DS 2022

Data Science Project

Team nr: 8	**Student 1:** Filipa Cotrim	**IST nr:** 95572
	Student 2: Leonor Barreiros	**IST nr:** 95618
	Student 3: Martim Santos	**IST nr:** 95638

CLASSIFICATION

1 DATA PROFILING

In set 1 we used the paper to process variables, eg only 9 important diag groups. In 2 we used domain knowledge, eg temperatures can be negative to know if negative/0/big values were wrong/missing.

Data Dimensionality

In both the nr records is enough for the nr variables: no dimensionality curse. In 1, data is mostly qualitative. In 2, data is mainly quantitative: suggests it's collected structurally (sensor data).

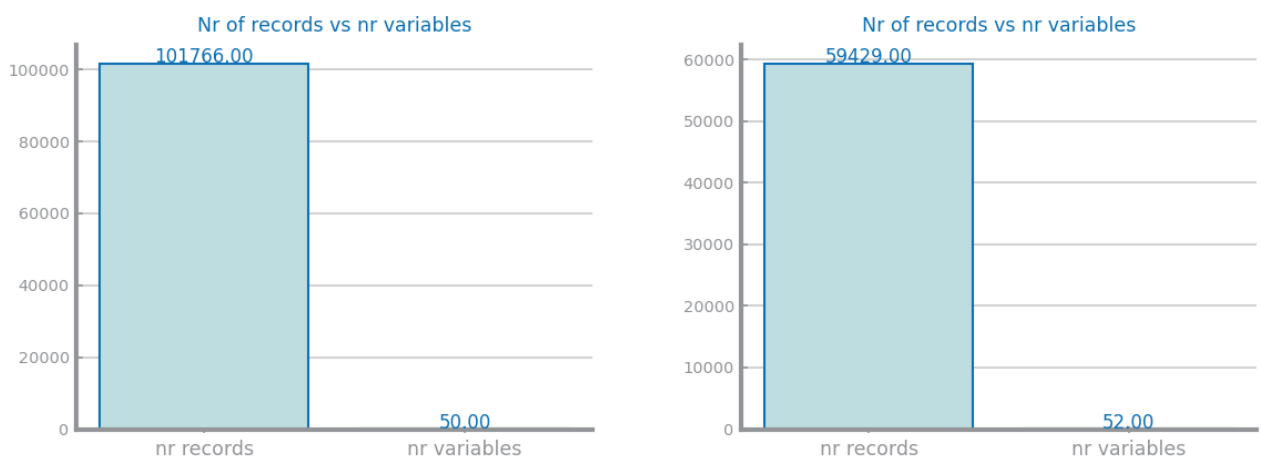


Figure 1: Nr Records x Nr variables for dataset 1 (left) and dataset 2 (right)

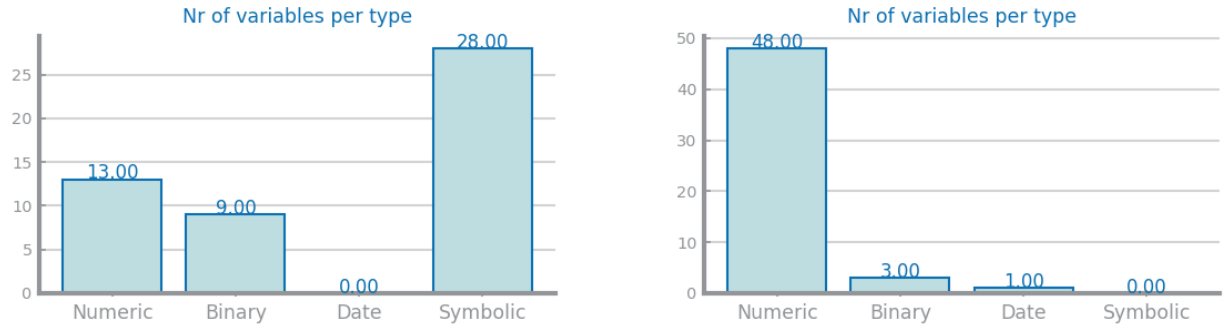


Figure 2: Nr variables per type for dataset 1 (left) and dataset 2 (right)

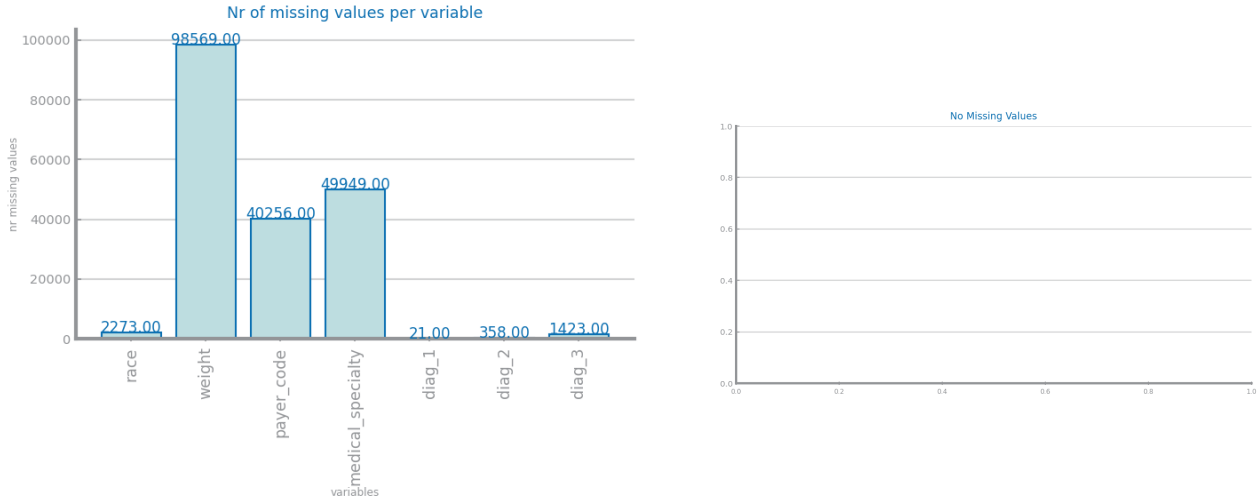
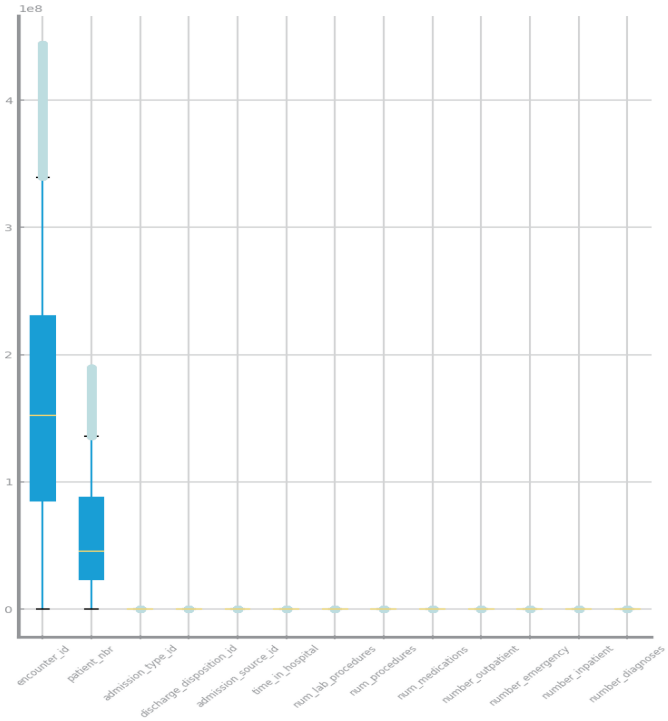


Figure 3: Nr missing values for dataset 1 (left) and dataset 2 (right)

Data Distribution

LogNorm didn't fit data so we only chose best distributions. For outliers, the 1.5IQR criteria is bad because some samples aren't outliers and are left out so we chose 3STDEV because it covered 80% of samples. We grouped ordinal variables according to their order (not bar size). We see class imbalance in dataset 1 - we must use recall and not accuracy especially since minority classes are the ones important to predict (this is for both datasets).



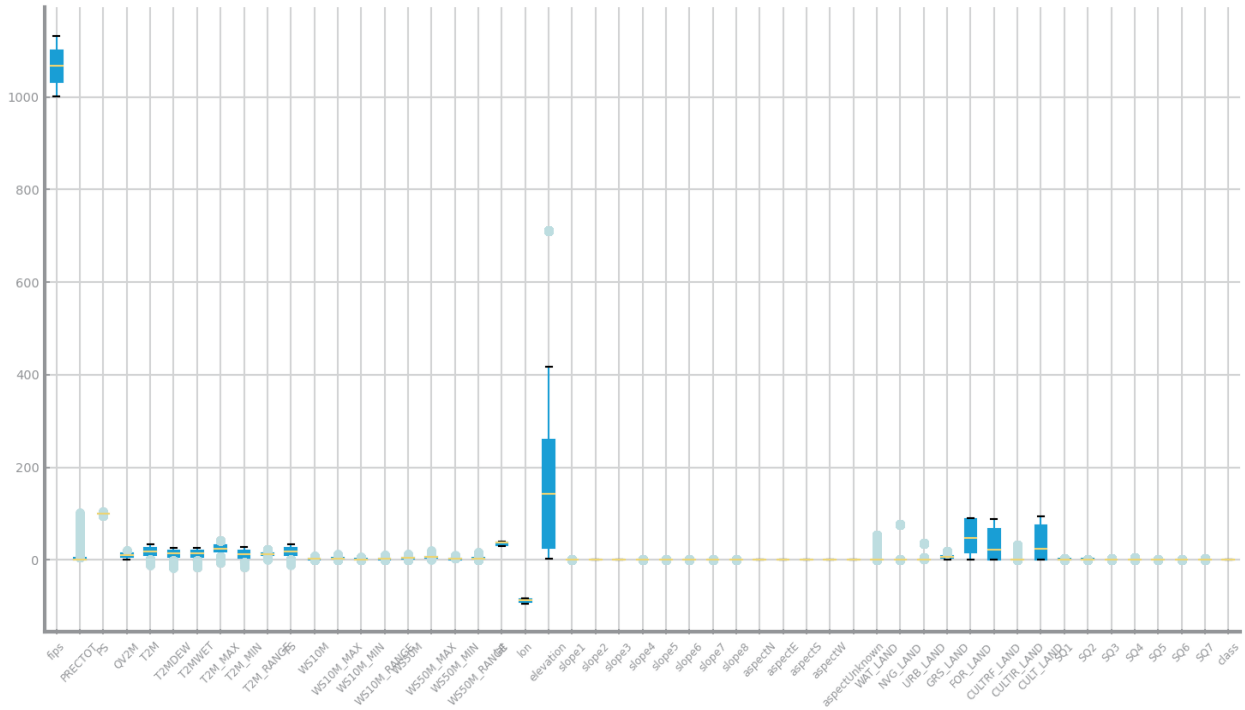


Figure 4: Global boxplots dataset 1 (up) and dataset 2 (down)

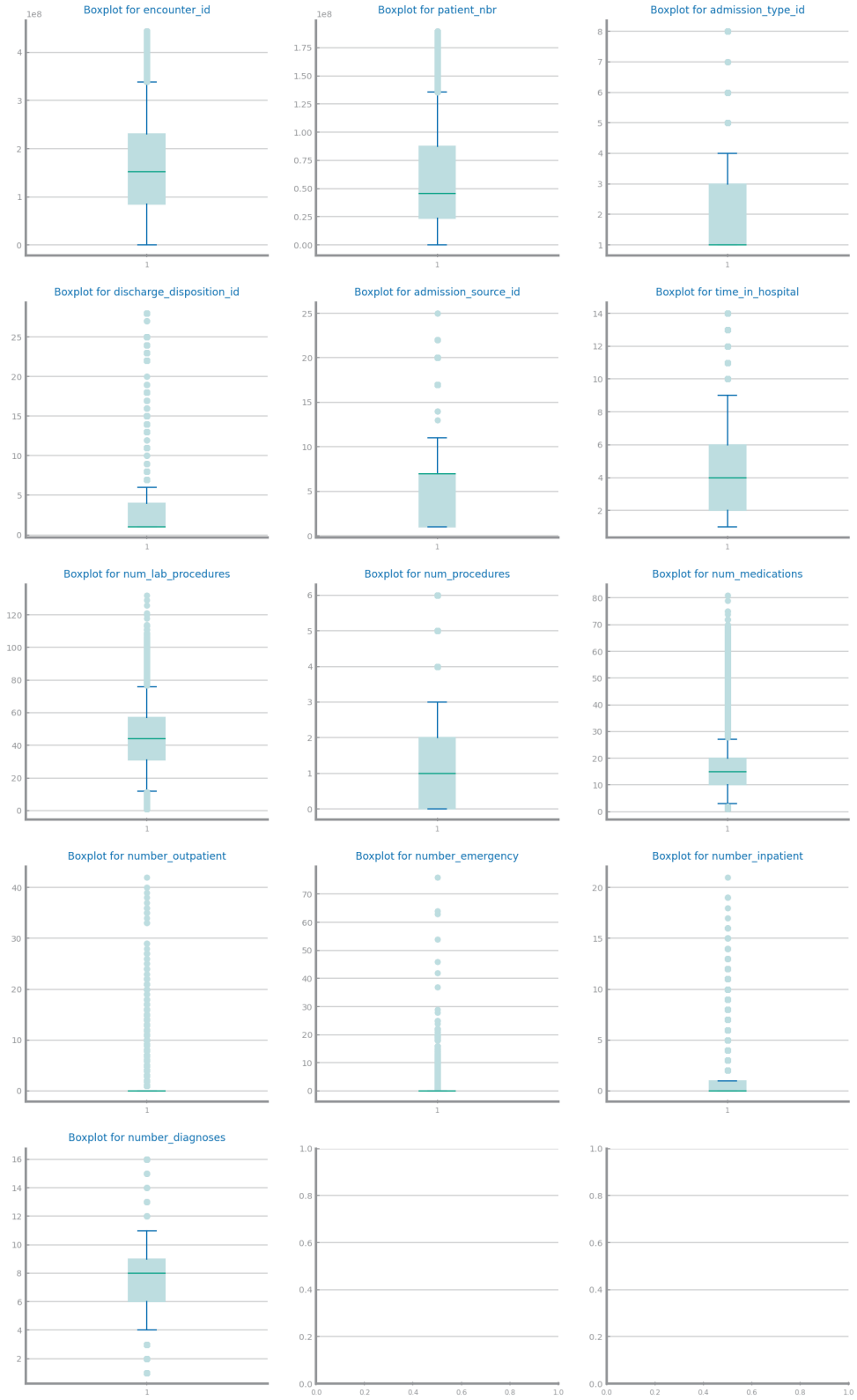
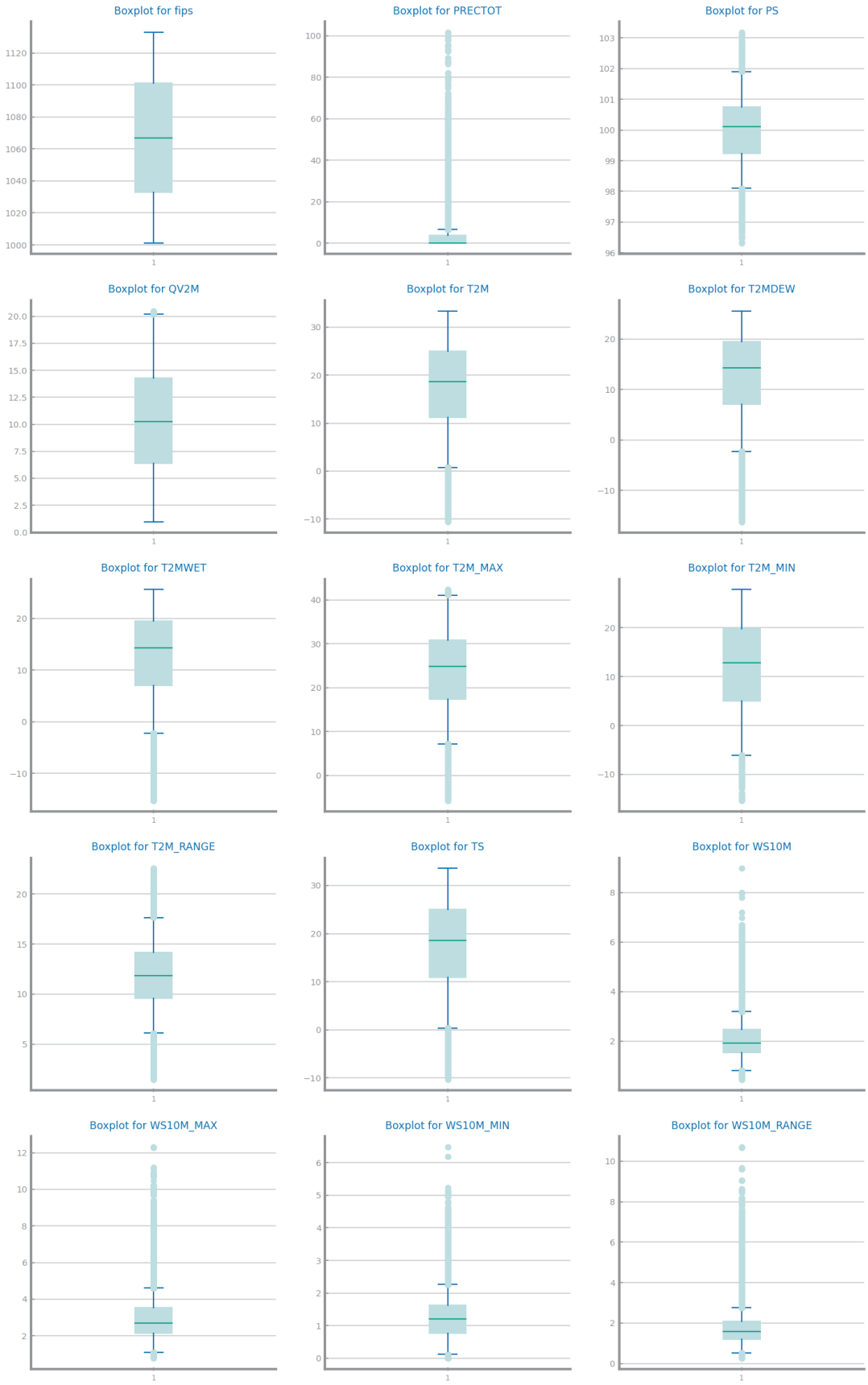
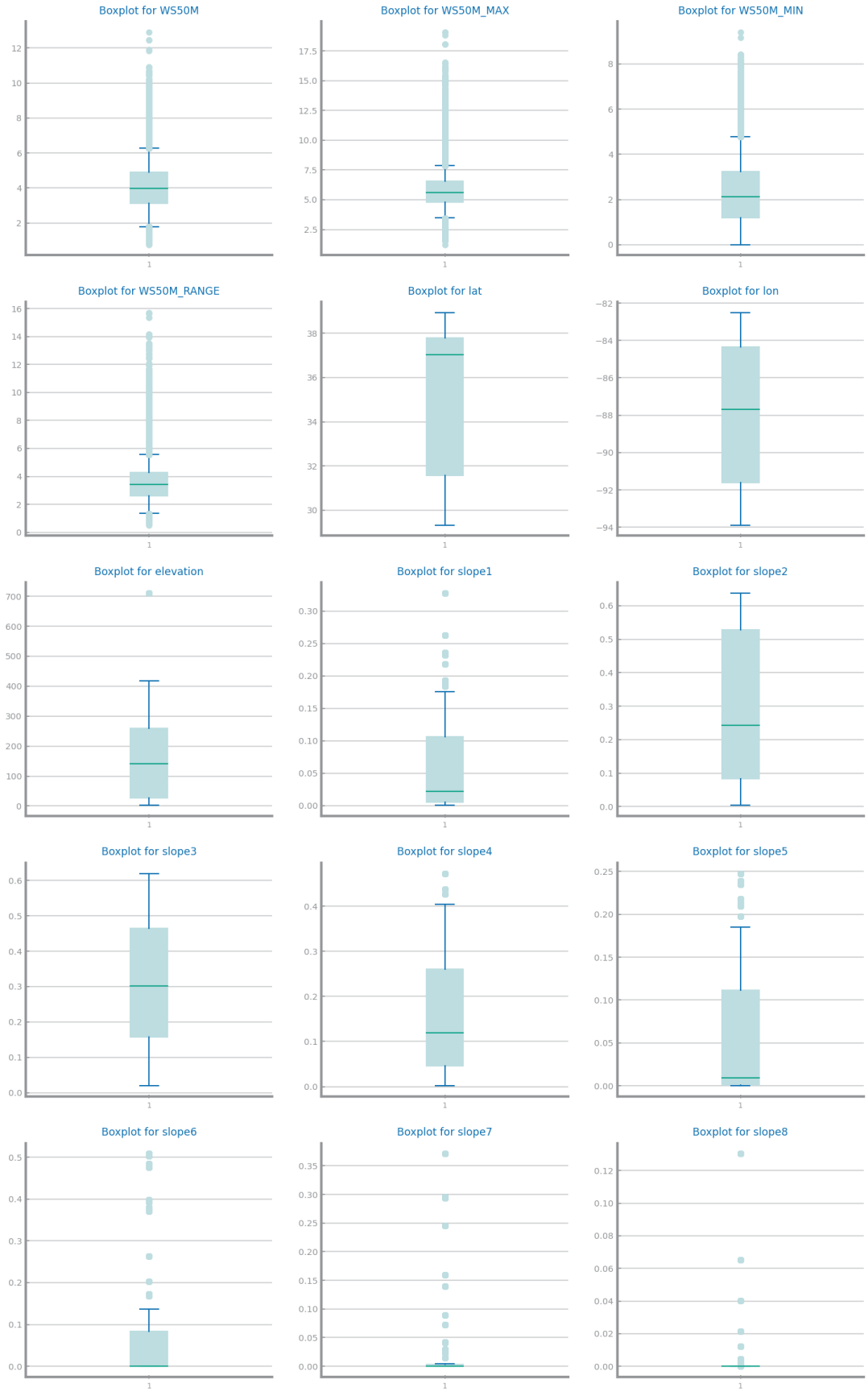
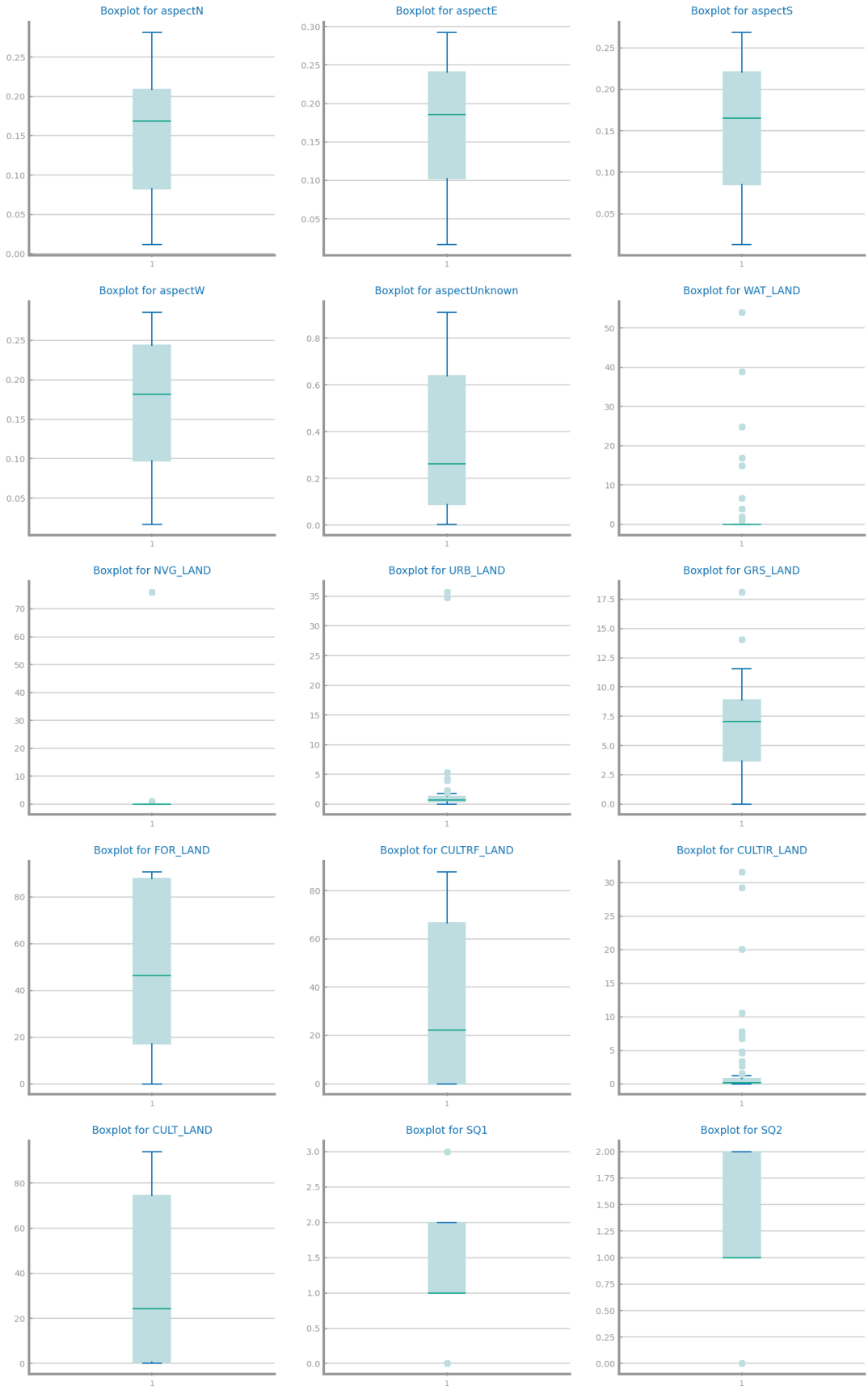


Figure 5: Single variable boxplots for dataset 1


Boxplot for fips
Boxplot for PRECTOT
Boxplot for PS
Boxplot for QV2M
Boxplot for T2M
Boxplot for T2MDEW
Boxplot for T2MWET
Boxplot for T2M_MAX
Boxplot for T2M_MIN
Boxplot for T2M_RANGE
Boxplot for TS
Boxplot for WS10M
Boxplot for WS10M_MAX
Boxplot for WS10M_MIN
Boxplot for WS10M_RANGE


Boxplot for WS50M
Boxplot for WS50M_MAX
Boxplot for WS50M_MIN
Boxplot for WS50M_RANGE
Boxplot for lat
Boxplot for lon
Boxplot for elevation
Boxplot for slope1
Boxplot for slope2
Boxplot for slope3
Boxplot for slope4
Boxplot for slope5
Boxplot for slope6
Boxplot for slope7
Boxplot for slope8


Boxplot for aspectN
Boxplot for aspectE
Boxplot for aspectS
Boxplot for aspectW
Boxplot for aspectUnknown
Boxplot for WAT_LAND
Boxplot for NVG_LAND
Boxplot for URB_LAND
Boxplot for GRS_LAND
Boxplot for FOR_LAND
Boxplot for CULTRF_LAND
Boxplot for CULTIR_LAND
Boxplot for CULT_LAND
Boxplot for SQ1
Boxplot for SQ2

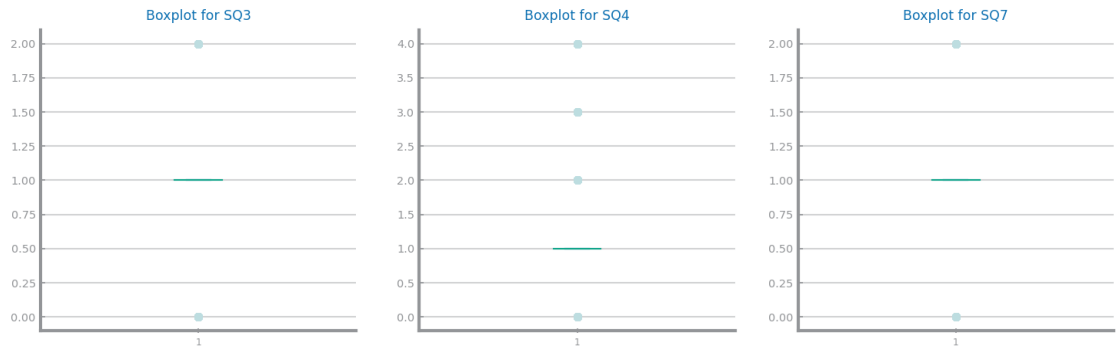


Figure 6: Single variable boxplots for dataset 2

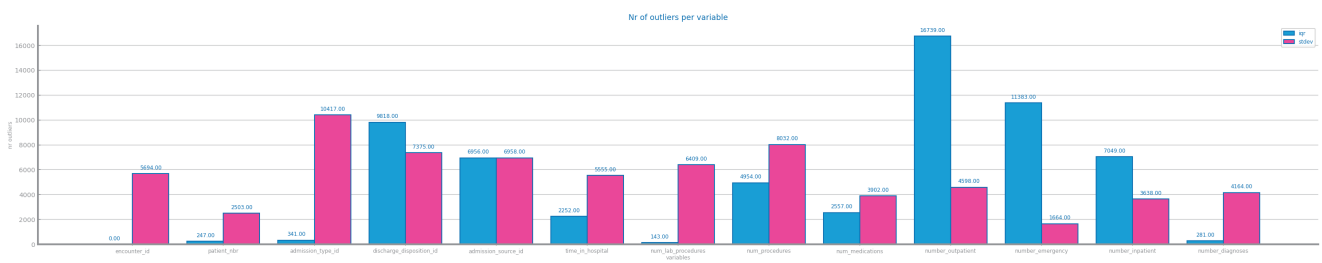


Figure 7: Outliers for dataset 1

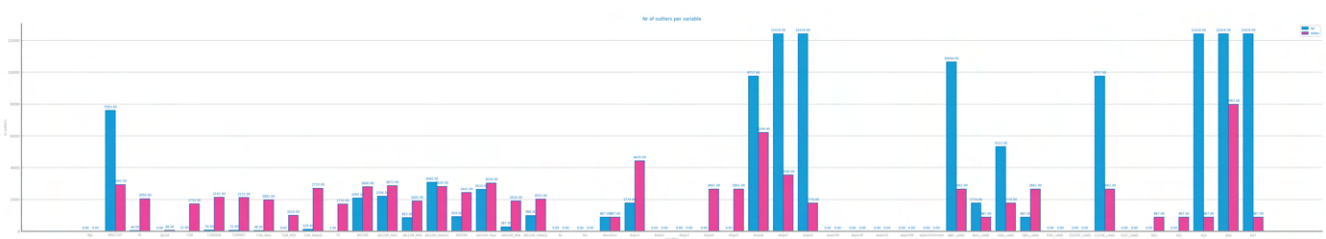


Figure 8: Outliers for dataset 2

Histogram for race
Histogram for gender
Histogram for age
Histogram for weight
Histogram for payer_code
Histogram for medical_specialty
Histogram for diag_1
Histogram for diag_2
Histogram for diag_3
Histogram for max_glu_serum
Histogram for A1Cresult
Histogram for metformin
Histogram for repaglinide
Histogram for nateglinide
Histogram for chlorpropamide

