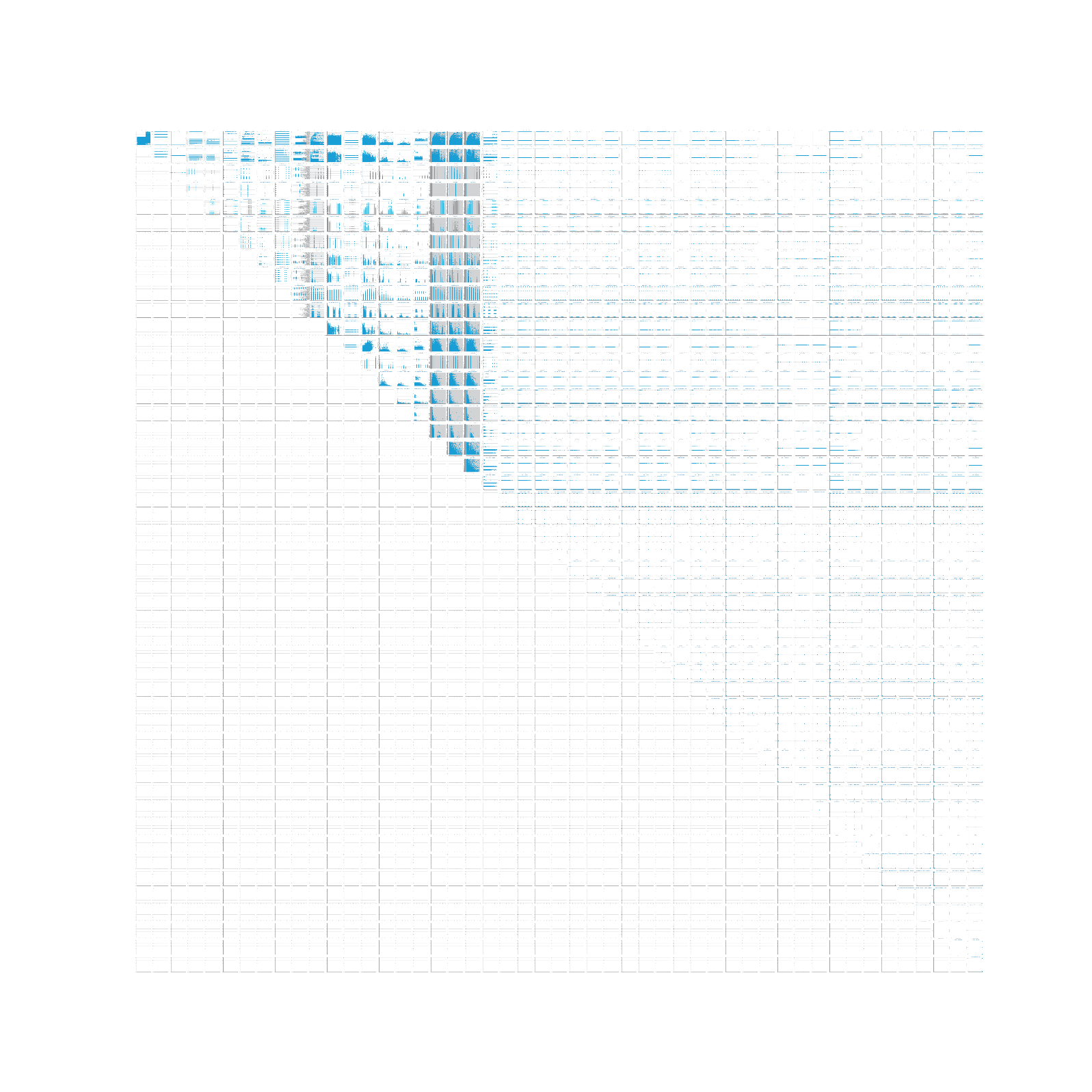


Figure 11 Sparsity analysis for dataset 1

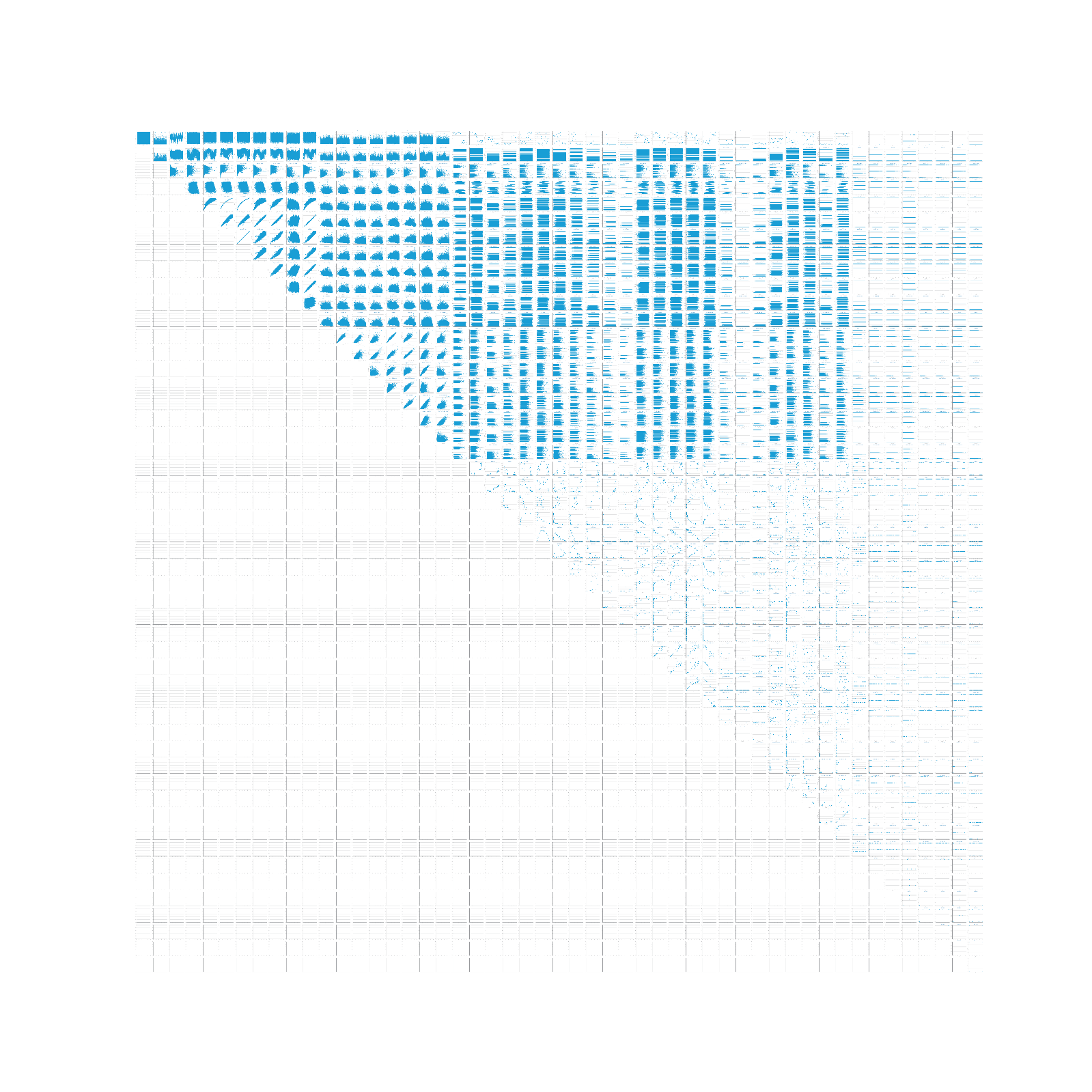


Figure 12 Sparsity analysis for dataset 2

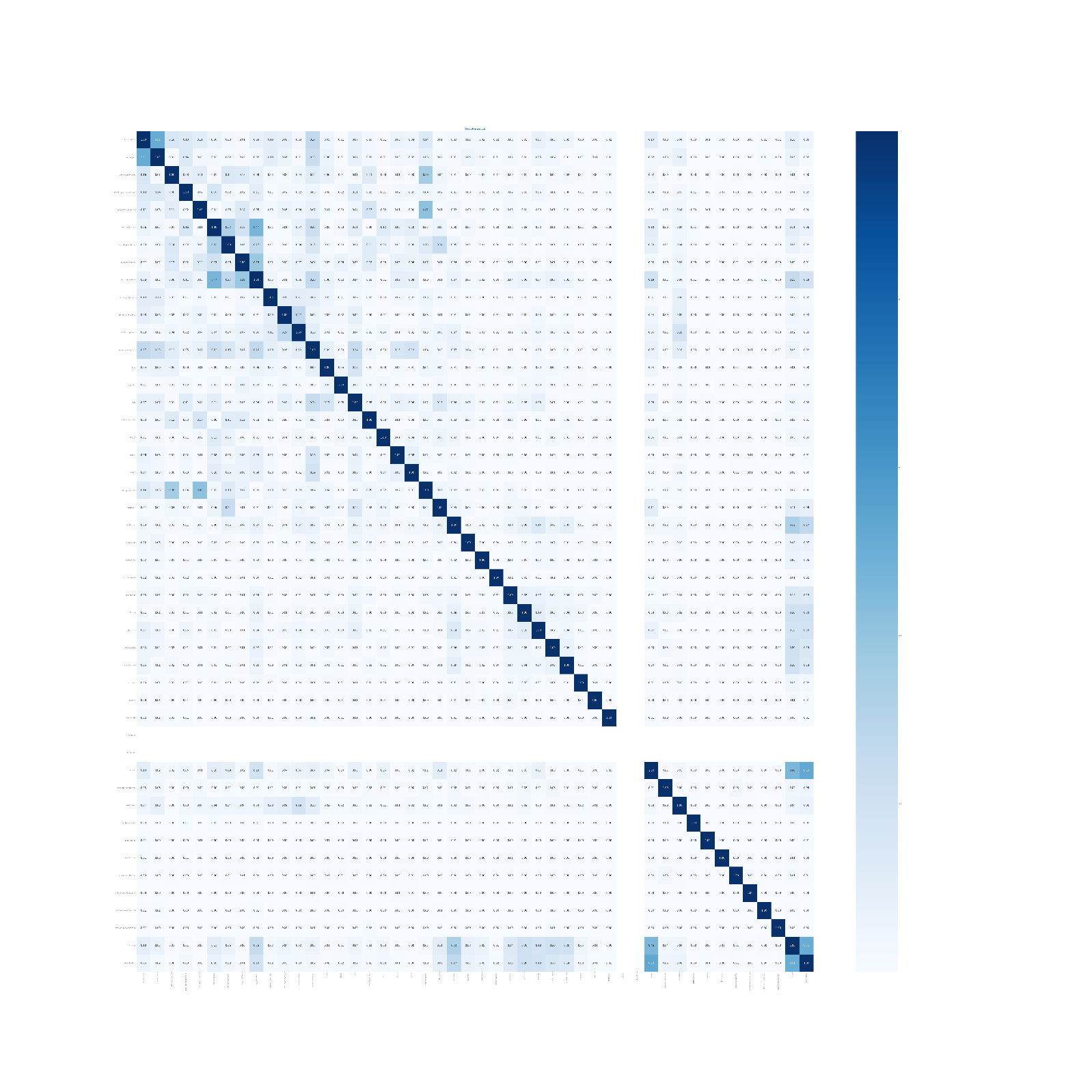


Figure 13 Correlation analysis for dataset 1

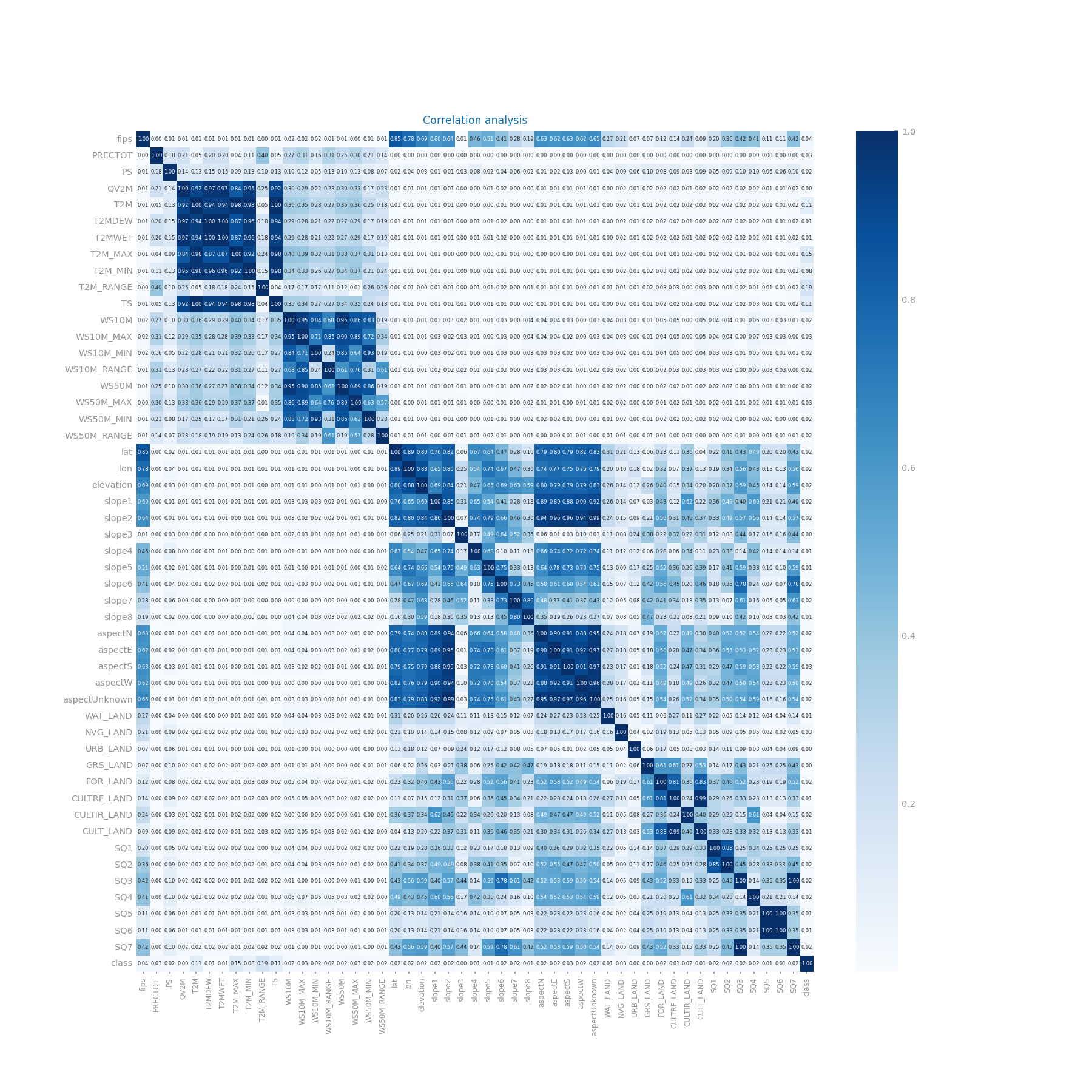


Figure 14 Correlation analysis for dataset 2

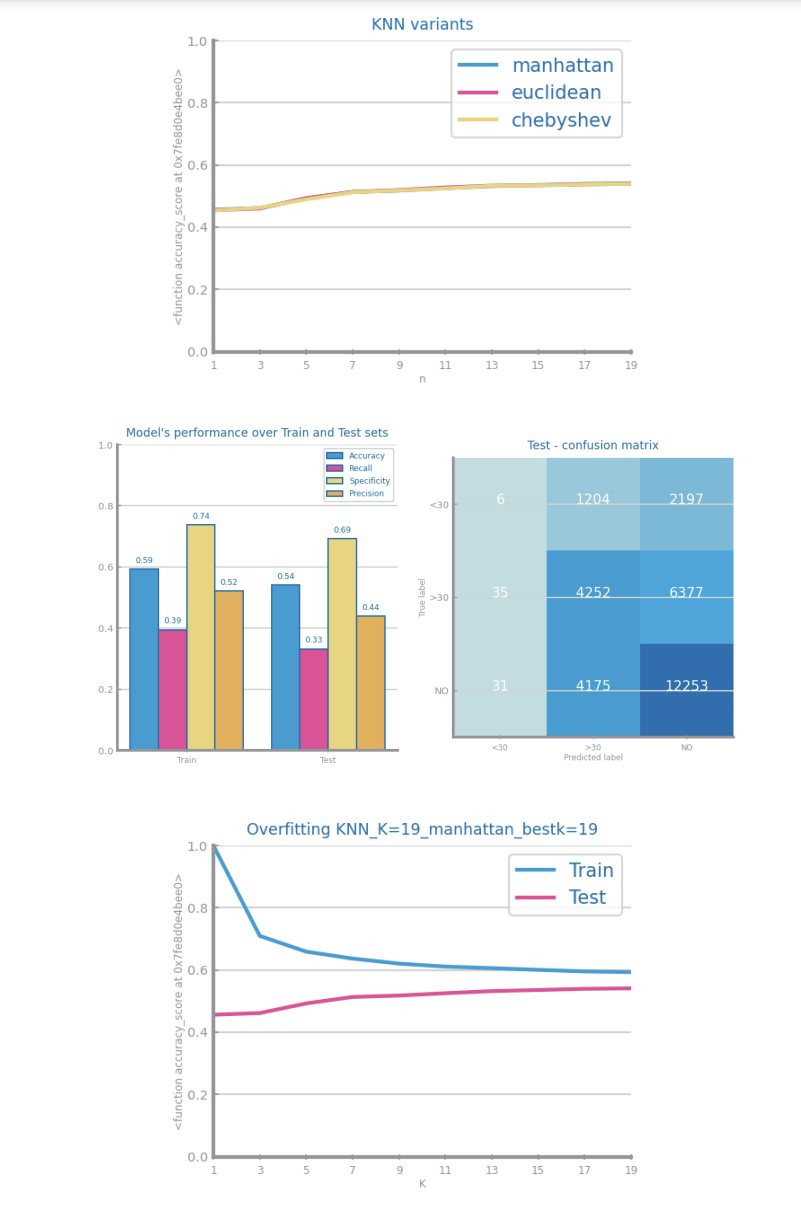
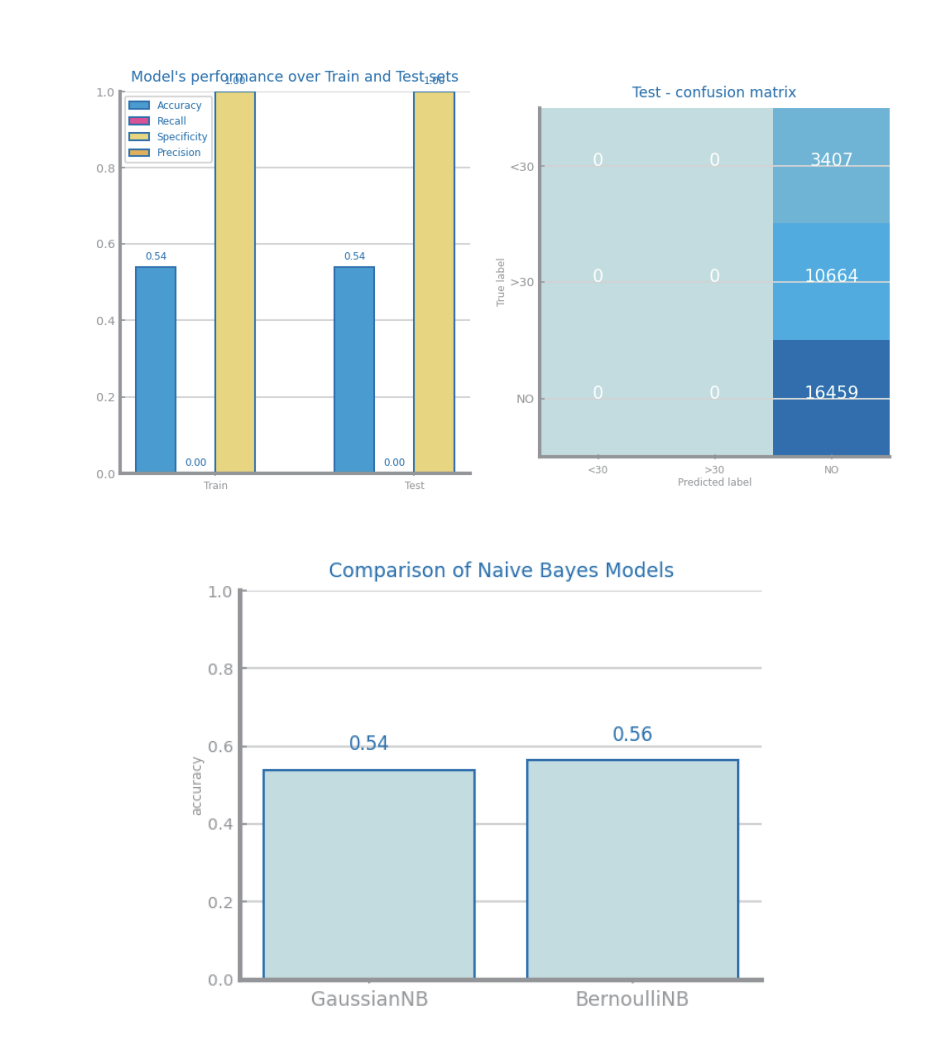
# Data Preparation

## Variables Encoding

For dataset 1 we changed most of the symbolic variables to encode in ordinal numbers example: [‘No’,’Down’,’Steady’,’Up’] is replaced by [‘0’,’1’,’2’,’3’] we did this because most of the variables have an intrinsic order to them. The only symbolic variables where we did not do this were on the diagnosis variables, here as they were already encoded numerically except for examples like ‘V57’ we just replaced ‘V’ to ‘-2.’ and ‘E’ to ‘-1.’ to preserve the variables. In dataset 2 we divided the date column into year, month and day columns so that this information could be used going forward.