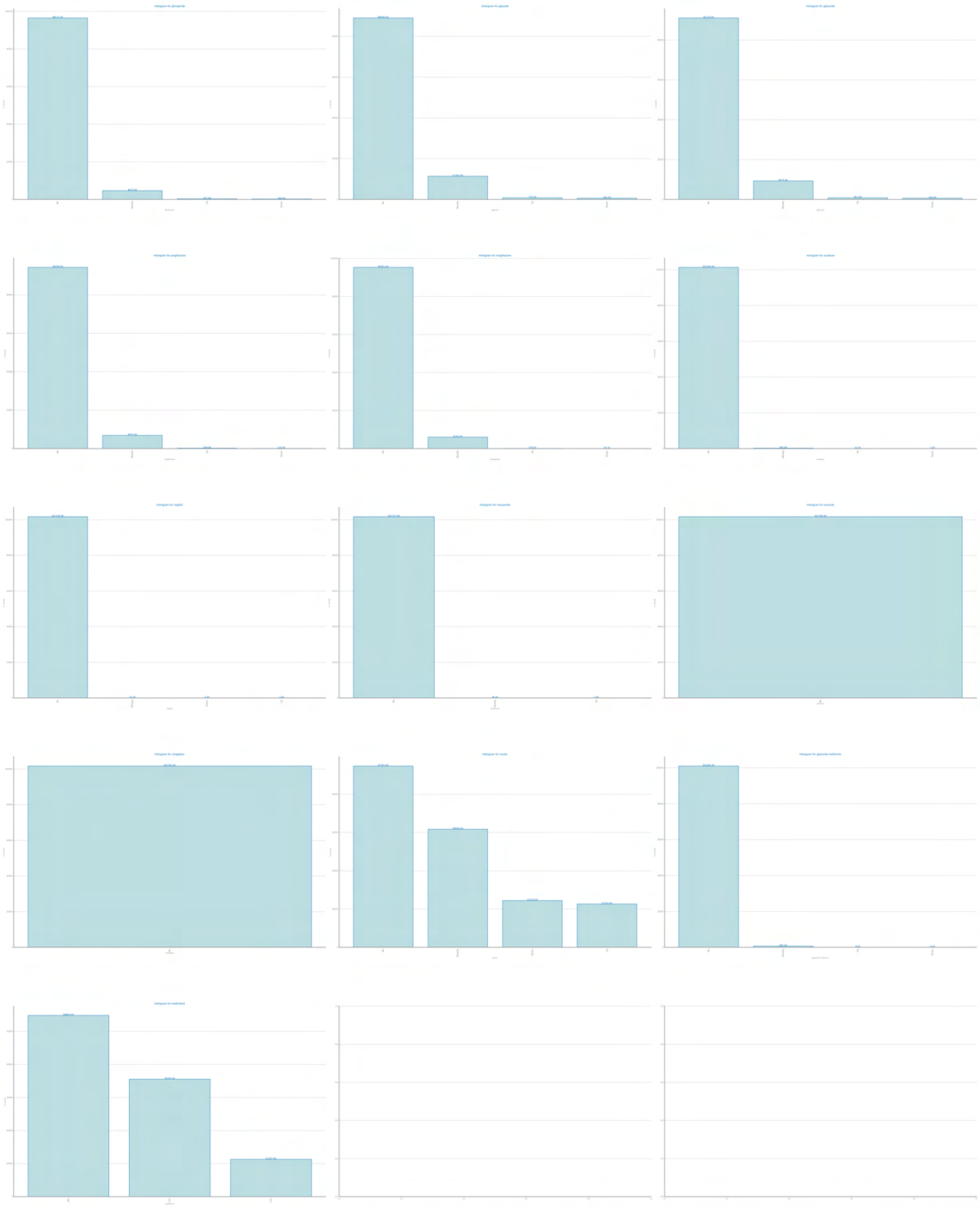


Figure 9: Histograms of symbolic variables for dataset 1

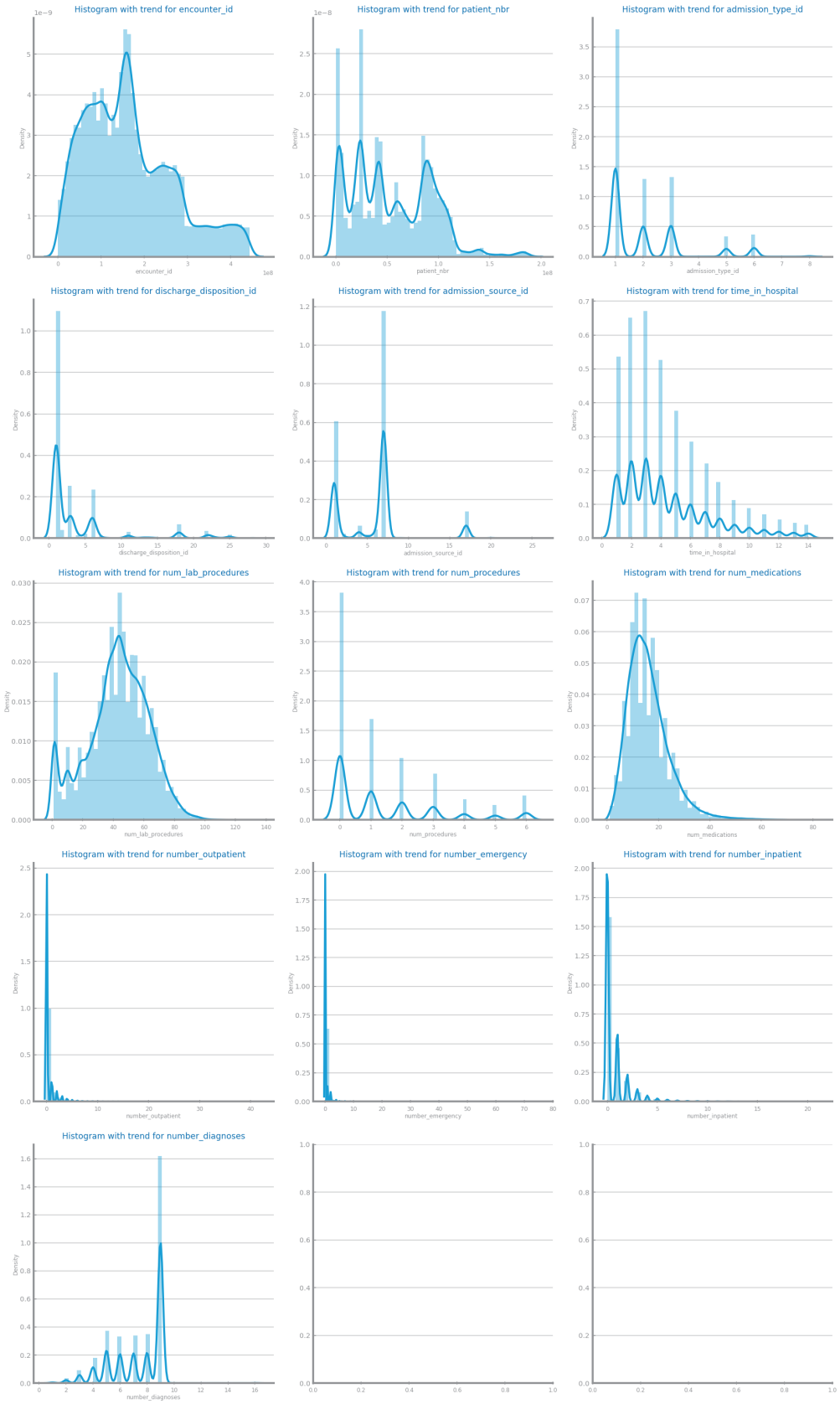


Figure 10: Histograms of numeric variables trend for dataset 1

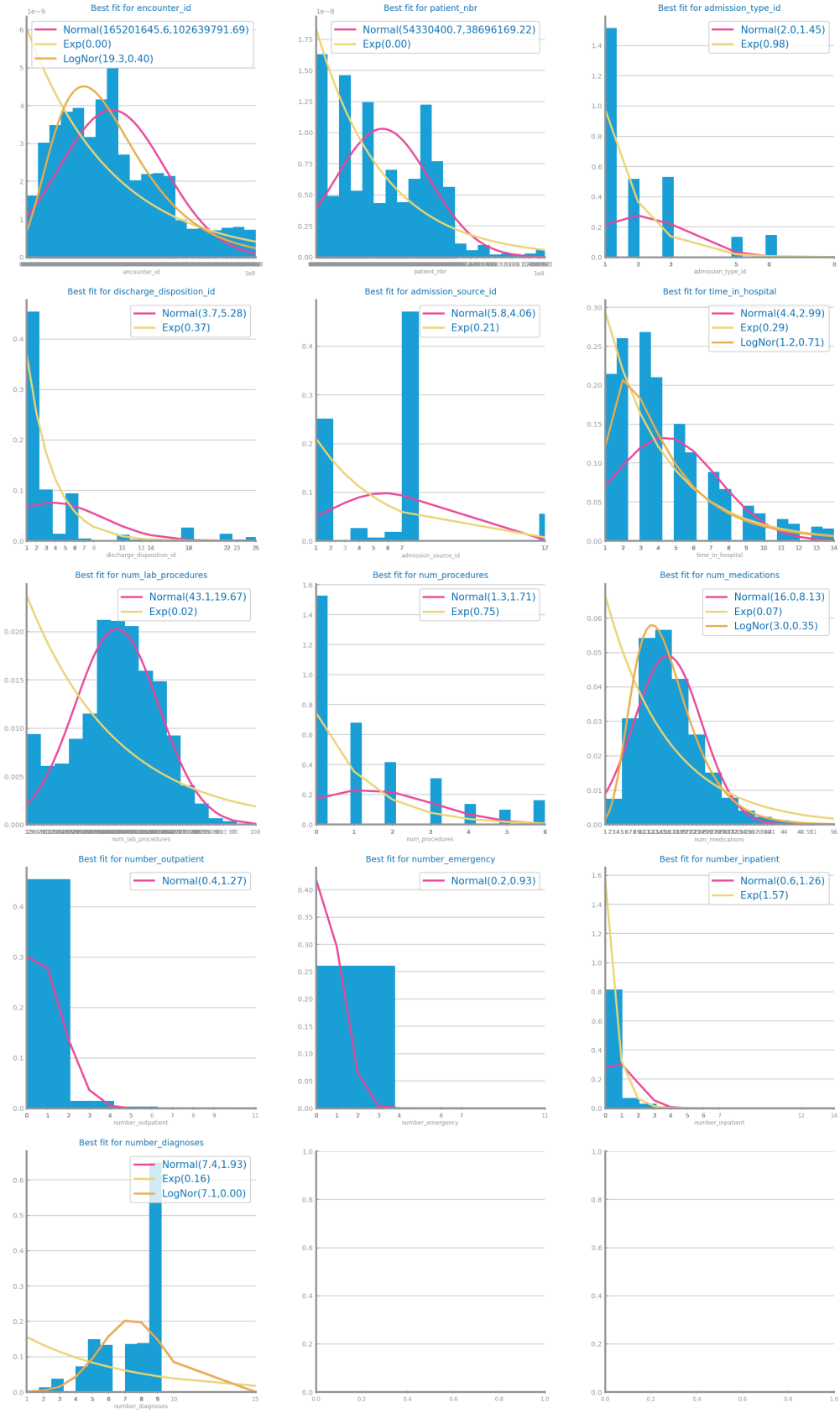


Figure 11: Histograms of numeric variables distributions for dataset 1

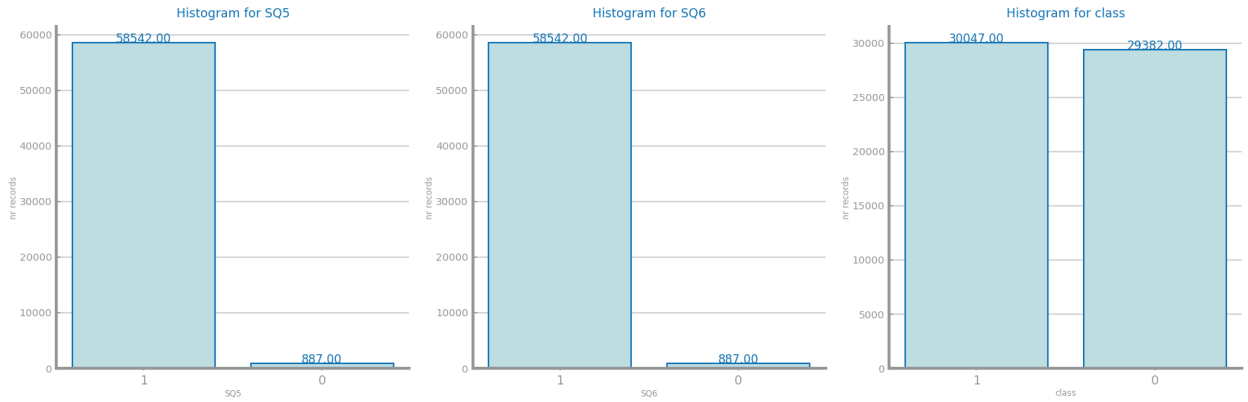
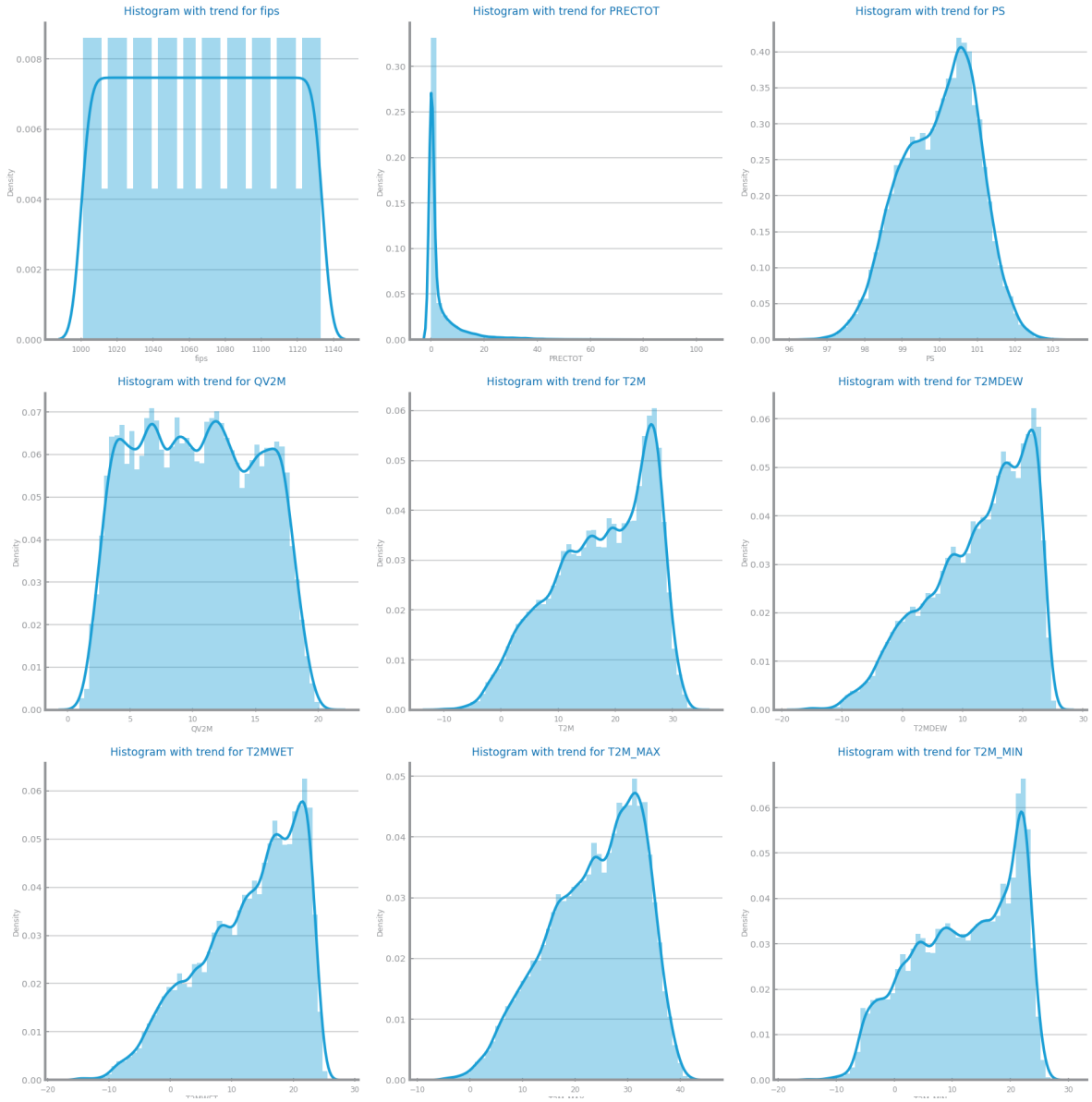
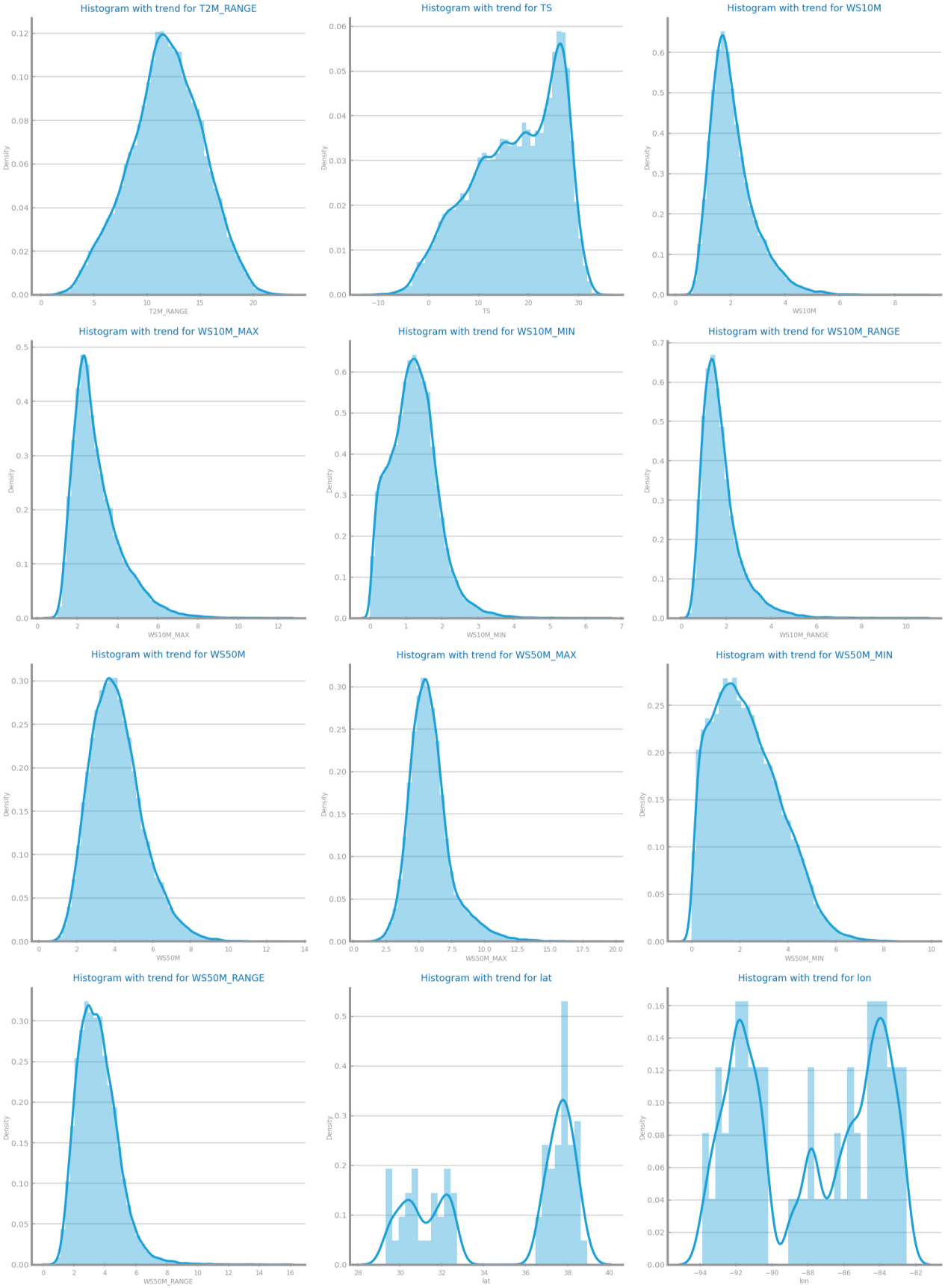


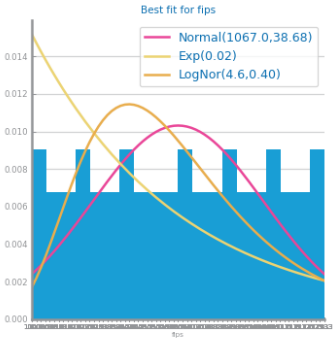
Figure 12: Histograms of symbolic variables for dataset 2

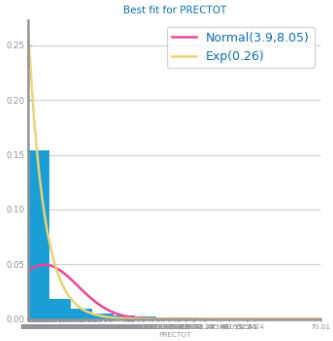


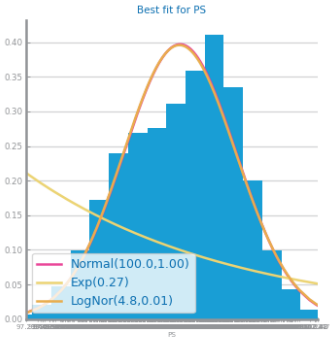

Histogram with trend for T2M_RANGE
Density
T2M_RANGE
Histogram with trend for TS
Density
TS
Histogram with trend for WS10M
Density
WS10M
Histogram with trend for WS10M_MAX
Density
WS10M_MAX
Histogram with trend for WS10M_MIN
Density
WS10M_MIN
Histogram with trend for WS10M_RANGE
Density
WS10M_RANGE
Histogram with trend for WS50M
Density
WS50M
Histogram with trend for WS50M_MAX
Density
WS50M_MAX
Histogram with trend for WS50M_MIN
Density
WS50M_MIN
Histogram with trend for WS50M_RANGE
Density
WS50M_RANGE
Histogram with trend for lat
Density
lat
Histogram with trend for lon
Density
lon

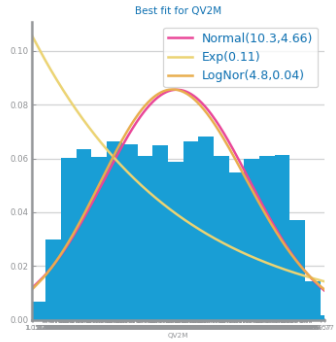
Histogram with trend for elevation
Histogram with trend for slope1
Histogram with trend for slope2
Histogram with trend for slope3
Histogram with trend for slope4
Histogram with trend for slope5
Histogram with trend for slope6
Histogram with trend for slope7
Histogram with trend for slope8
Histogram with trend for aspectN
Histogram with trend for aspectE
Histogram with trend for aspectS
Histogram with trend for aspectW
Histogram with trend for aspectUnknown
Histogram with trend for WAT_LAND
Density
elevation
slope1
slope2
slope3
slope4
slope5
slope6
slope7
slope8
aspectN
aspectE
aspectS
aspectW
aspectUnknown
WAT_LAND

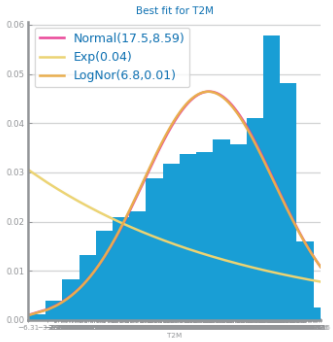
Figure 13: Histograms of numeric variables trend for dataset 2

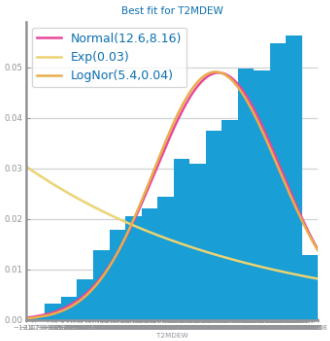

Best fit for fips
Normal(1067.0,38.68)
Exp(0.02)
LogNor(4.6,0.40)

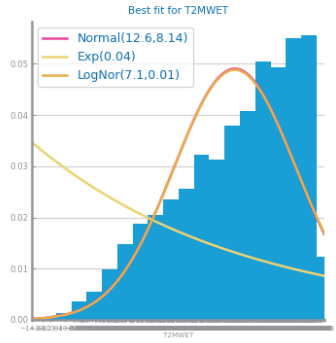

Best fit for PRECTOT
Normal(3.9,8.05)
Exp(0.26)

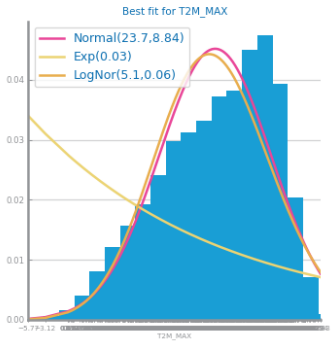

Best fit for PS
Normal(100.0,1.00)
Exp(0.27)
LogNor(4.8,0.01)

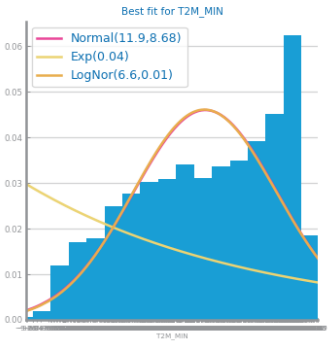

Best fit for QV2M
Normal(10.3,4.66)
Exp(0.11)
LogNor(4.8,0.04)

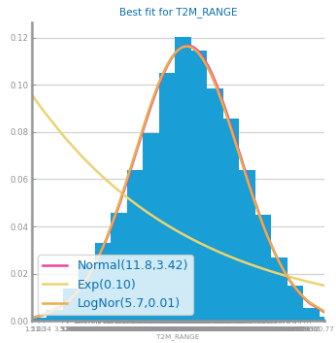

Best fit for T2M
Normal(17.5,8.59)
Exp(0.04)
LogNor(6.8,0.01)

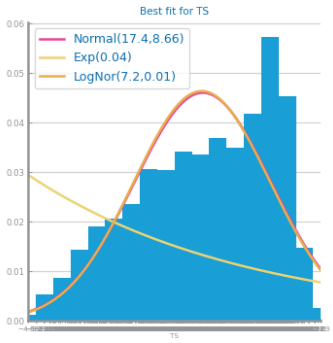

Best fit for T2MDEW
Normal(12.6,8.16)
Exp(0.03)
LogNor(5.4,0.04)

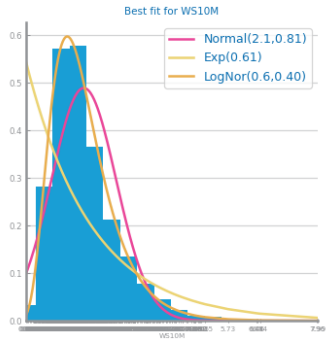

Best fit for T2MWET
Normal(12.6,8.14)
Exp(0.04)
LogNor(7.1,0.01)

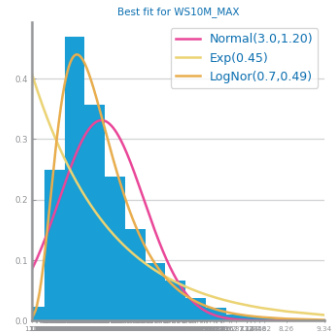

Best fit for T2M_MAX
Normal(23.7,8.84)
Exp(0.03)
LogNor(5.1,0.06)

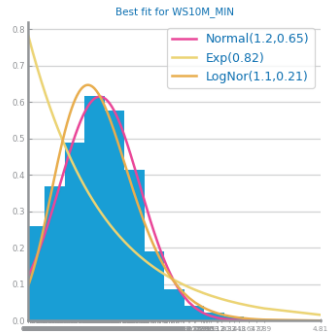

Best fit for T2M_MIN
Normal(11.9,8.68)
Exp(0.04)
LogNor(6.6,0.01)

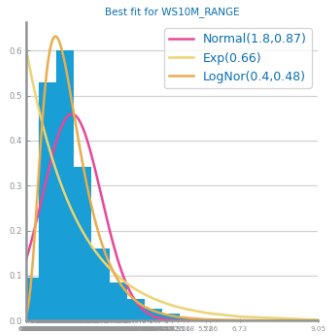

Best fit for T2M_RANGE
Normal(11.8,3.42)
Exp(0.10)
LogNor(5.7,0.01)

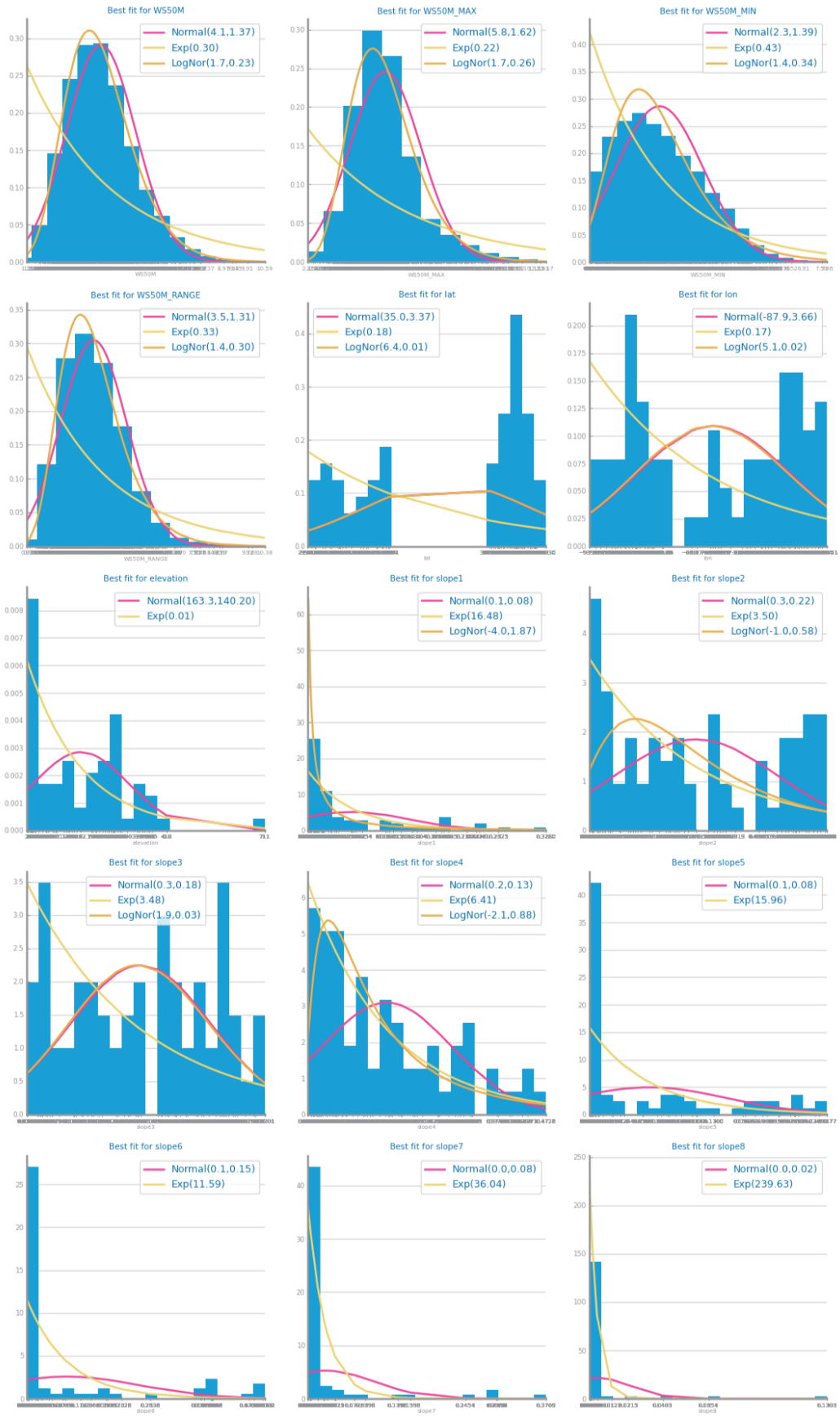

Best fit for TS
Normal(17.4,8.66)
Exp(0.04)
LogNor(7.2,0.01)


Best fit for WS10M
Normal(2.1,0.81)
Exp(0.61)
LogNor(0.6,0.40)


Best fit for WS10M_MAX
Normal(3.0,1.20)
Exp(0.45)
LogNor(0.7,0.49)


Best fit for WS10M_MIN
Normal(1.2,0.65)
Exp(0.82)
LogNor(1.1,0.21)


Best fit for WS10M_RANGE
Normal(1.8,0.87)
Exp(0.66)
LogNor(0.4,0.48)


Best fit for WS50M
Normal(4.1,1.37)
Exp(0.30)
LogNor(1.7,0.23)
Best fit for WS50M_MAX
Normal(5.8,1.62)
Exp(0.22)
LogNor(1.7,0.26)
Best fit for WS50M_MIN
Normal(2.3,1.39)
Exp(0.43)
LogNor(1.4,0.34)
Best fit for WS50M_RANGE
Normal(3.5,1.31)
Exp(0.33)
LogNor(1.4,0.30)
Best fit for lat
Normal(35.0,3.37)
Exp(0.18)
LogNor(6.4,0.01)
Best fit for lon
Normal(-87.9,3.66)
Exp(0.17)
LogNor(5.1,0.02)
Best fit for elevation
Normal(163.3,140.20)
Exp(0.01)
Best fit for slope1
Normal(0.1,0.08)
Exp(16.48)
LogNor(-4.0,1.87)
Best fit for slope2
Normal(0.3,0.22)
Exp(3.50)
LogNor(-1.0,0.58)
Best fit for slope3
Normal(0.3,0.18)
Exp(3.48)
LogNor(1.9,0.03)
Best fit for slope4
Normal(0.2,0.13)
Exp(6.41)
LogNor(-2.1,0.88)
Best fit for slope5
Normal(0.1,0.08)
Exp(15.96)
Best fit for slope6
Normal(0.1,0.15)
Exp(11.59)
Best fit for slope7
Normal(0.0,0.08)
Exp(36.04)
Best fit for slope8
Normal(0.0,0.02)
Exp(239.63)

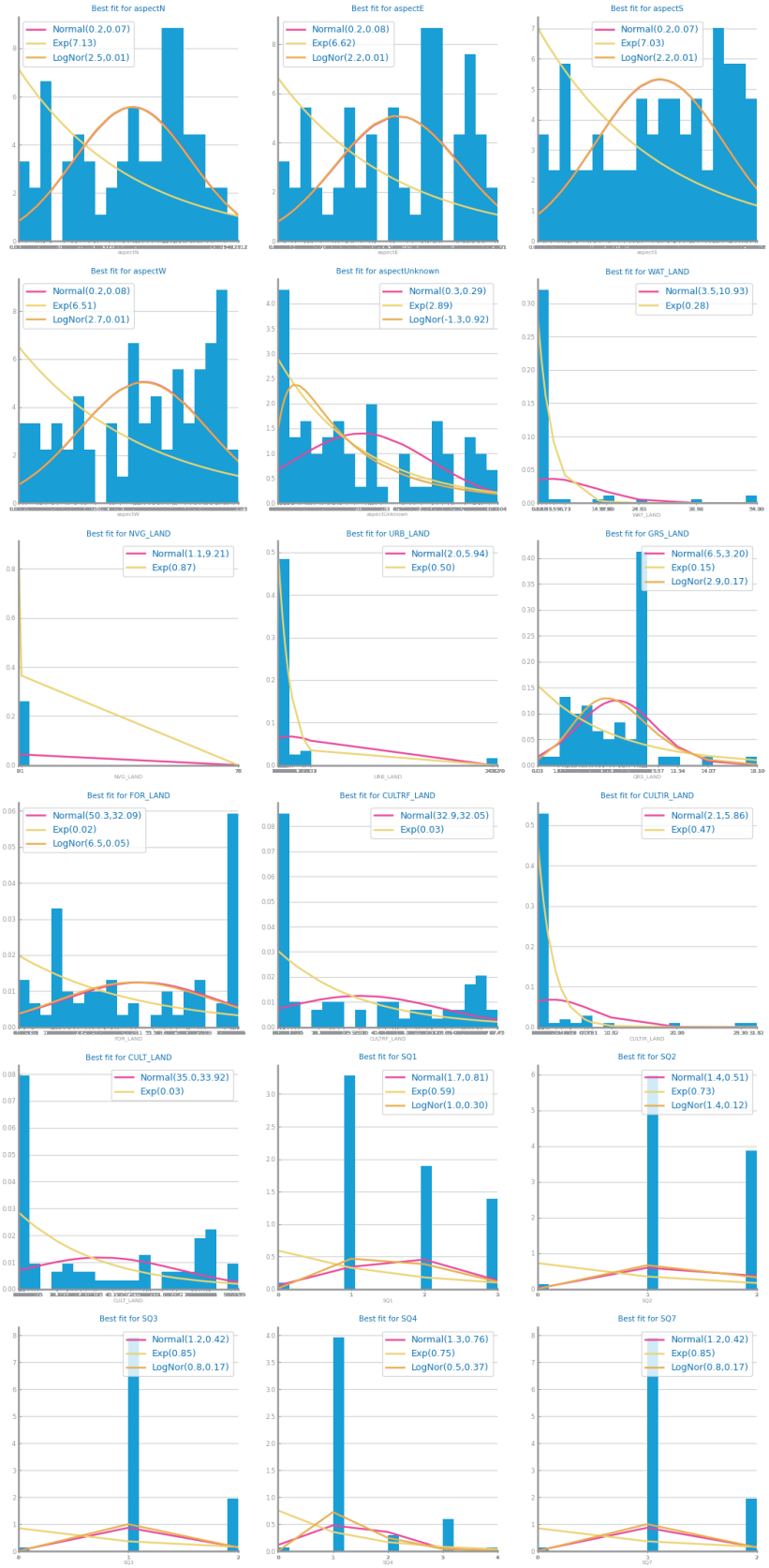


Figure 14: Histograms of numeric variables distributions for dataset 2

Data Granularity

With more bins (small intervals) we can find fine-grained patterns/outliers. In 1, we grouped symbolic vars using the granularity the paper suggested. In 2, we analysed date by days, months and years.

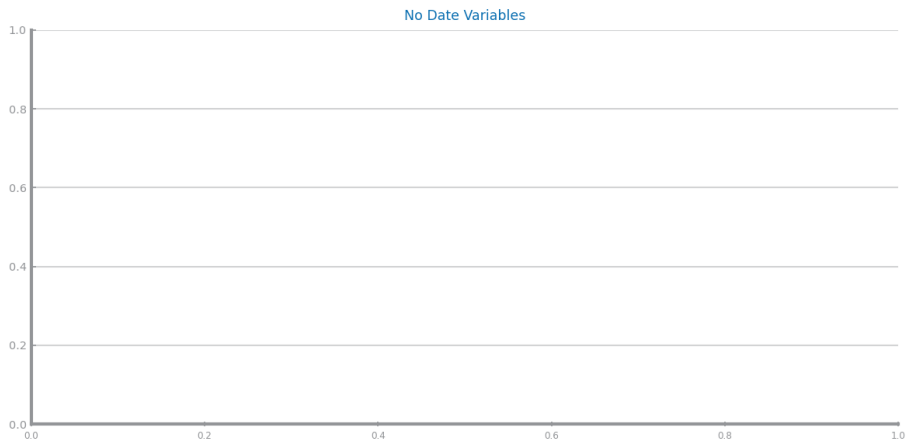
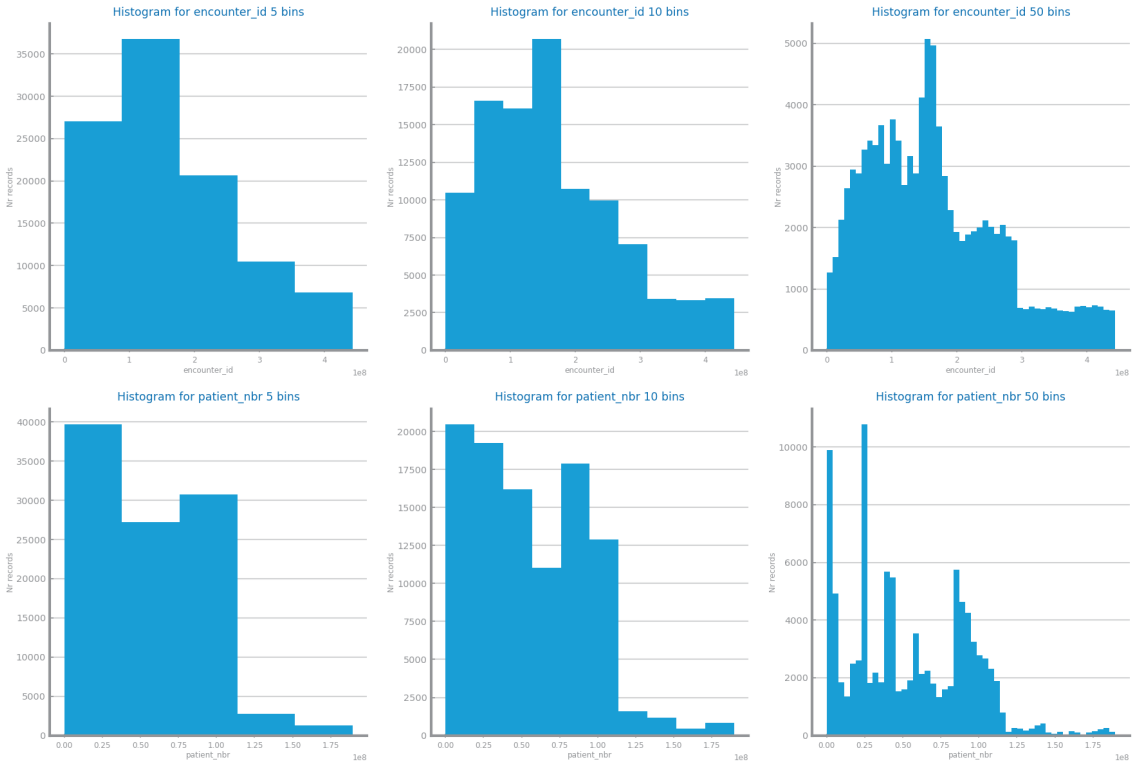
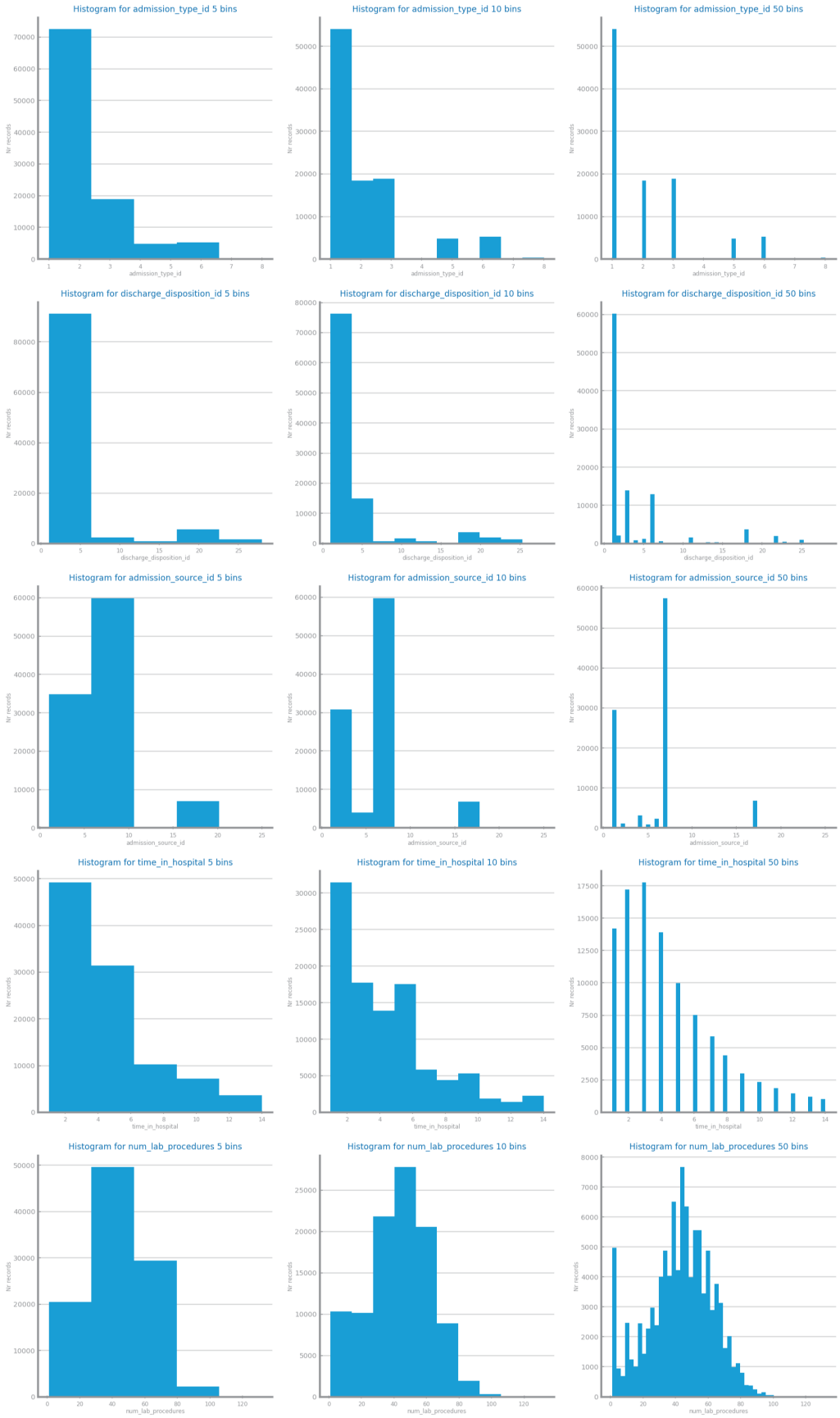


Figure 15: Date granularity analysis for dataset 1




Histogram for admission_type_id 5 bins
Histogram for admission_type_id 10 bins
Histogram for admission_type_id 50 bins
Histogram for discharge_disposition_id 5 bins
Histogram for discharge_disposition_id 10 bins
Histogram for discharge_disposition_id 50 bins
Histogram for admission_source_id 5 bins
Histogram for admission_source_id 10 bins
Histogram for admission_source_id 50 bins
Histogram for time_in_hospital 5 bins
Histogram for time_in_hospital 10 bins
Histogram for time_in_hospital 50 bins
Histogram for num_lab_procedures 5 bins
Histogram for num_lab_procedures 10 bins
Histogram for num_lab_procedures 50 bins
Nr records
admission_type_id
discharge_disposition_id
admission_source_id
time_in_hospital
num_lab_procedures