## Missing Value Imputation

For dataset 1 we used two approaches and in both we dropped the weight and payer\_code variables. In the first approach we replaced the missing values with a fixed value. And in the second we replaced them with the most frequent value on that variable. On dataset 2 there were no missing values so we did not do anything in this regard.



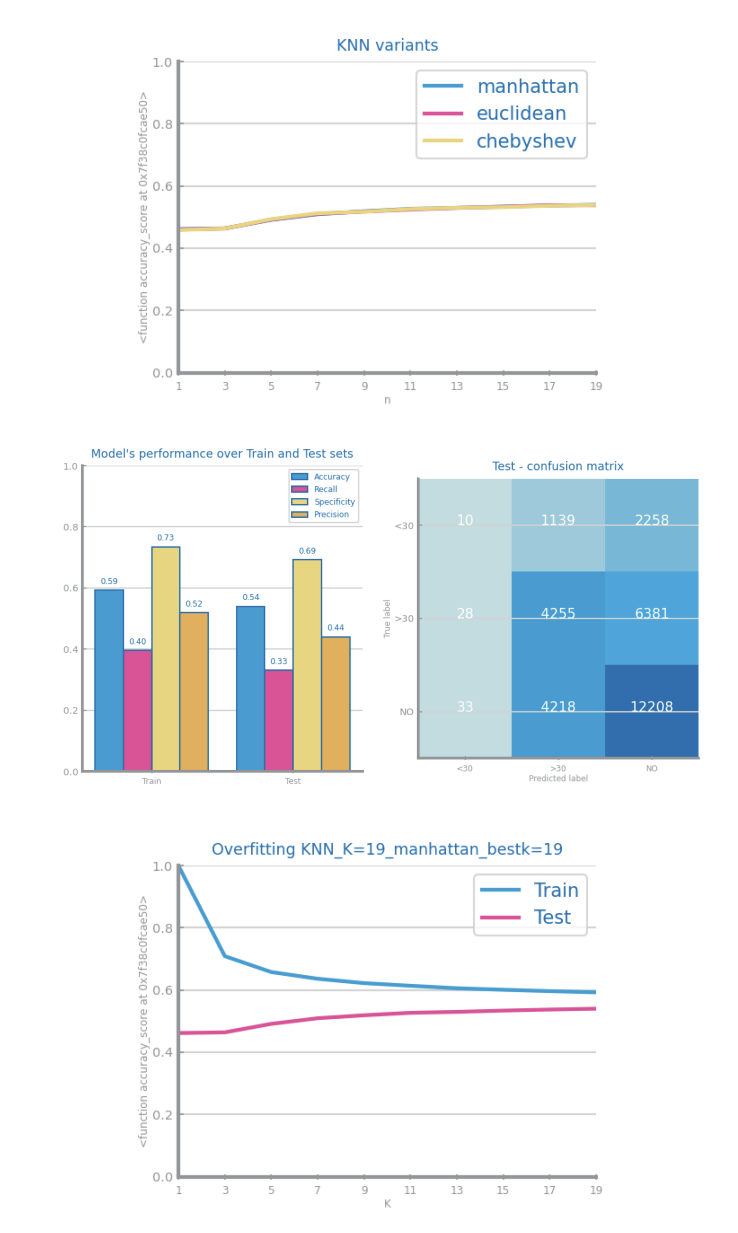
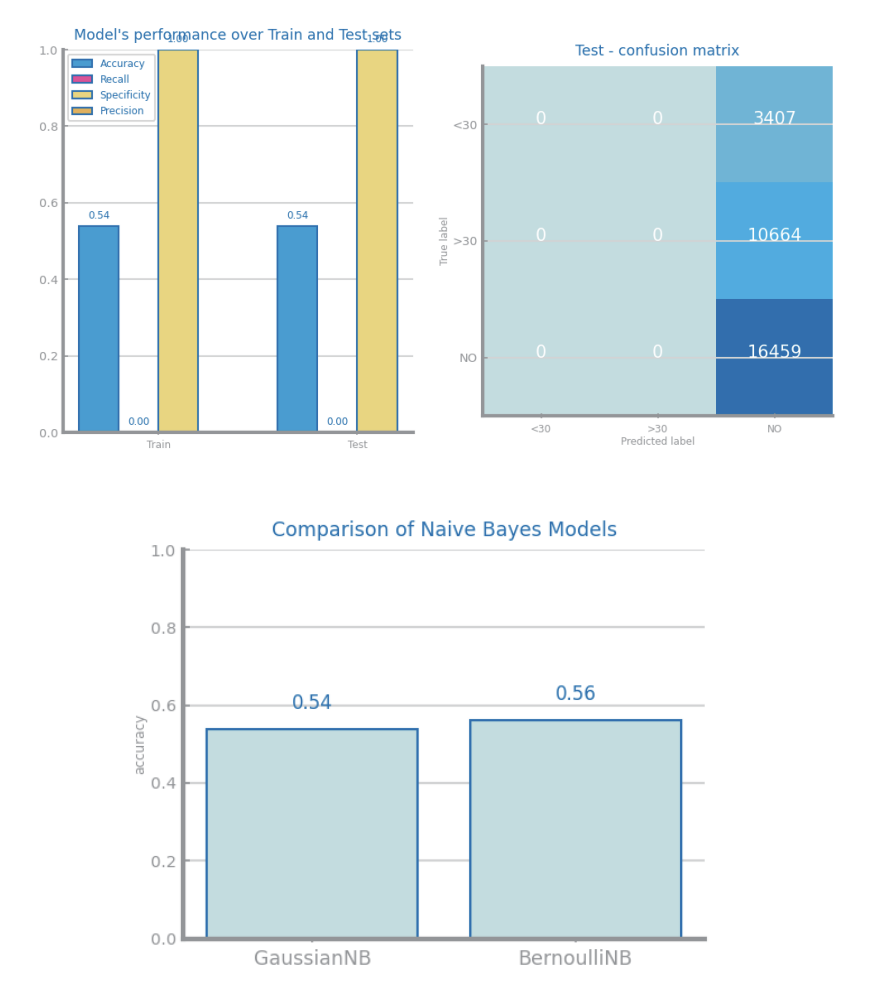


Figure 15 Missing values imputation results with different approaches for dataset 1

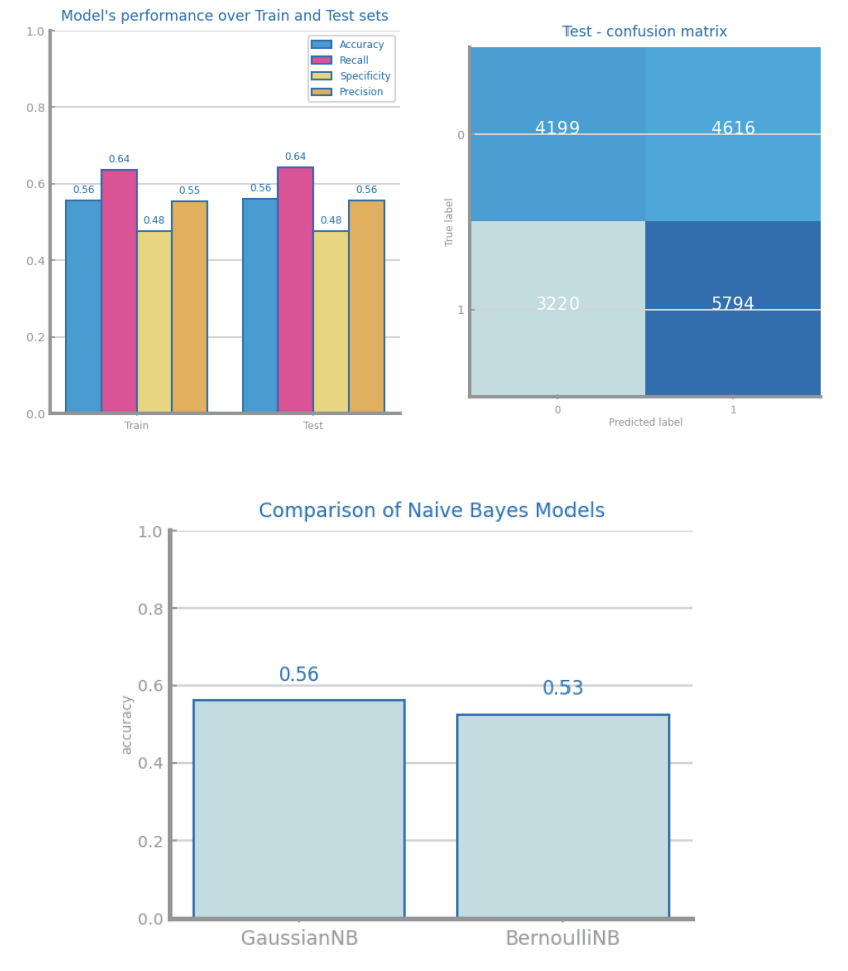


Figure 16 Missing values imputation results with different approaches for dataset 2

## Outliers Treatment

In both datasets we detected the outliers using the standard deviation method with 3 stdev as parameter, then we tested dropping the outliers (left) and truncating them (right). We chose dropping them going forward.

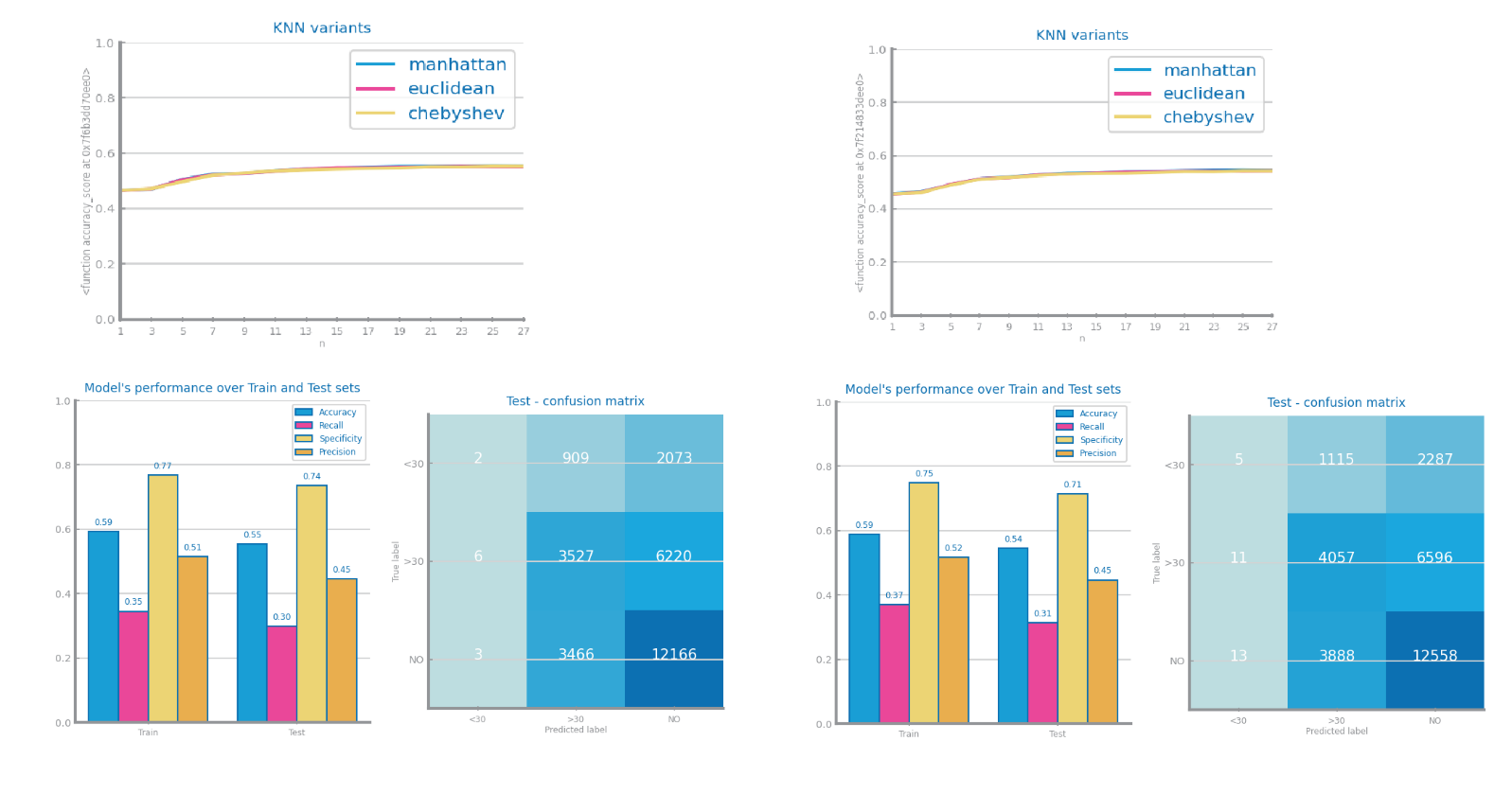


Figure 17 Outliers imputation results with different approaches for dataset 1

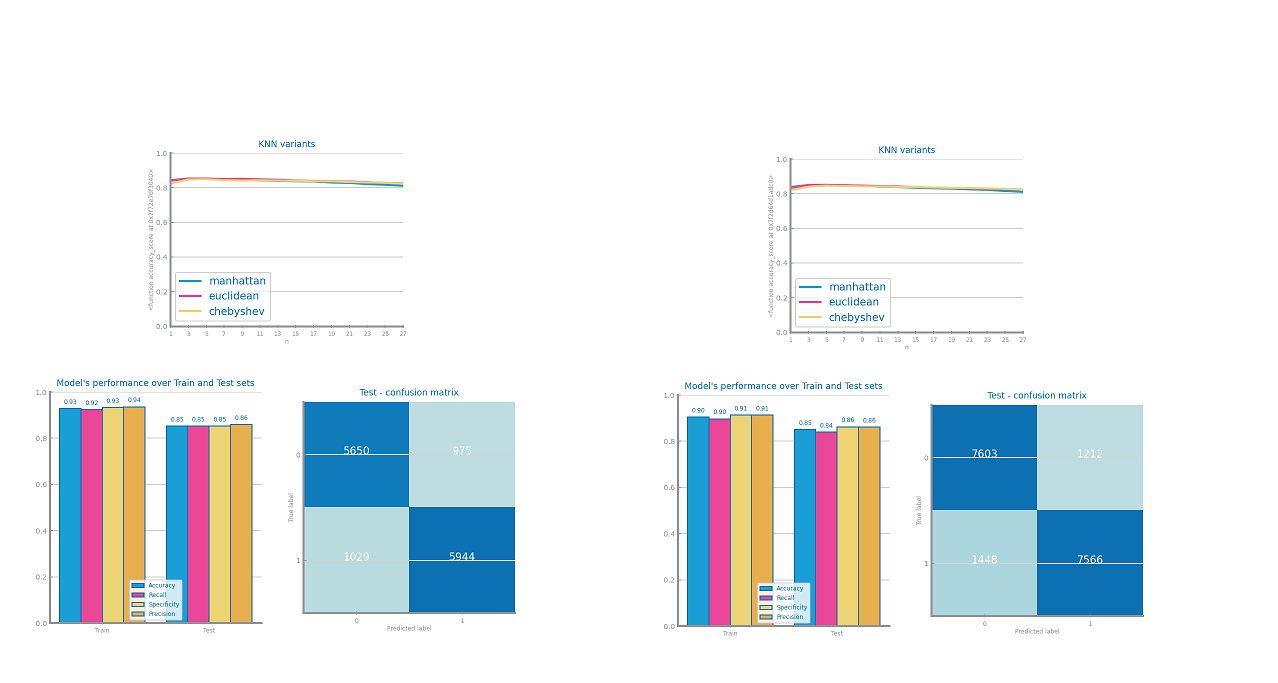


Figure 18 Outliers imputation results with different approaches for dataset 2

## Scaling

We tested both the minmax (on the left) and Z-score (on the right) on both datasets, using knn to compare the results. After comparing the accuracy we chose Z-score going forward.



Figure 19 Scaling results with different approaches for dataset 1

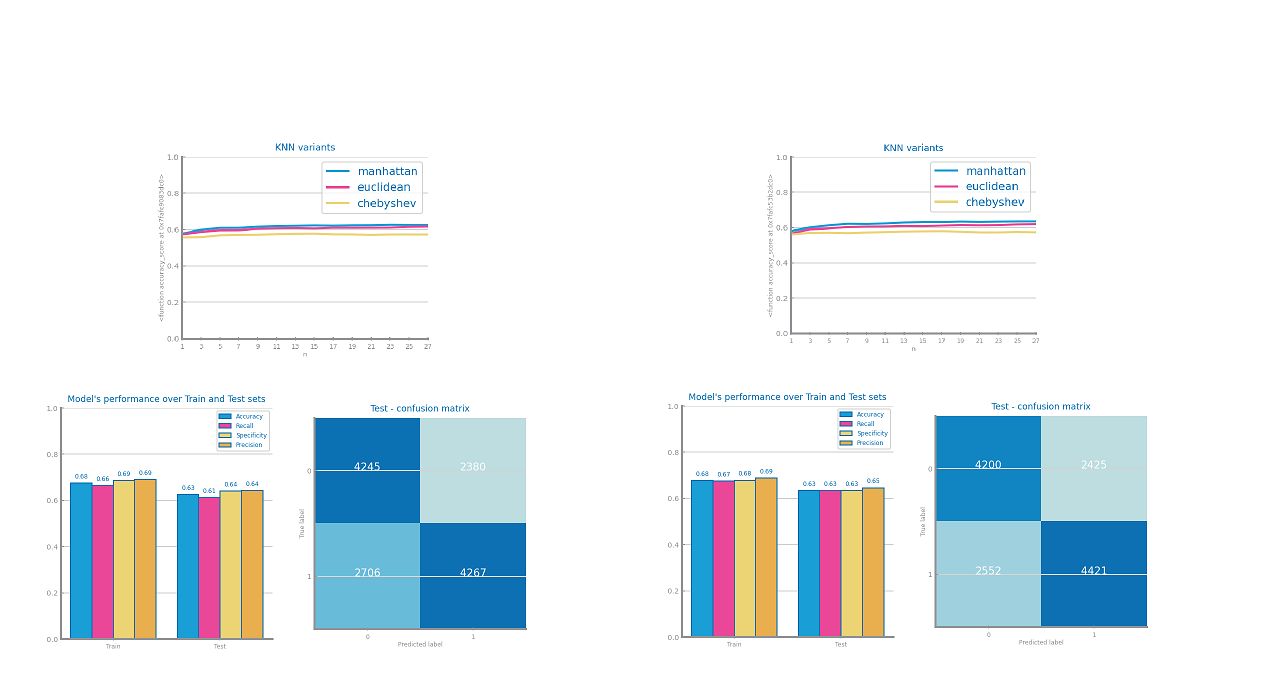


Figure 20 Scaling results with different approaches for dataset 2

## Balancing

We tested the three balancing methods on both datasets, over(right), under(center) and smote(left). After comparing the knn results we decided to choose the over method because it’s accuracy was greater than the others.