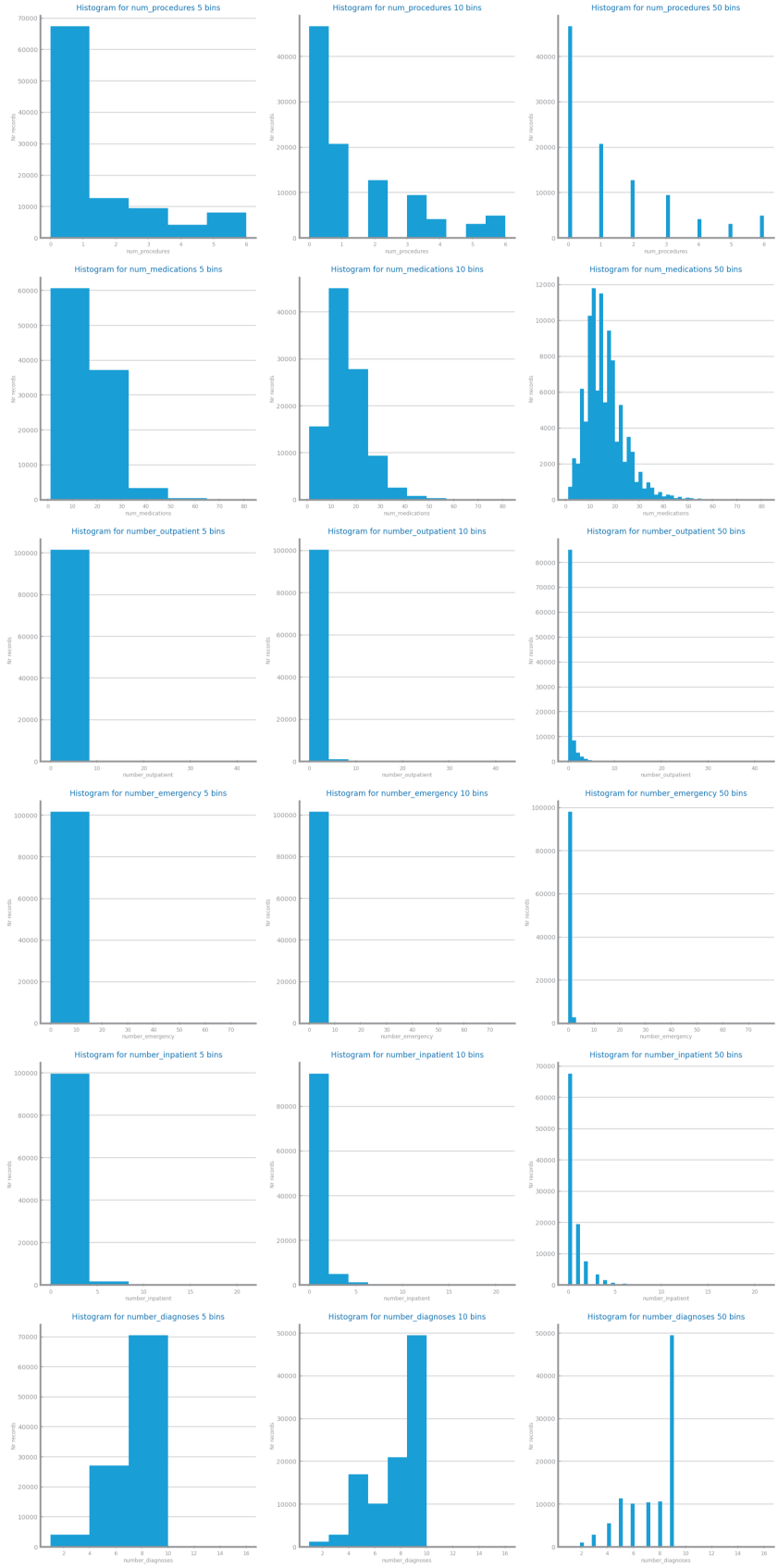
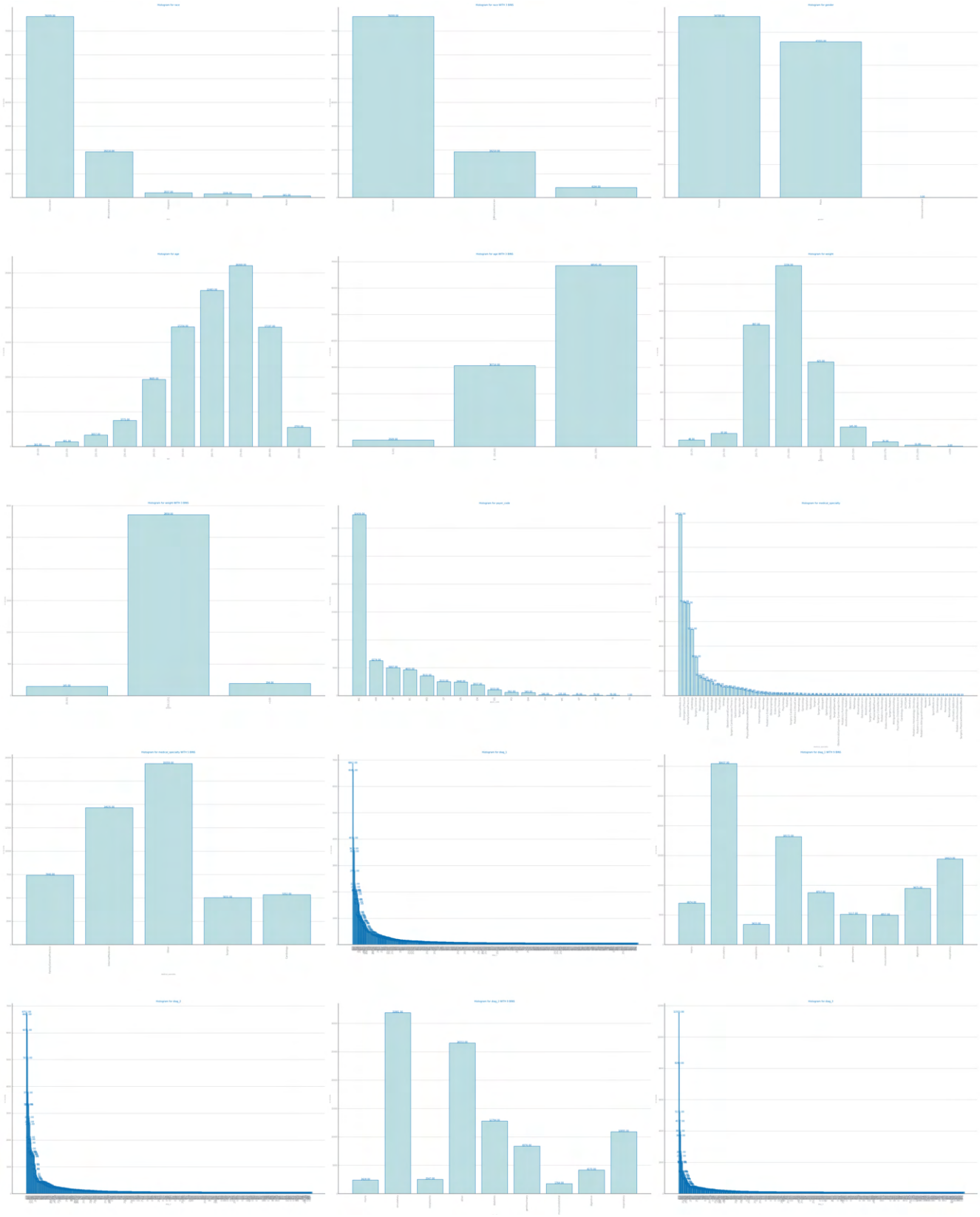


Figure 16: Numeric granularity analysis for dataset 1


Histogram for race
Histogram for race BIN 2 BINS
Histogram for gender
Histogram for age
Histogram for age BIN 2 BINS
Histogram for weight
Histogram for weight BIN 3 BINS
Histogram for payer_code
Histogram for medical_specialty
Histogram for medical_specialty BIN 5 BINS
Histogram for diag_1
Histogram for diag_1 BIN 9 BINS
Histogram for diag_2
Histogram for diag_2 BIN 9 BINS
Histogram for diag_3

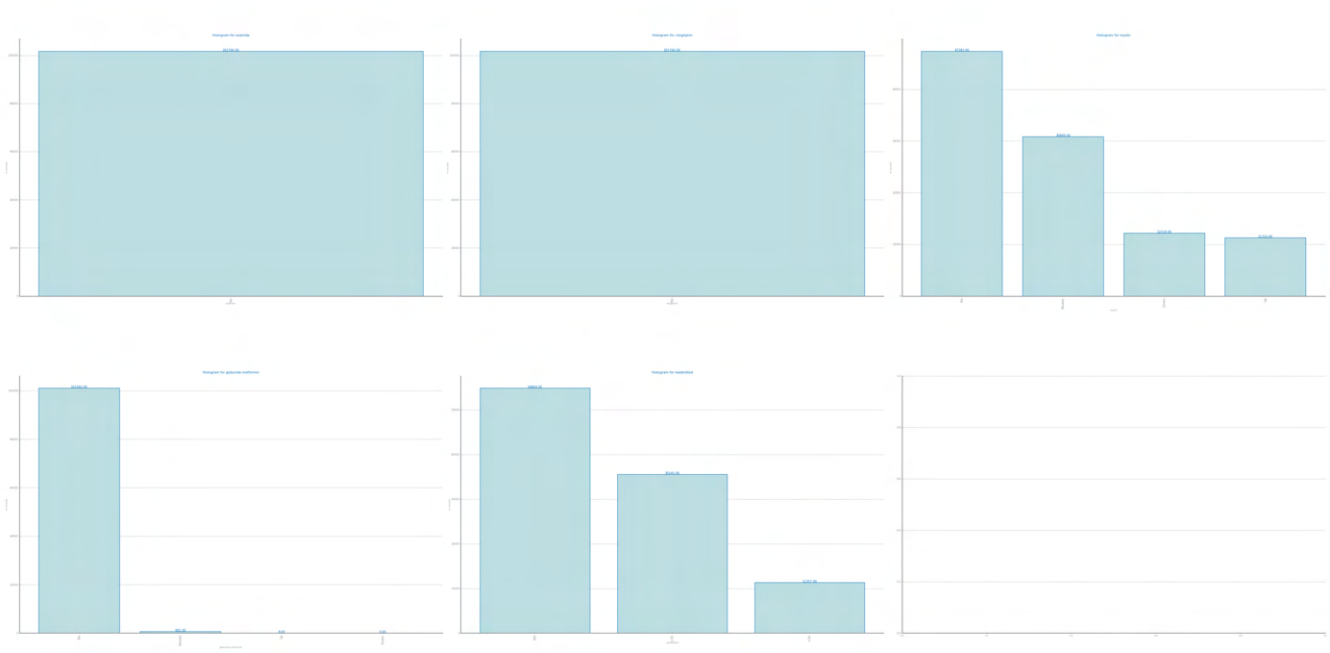


Figure 17: Symbolic granularity analysis for dataset 1

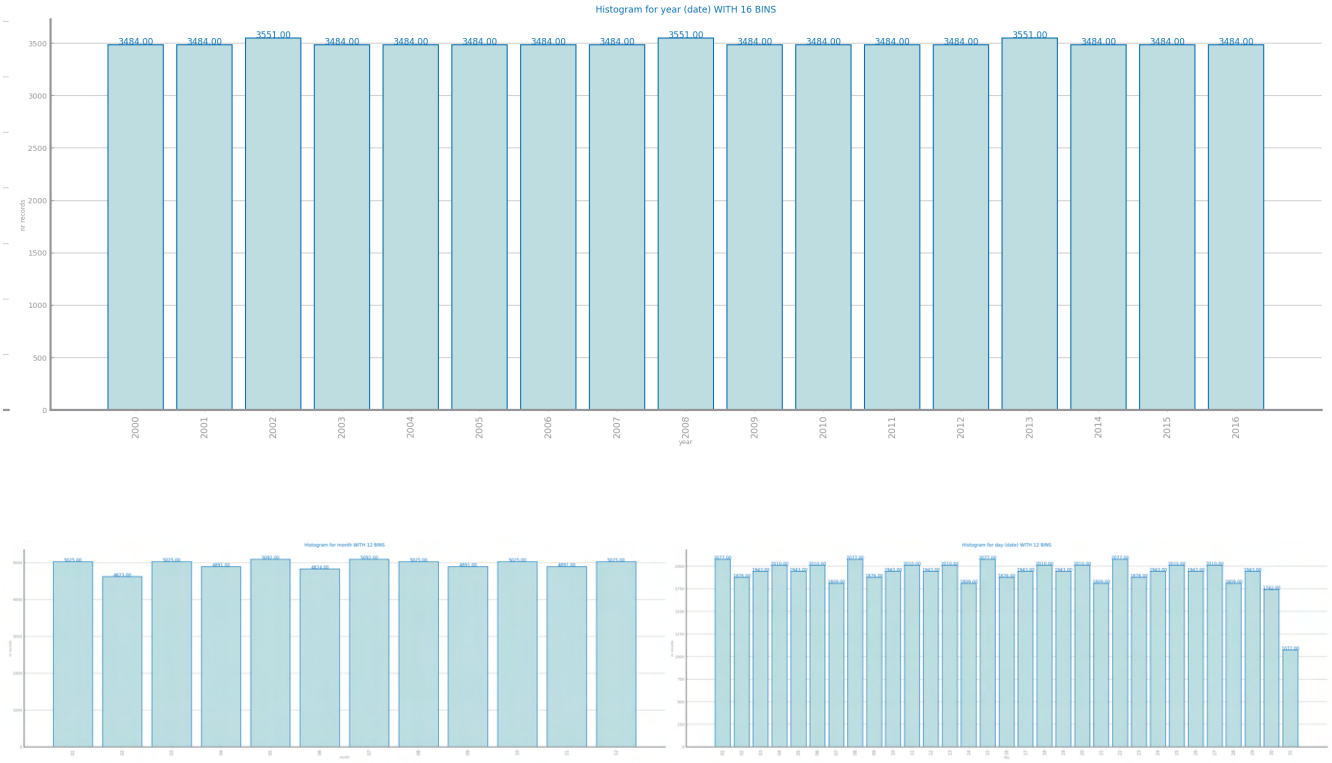
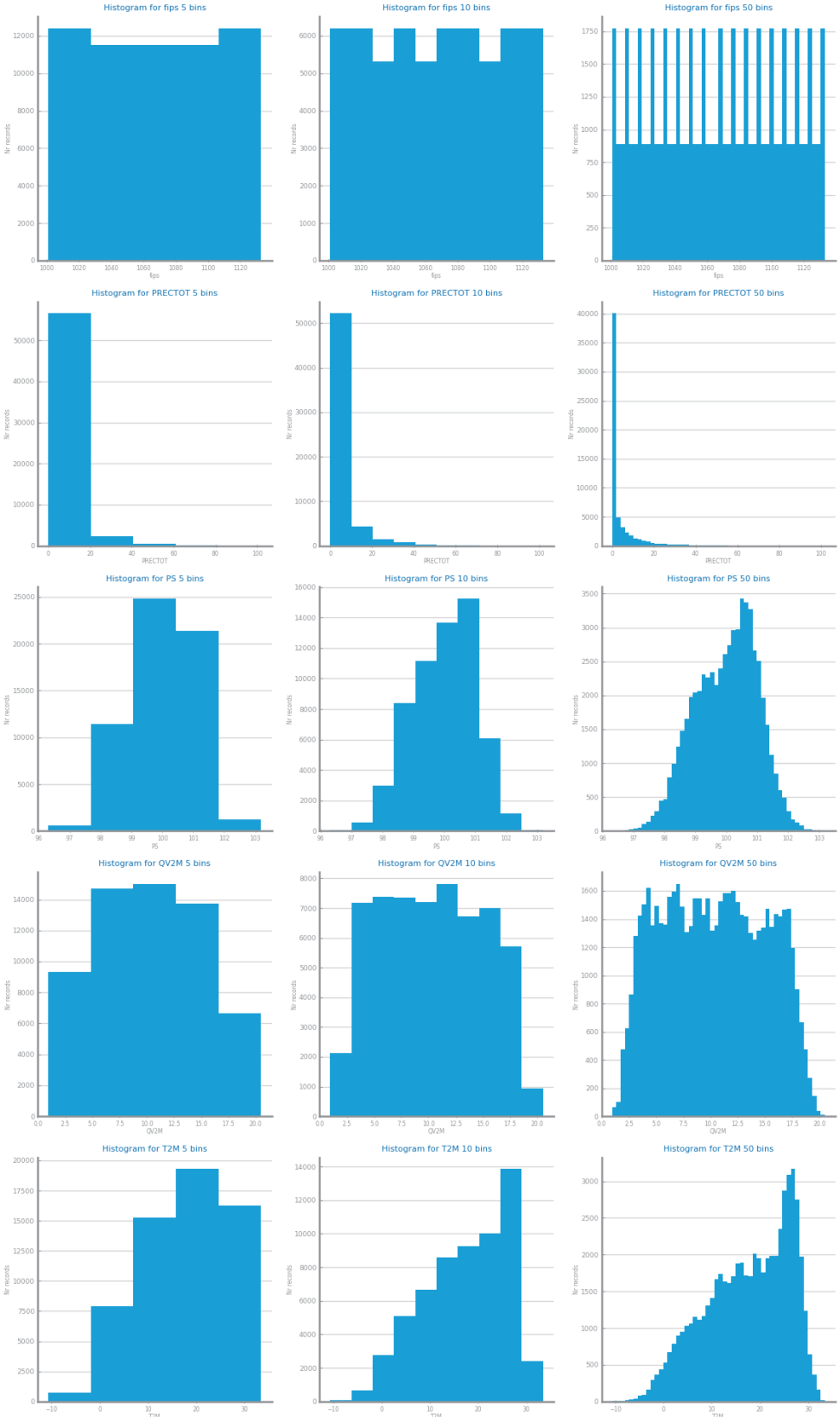
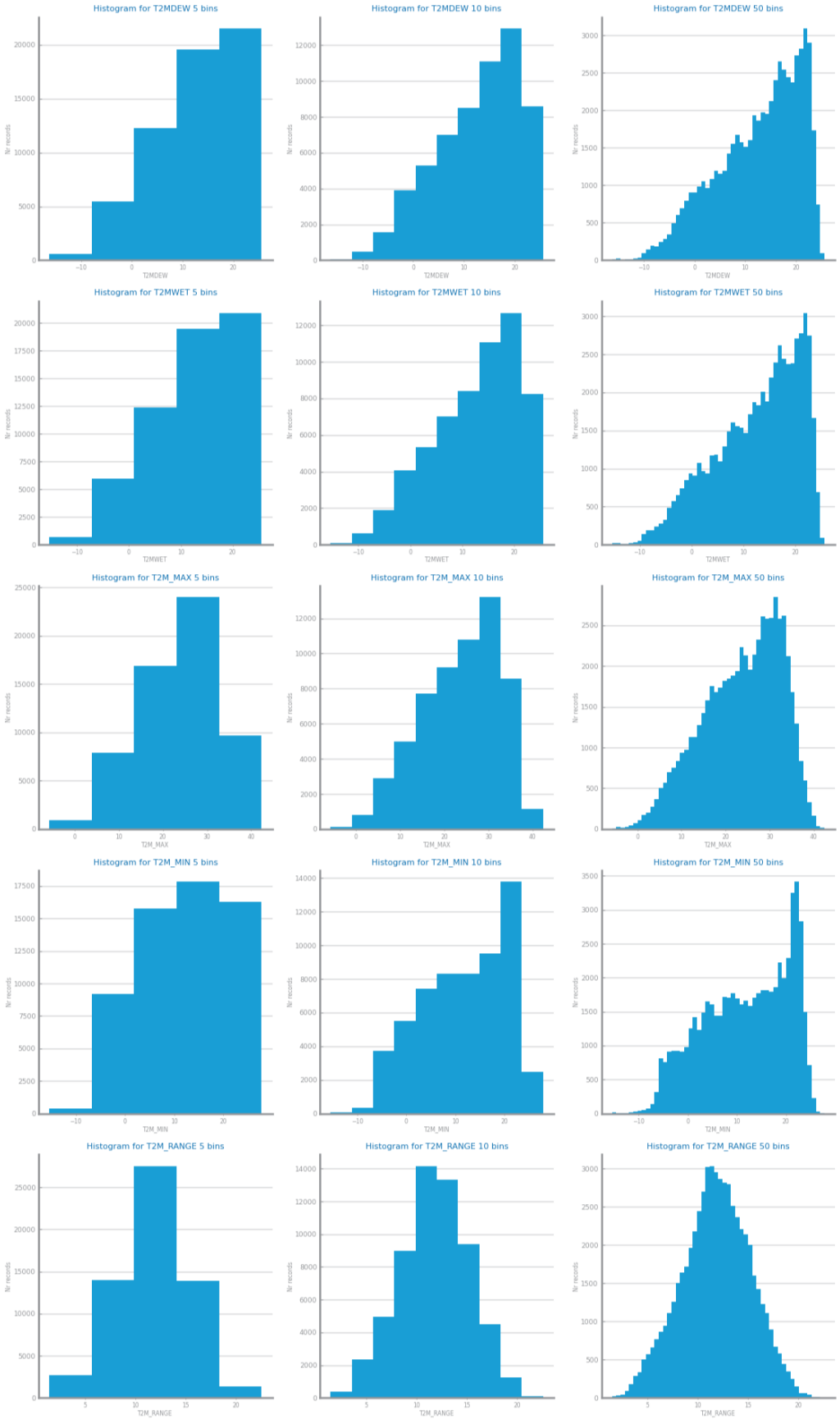
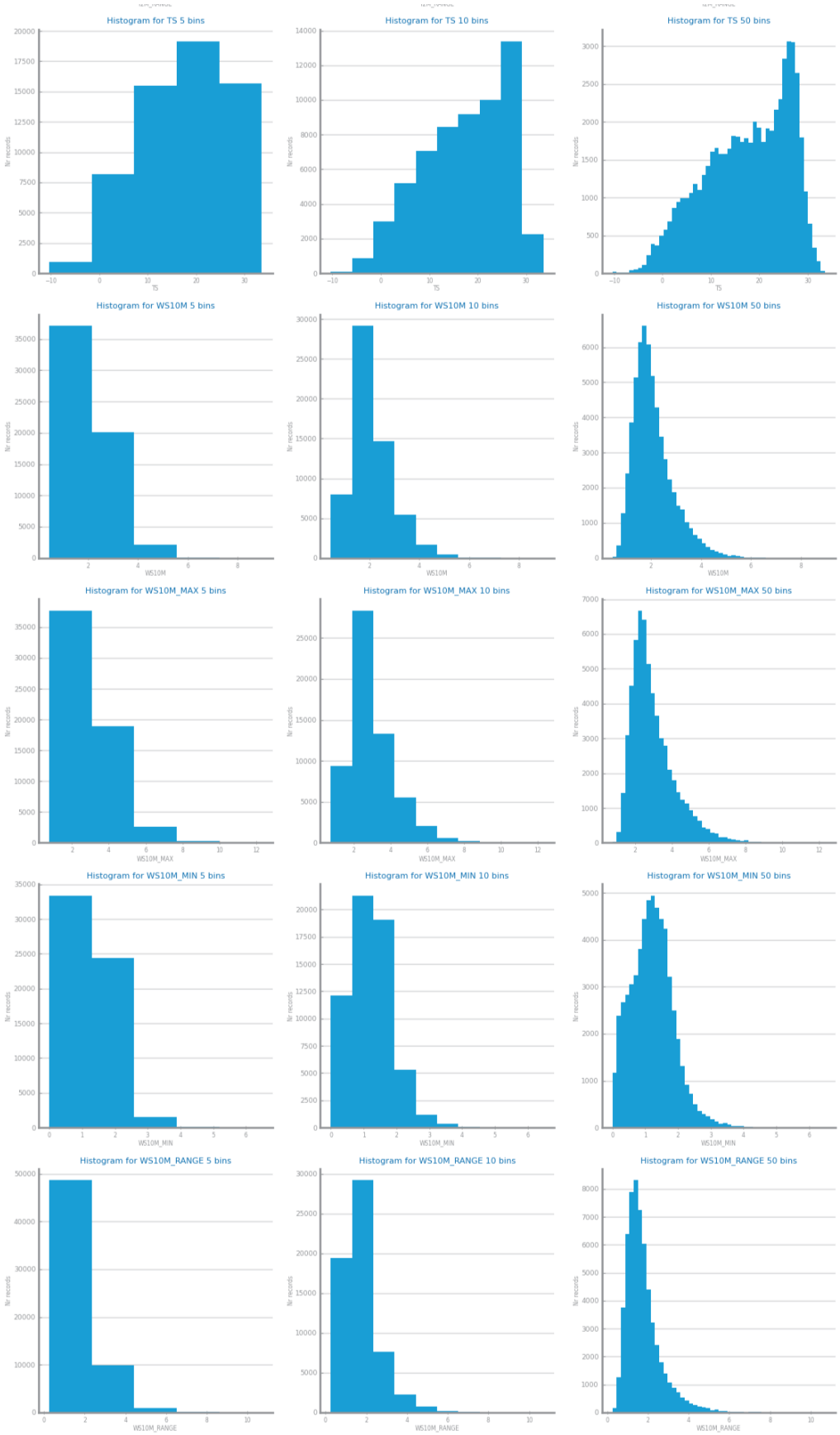
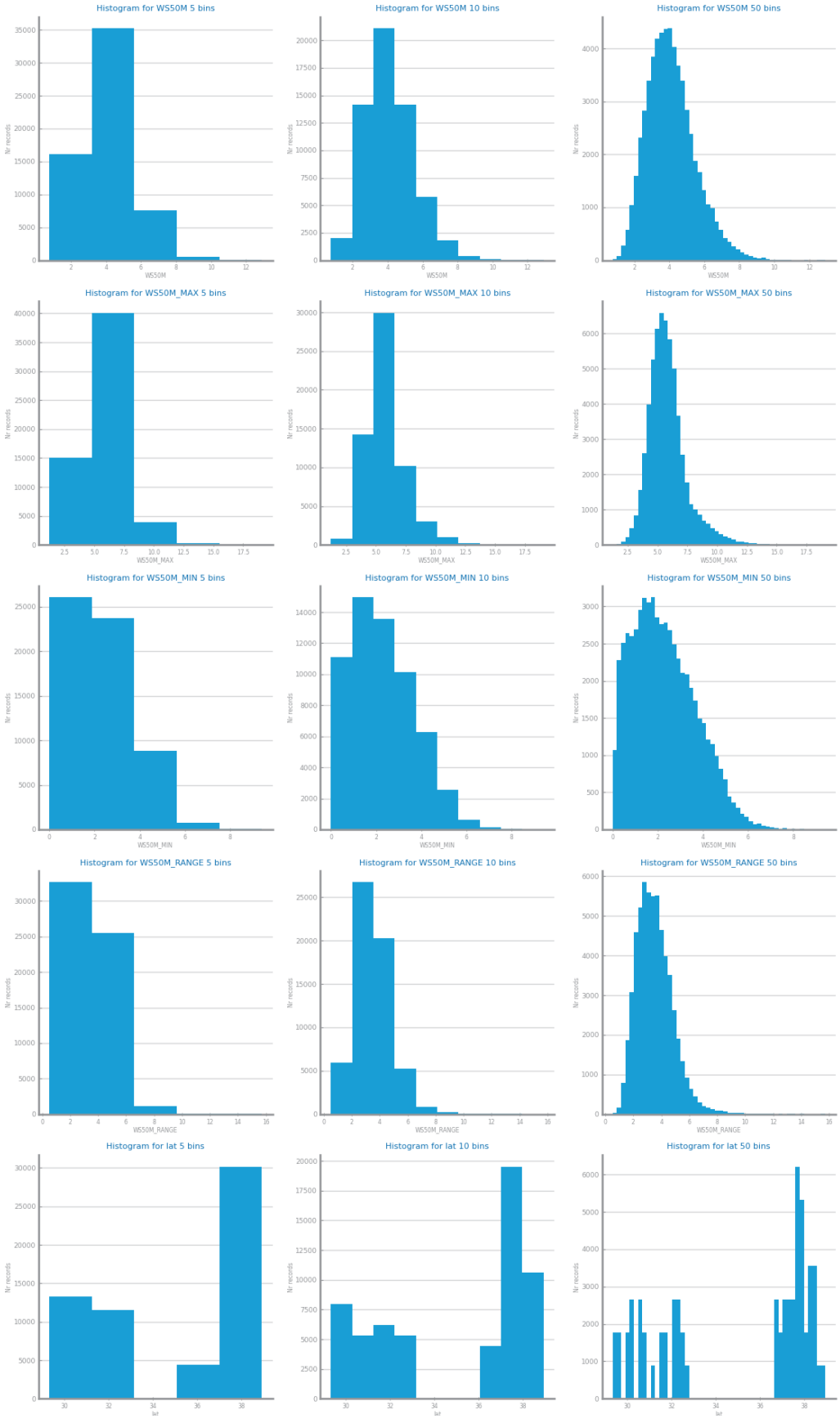


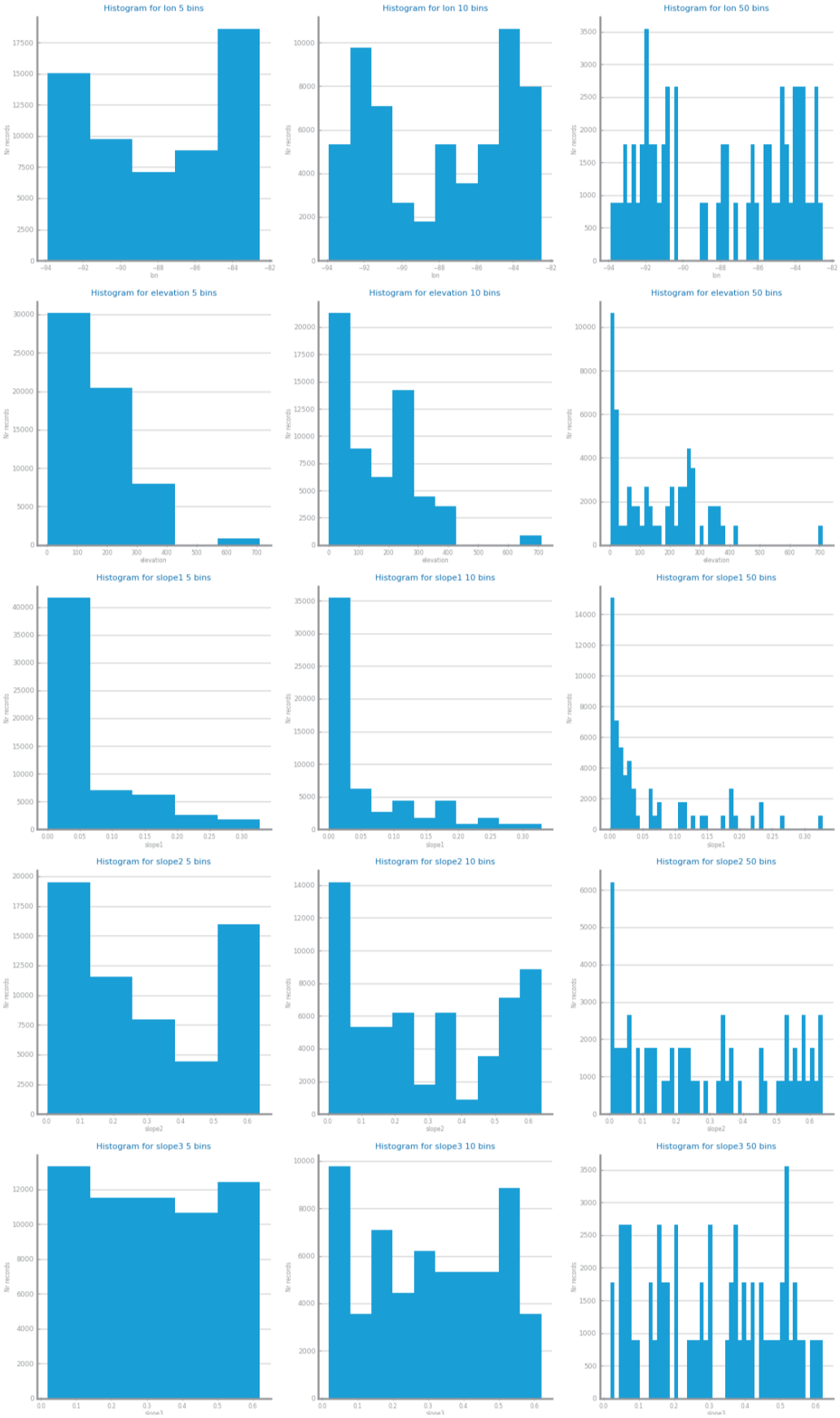
Figure 18: Date granularity analysis for dataset 2

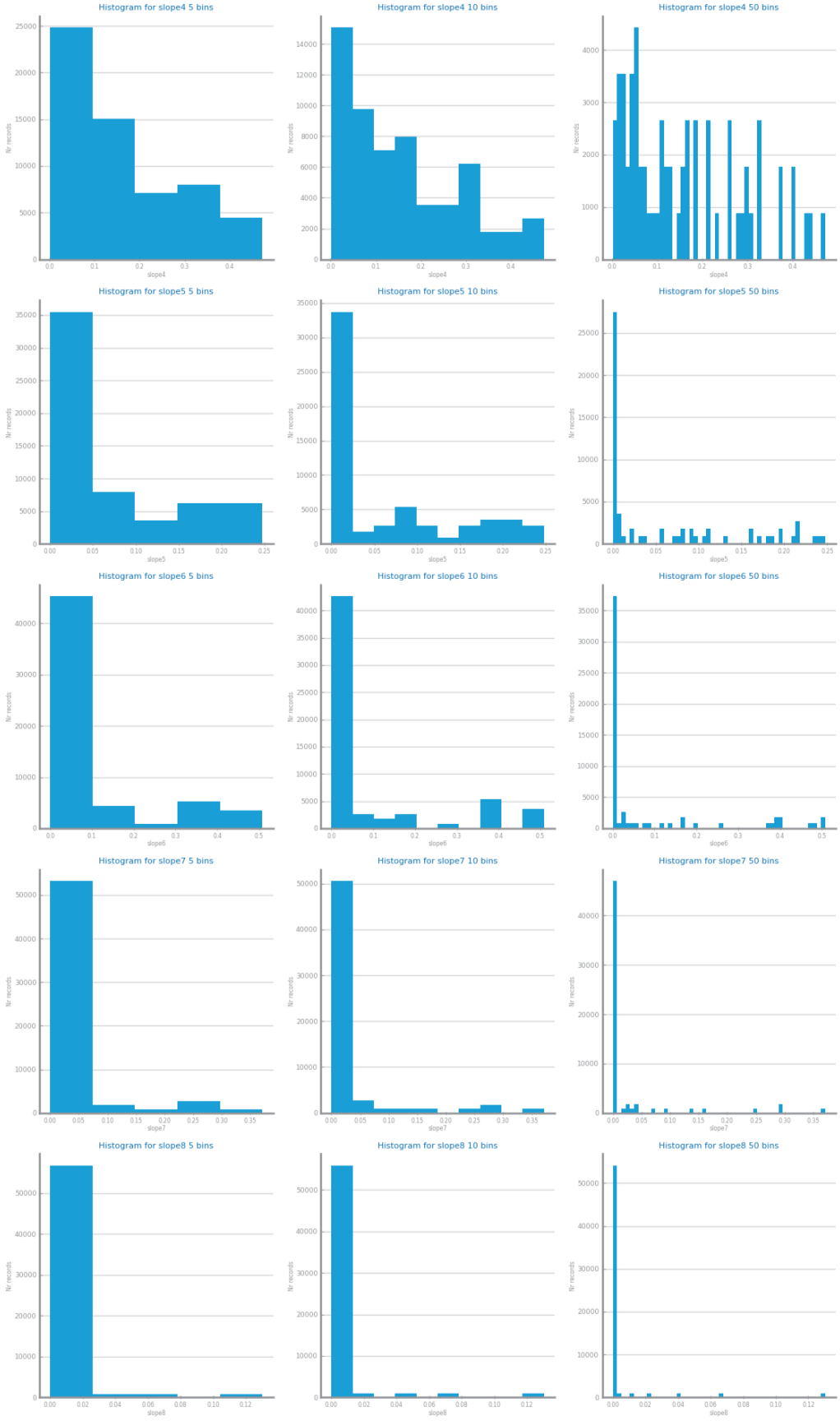

Histogram for fips 5 bins
Histogram for fips 10 bins
Histogram for fips 50 bins
Histogram for PRECTOT 5 bins
Histogram for PRECTOT 10 bins
Histogram for PRECTOT 50 bins
Histogram for PS 5 bins
Histogram for PS 10 bins
Histogram for PS 50 bins
Histogram for QV2M 5 bins
Histogram for QV2M 10 bins
Histogram for QV2M 50 bins
Histogram for T2M 5 bins
Histogram for T2M 10 bins
Histogram for T2M 50 bins
Nr records
fips
PRECTOT
PS
QV2M
T2M

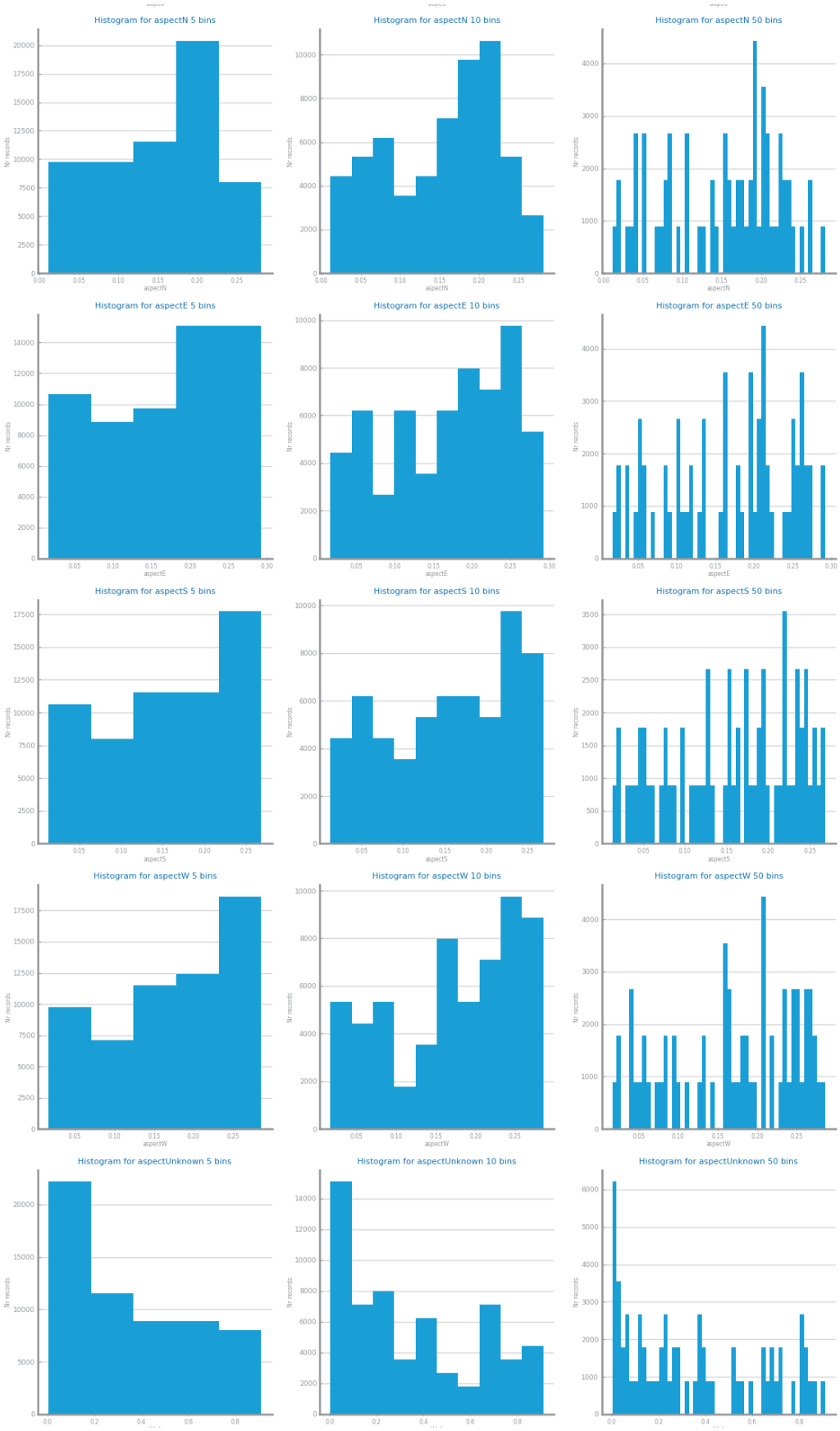

Histogram for T2MDEW 5 bins
Histogram for T2MDEW 10 bins
Histogram for T2MDEW 50 bins
Histogram for T2MWET 5 bins
Histogram for T2MWET 10 bins
Histogram for T2MWET 50 bins
Histogram for T2M_MAX 5 bins
Histogram for T2M_MAX 10 bins
Histogram for T2M_MAX 50 bins
Histogram for T2M_MIN 5 bins
Histogram for T2M_MIN 10 bins
Histogram for T2M_MIN 50 bins
Histogram for T2M_RANGE 5 bins
Histogram for T2M_RANGE 10 bins
Histogram for T2M_RANGE 50 bins
Nr records
T2MDEW
T2MWET
T2M_MAX
T2M_MIN
T2M_RANGE