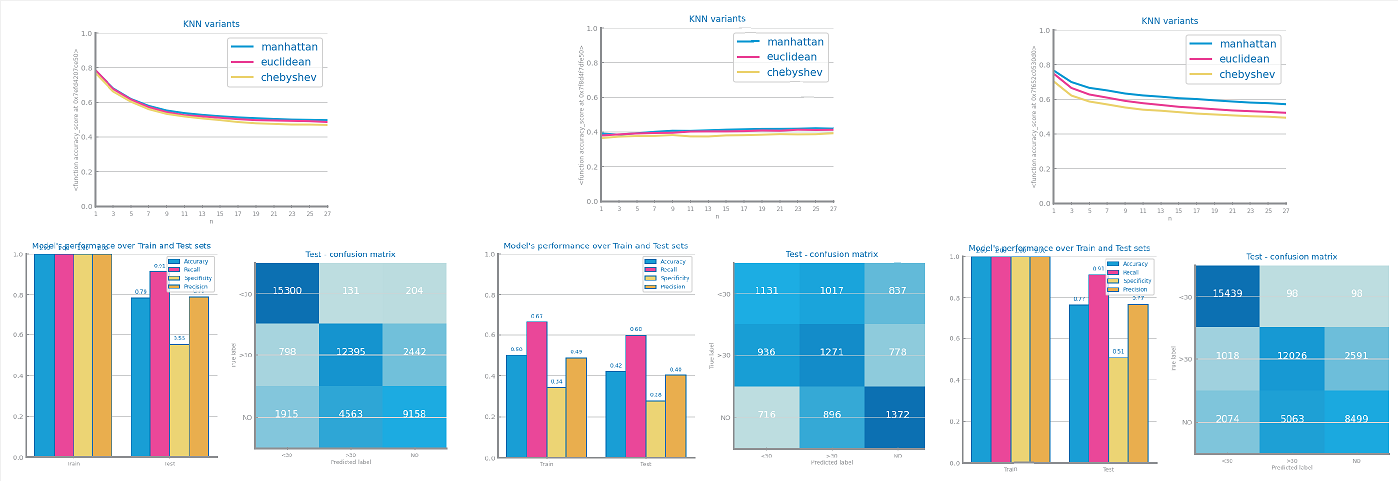


Figure 21 Balancing results with different approaches for dataset 1

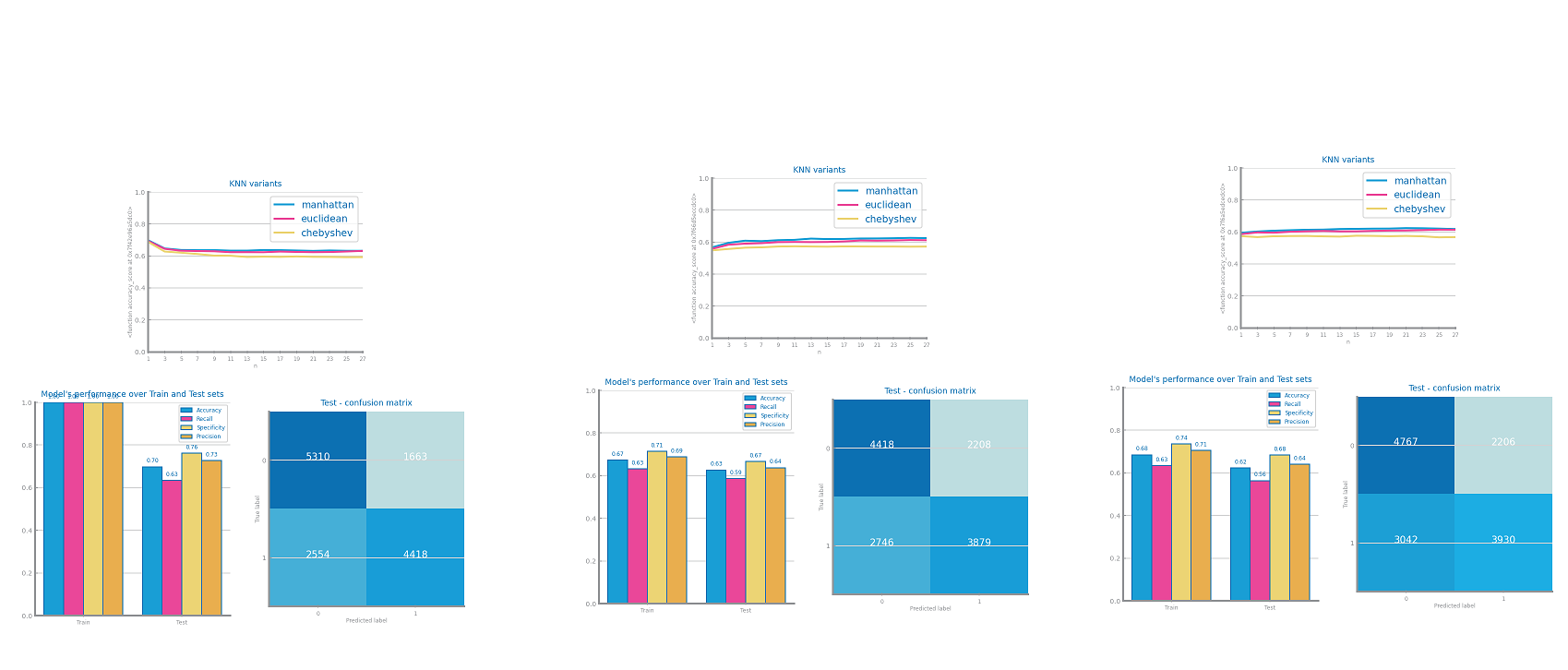
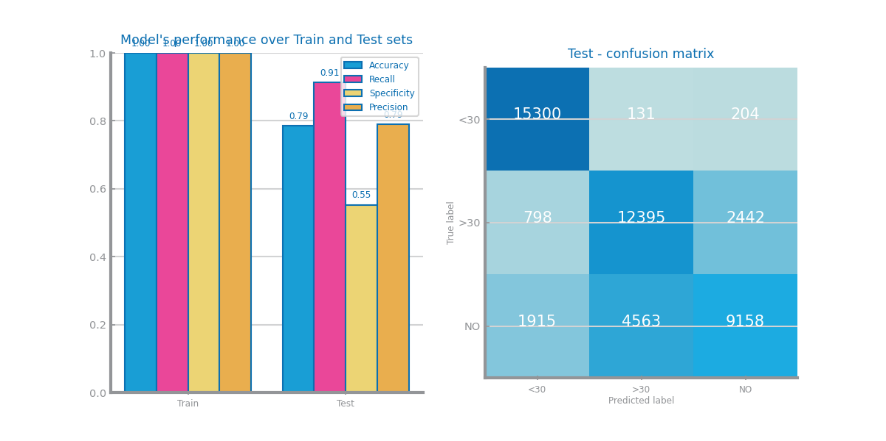
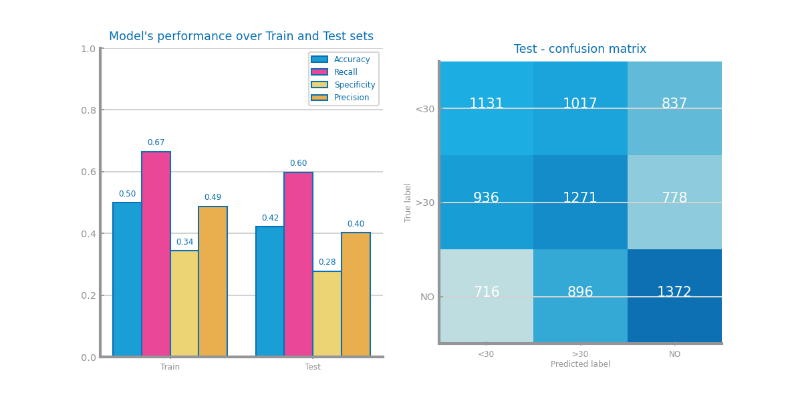


Figure 22 Balancing results with different approaches for dataset 2

## Feature Selection

For dataset 1 we dropped features that only had the same value in all records we did not drop based on correlation because there was no correlation over 90% on dataset 1. On dataset 2 we dropped features that had a correlation over 90%.



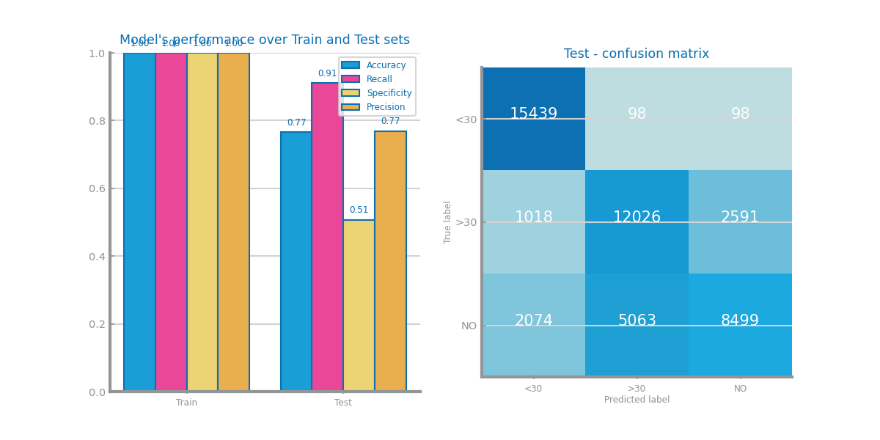


Figure 23 Feature selection of redundant variables results with different parameters for dataset 1

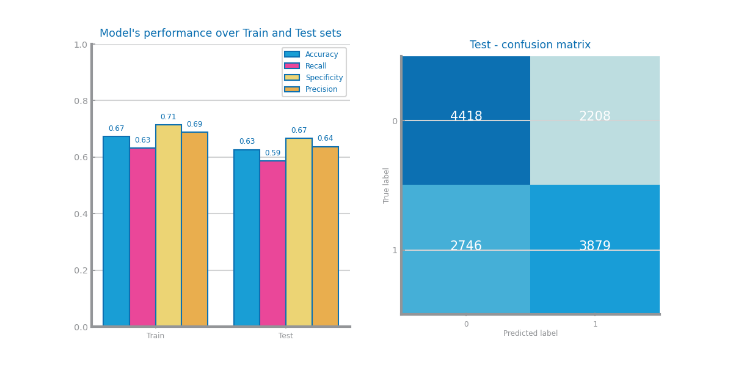
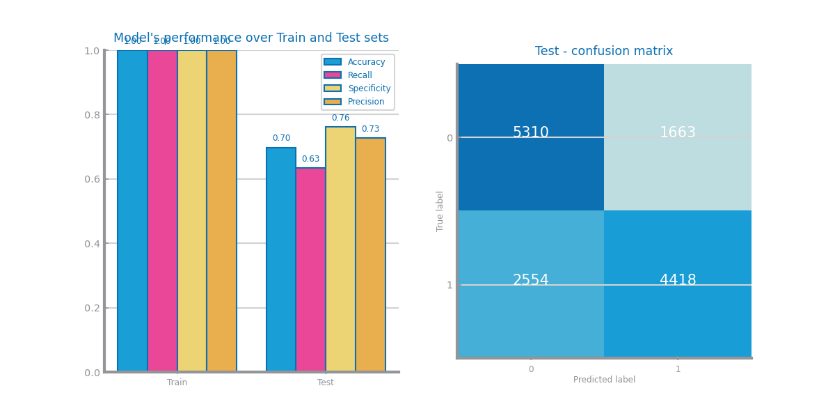