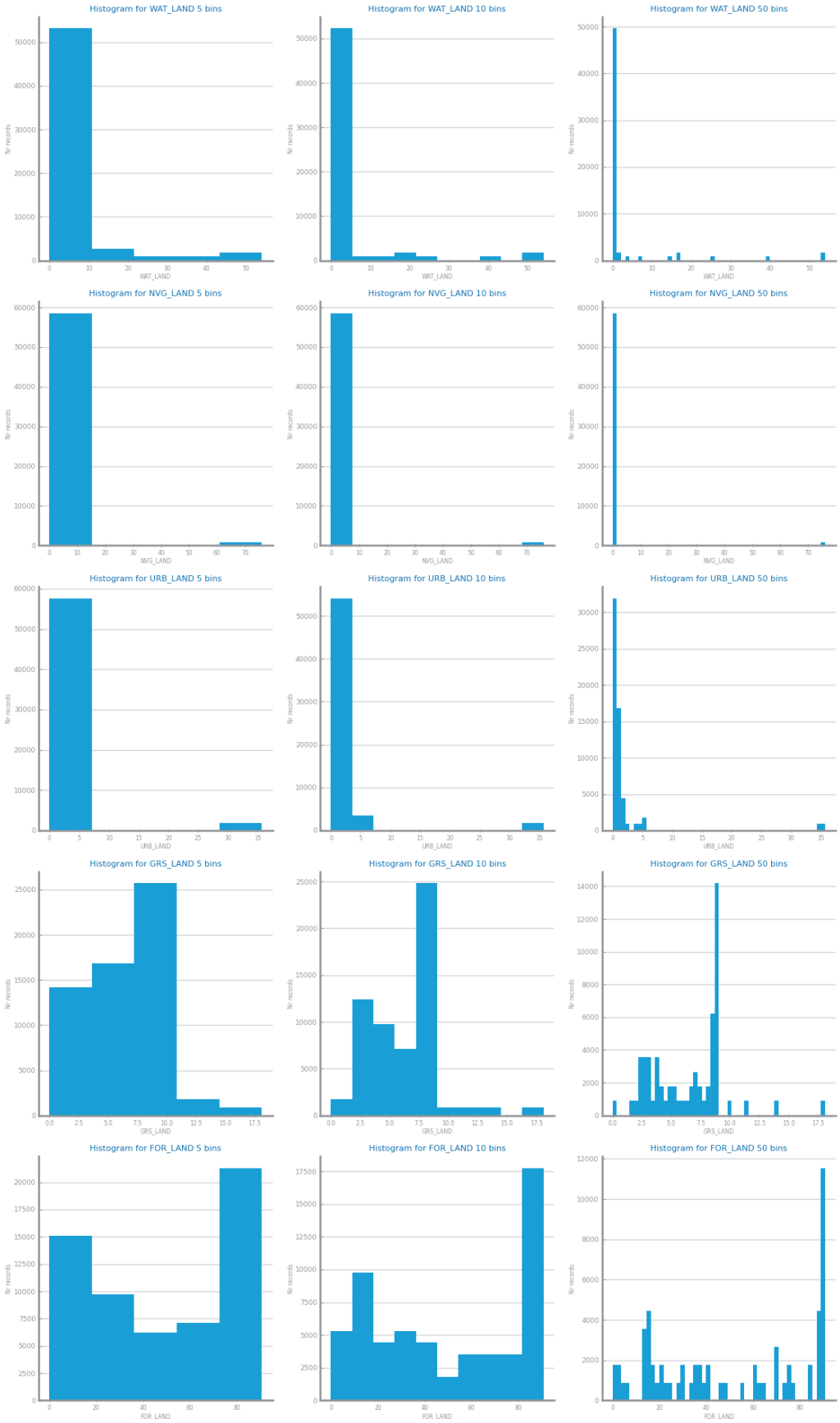

Histogram for TS 5 bins
Histogram for TS 10 bins
Histogram for TS 50 bins
Histogram for WS10M 5 bins
Histogram for WS10M 10 bins
Histogram for WS10M 50 bins
Histogram for WS10M_MAX 5 bins
Histogram for WS10M_MAX 10 bins
Histogram for WS10M_MAX 50 bins
Histogram for WS10M_MIN 5 bins
Histogram for WS10M_MIN 10 bins
Histogram for WS10M_MIN 50 bins
Histogram for WS10M_RANGE 5 bins
Histogram for WS10M_RANGE 10 bins
Histogram for WS10M_RANGE 50 bins
Nr records
TS
WS10M
WS10M_MAX
WS10M_MIN
WS10M_RANGE

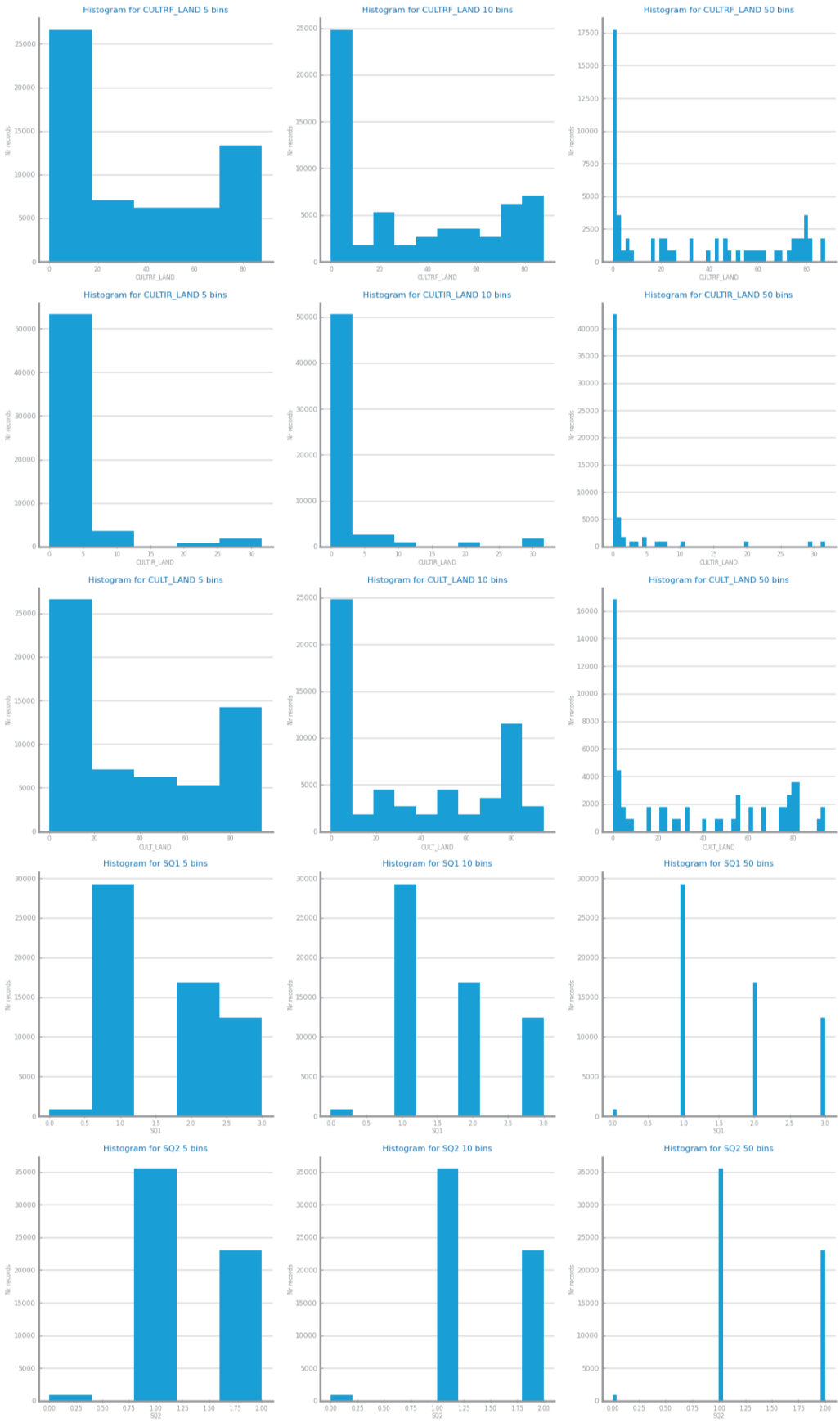

Histogram for WS50M 5 bins
Histogram for WS50M 10 bins
Histogram for WS50M 50 bins
Histogram for WS50M_MAX 5 bins
Histogram for WS50M_MAX 10 bins
Histogram for WS50M_MAX 50 bins
Histogram for WS50M_MIN 5 bins
Histogram for WS50M_MIN 10 bins
Histogram for WS50M_MIN 50 bins
Histogram for WS50M_RANGE 5 bins
Histogram for WS50M_RANGE 10 bins
Histogram for WS50M_RANGE 50 bins
Histogram for lat 5 bins
Histogram for lat 10 bins
Histogram for lat 50 bins
Nr records
WS50M
WS50M_MAX
WS50M_MIN
WS50M_RANGE
lat


Histogram for lon 5 bins
Histogram for lon 10 bins
Histogram for lon 50 bins
Histogram for elevation 5 bins
Histogram for elevation 10 bins
Histogram for elevation 50 bins
Histogram for slope1 5 bins
Histogram for slope1 10 bins
Histogram for slope1 50 bins
Histogram for slope2 5 bins
Histogram for slope2 10 bins
Histogram for slope2 50 bins
Histogram for slope3 5 bins
Histogram for slope3 10 bins
Histogram for slope3 50 bins
Nr records
lon
elevation
slope1
slope2
slope3


Histogram for slope4 5 bins
Histogram for slope4 10 bins
Histogram for slope4 50 bins
Histogram for slope5 5 bins
Histogram for slope5 10 bins
Histogram for slope5 50 bins
Histogram for slope6 5 bins
Histogram for slope6 10 bins
Histogram for slope6 50 bins
Histogram for slope7 5 bins
Histogram for slope7 10 bins
Histogram for slope7 50 bins
Histogram for slope8 5 bins
Histogram for slope8 10 bins
Histogram for slope8 50 bins
Nr records
slope4
slope5
slope6
slope7
slope8


Histogram for aspectN 5 bins
Histogram for aspectN 10 bins
Histogram for aspectN 50 bins
Histogram for aspectE 5 bins
Histogram for aspectE 10 bins
Histogram for aspectE 50 bins
Histogram for aspectS 5 bins
Histogram for aspectS 10 bins
Histogram for aspectS 50 bins
Histogram for aspectW 5 bins
Histogram for aspectW 10 bins
Histogram for aspectW 50 bins
Histogram for aspectUnknown 5 bins
Histogram for aspectUnknown 10 bins
Histogram for aspectUnknown 50 bins
Nr records
aspectN
aspectE
aspectS
aspectW


Histogram for WAT_LAND 5 bins
Histogram for WAT_LAND 10 bins
Histogram for WAT_LAND 50 bins
Histogram for NVG_LAND 5 bins
Histogram for NVG_LAND 10 bins
Histogram for NVG_LAND 50 bins
Histogram for URB_LAND 5 bins
Histogram for URB_LAND 10 bins
Histogram for URB_LAND 50 bins
Histogram for GRS_LAND 5 bins
Histogram for GRS_LAND 10 bins
Histogram for GRS_LAND 50 bins
Histogram for FOR_LAND 5 bins
Histogram for FOR_LAND 10 bins
Histogram for FOR_LAND 50 bins
Nr records
WAT_LAND
NVG_LAND
URB_LAND
GRS_LAND
FOR_LAND


Histogram for CULTRF_LAND 5 bins
Histogram for CULTRF_LAND 10 bins
Histogram for CULTRF_LAND 50 bins
Histogram for CULTIR_LAND 5 bins
Histogram for CULTIR_LAND 10 bins
Histogram for CULTIR_LAND 50 bins
Histogram for CULT_LAND 5 bins
Histogram for CULT_LAND 10 bins
Histogram for CULT_LAND 50 bins
Histogram for SQ1 5 bins
Histogram for SQ1 10 bins
Histogram for SQ1 50 bins
Histogram for SQ2 5 bins
Histogram for SQ2 10 bins
Histogram for SQ2 50 bins
Nr records
CULTRF_LAND
CULTIR_LAND
CULT_LAND
SQ1
SQ2

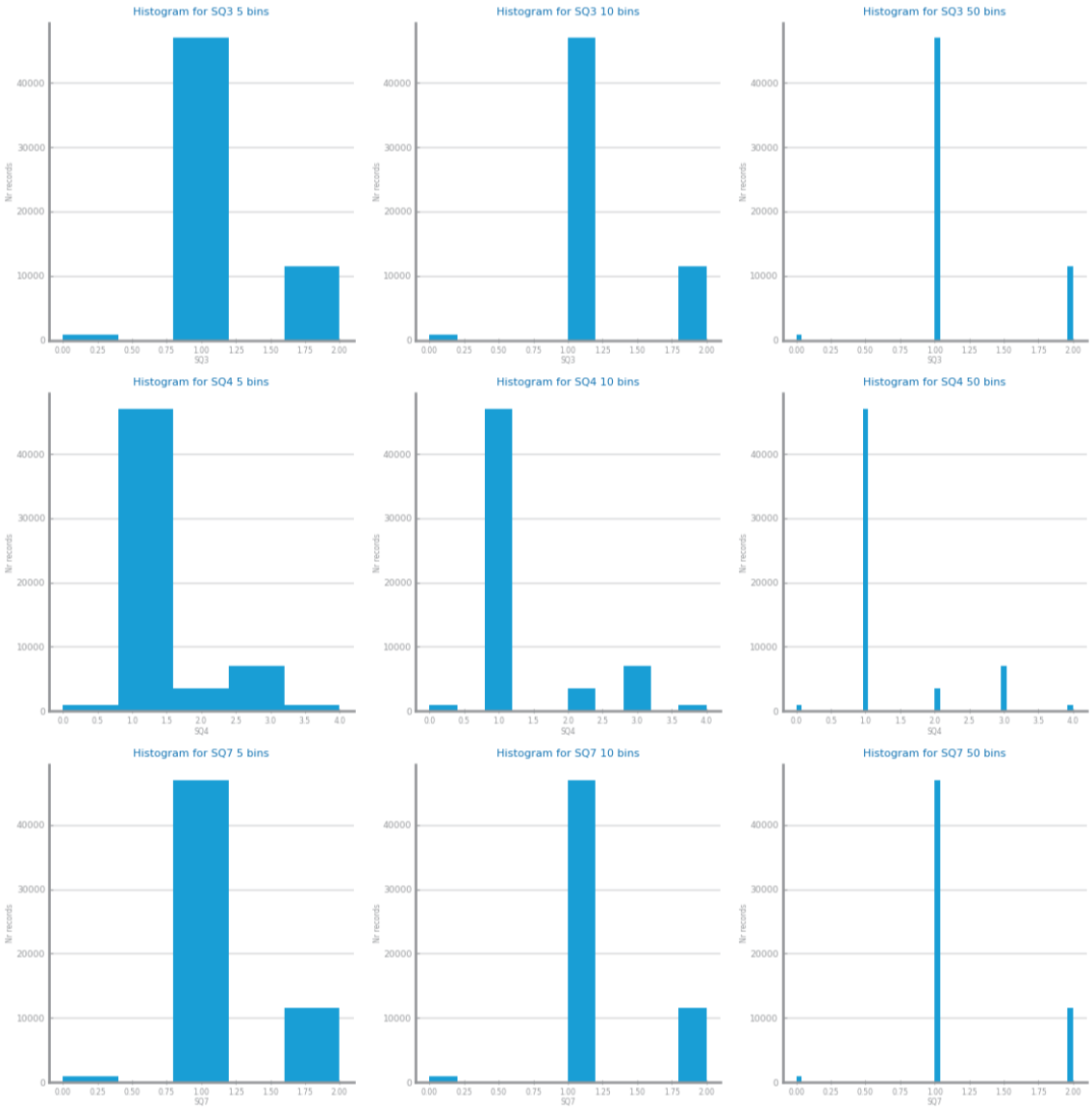
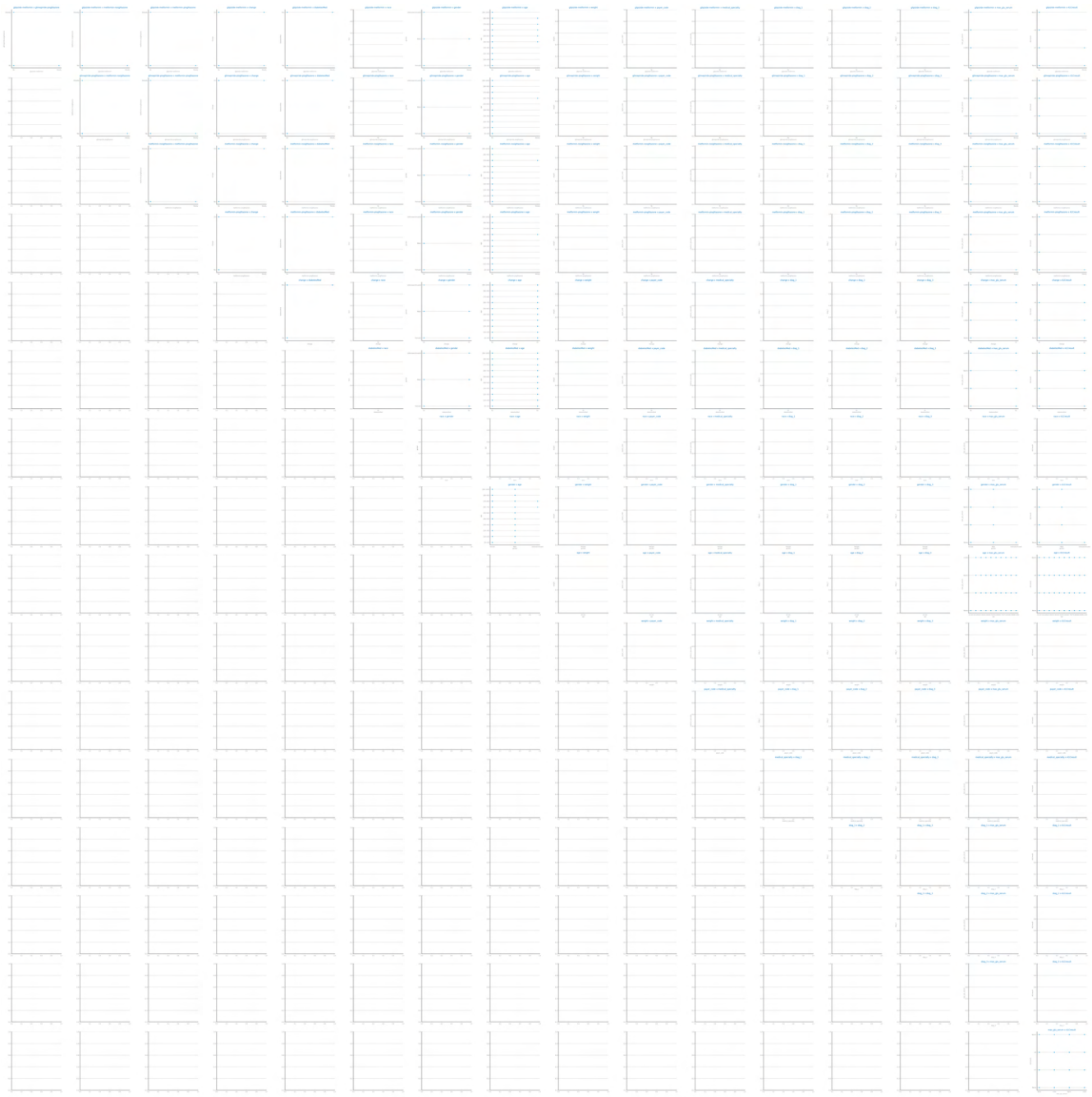


Figure 19: Numeric granularity analysis for dataset 2

Figure 20: Symbolic granularity analysis for dataset 2

Data Sparsity

In both the domain seems fairly covered (considering possible values): no need for a broader sample. In 1 variables aren't very correlated so they should complement each other in prediction. In 2 some are highly correlated which means their coexistence will make the KDD process longer and not richer.



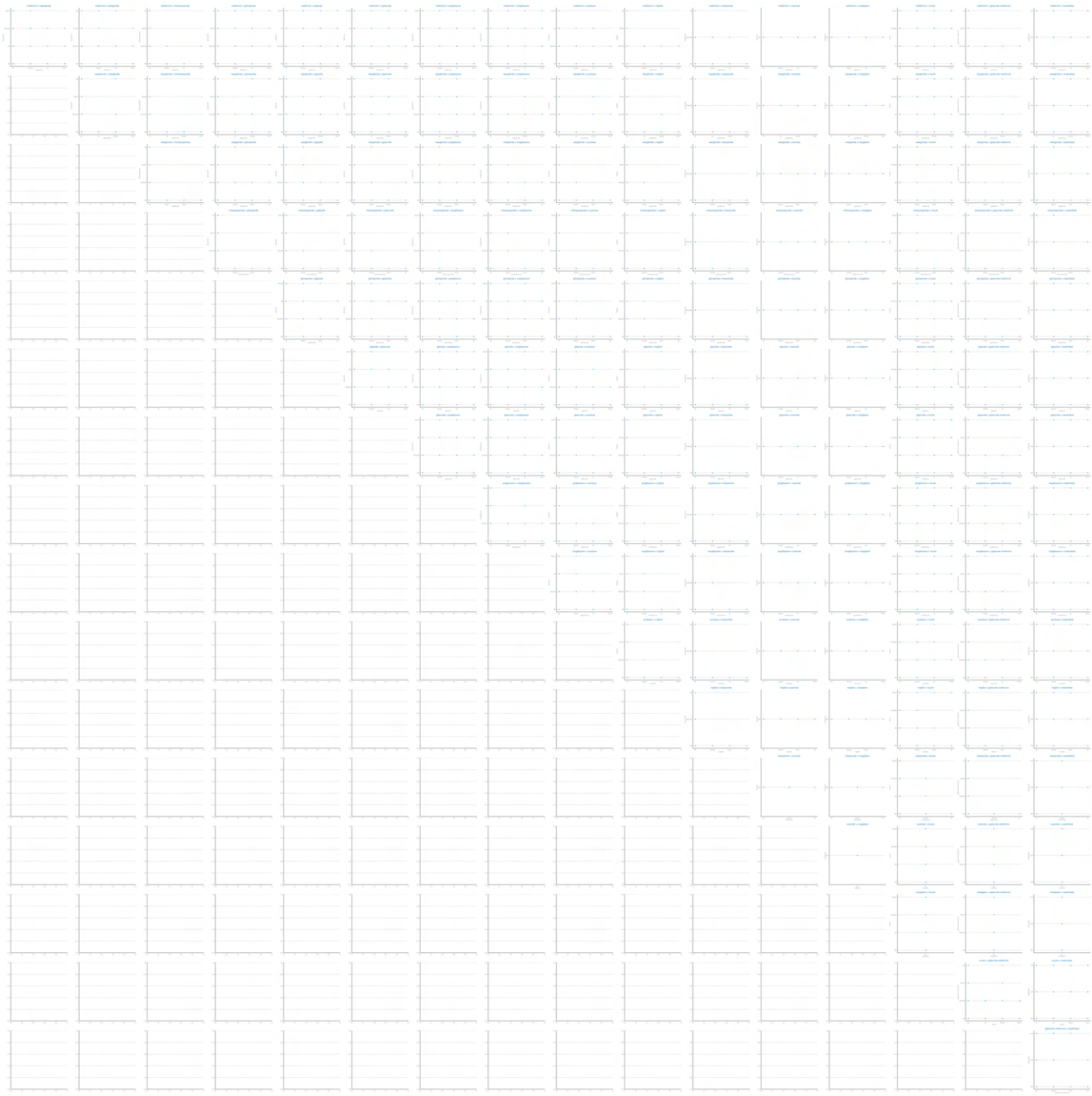
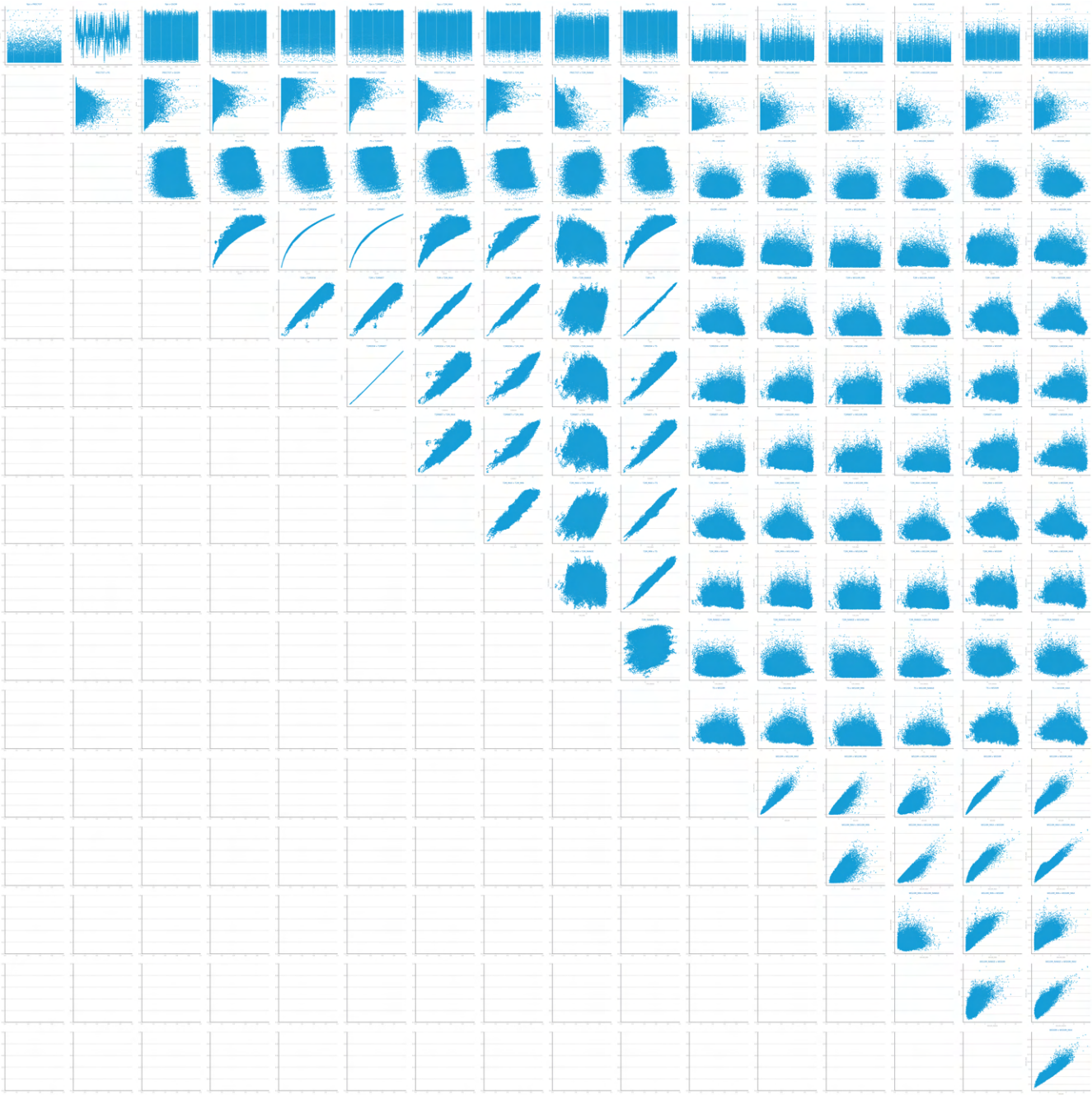
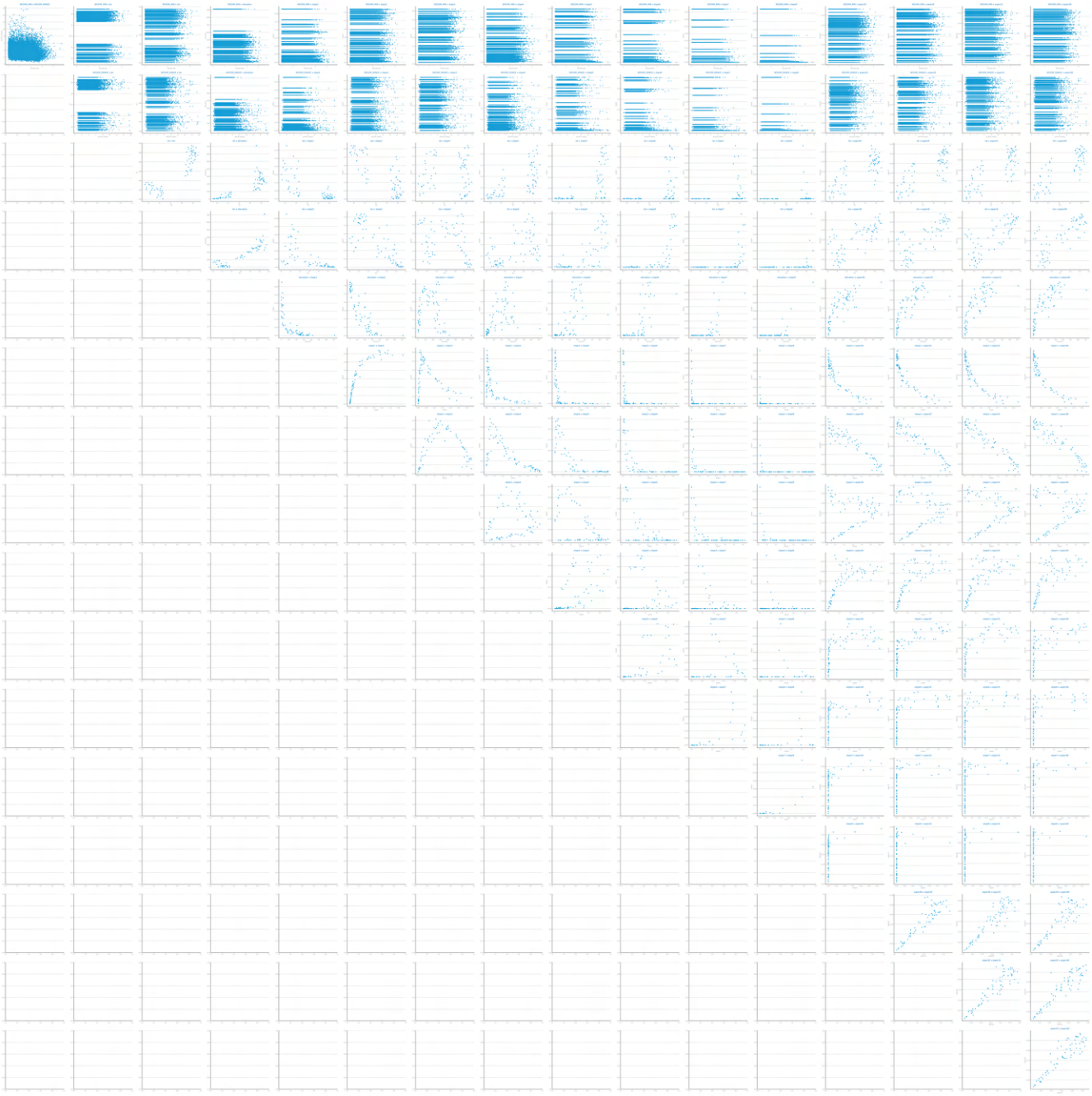


Figure 21: Sparsity analysis for dataset 1





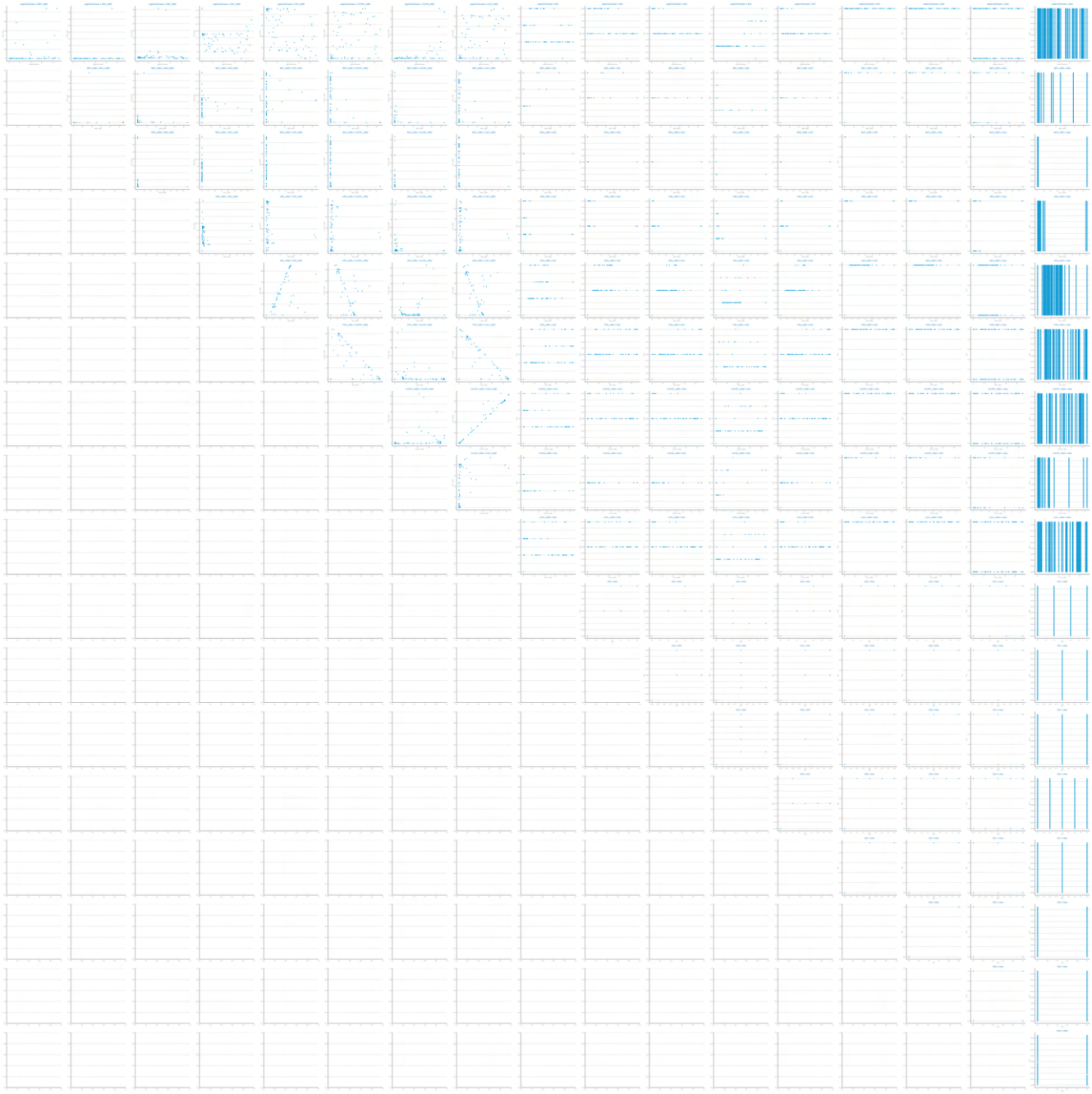


Figure 22: Sparsity analysis for dataset 2

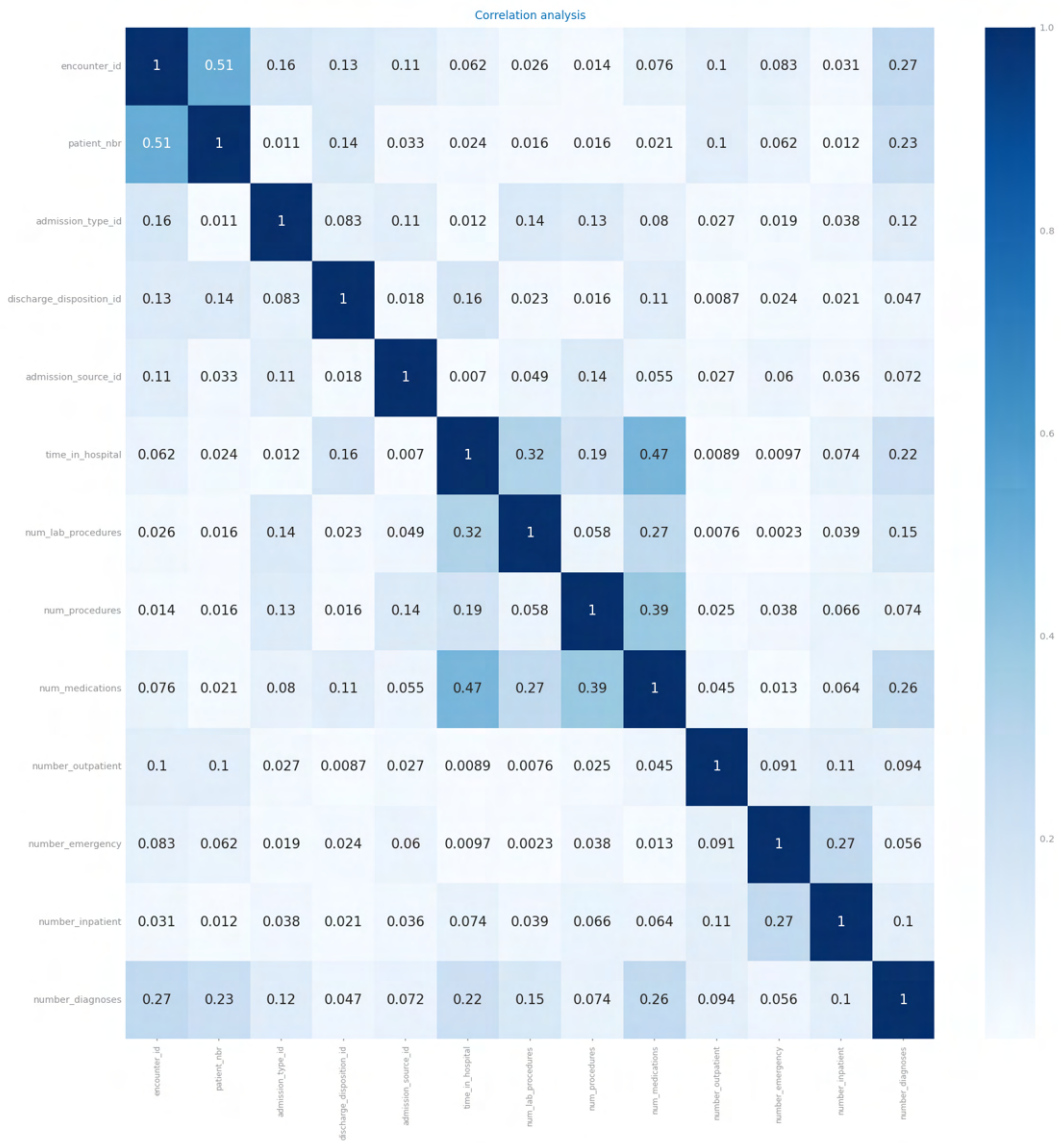


Figure 23: Correlation analysis for dataset 1

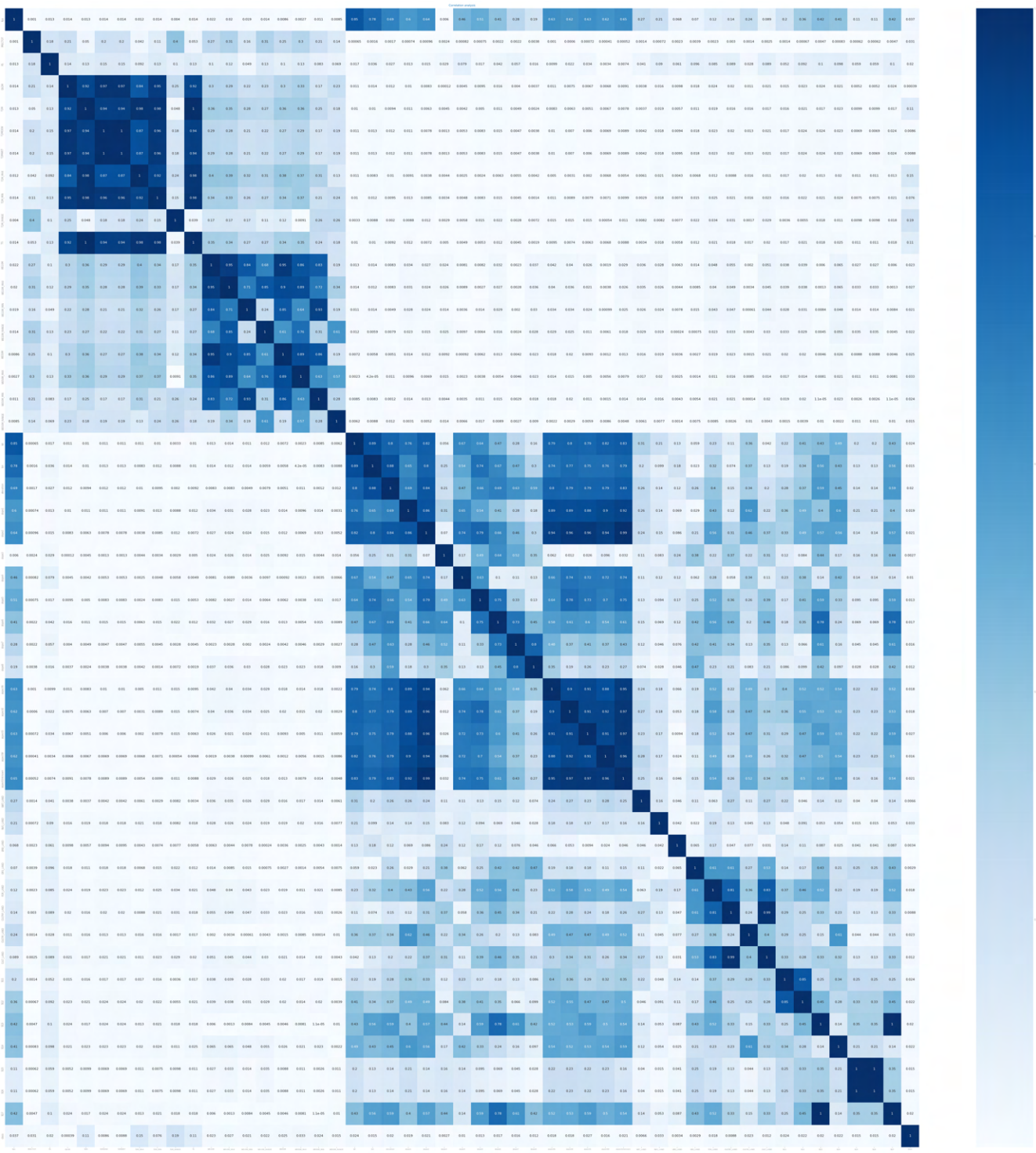


Figure 24: Correlation analysis for dataset 2

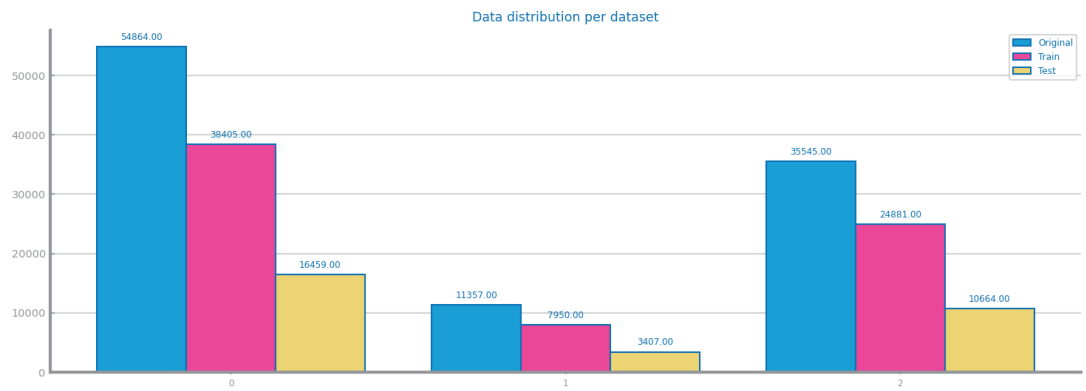
2 DATA PREPARATION

Variables Encoding

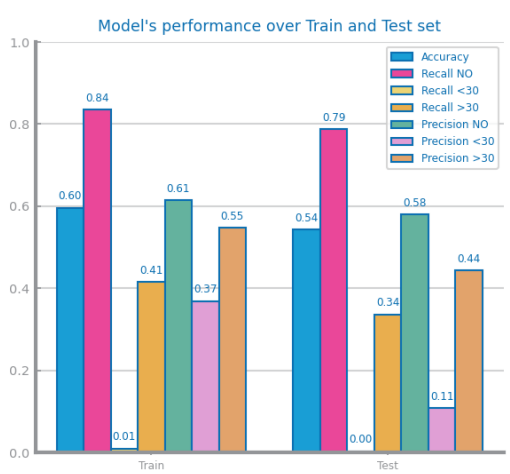
We avoided dummification because that puts us in risk for the dimensionality curse and we might lose information (eg if we dummified date as some months are drier). In 1 we put an ordinal encoding in all symbolic vars. Diagnoses, races and medical specialties were mapped to their group (given in the paper). The order was ascending according to % of readmitted patients with that characteristic. In 2 we encoded the date into year and month (month is cyclical so we used the trigonometric functions).

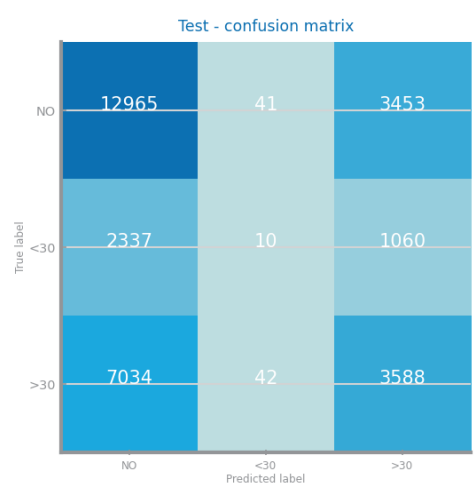
Missing Value Imputation

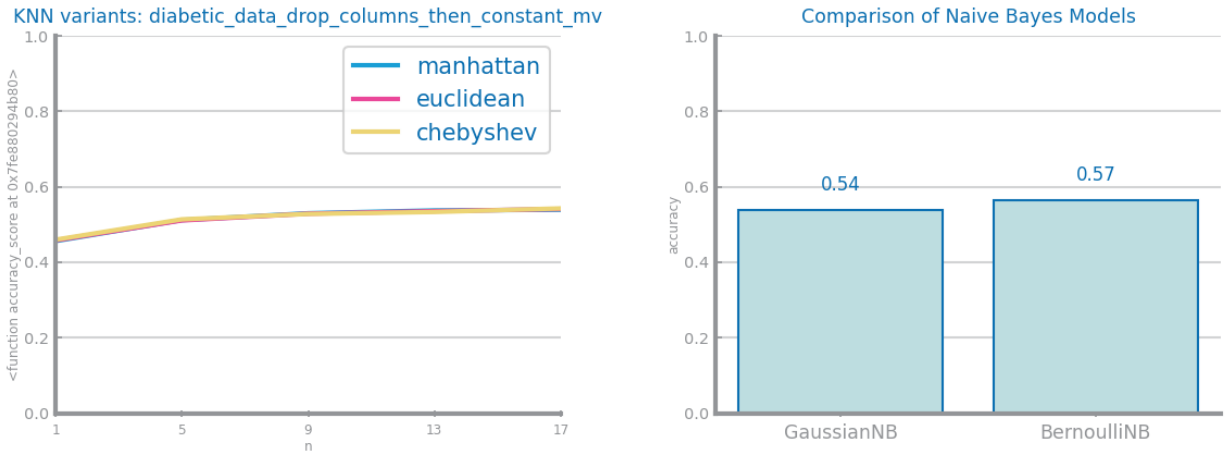
In 1 we dropped *weight* (>90% mv). We compared filling with 0 & most frequent value. We chose the latter as it keeps distribution (theoretically better) but results were similar. 2 had no mv.



KNN Best Results: K=17 and distance=chebyshev







NB Best Results: Bernoulli

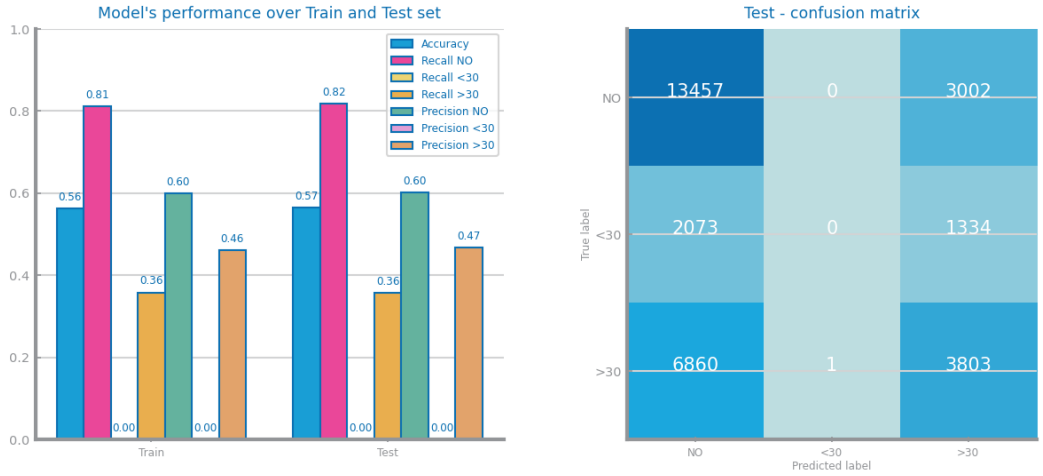
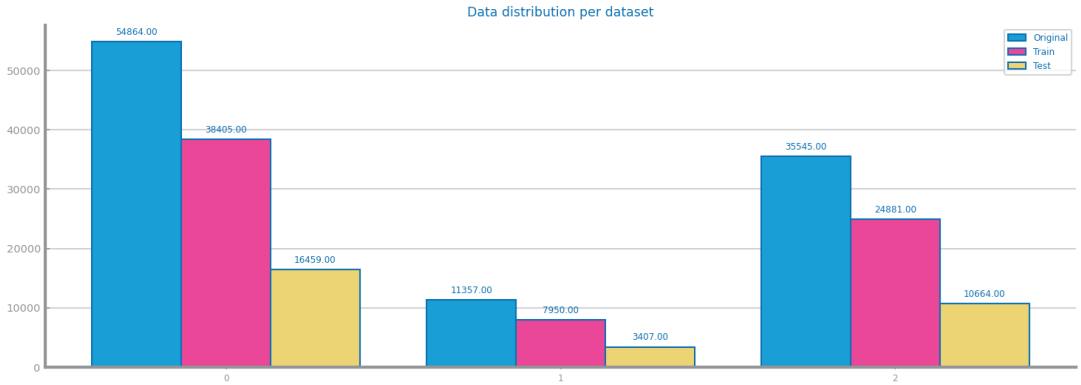
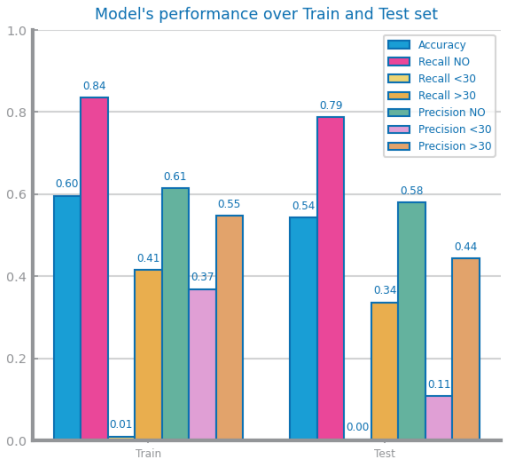
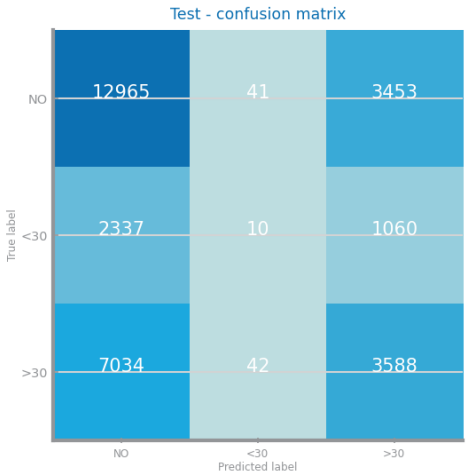


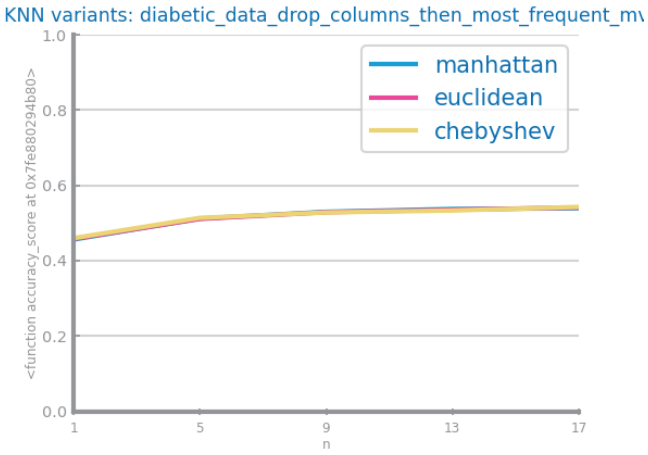
Figure 26: Missing values imputation results for the first approach for dataset 1

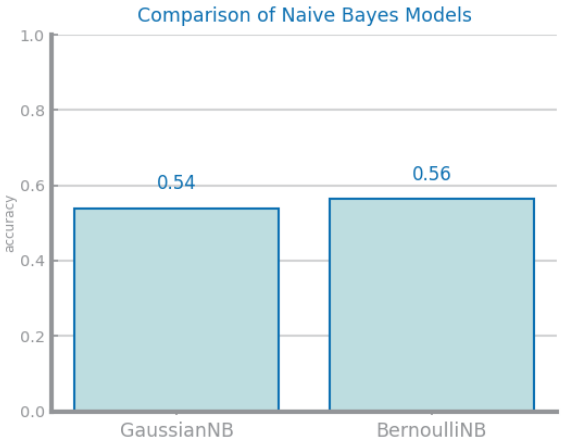


KNN Best Results: K=17 and distance=chebyshev









NB Best Results: Bernoulli

Figure 28: Missing values imputation results for the second approach for dataset 1

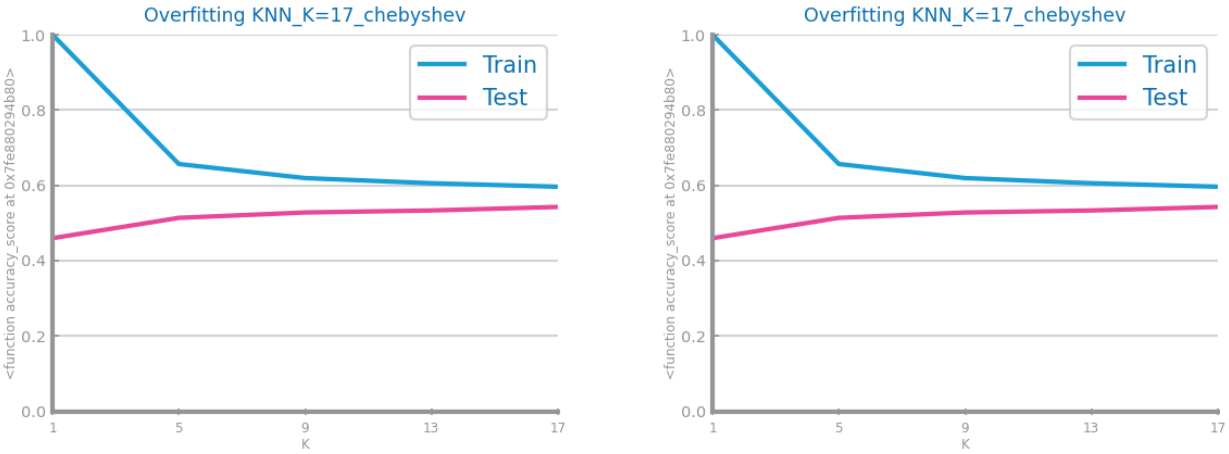


Figure 29: Overfitting results for the first approach (left) and second approach (right) for dataset 1

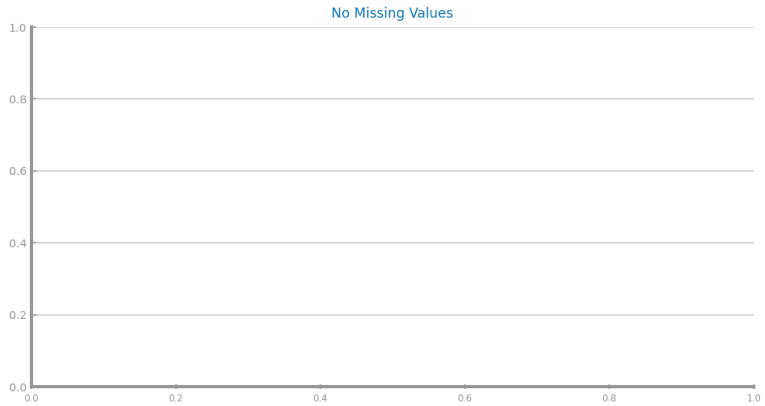
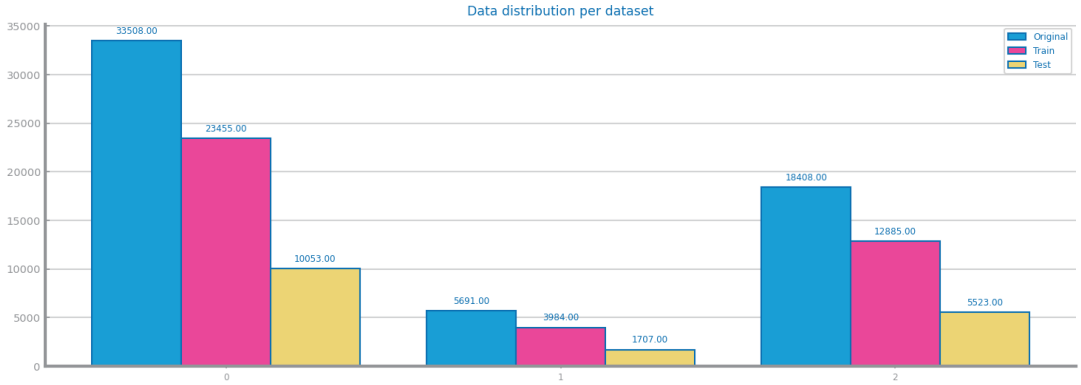


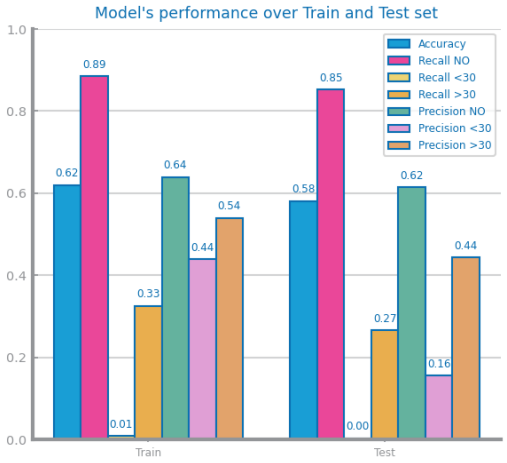
Figure 30: Missing values imputation results with different approaches for dataset 2

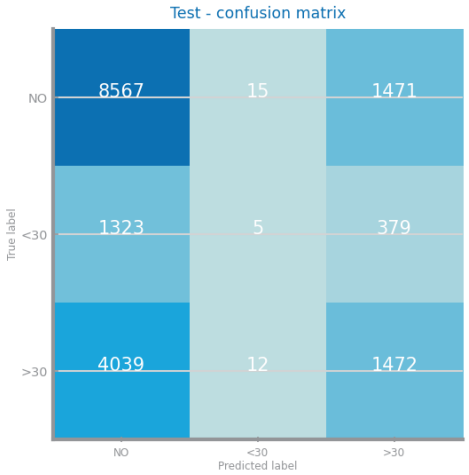
Outliers Treatment

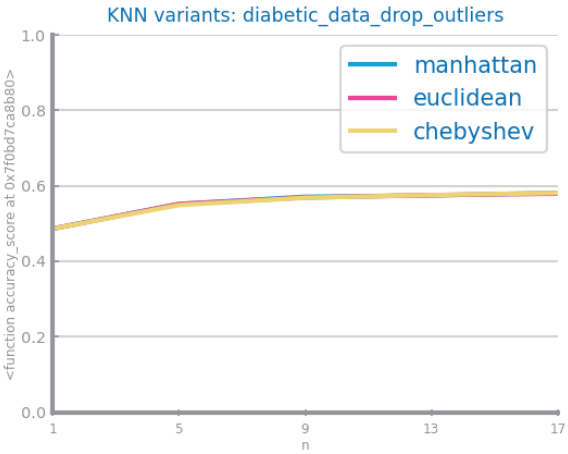
In 1 we used drop/truncate outliers on selected columns- domain knowledge. No improvement: kept previous set. In 2 we used the same and chose drop because scaling was better with it (delayed decision).

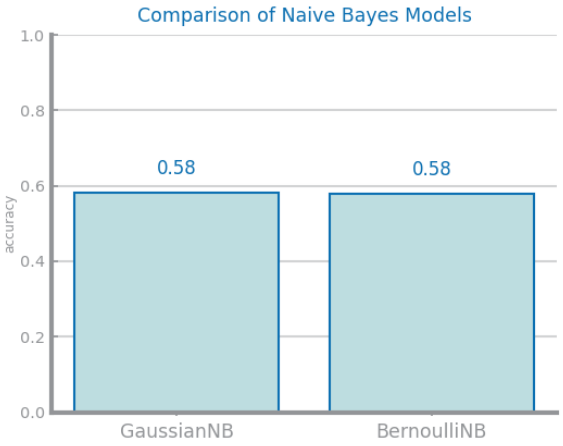


KNN Best Results: K=17 and distance=manhattan









NB Best Results: Bernoulli/Gaussian

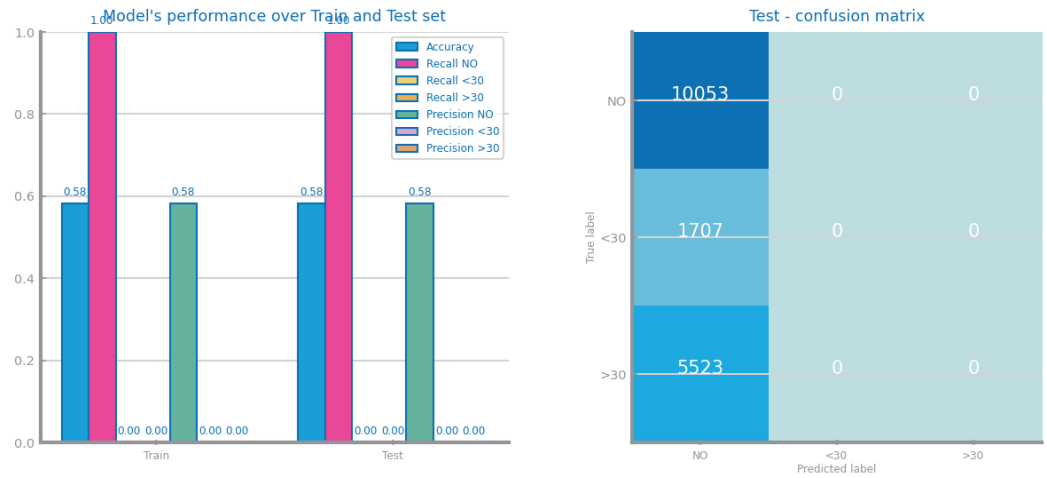
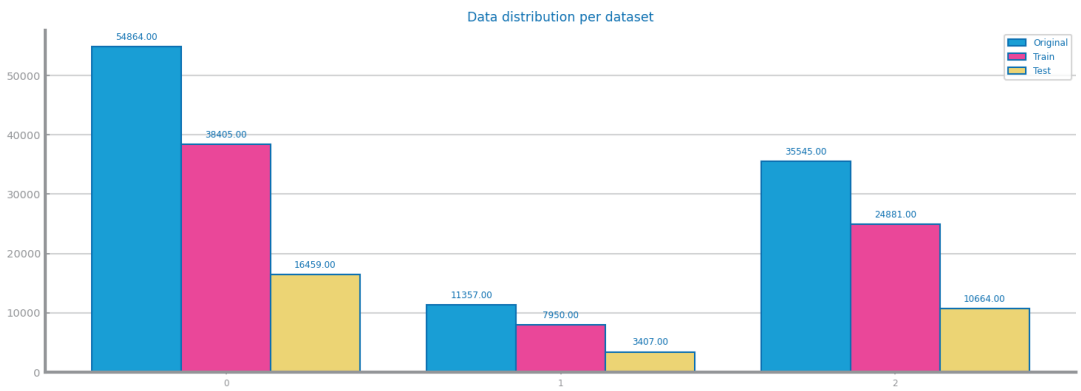
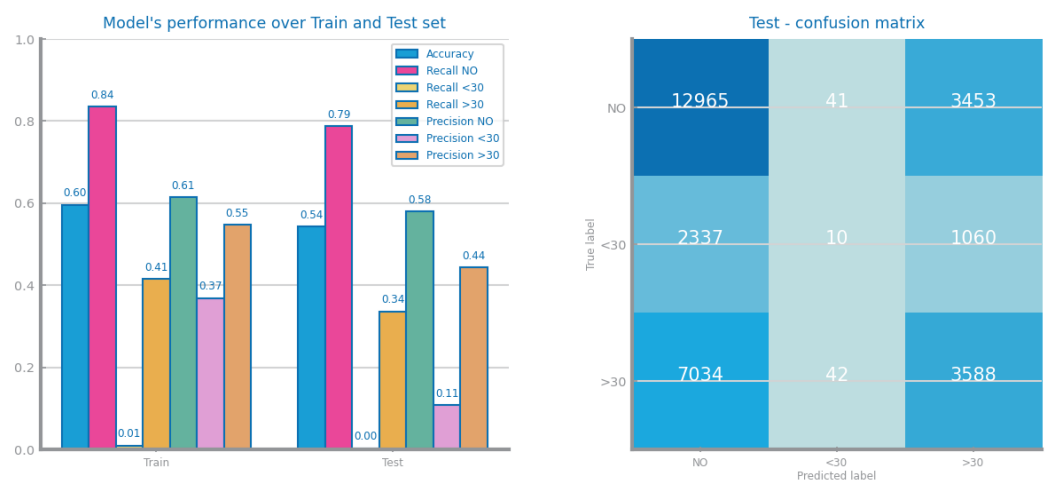
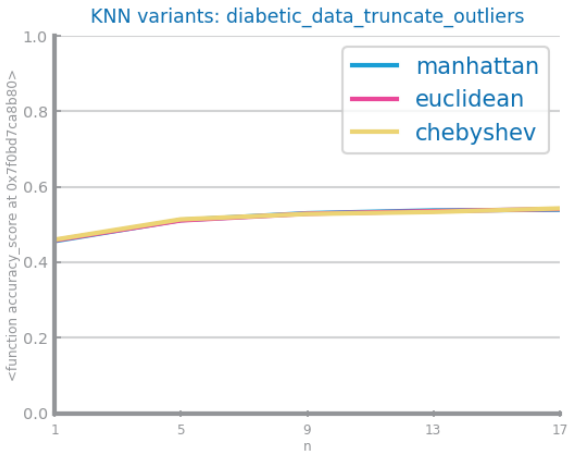


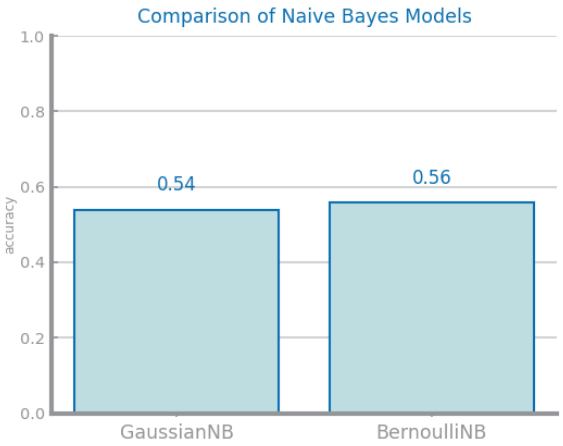
Figure 32: Outliers treatment results for the first approach for dataset 1



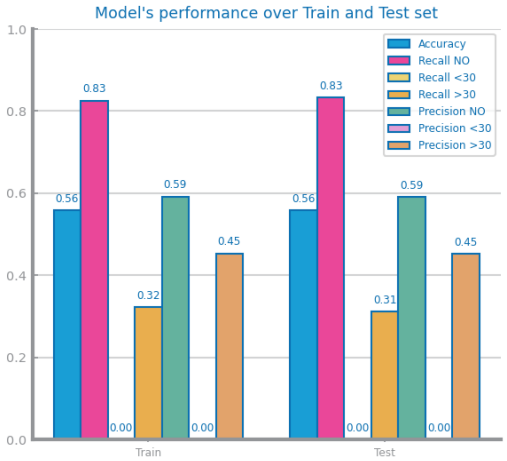
KNN Best Results: K=17 and distance=chebyshev







NB Best Results: Bernoulli



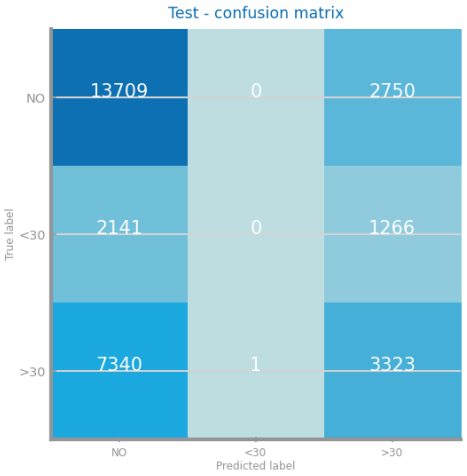


Figure 34: Outliers treatment results for the second approach for dataset 1

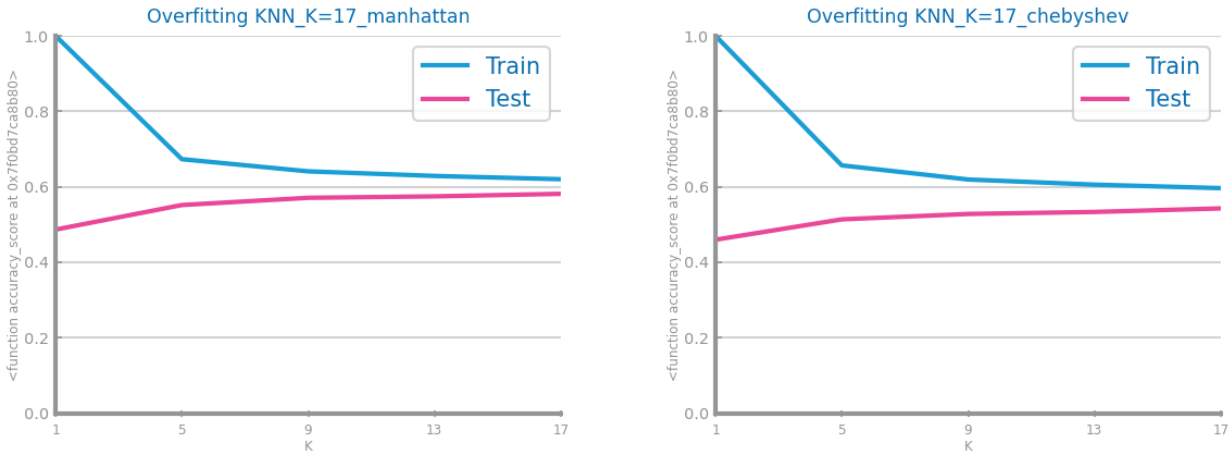
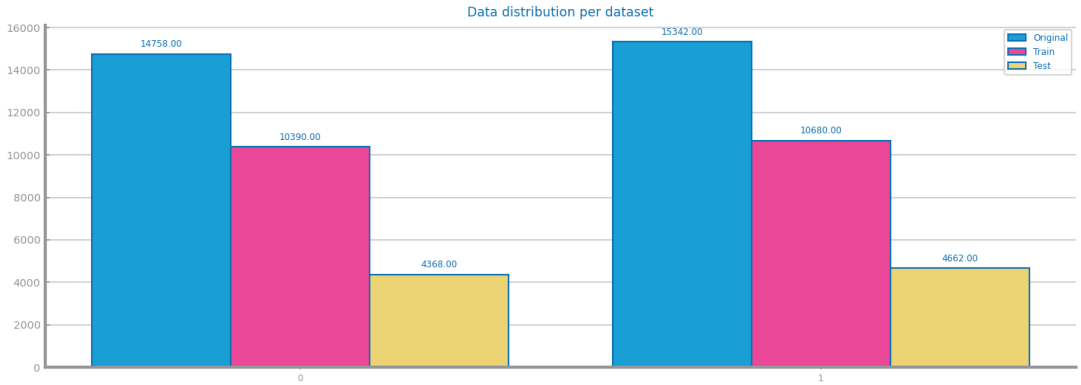
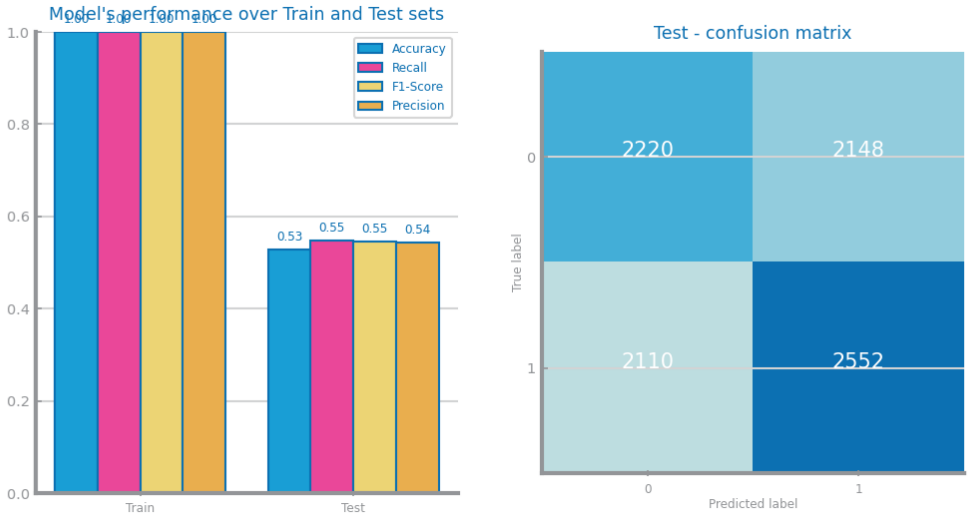
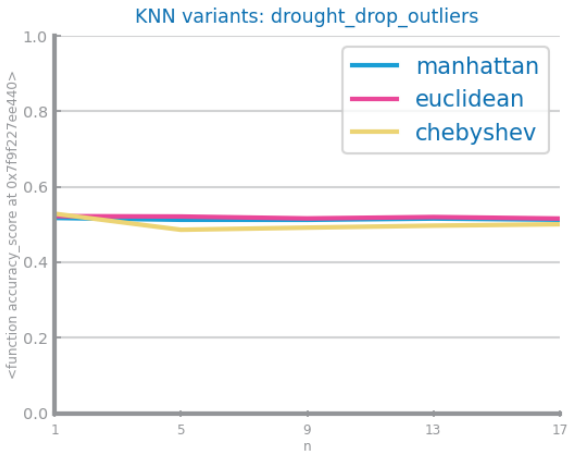


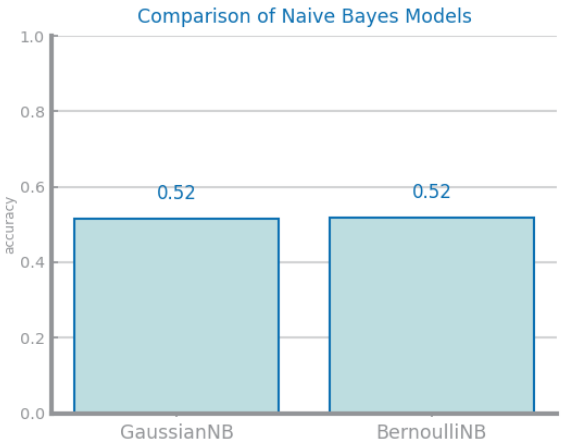
Figure 35: Overfitting results for the first approach (left) and second approach (right) for dataset 1



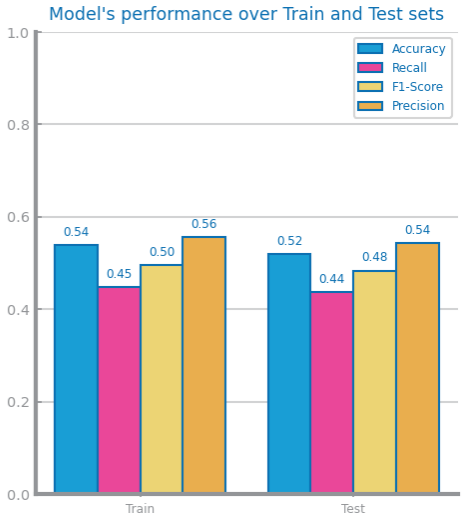
KNN Best Results: K=17 and distance=euclidean







NB Best Results: Bernoulli/Gaussian



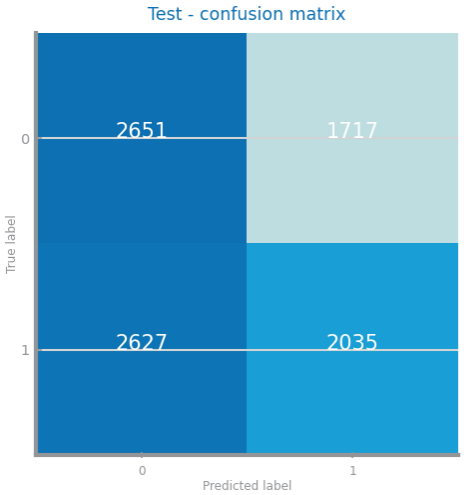
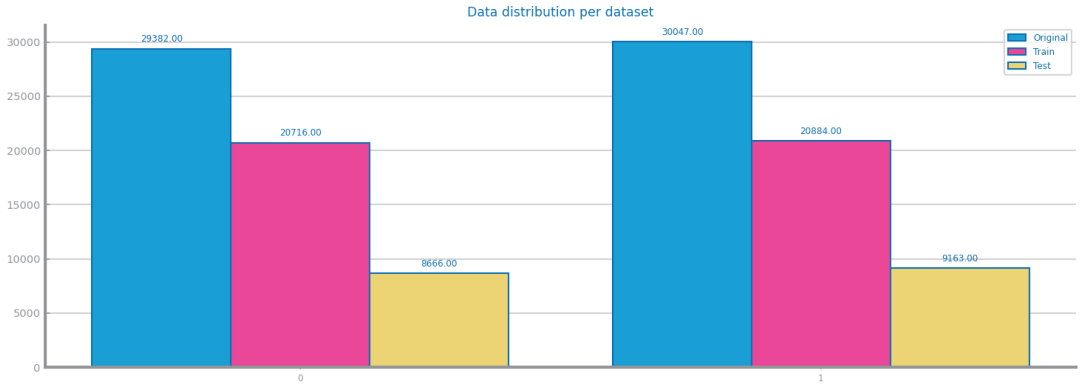
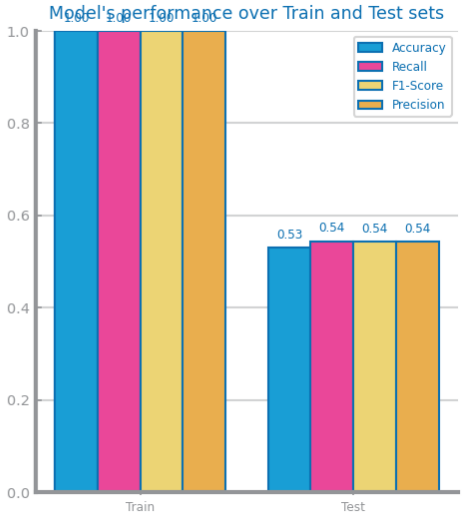
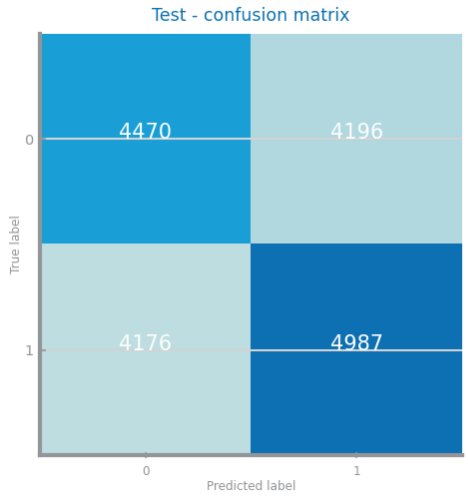