Figure 24 Feature selection of redundant variables results with different parameters for dataset 2

## Feature Extraction (optional)

Shall 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 25 Principal components analysis and feature extraction results for dataset 1

Figure 26 Principal components analysis and feature extraction results for dataset 2

## Feature Generation (optional)

Shall contain all relevant information and charts respecting to feature generation. The different choices and their impact on the modelling results shall be presented and explained. Shall summarize all variables generated and the formula used to derive them (in a table). **Shall not exceed 300 characters.**

Figure 27 Feature generation results for dataset 1

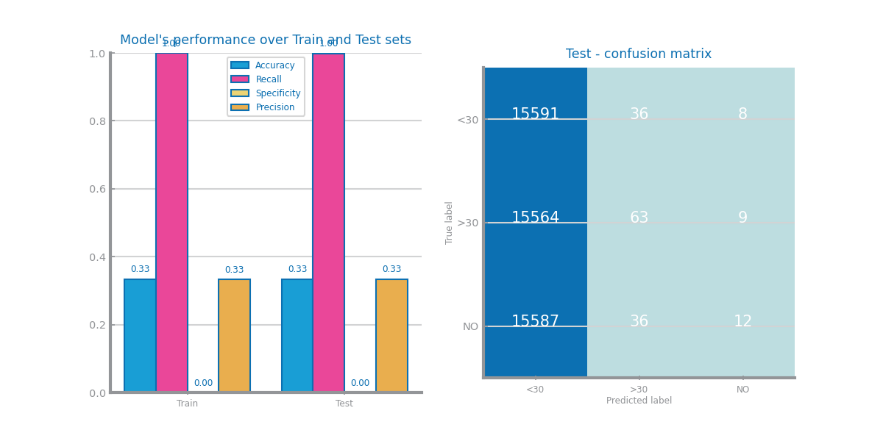
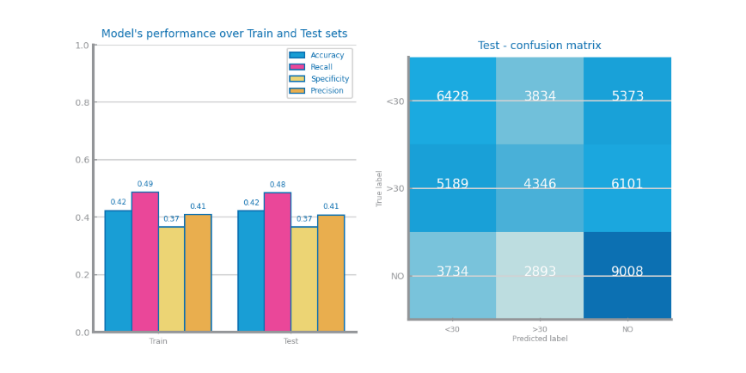
Figure 28 Feature generation results for dataset 2

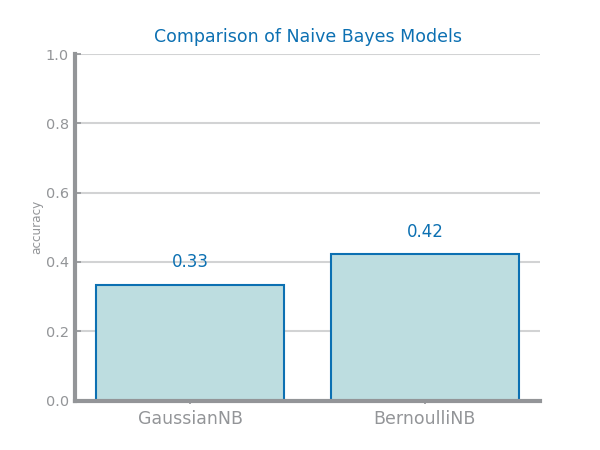
# Models’ Evaluation

In the model’s evaluation we used the train-test split technique with 70% of the data to train and 30% to test. To evaluate the results we used accuracy, recall, specificity and precision. For dataset 1 we considered ‘<30’ and ‘>30’ has positive values and ‘NO’ has negative values in the calculation of the evaluation metrics.

## Naïve Bayes

For both datasets we only used the Gaussian(left) and Bernoulli (right) Naïve Bayes because we had negative values in the datasets so we could not use Multinomial NB. In dataset 1 Bernoulli was better than Gaussian by a large margin in dataset 2 the two methods were almost equal.





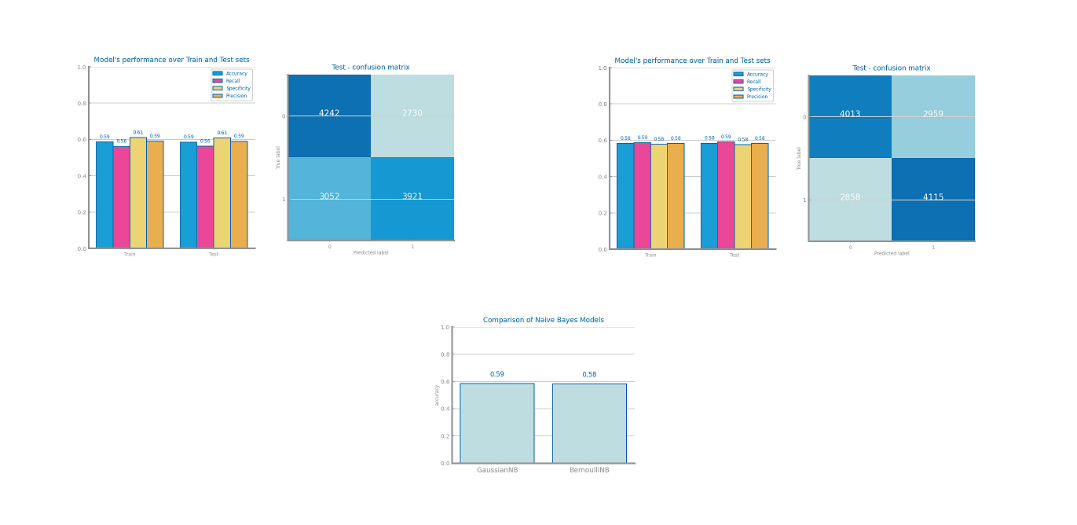
Figure 29 Naïve Bayes alternatives comparison for dataset 1

Figure 30 Naïve Bayes alternative comparison for dataset 2

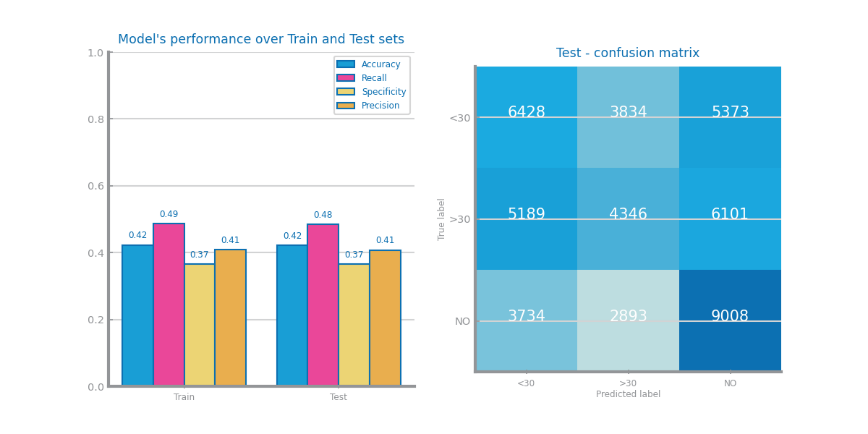
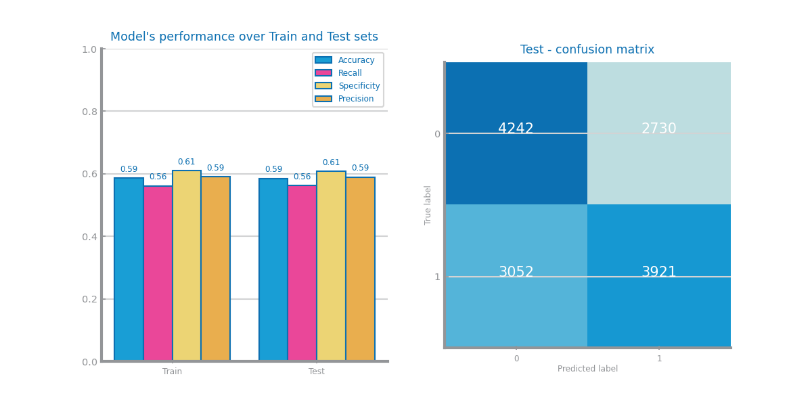
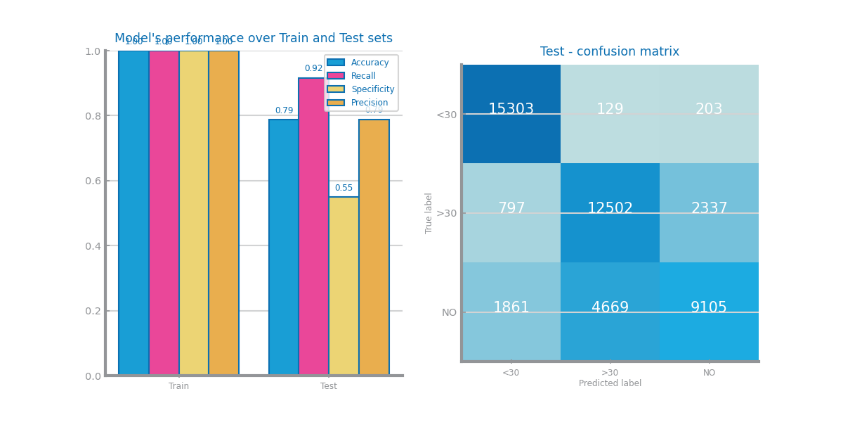
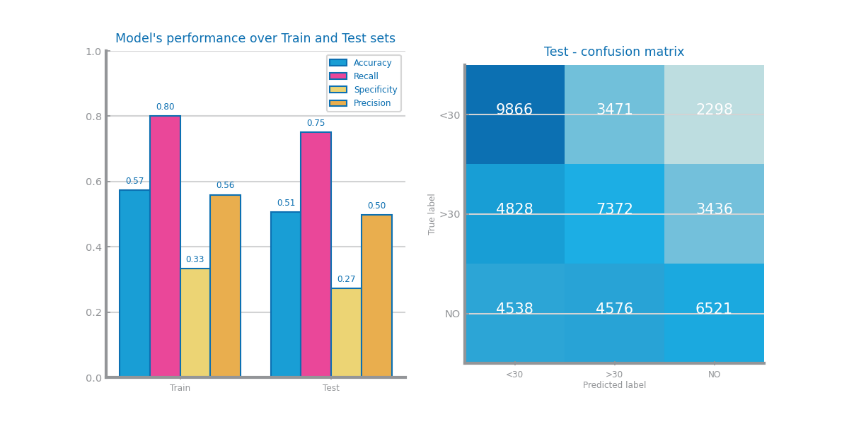


Figure 31 Naïve Bayes best model results for dataset 1 (left) and dataset 2 (right)

## KNN

In both datasets we tested with K’s in [1,..,27] and three distance methods as seen in the multiple line chart (down-right) and we show the evaluation metrics and confusion matrix with the best distance method for K=1 (up-left), K=19 (up-right) and K=27 (down-left). As we can see from the results for both datasets the best K was 1 neighbor with manhattan distance however this also makes the model with 100% accuracy to the train set which may be caused by the over balancing method. Despite this, we chose the K=1 with Manhattan has the best model because it had the bigger accuracy for the test set.