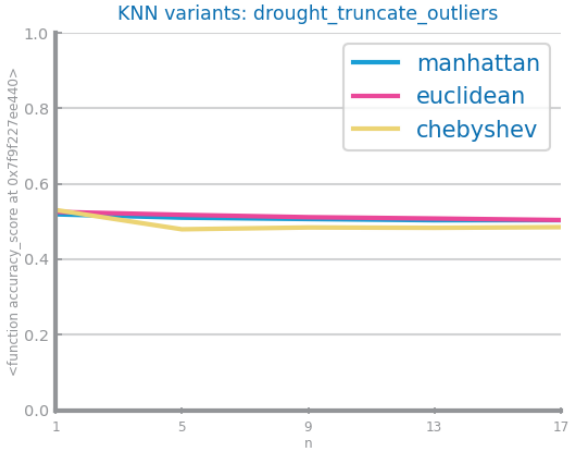
Figure 37: Outliers treatment results for the first approach for dataset 2

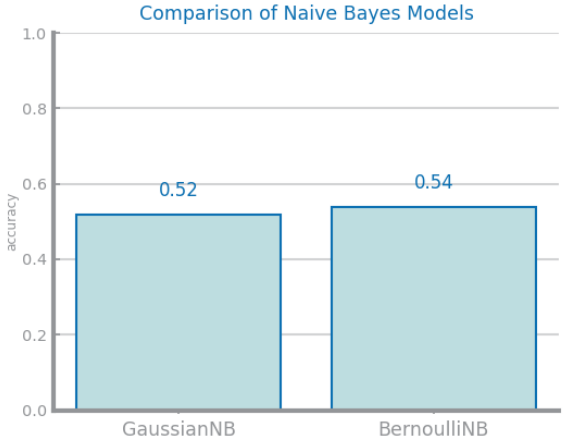


KNN Best Results: K=17 and distance=euclidean









NB Best Results: Bernoulli

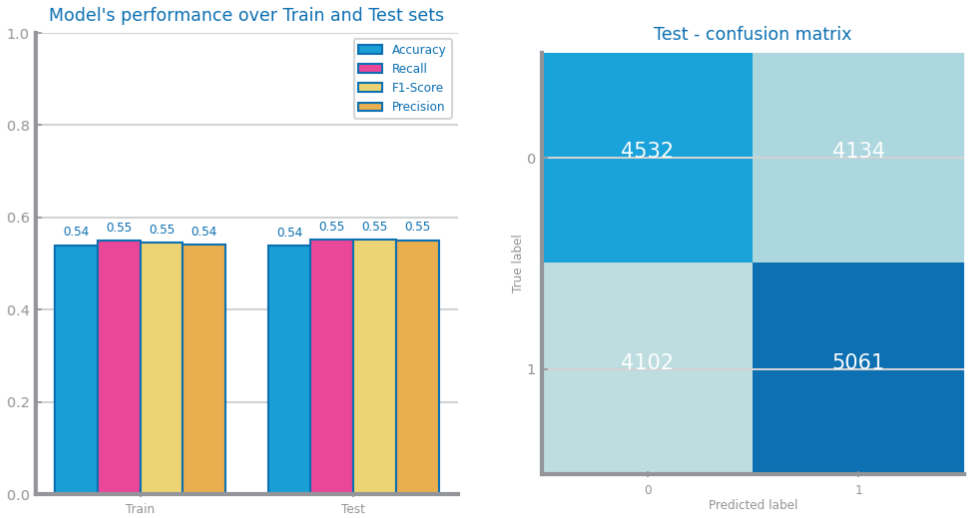


Figure 39: Outliers treatment results for the second approach for dataset 2

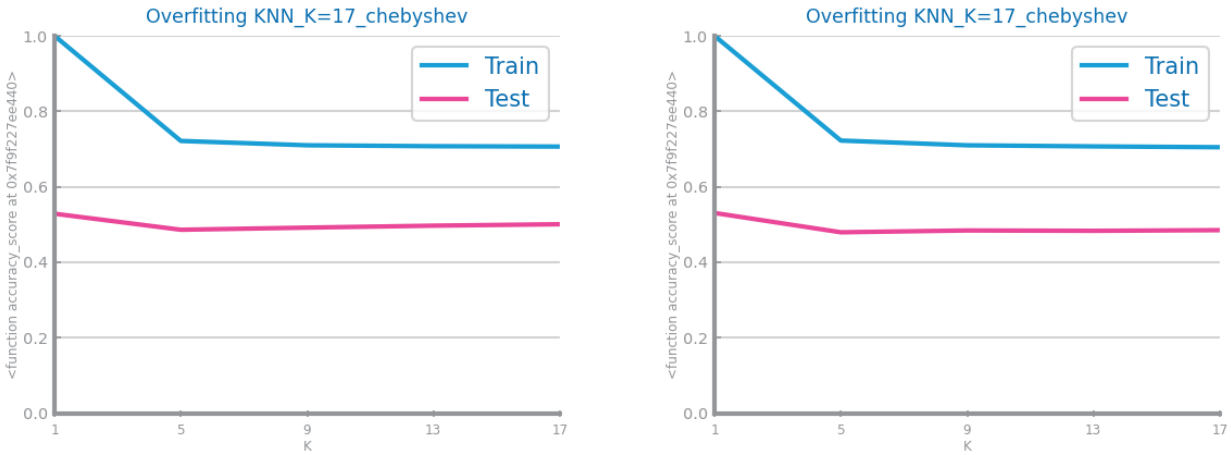


Figure 40: Overfitting results for the first approach (left) and second approach (right) for dataset 2

Scaling

Tested MinMax & Zscore in both. In 1 we chose zscore. In 2 results were alike in NB but KNN was a bit better for minmax. Zscore is good for norm and MinMax is good to preserve original distributions.

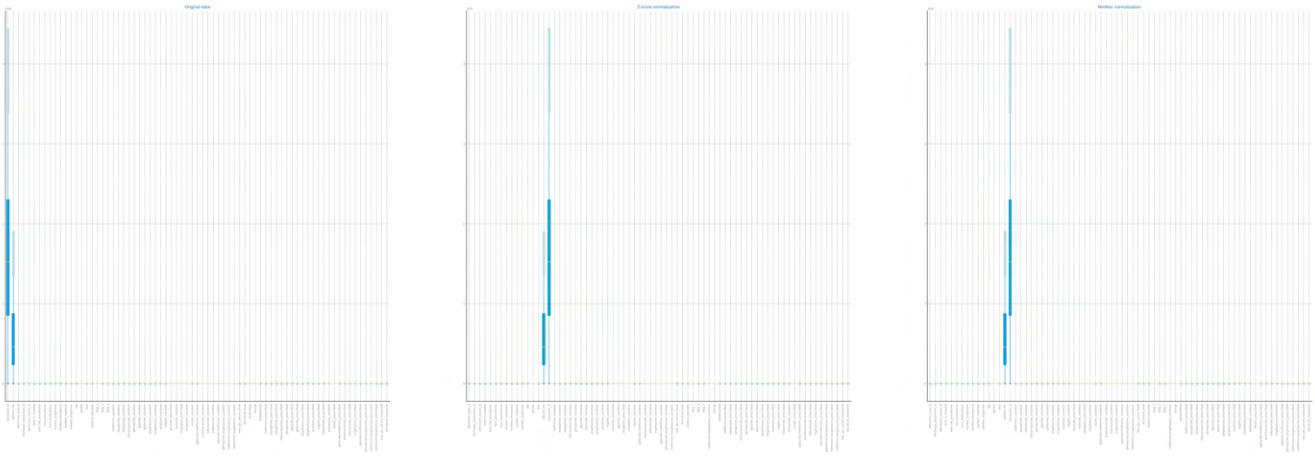
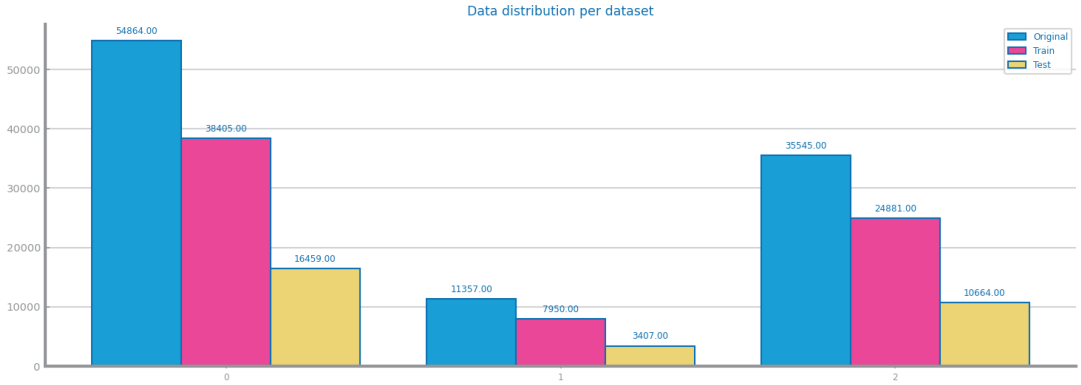
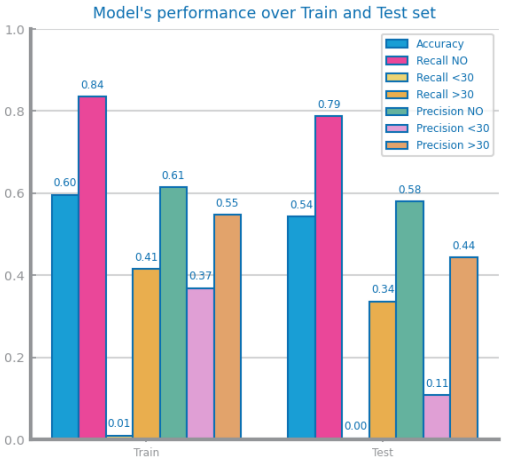
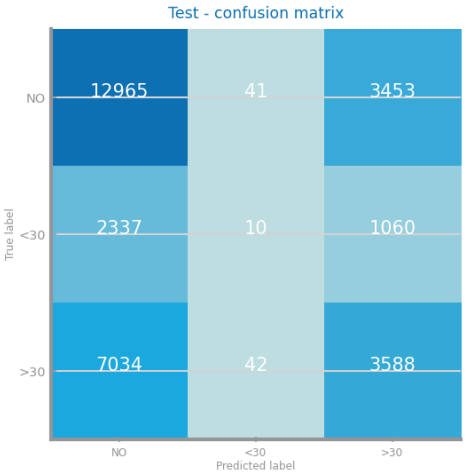


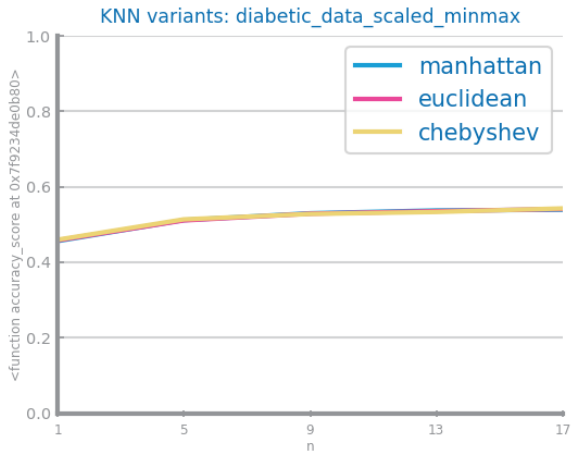
Figure 41: Scale comparison for dataset 1

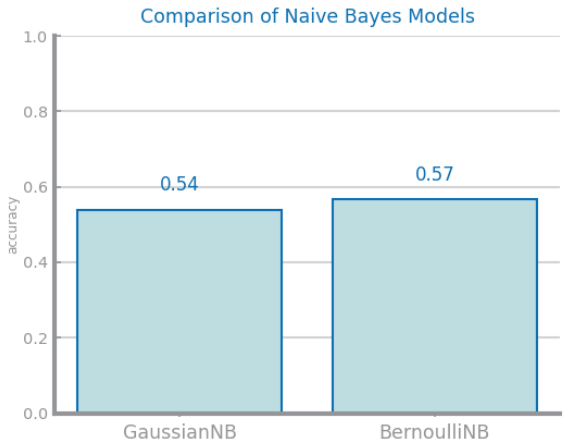


KNN Best Results: K=17 and distance=chebyshev

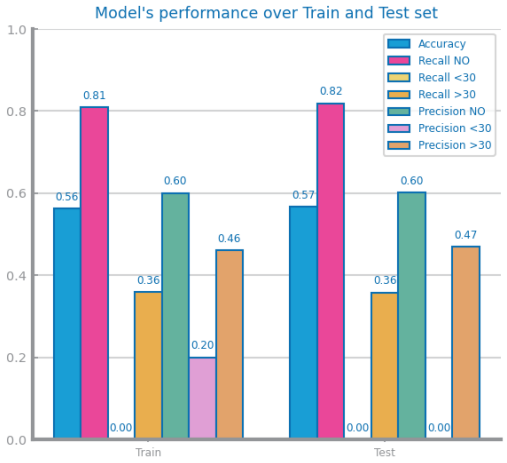








NB Best Results: Bernoulli



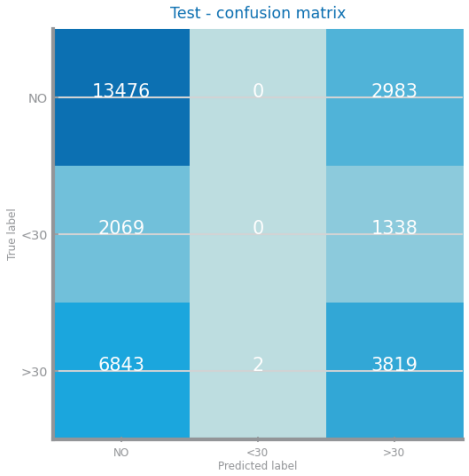
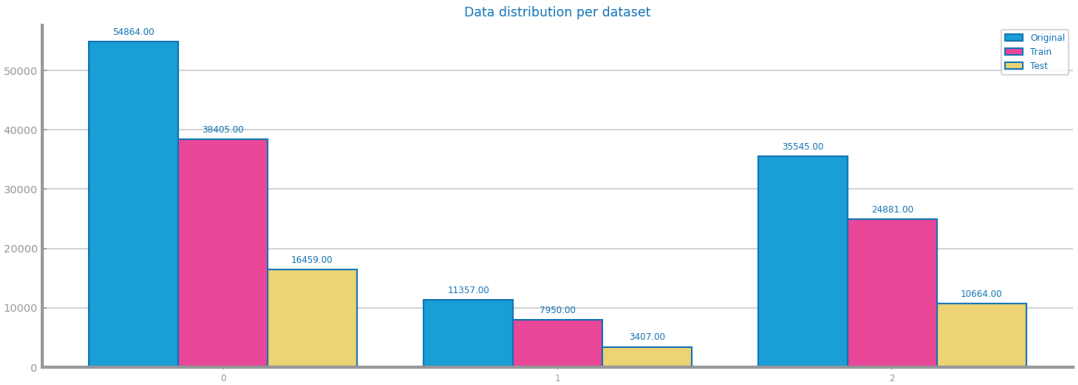
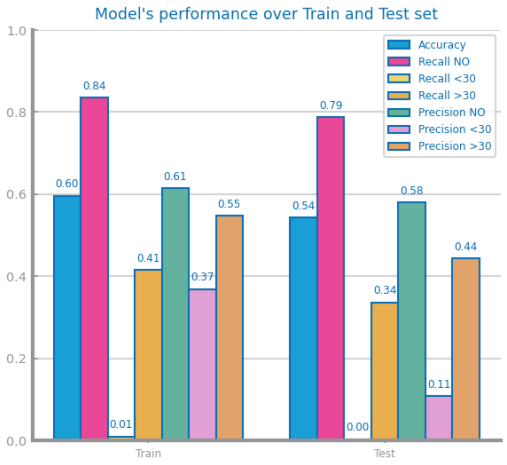
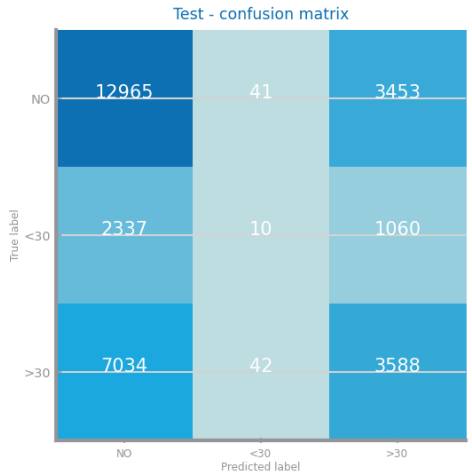


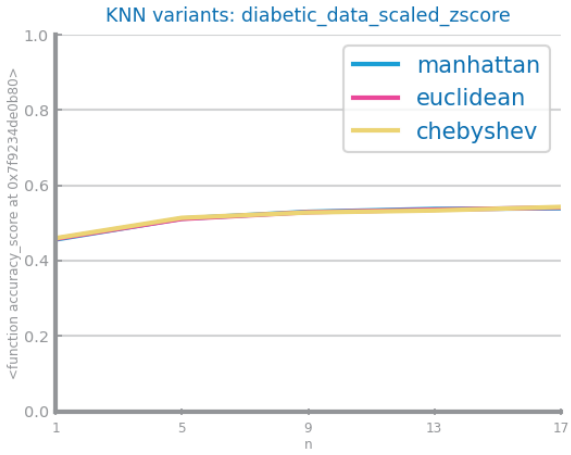
Figure 43: Scaling results for the first approach (Min-Max scaler) for dataset 1

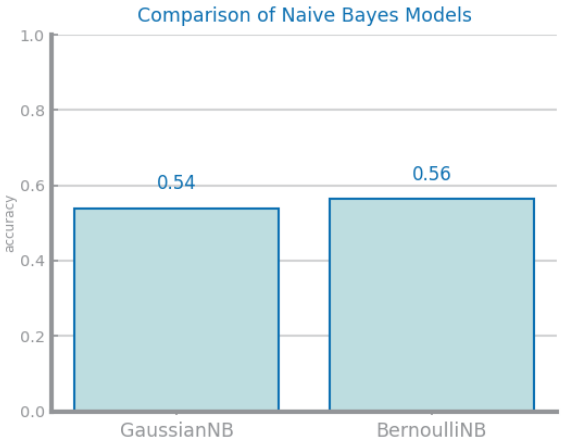


KNN Best Results: K=17 and distance=chebyshev









NB Best Results: Bernoulli

Figure 45: Scaling results for the second approach (Z-score) for dataset 1

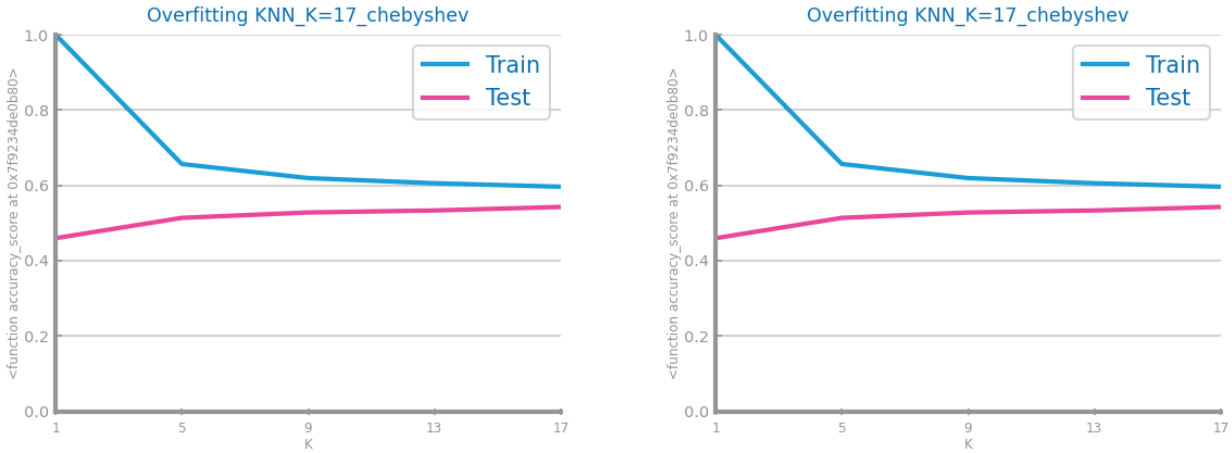


Figure 46: Overfitting results for the first approach (left) and second approach (right) for dataset 1

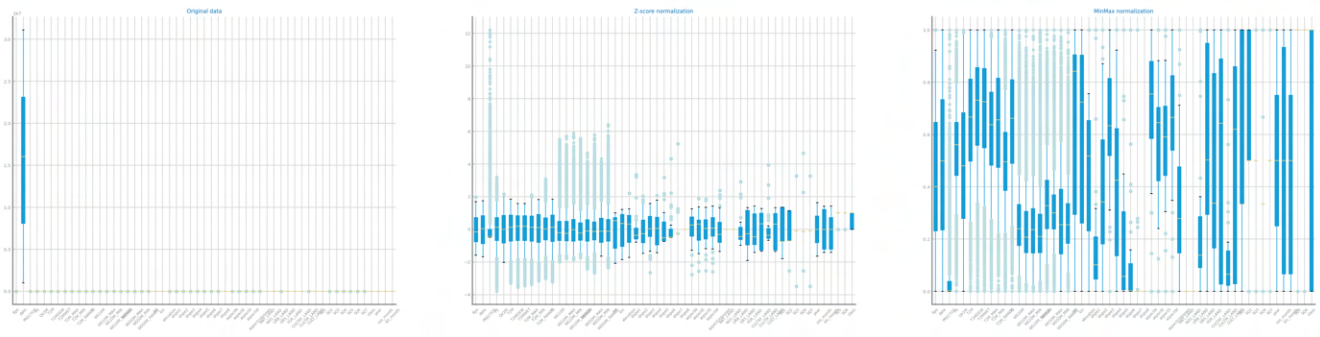
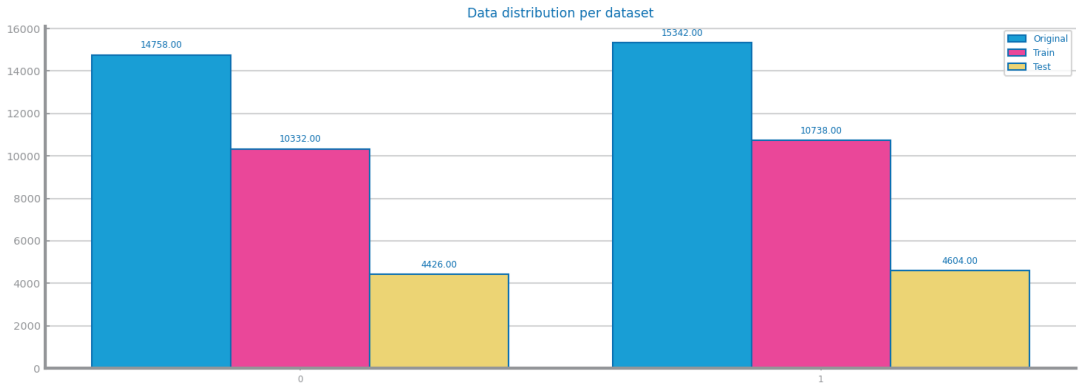
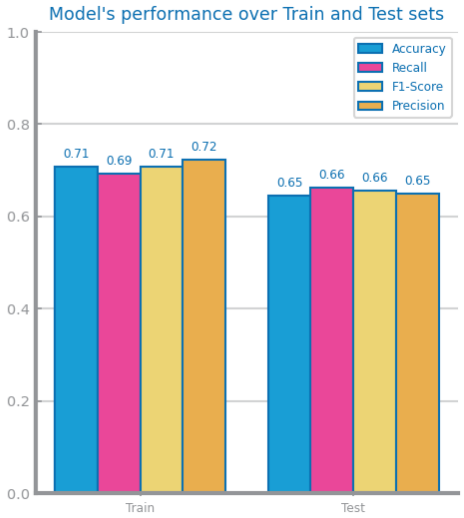
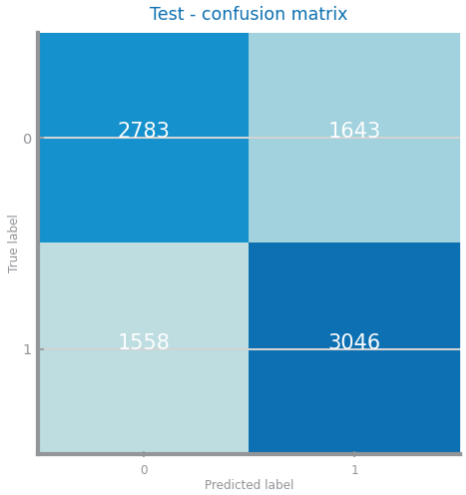


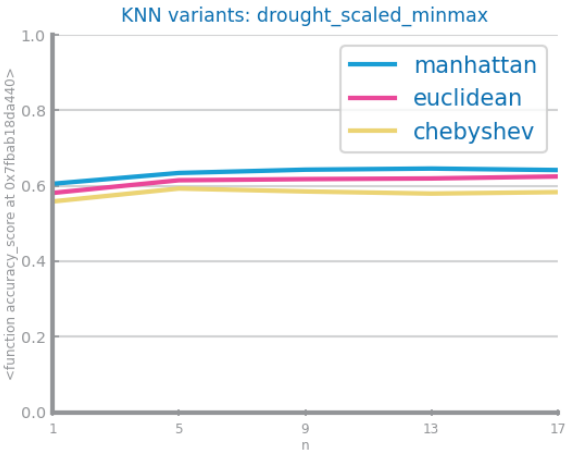
Figure 47: Scale comparison for dataset 2

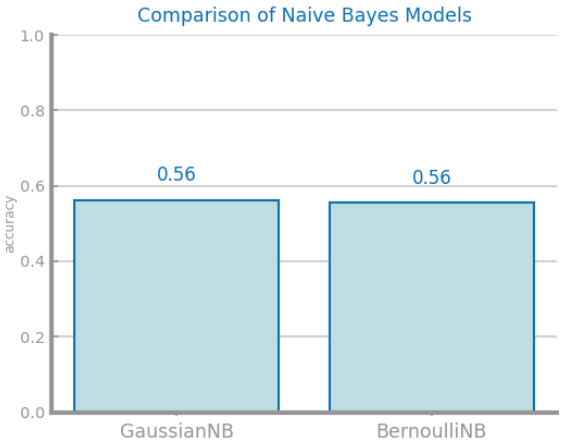


KNN Best Results: K=17 and distance=manhattan









NB Best Results: Bernoulli/Gaussian

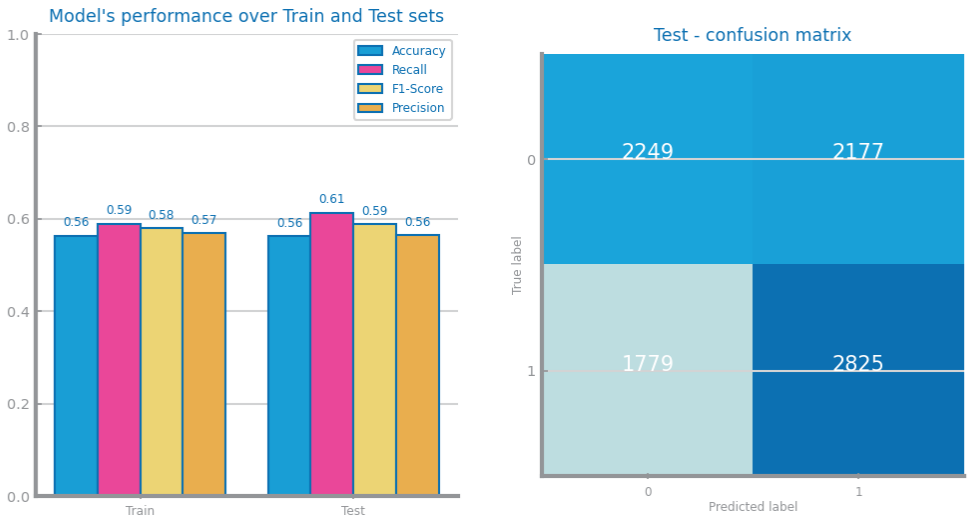
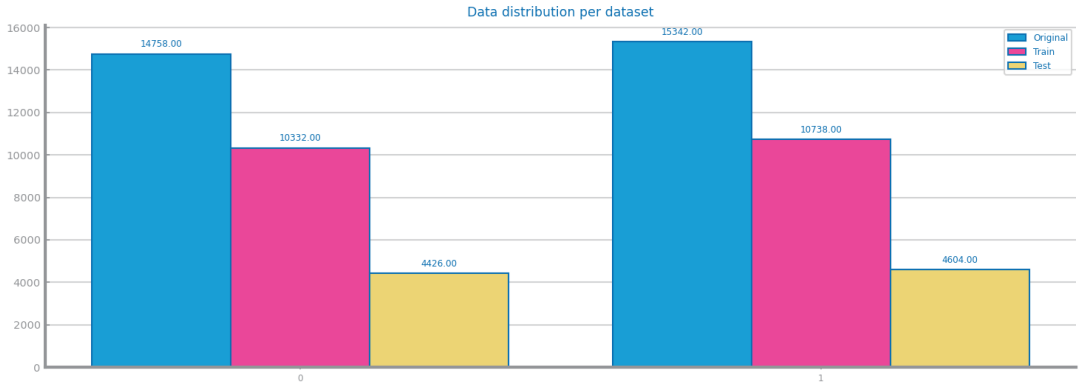
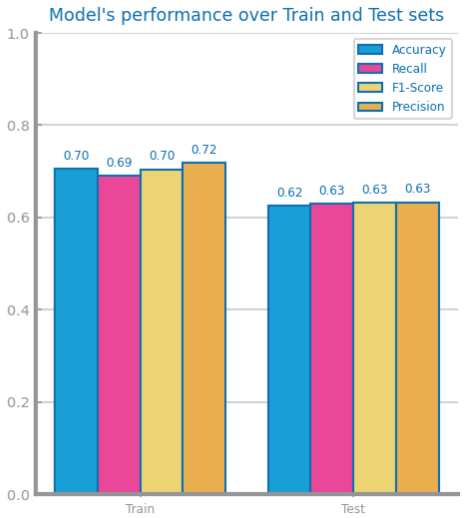
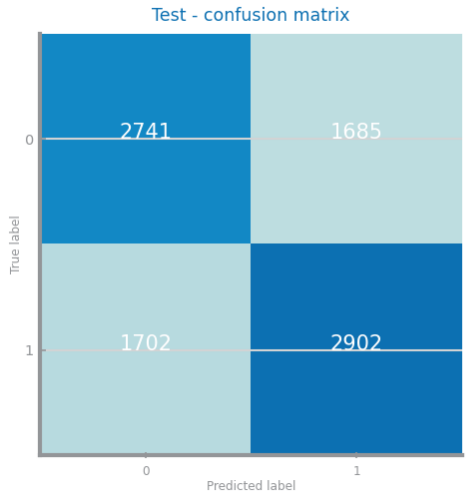


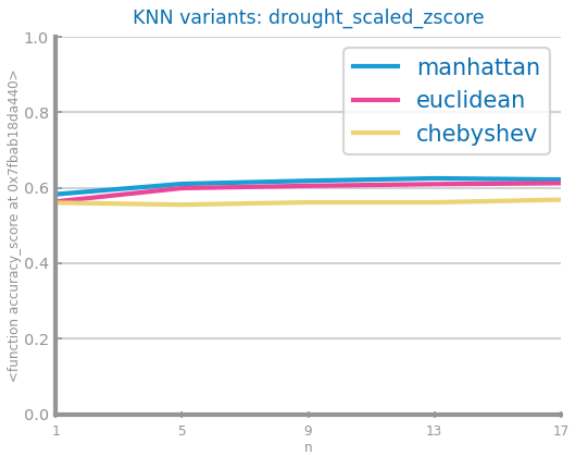
Figure 49: Scaling results for the first approach (Min-Max scaler) for dataset 2

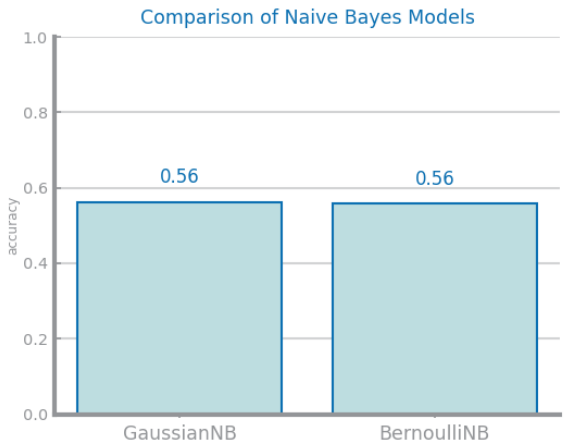


KNN Best Results: K=17 and distance=manhattan









NB Best Results: Bernoulli/Gaussian

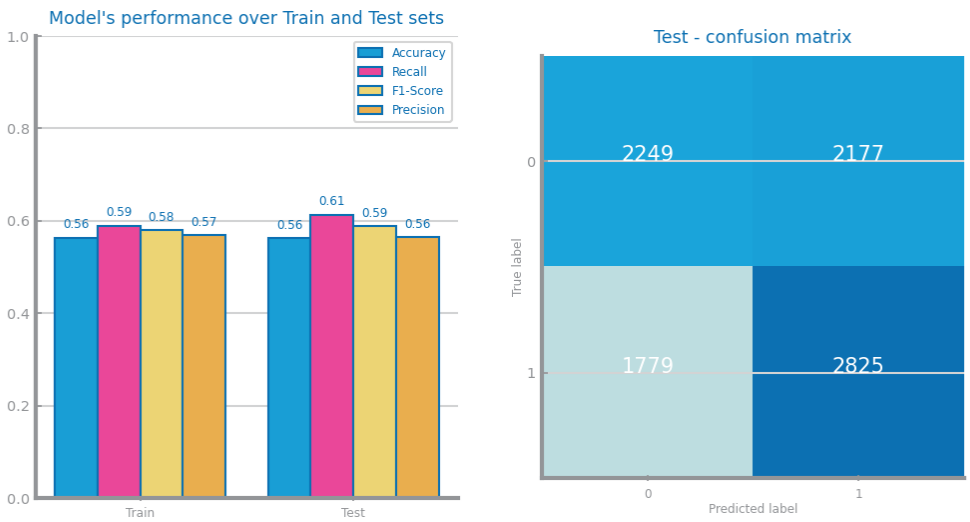


Figure 51: Scaling results for the second approach (Z-score) for dataset 2

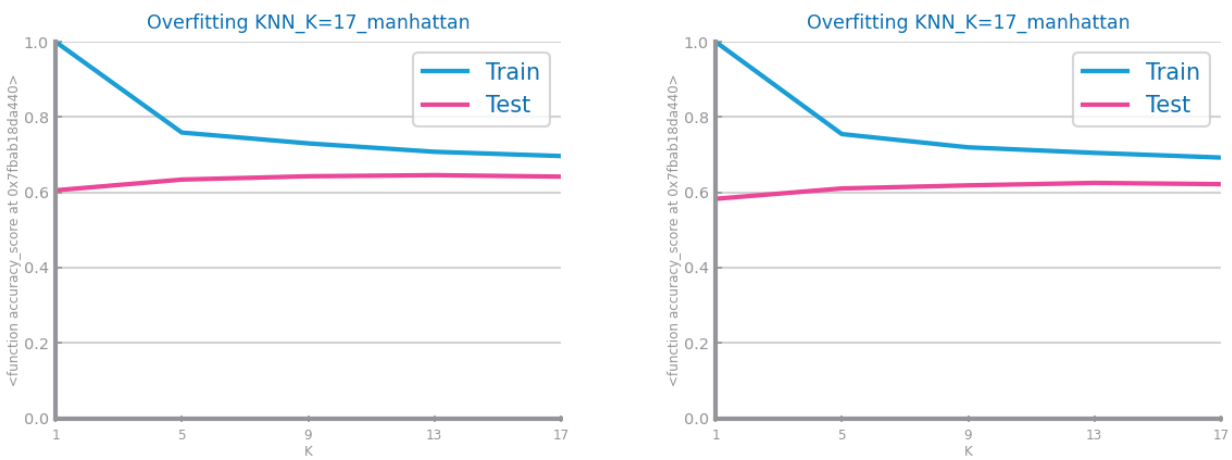


Figure 52: Overfitting results for the first approach (left) and second approach (right) for dataset 2

Balancing

We tried SMOTE, under, over & a mixed sampling for both. For 1, under was chosen since it had improved metrics for the most important under-represented classes. For 2, SMOTE had better results.

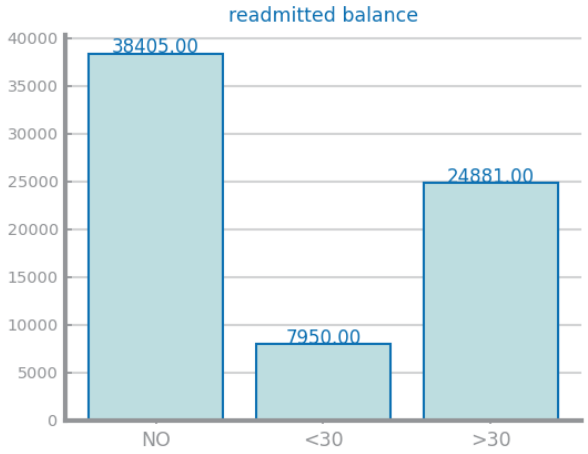
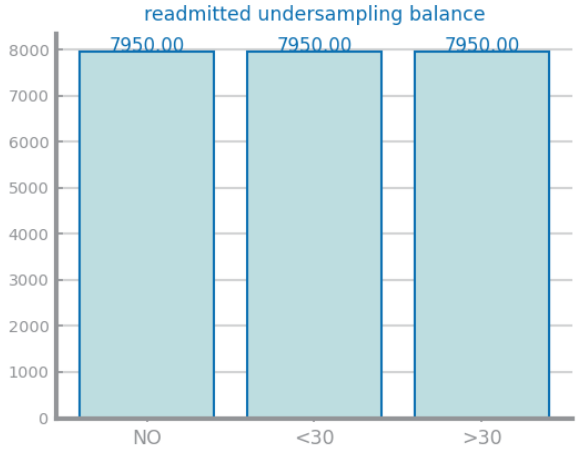
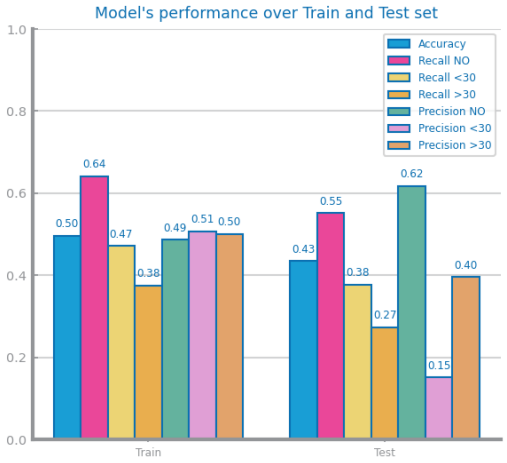
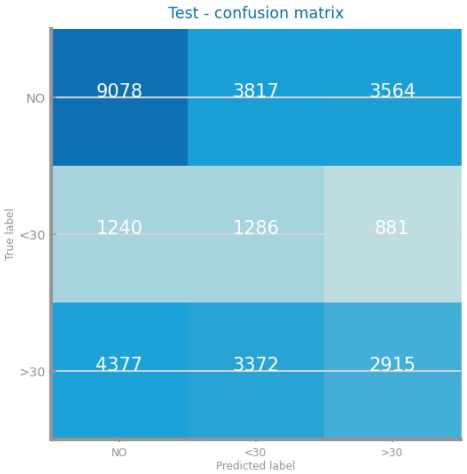


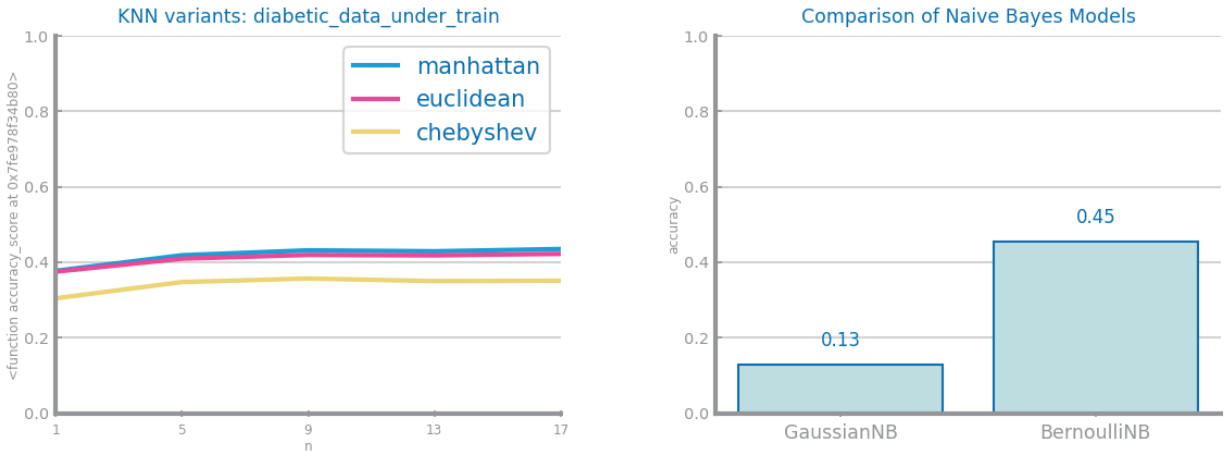
Figure 53: Data balance before balancing techniques



KNN Best Results: K=17 and distance=manhattan







NB Best Results: Bernoulli

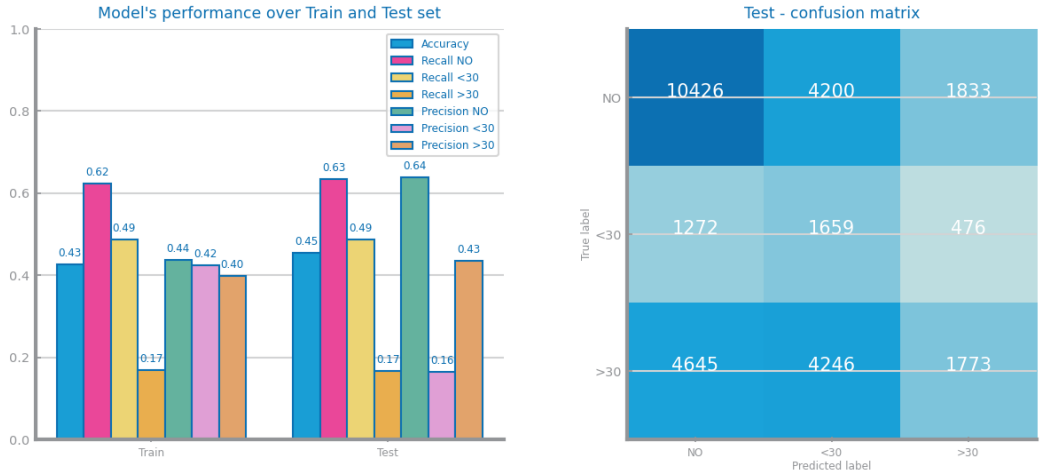
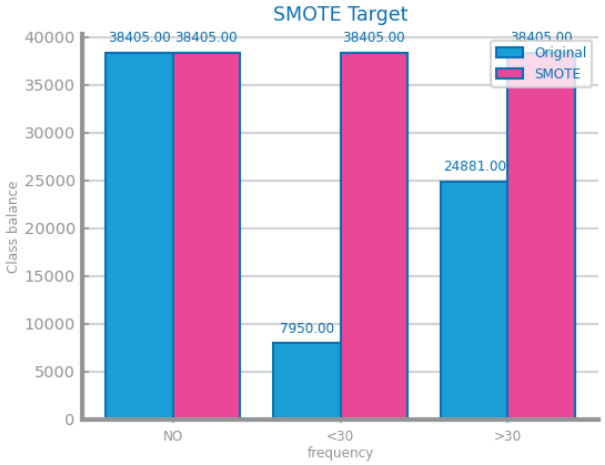
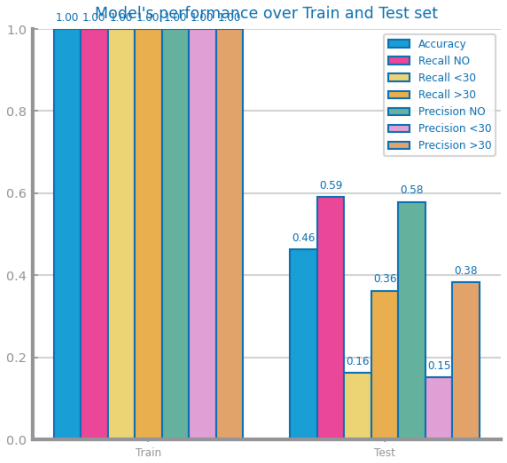
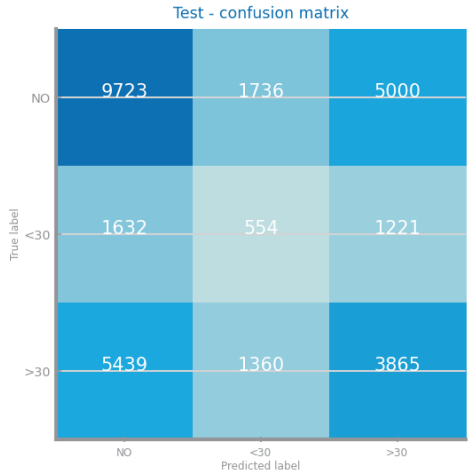


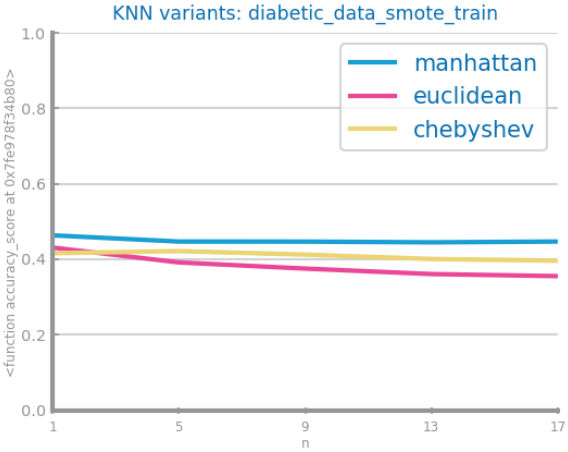
Figure 55: Balancing results for the first approach (Undersampling) for dataset 1

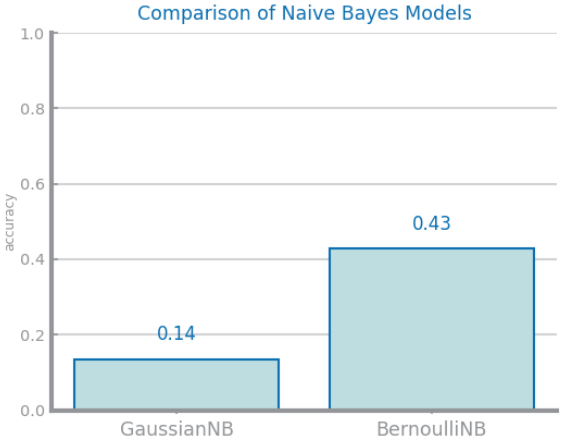


KNN Best Results: K=17 and distance=manhattan









NB Best Results: Bernoulli

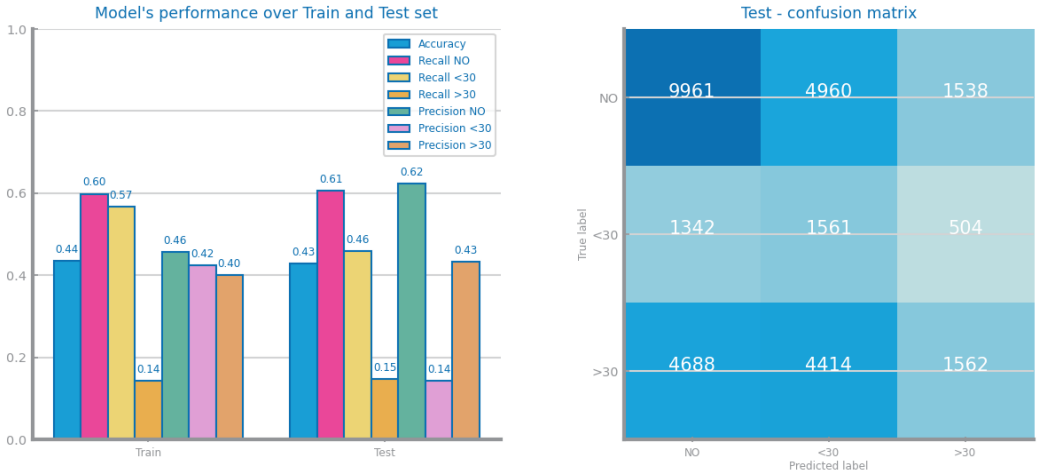


Figure 57: Balancing results for the second approach (SMOTE) for dataset 1

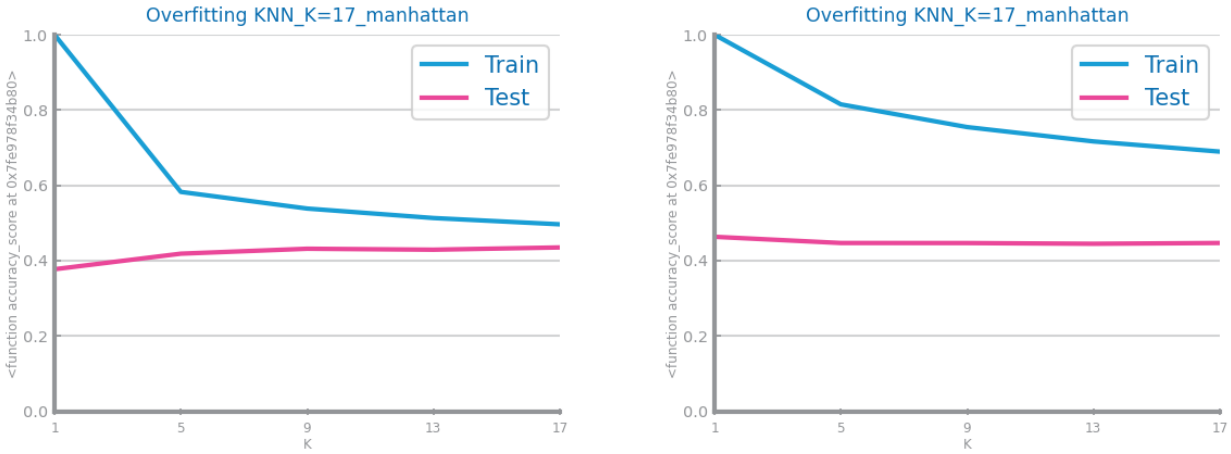
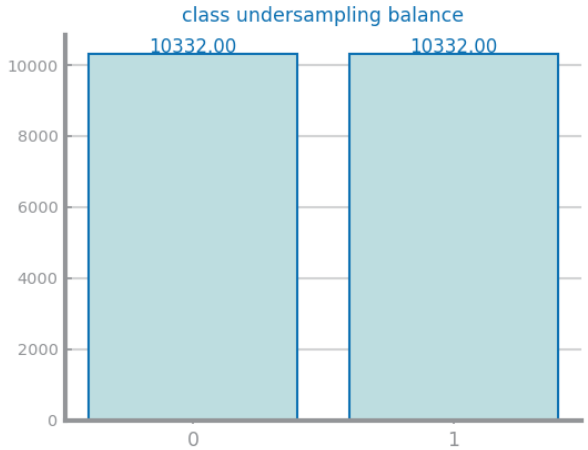
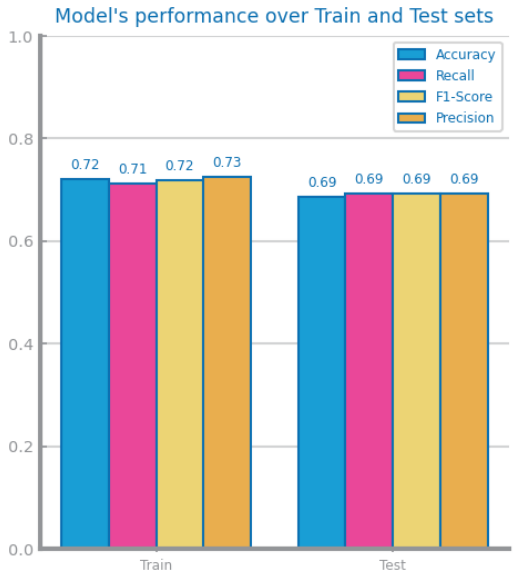
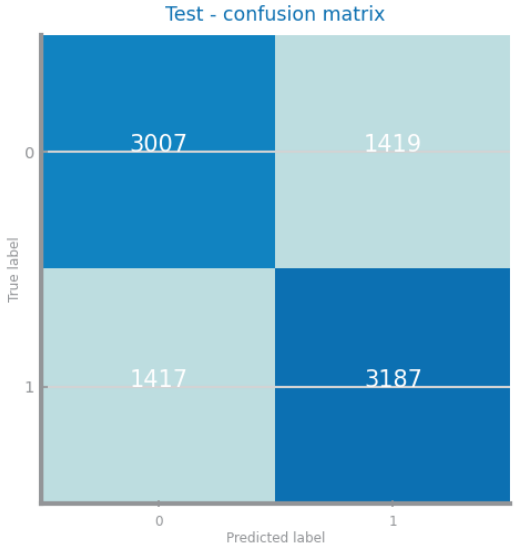


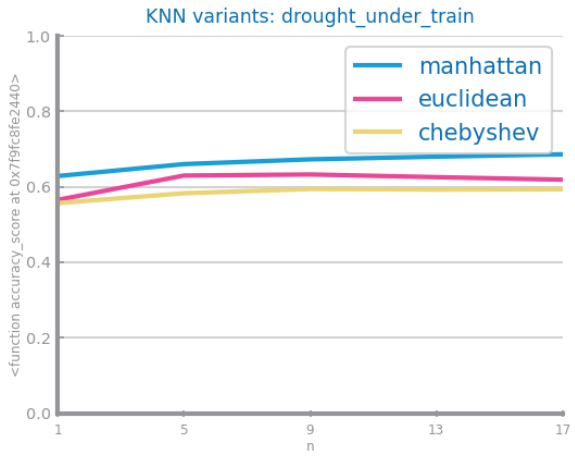
Figure 58: Overfitting results for the first approach (left) and second approach (right) for dataset 1

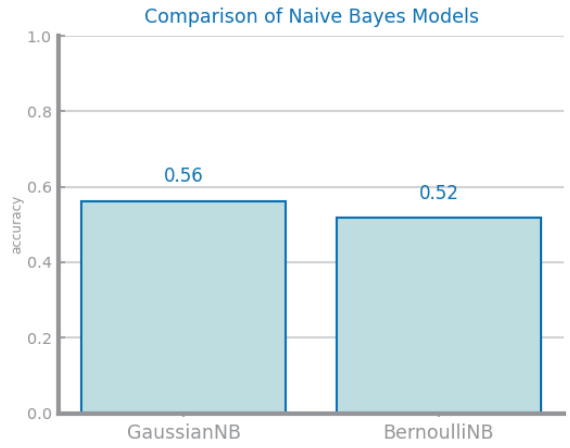


KNN Best Results: K=17 and distance=manhattan

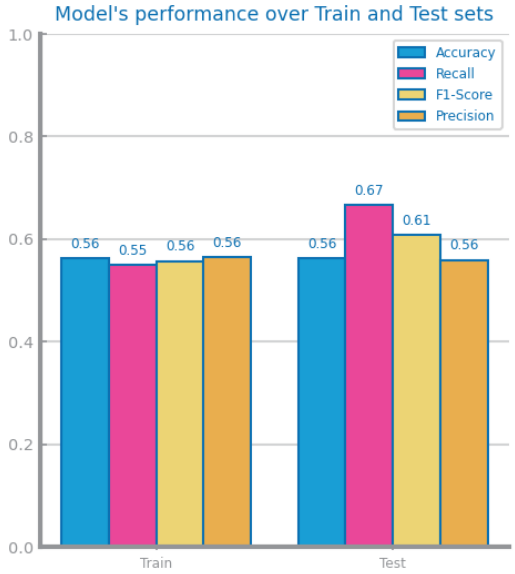








NB Best Results: Gaussian



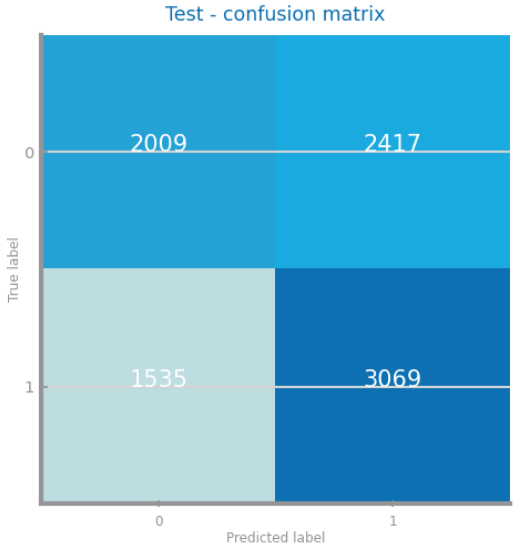
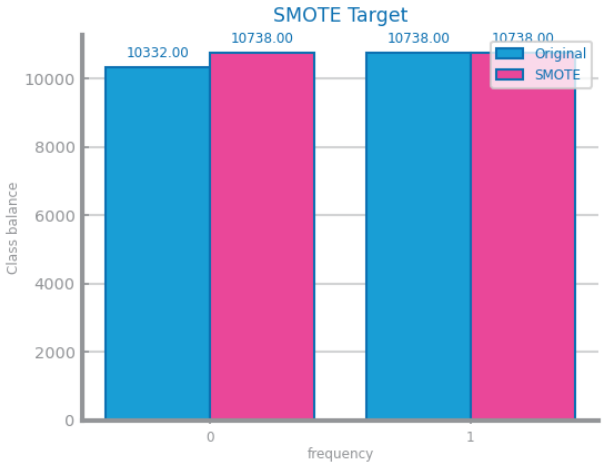
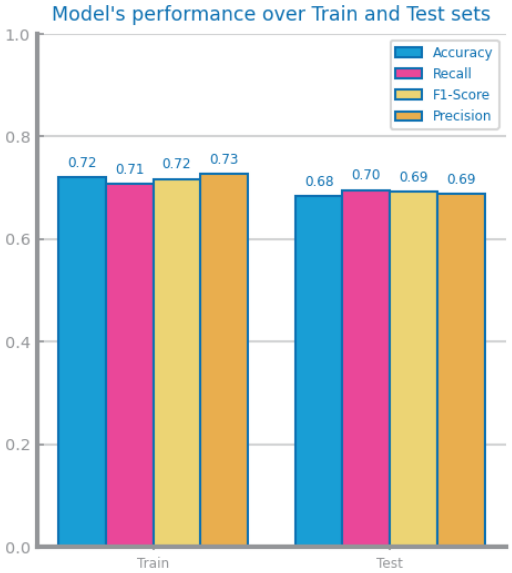
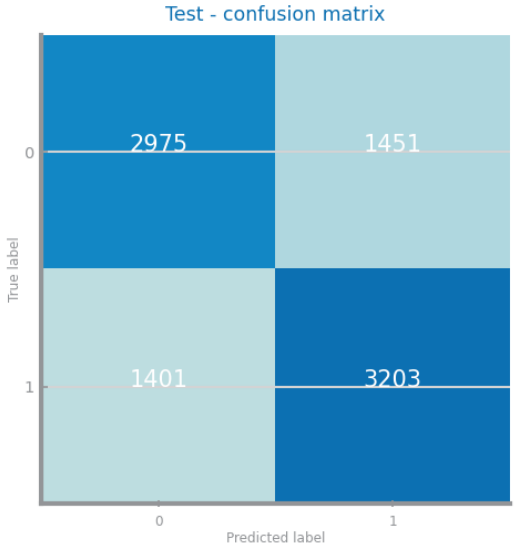


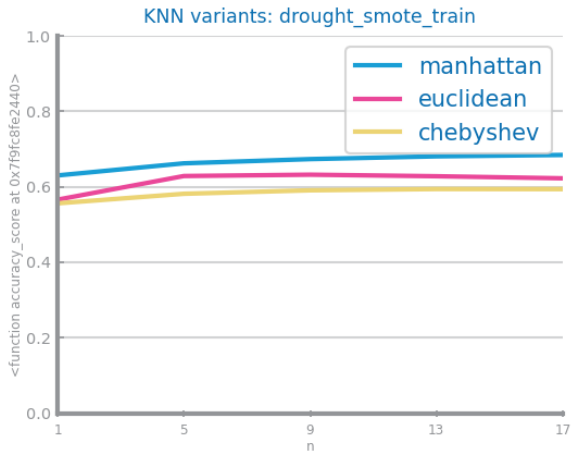
Figure 60: Balancing results for the first approach (Undersampling) for dataset 2

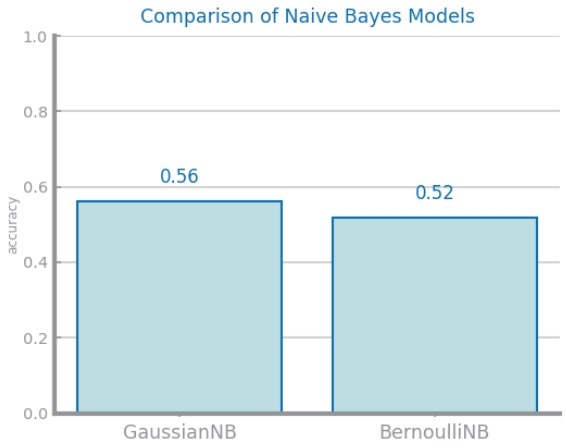


KNN Best Results: K=17 and distance=manhattan

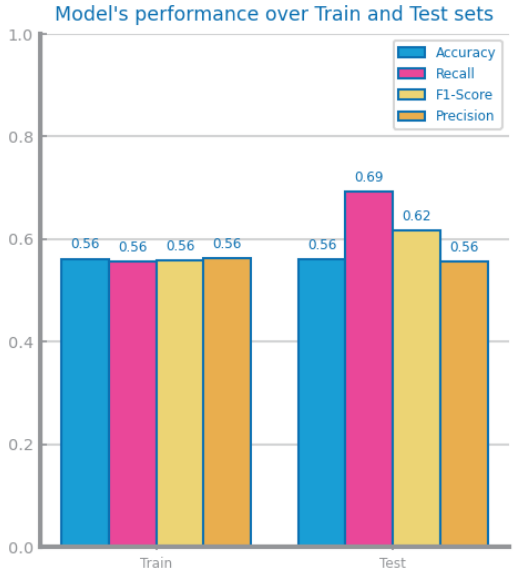








NB Best Results: Gaussian



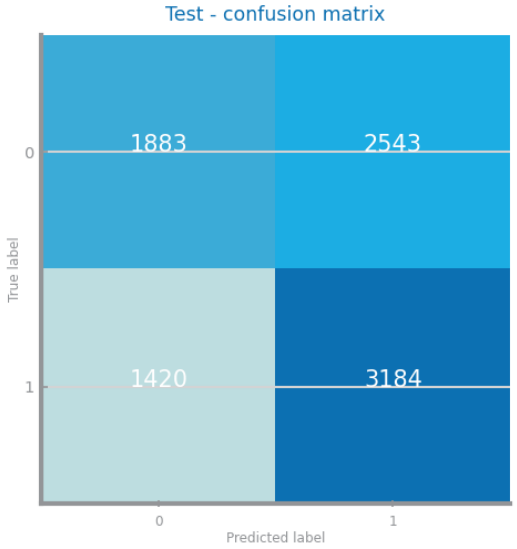


Figure 62: Balancing results for the second approach (SMOTE) for dataset 2

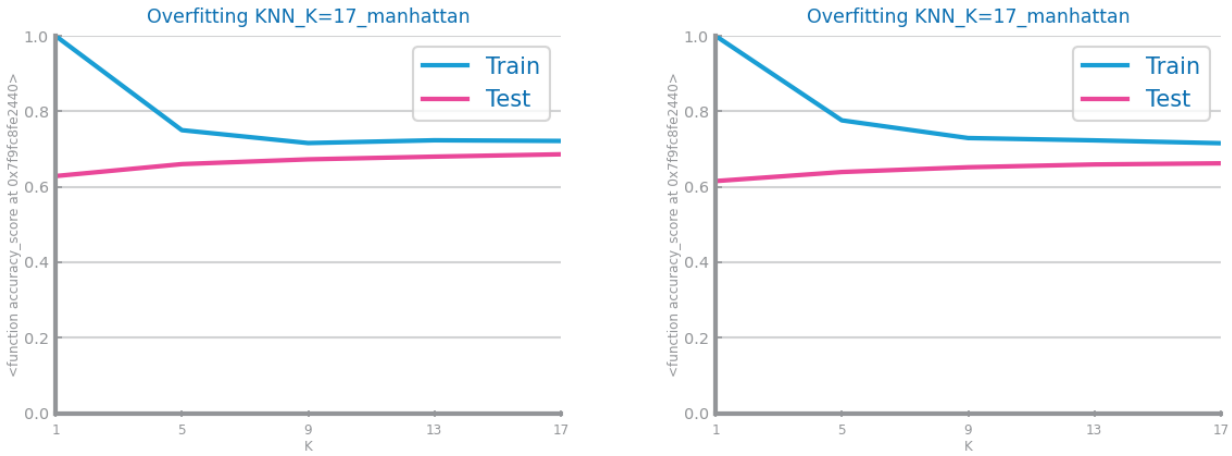
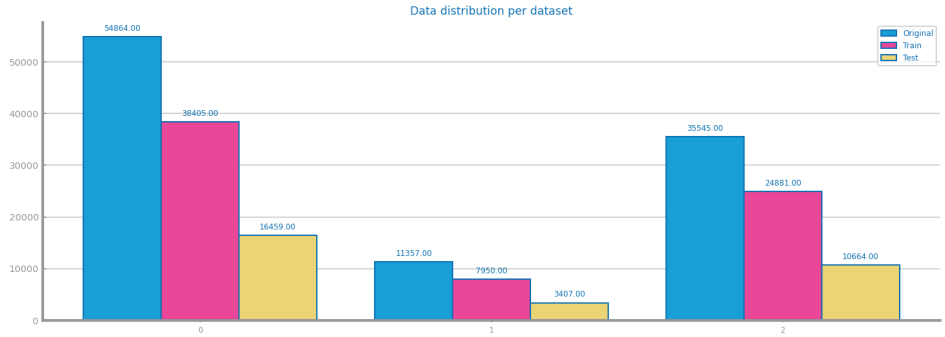
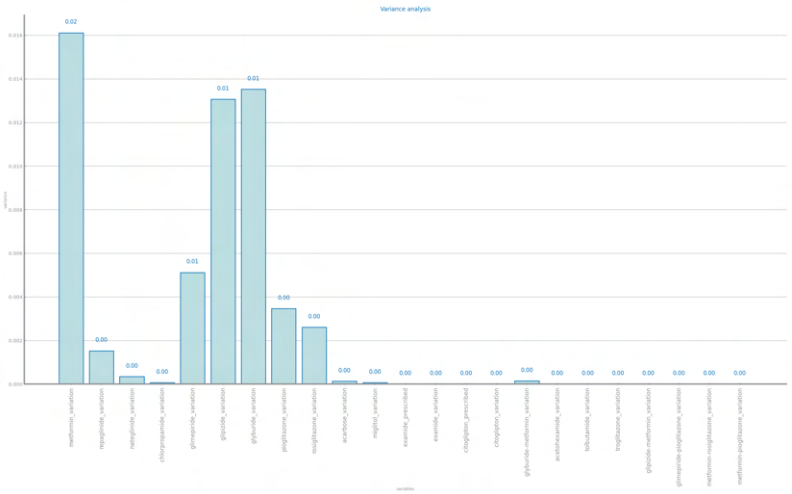


Figure 63: Overfitting results for the first approach (left) and second approach (right) for dataset 2

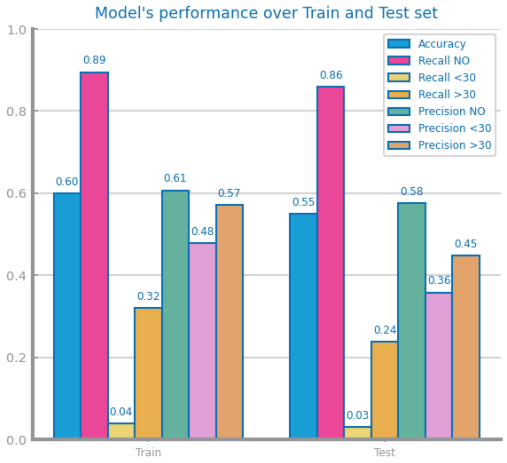
Feature Selection

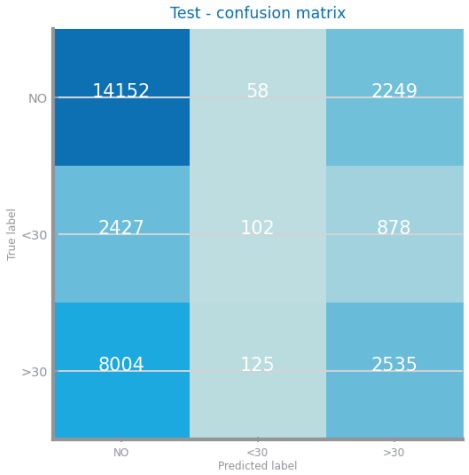
In 1 no vars are redundant; vars with <0.05VAR were irrelevant. We dropped all IDs as they can lead to overfitting. In 2 variables with >90%CORR are redundant; same criteria used for irrelevance.

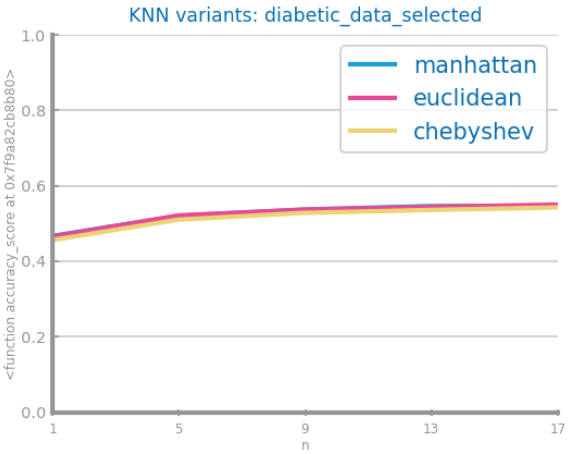


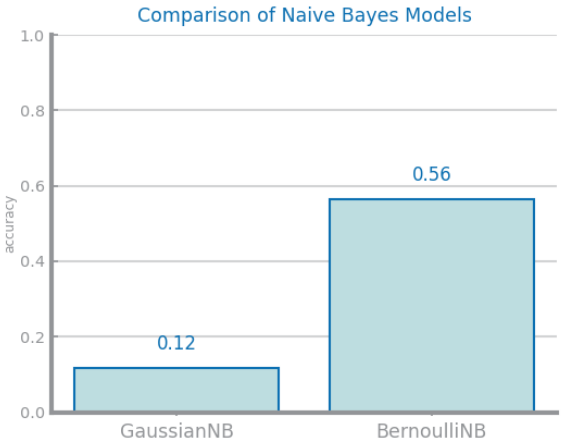


KNN Best Results: K=17 and distance=euclidean









NB Best Results: Bernoulli

Figure 65: Feature selection of redundant and irrelevant variables results with different parameters for dataset 1

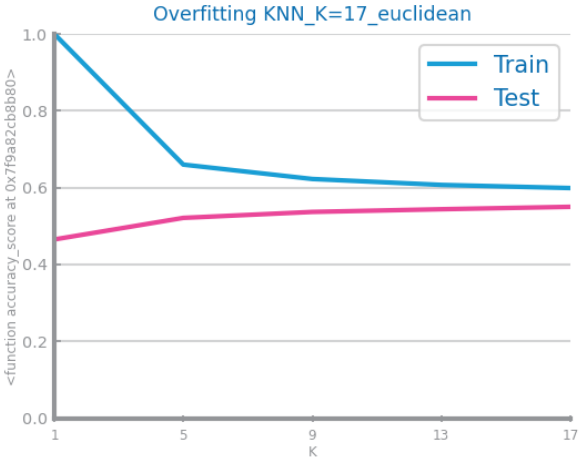
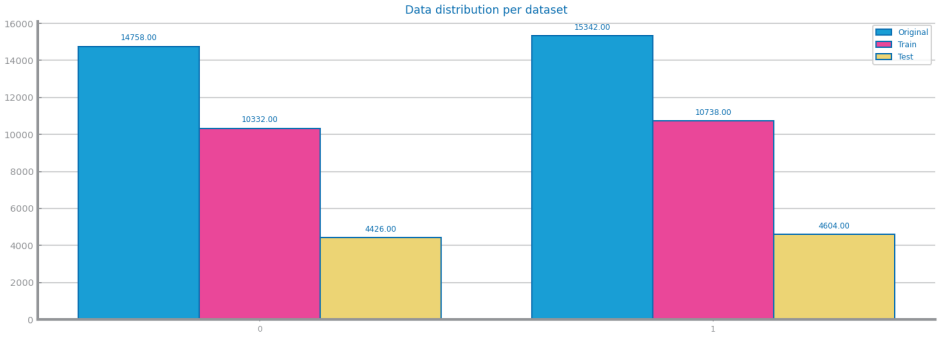
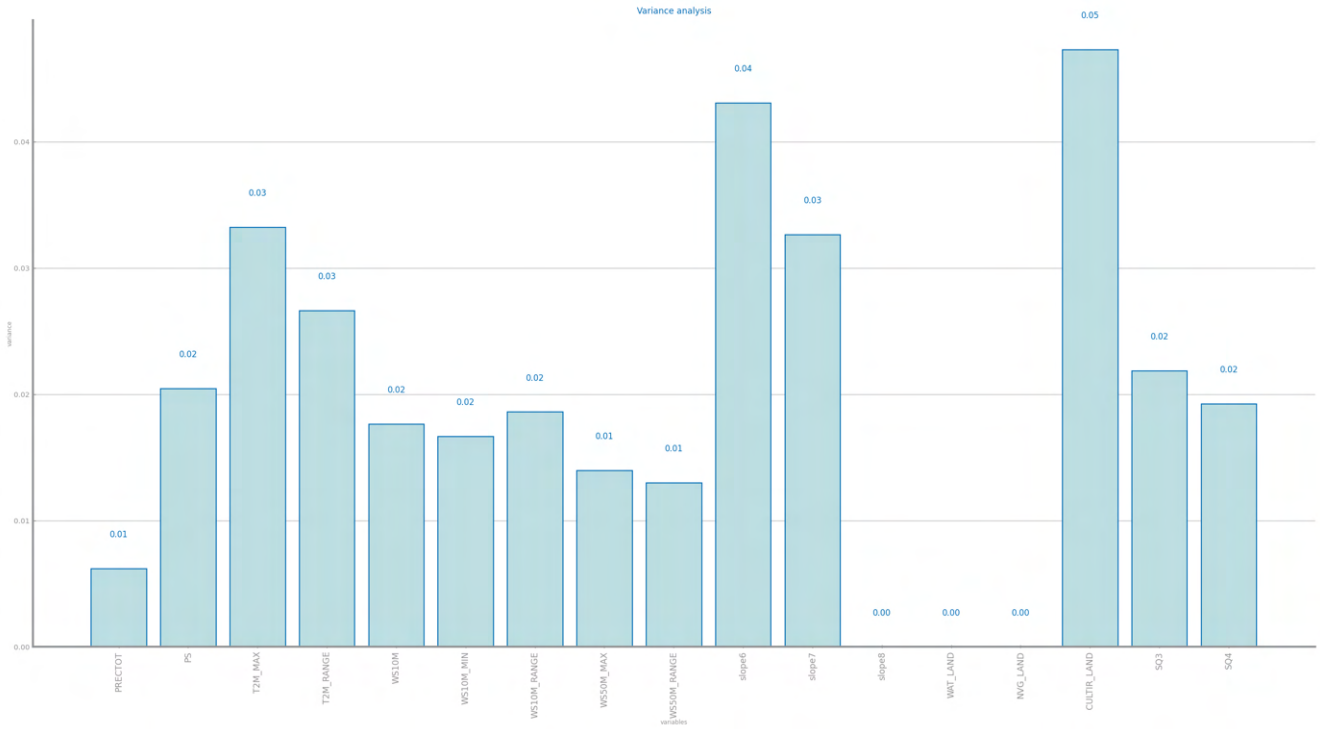


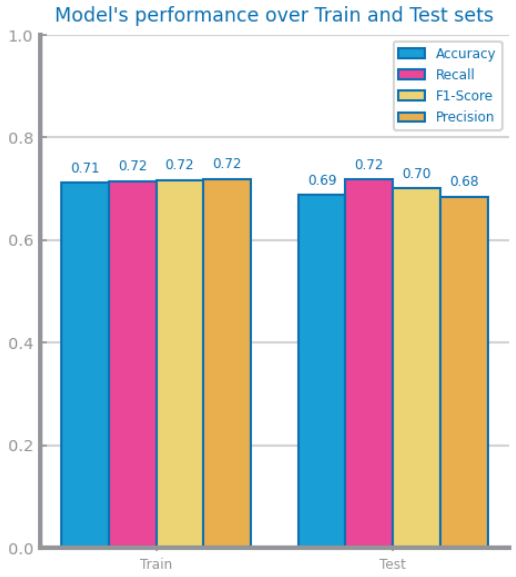
Figure 66: Overfitting results for dataset 1 without redundant and irrelevant variables