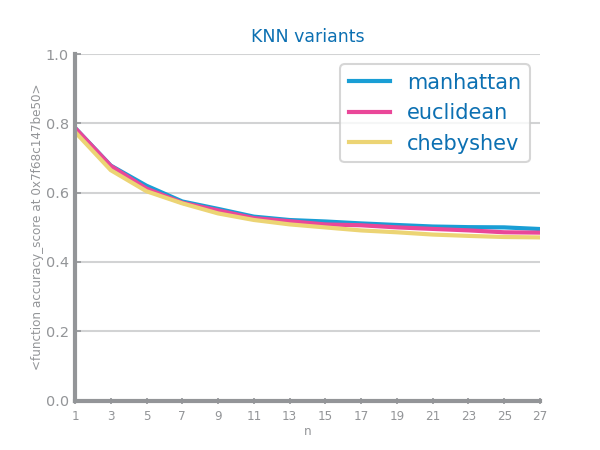
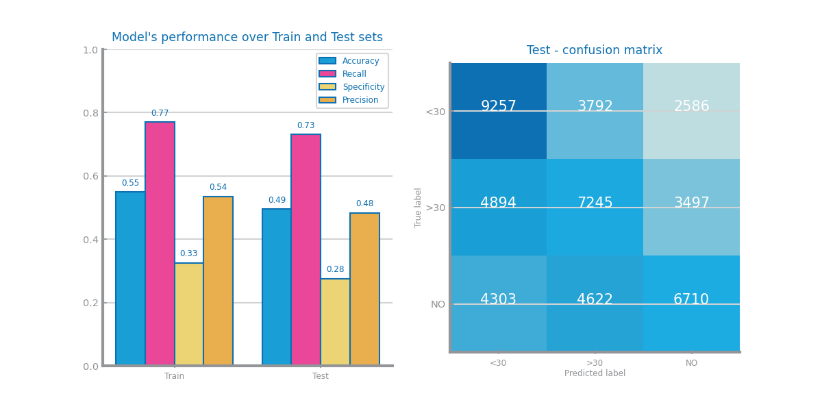


Figure 32 KNN different parameterizations comparison for dataset 1

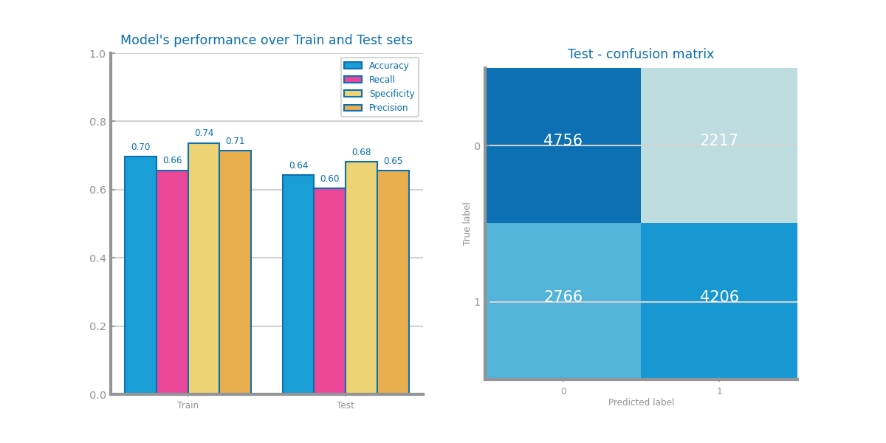
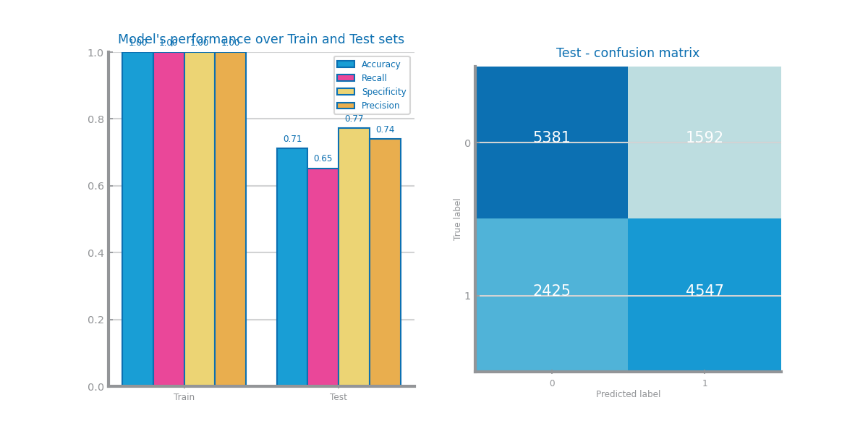




Figure 33 KNN different parameterizations comparison for dataset 2

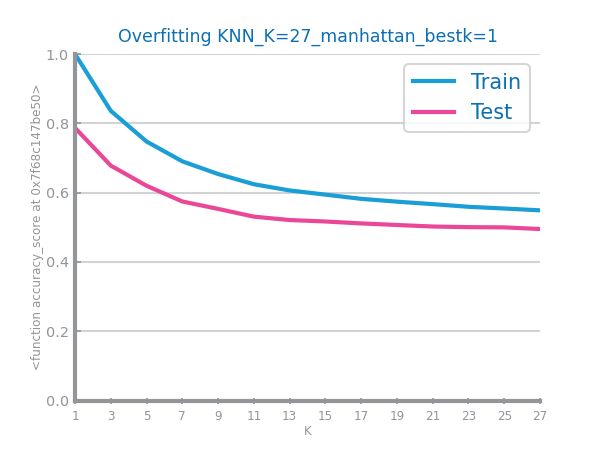
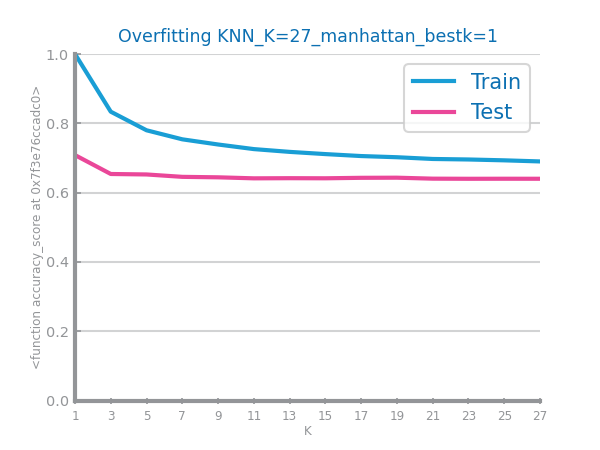


Figure 34 KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

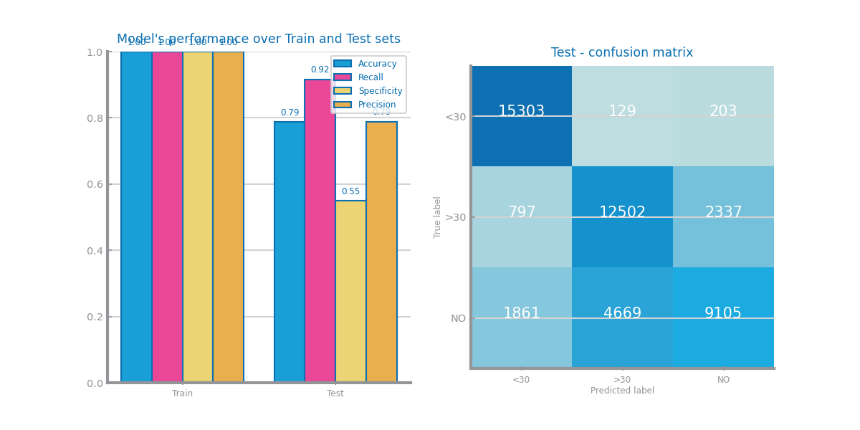
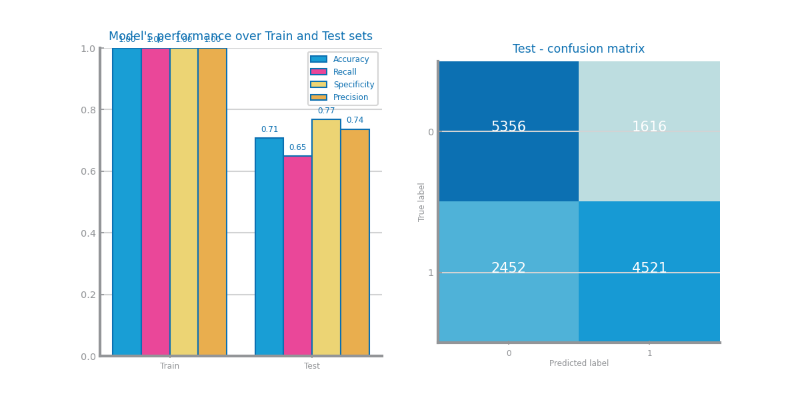


Figure 35 KNN best model results for dataset 1 (left) and dataset 2 (right)

## Decision Trees

Shall be used to present the results achieved through different parameterizations 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 parameterizations comparison for dataset 1

Figure 37 Decision Trees different parameterizations 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 trees for dataset 2

## Random Forests

Shall be used to present the results achieved through different parameterizations for the train of random forests. The results shall be compared and explanations for them shall be presented. Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. May be used to present the most important variables in the model. **Shall not exceed 500 characters**

Figure 42 Random Forests different parameterizations comparison for dataset 1

Figure 43 Random Forests different parameterizations comparison for dataset 2

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 parameterizations for the train of gradient boosting. The results shall be compared and explanations for them shall be presented. Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. May be used to present the most important variables in the model. **Shall not exceed 500 characters**

Figure 47 Gradient boosting different parameterizations comparison for dataset 1

Figure 48 Gradient boosting different parameterizations comparison for dataset 2

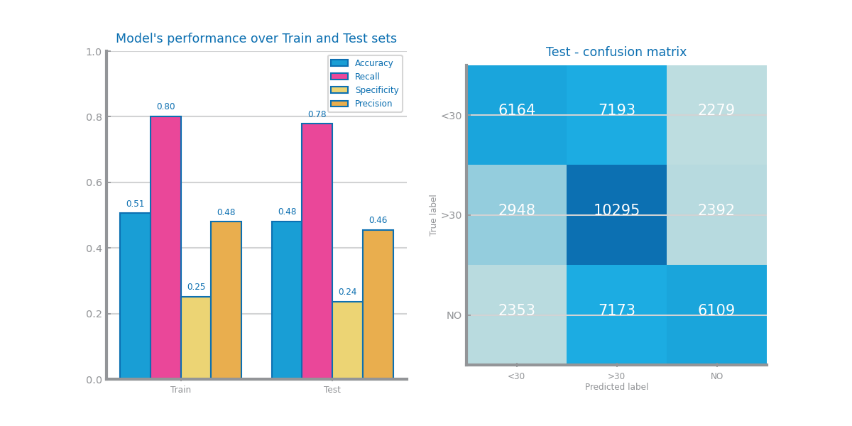
Figure 49 Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

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 parameterizations for the train of MLPs. The results shall be compared and explanations for them shall be presented. Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it. In particular by analysing the loss\_curve\_ available at the end of each train. Shall be used to present the evaluation of the best model achieved. **Shall not exceed 500 characters**

Chart, bar chart

Description automatically generated

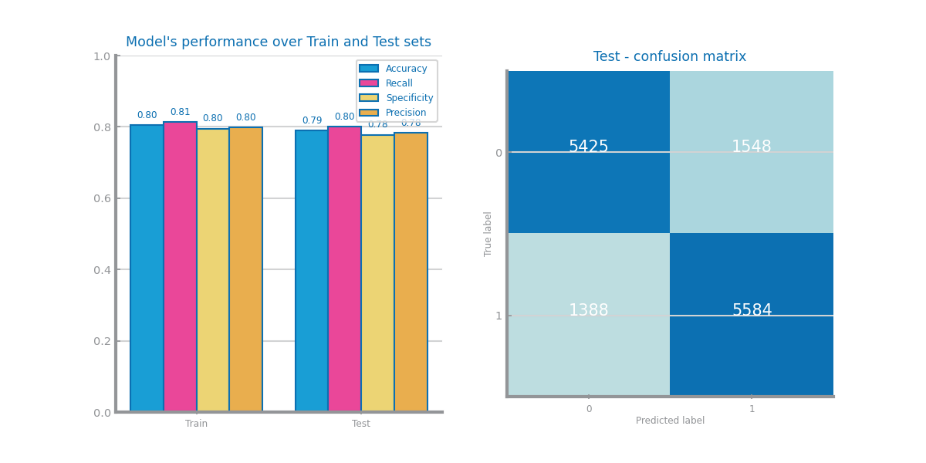
Chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 52 MLP different parameterizations comparison for dataset 1

Chart

Description automatically generatedChart, bar chart

Description automatically generated

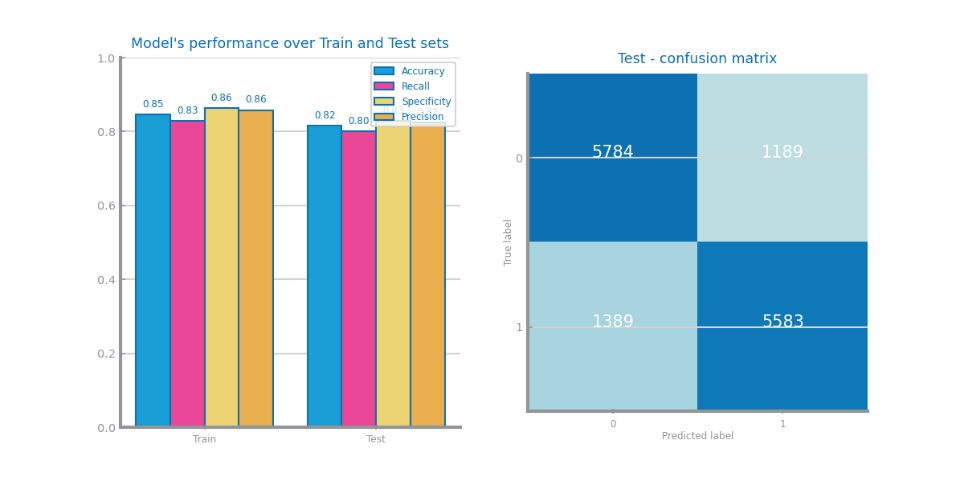


Figure 53 MLP different parameterizations comparison for dataset 2

Graphical user interface, chart, line chart

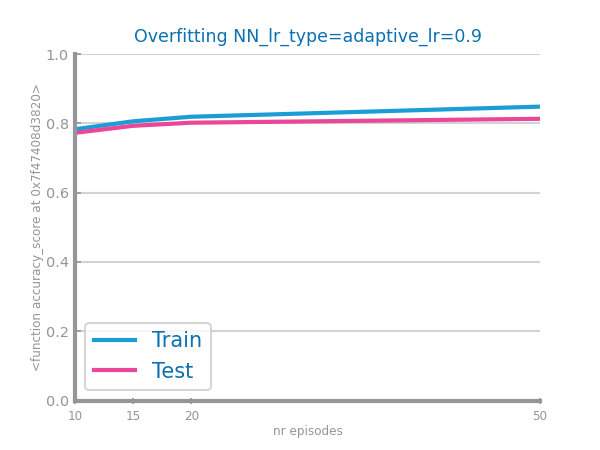
Description automatically generated

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

Figure 55 Loss curves analysis for dataset 1 (left) and dataset 2 (right)

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

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

# Critical Analysis

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

A cross-analysis of the different models 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 Forecasting

# Data Profiling

## Data Granularity

May be used to identify the most atomic granularity and two other different granularities to consider. **Shall not exceed 300 characters.**

Figure 57 Time series 1 at the most granular detail

Figure 58 Time series 1 at the second chosen granularity

Figure 59 Time series 1 at the third chosen granularity

Figure 60 Time series 2 at the most granular detail

Figure 61 Time series 2 at the second chosen granularity

Figure 62 Time series 2 at the third chosen granularity

## Data Distribution and Stationarity

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

Figure 63 Boxplot(s) for time series 1

Figure 64 Boxplot(s) for time series 2

Figure 65 Histogram(s) for time series 1

Figure 66 Histogram(s) for time series 2

Figure 67 Stationarity study for time series 1

Figure 68 Stationarity study for time series 2

# Data Transformation

## Aggregation

Shall describe the results of applying the persistence model over the three different aggregations over both datasets, and identifying the granularity chosen to proceed. **Shall not exceed 200 characters.**

Figure 69 Forecasting plots after different aggregations on time series 1

Figure 70 Forecasting results after different aggregations on time series 1

Figure 71 Forecasting plots after different aggregations on time series 2

Figure 72 Forecasting results after different aggregations on time series 2

## Smoothing

Shall describe the results of applying the persistence model over different smoothing transformations over both datasets, and identifying the best result to proceed. **Shall not exceed 200 characters.**

Figure 73 Forecasting plots after different smoothing parameterizations on time series 1

Figure 74 Forecasting results after different smoothing parameterizations on time series 1

Figure 75 Forecasting plots after different smoothing parameterizations on time series 2

Figure 76 Forecasting results after different smoothing parameterizations on time series 2

## Differentiation

Shall describe the results of applying the persistence model over two consecutive differentiation of both datasets, and identifying the best result to proceed. **Shall not exceed 200 characters.**

Figure 77 Forecasting plots after first and second differentiation of time series 1

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

Figure 79 Forecasting plots after first and second differentiation of time series 2

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

# Models’ Evaluation

Shall be used to summarize the transformations done over the original time series. **Shall not exceed 200 characters.**

## Simple Average Model

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

Figure 81 Forecasting plots obtained with Simple Average model over time series 1

Figure 82 Forecasting plots obtained with Simple Average model over time series 2

## Persistence Model

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

Figure 83 Forecasting plots obtained with Persistence model over time series 1

Figure 84 Forecasting plots obtained with Persistence model over time series 2

## Rolling Mean Model

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

Figure 85 Forecasting study over different parameterizations of the rolling mean algorithm over time series 1

Figure 86 Forecasting plots obtained with the best parameterization of rolling mean algorithm, over time series 1

Figure 87 Forecasting results obtained with the best parameterization of rolling mean algorithm, over time series 1

Figure 88 Forecasting study over different parameterizations of the rolling mean algorithm over time series 2

Figure 89 Forecasting plots obtained with the best parameterization of rolling mean algorithm, over time series 2

Figure 90 Forecasting results obtained with the best parameterization of rolling mean algorithm, over time series 2

## ARIMA Model

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

Figure 91 Forecasting study over different parameterizations of the ARIMA algorithm over time series 1

Figure 92 Forecasting plots obtained with the best parameterization of ARIMA algorithm, over time series 1

Figure 93 Forecasting results obtained with the best parameterization of ARIMA algorithm, over time series 1

Figure 94 Forecasting study over different parameterizations of the ARIMA algorithm over time series 2

Figure 95 Forecasting plots obtained with the best parameterization of ARIMA algorithm, over time series 2

Figure 96 Forecasting results obtained with the best parameterization of ARIMA algorithm, over time series 2

## LSTMs Model

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

Figure 97 Forecasting study over different parameterizations of LSTMs over time series 1

Figure 98 Forecasting plots obtained with the best parameterization of LSTMs, over time series 1

Figure 99 Forecasting results obtained with the best parameterization of LSTMs, over time series 1

Figure 100 Forecasting study over different parameterizations of the LSTMs over time series 2

Figure 101 Forecasting plots obtained with the best parameterization of LSTMs, over time series 2

Figure 102 Forecasting results obtained with the best parameterization of LSTMs, over time series 2

# Critical Analysis

Shall 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.**