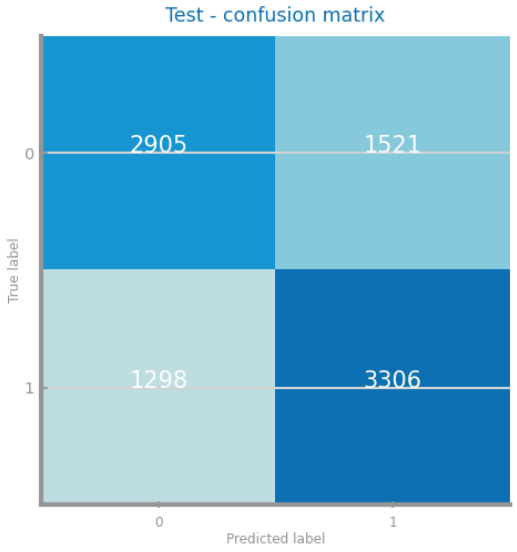
---

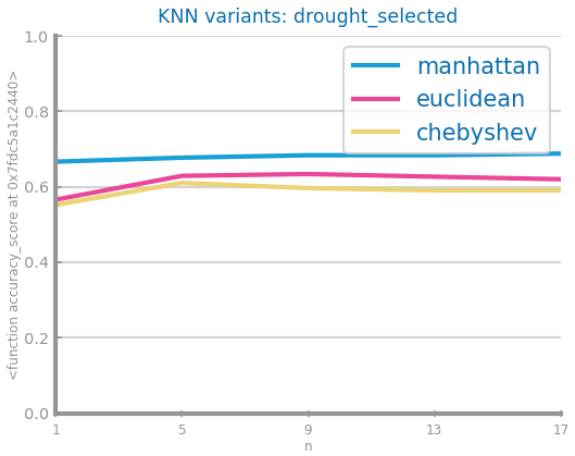


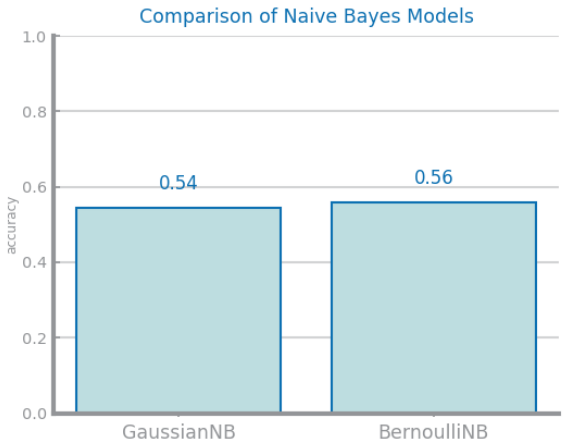


KNN Best Results: K=17 and distance=manhattan

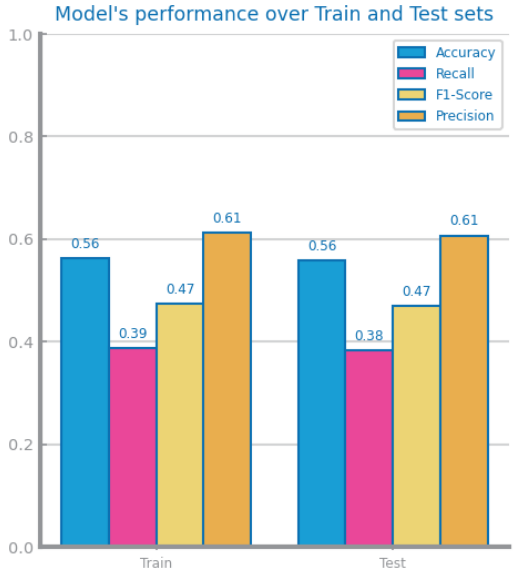








NB Best Results: Bernoulli



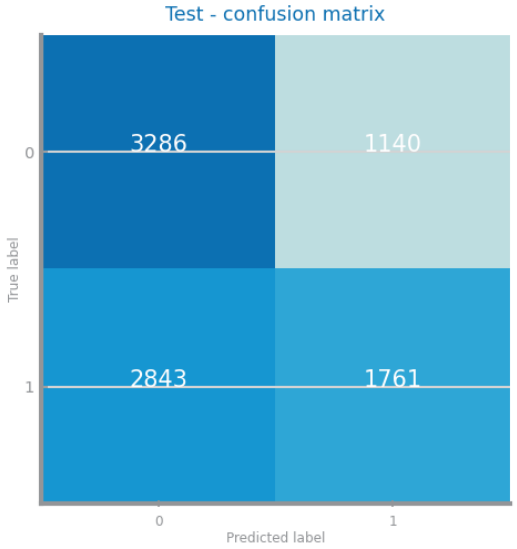


Figure 68: Feature selection of redundant and irrelevant variables results with different parameters for dataset 2

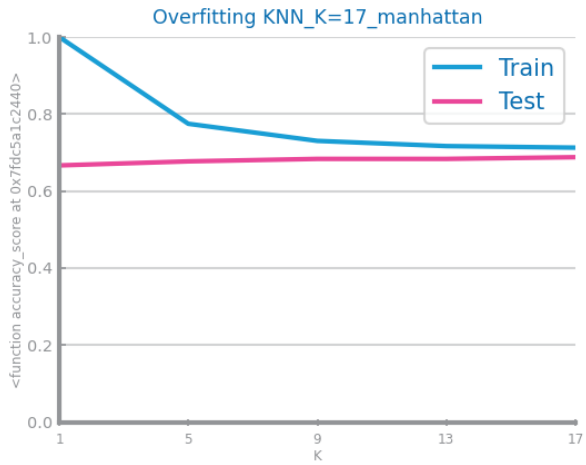
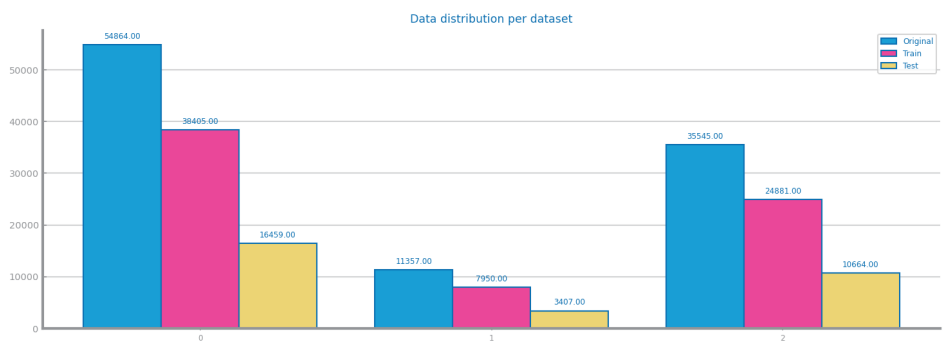
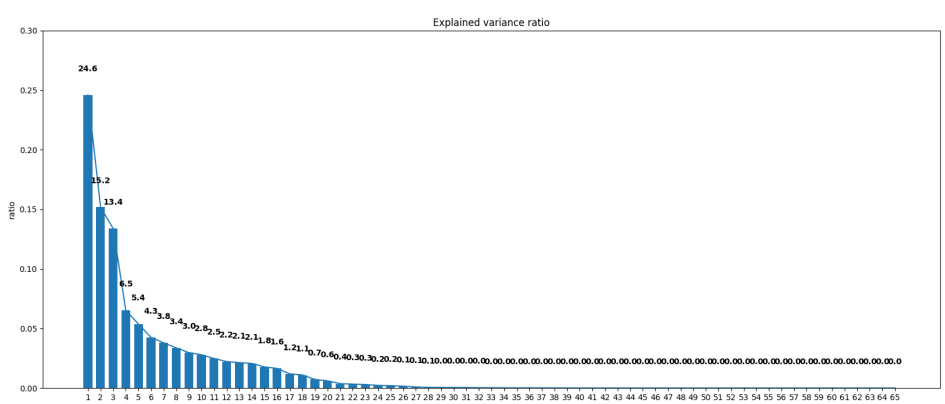


Figure 69: Overfitting results for dataset 2 without redundant and irrelevant variables

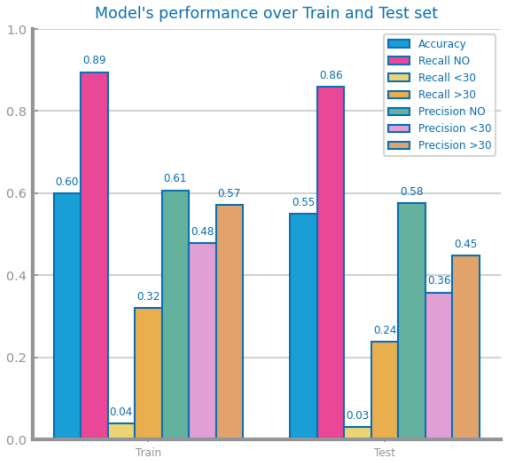
Feature Extraction

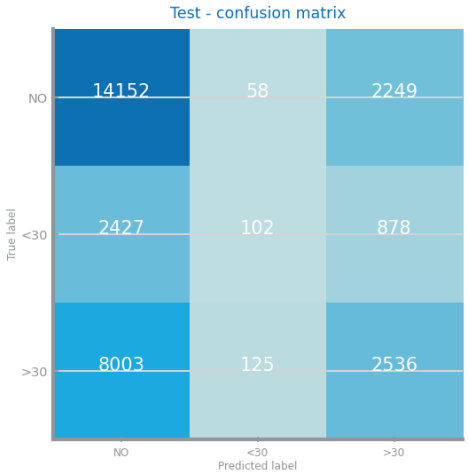
We only did evaluation with the resulting set in set 2 because it's more numerical in nature. It's better for interpretation that set 1 has the original vars.

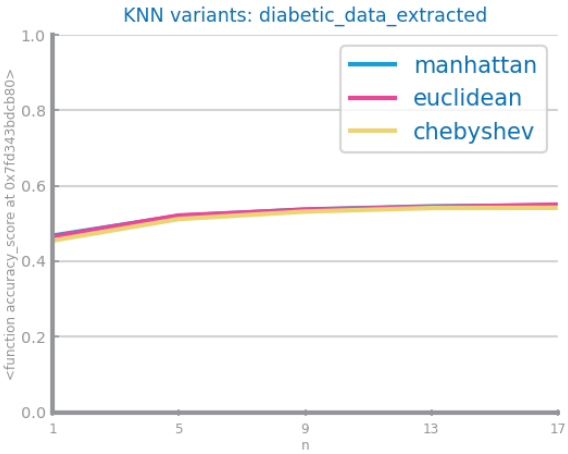


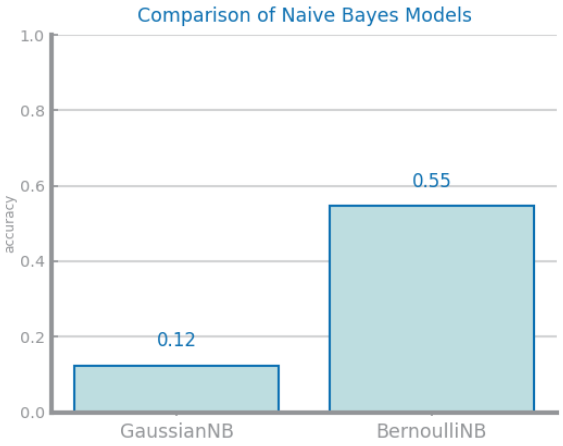


KNN Best Results: K=17 and distance=euclidean









NB Best Results: Bernoulli

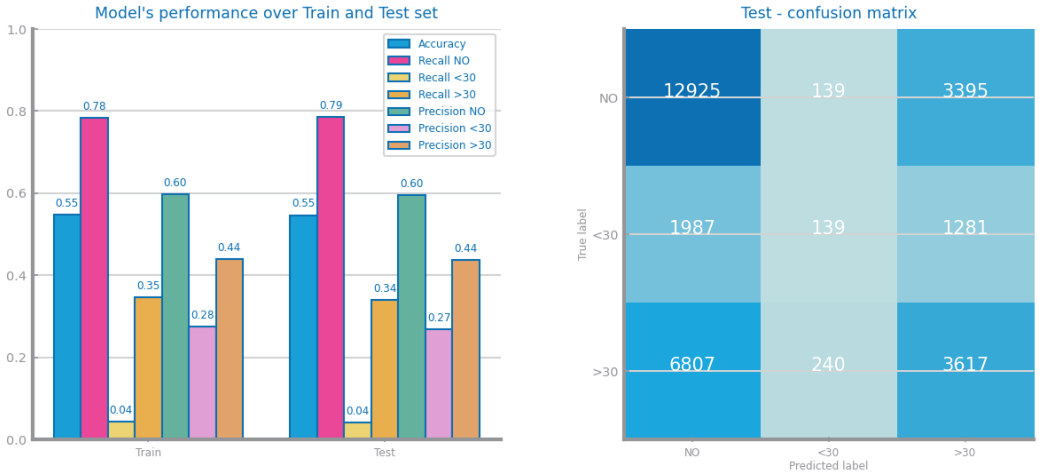


Figure 71: Principal components analysis and feature extraction results for dataset 1

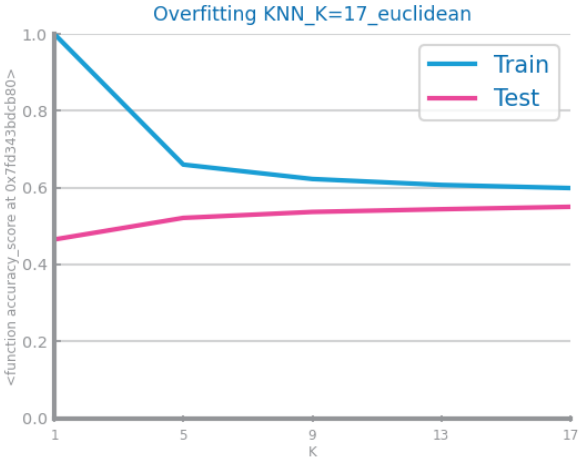
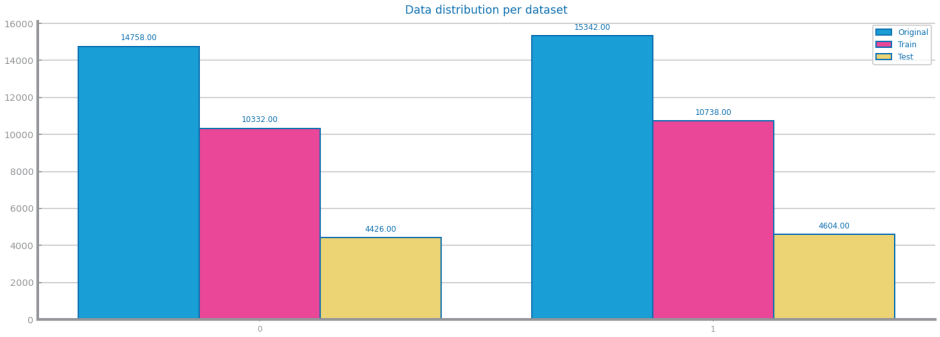
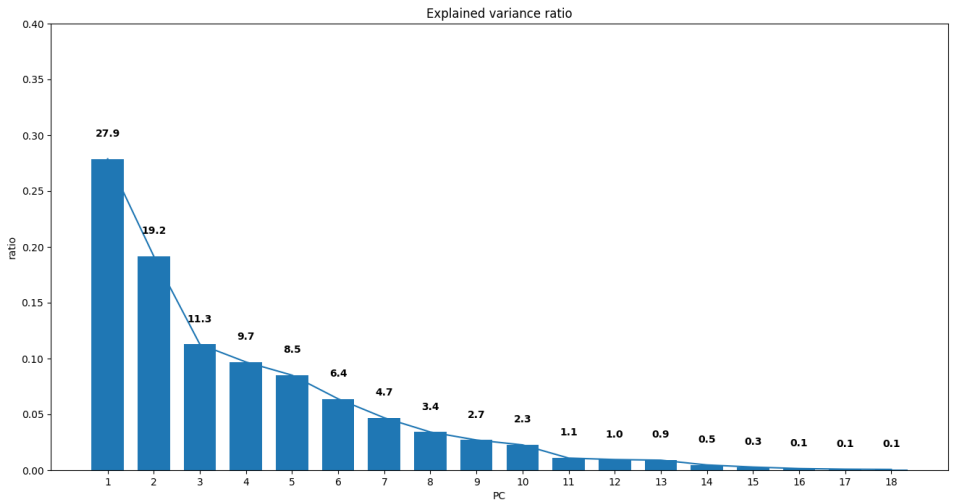
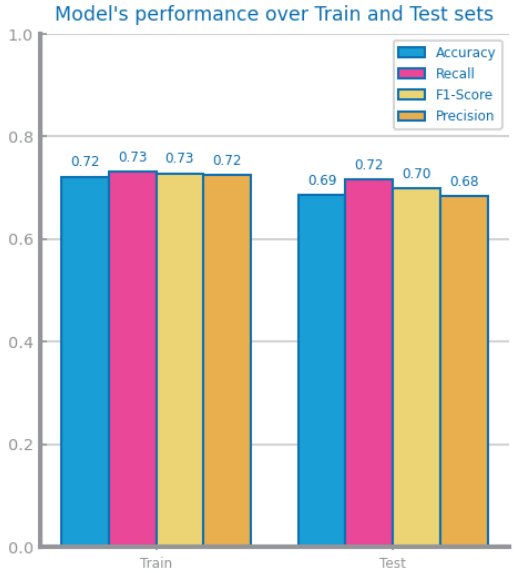


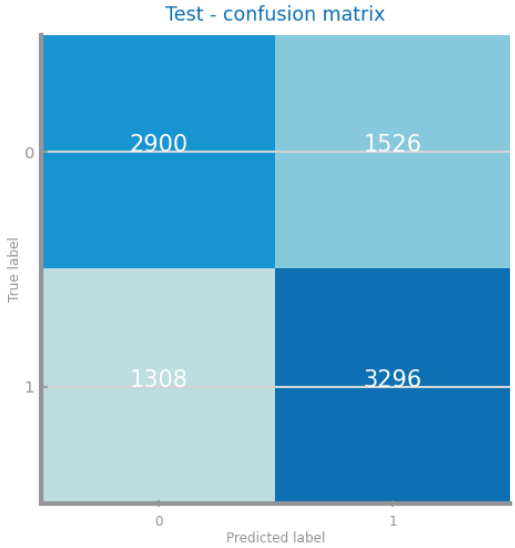
Figure 72: Principal components analysis and feature extraction overfitting for dataset 1

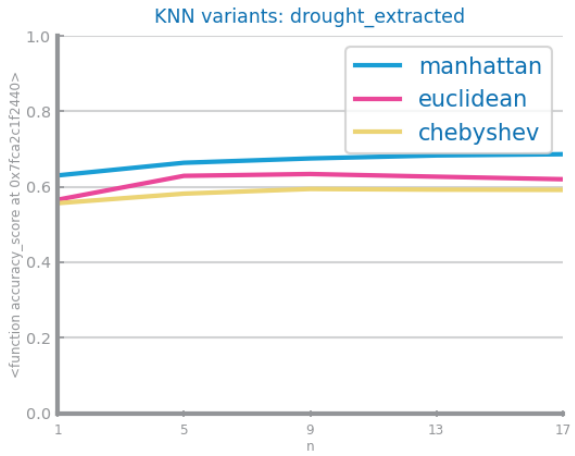


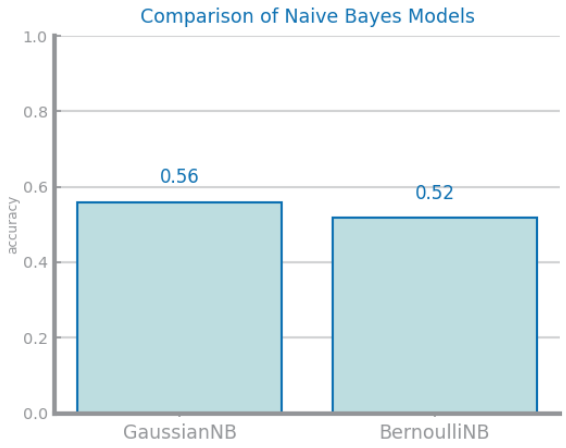


KNN Best Results: K=17 and distance=manhattan

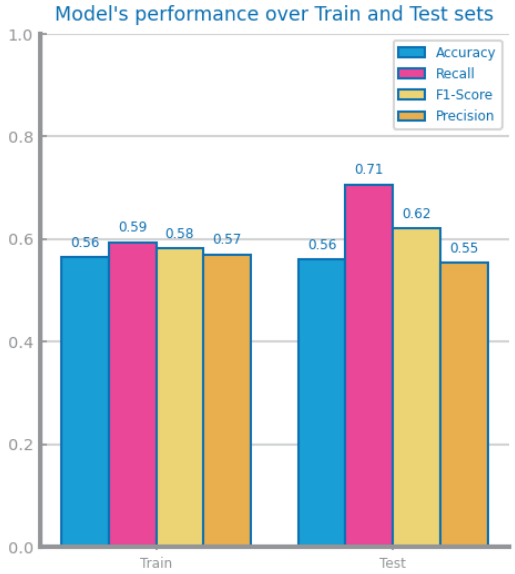








NB Best Results: Gaussian



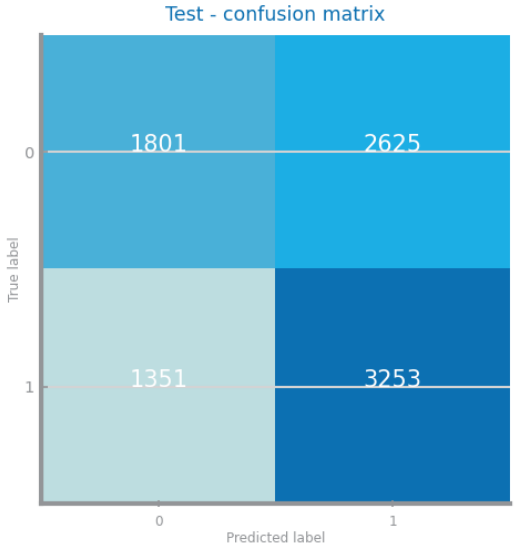


Figure 74: Principal components analysis and feature extraction results for dataset 2

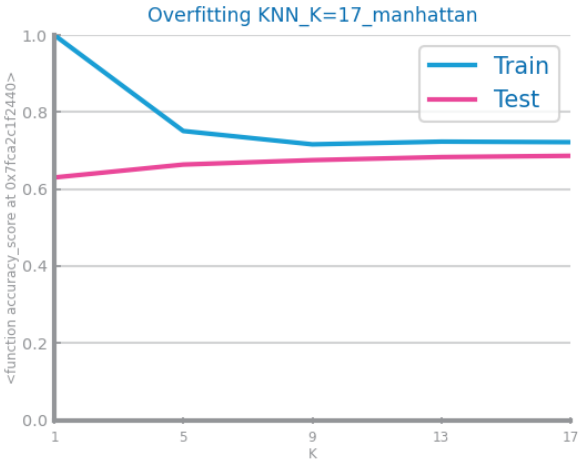
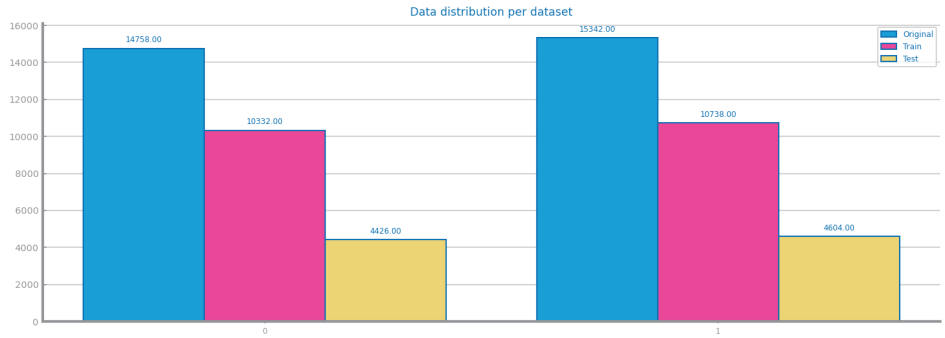


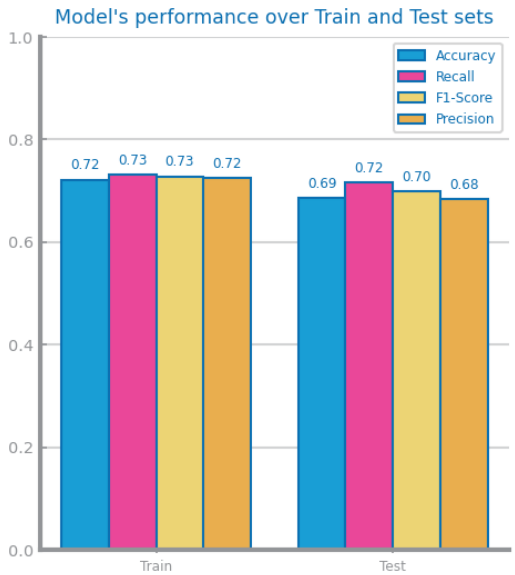
Figure 75: Principal components analysis and feature extraction overfitting for dataset 2

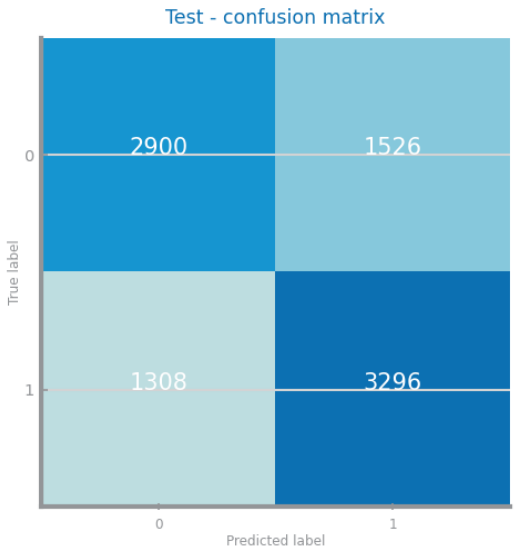
Feature Generation

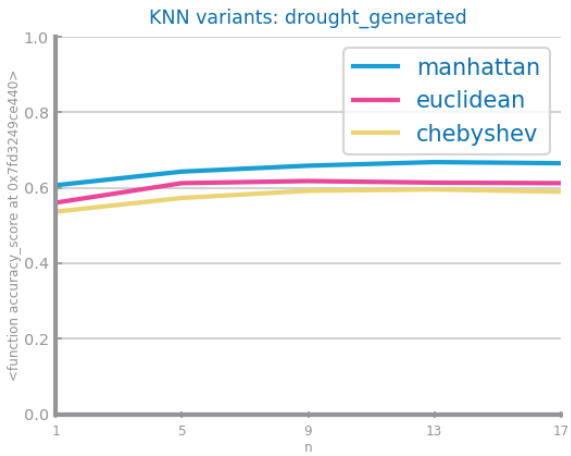
In set 2, we decided to generate a new variable *season* using the date column since it could impact the droughts. This method however did not improve the results so we discarded it.

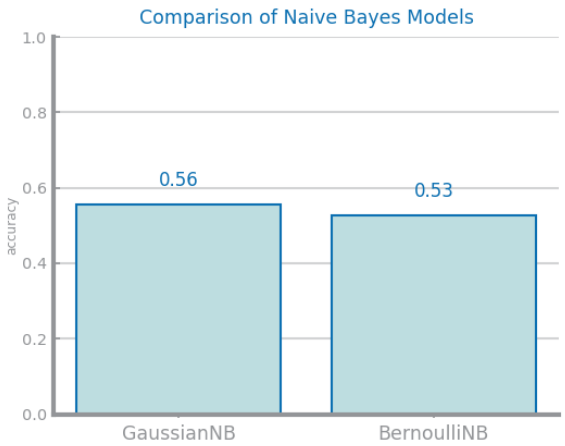


KNN Best Results: K=17 and distance=manhattan









NB Best Results: Gaussian

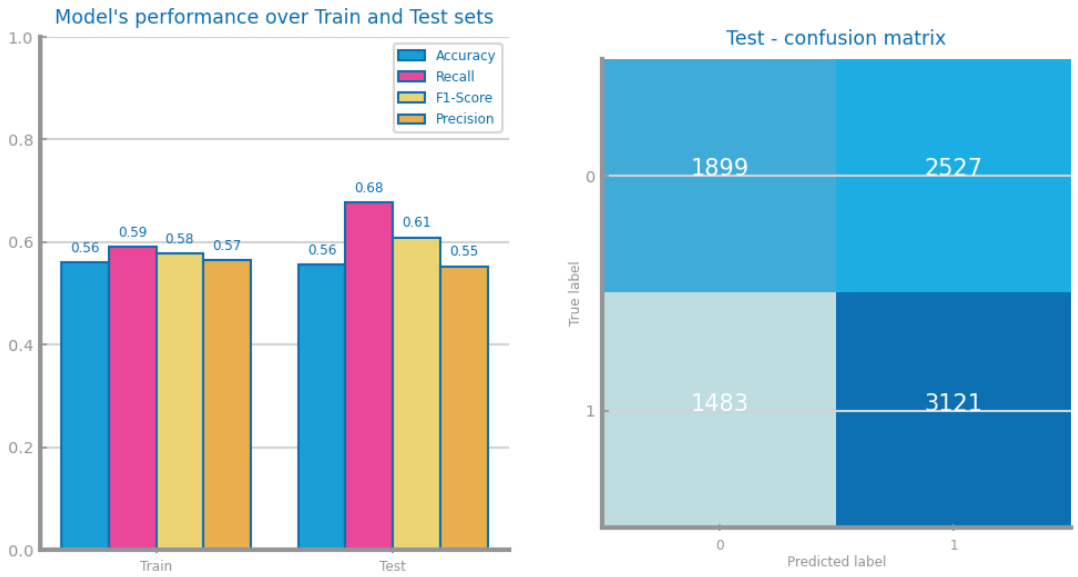


Figure 77: Feature generation results for dataset 2

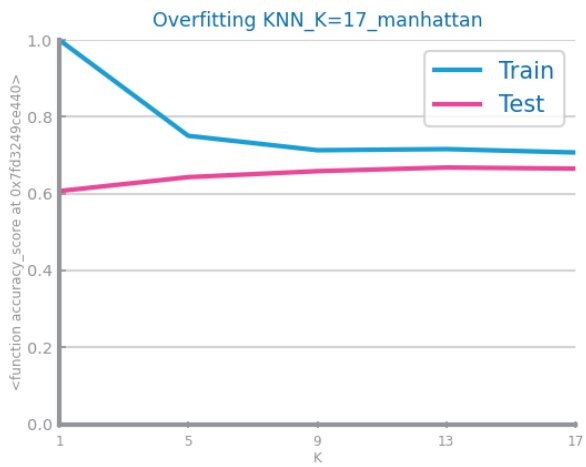


Figure 78: Feature generation overfitting for dataset 2

3 MODELS' EVALUATION

In set 1, to better evaluate the models and compare results, we decided to split the data always using the same seed. In set 2, the data was split chronologically, i.e. using the older data to train and the most recent data to test.

Naïve Bayes

Categorical and Multinomial NB were not used on the data that had negative and numerical values. We can also observe that Naïve Bayes does not have good recall and precision results for the important classes hence it is not able to predict them efficiently.

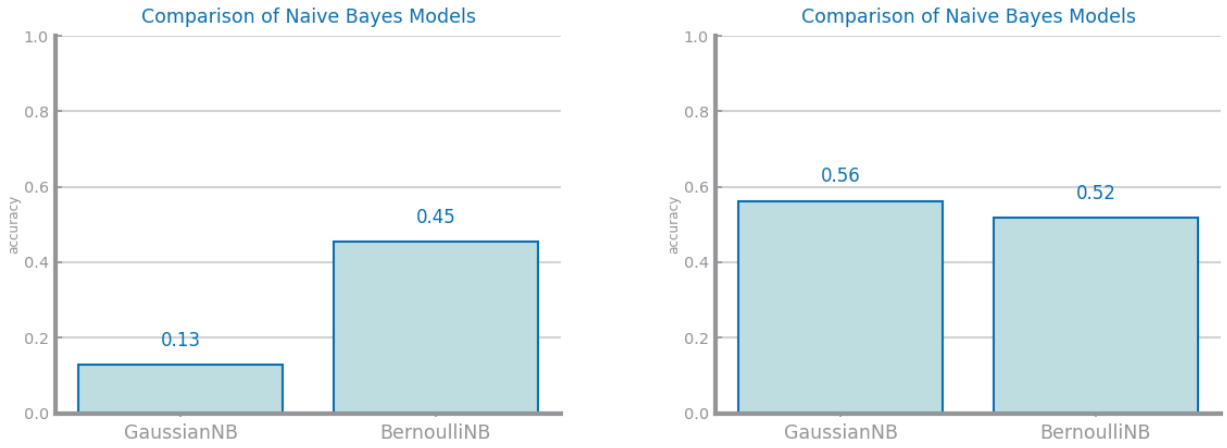


Figure 79: Naïve Bayes alternatives comparison for dataset 1 (left) and dataset 2 (right)

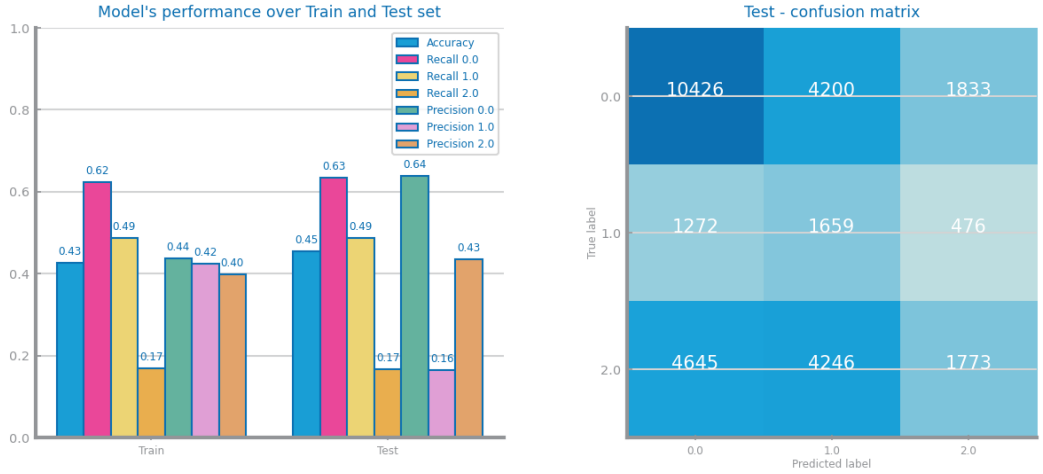


Figure 80: Naïve Bayes best model results for dataset 1

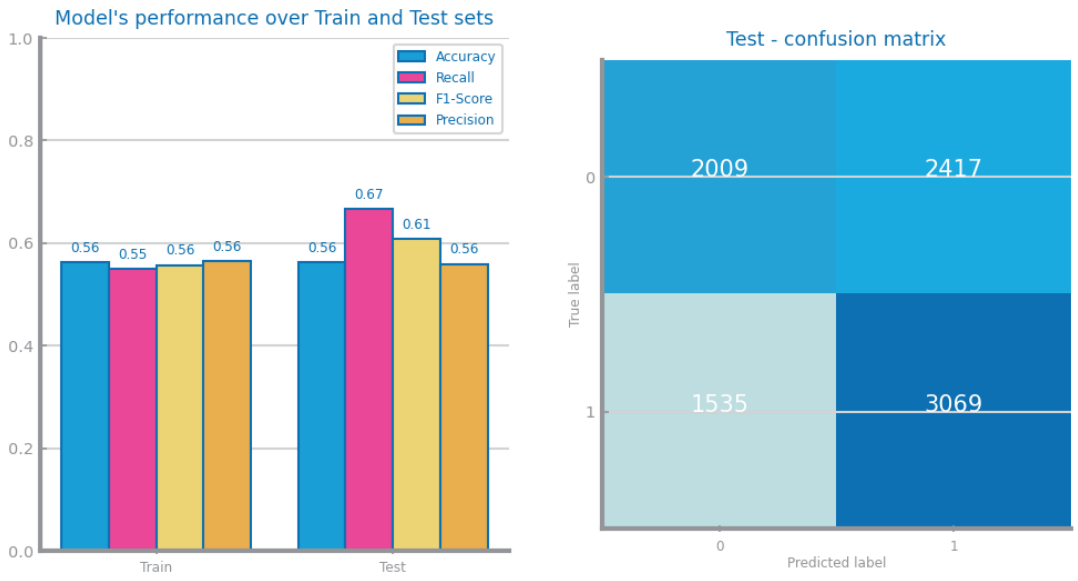


Figure 81: Naïve Bayes best model results for dataset 2

KNN

In 1, we chose K=174 as it's the square root of the nº of records. In 2, the best K is near the value where the trends for train and test start decreasing (1̃5). The models overfit for lower values of k: where the train rises and the test drops (if we read the chart R-L).

The results show a better recall and precision of the important classes on set 1 and overall good performance on set 2.

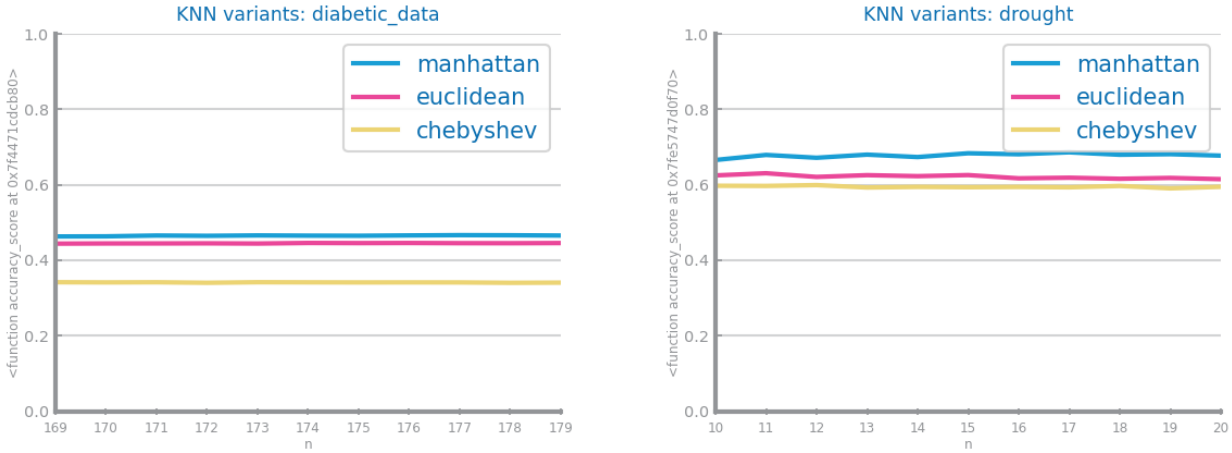


Figure 82: KNN different parameterizations comparison for dataset 1 (left) and dataset 2 (right)

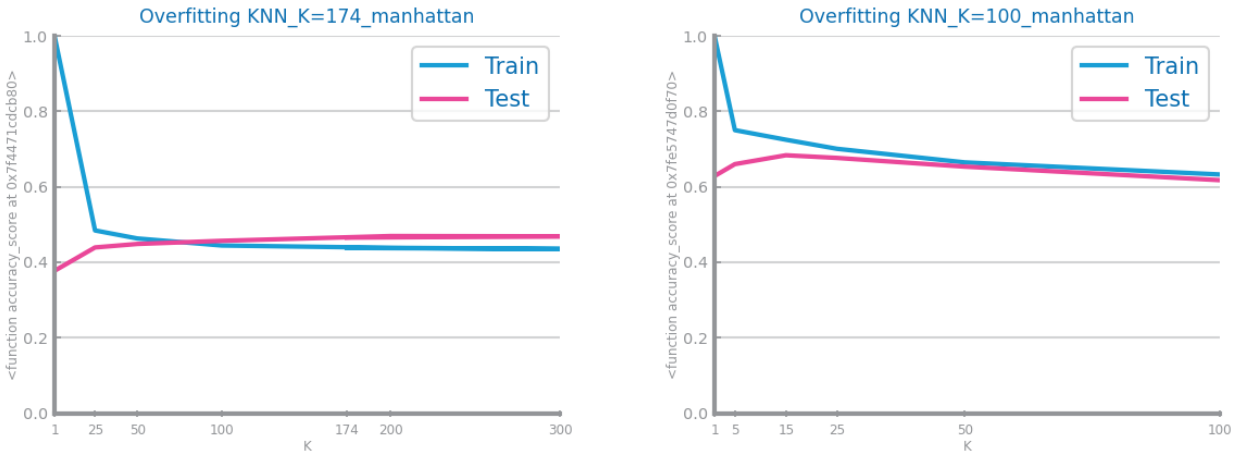


Figure 83: KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

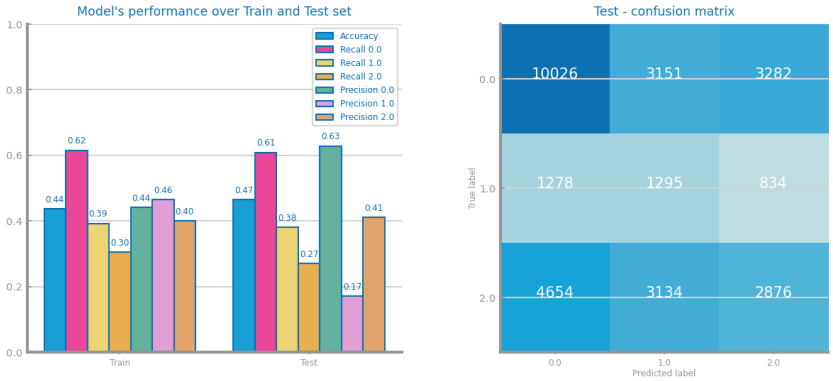


Figure 84: KNN best model results for dataset 1

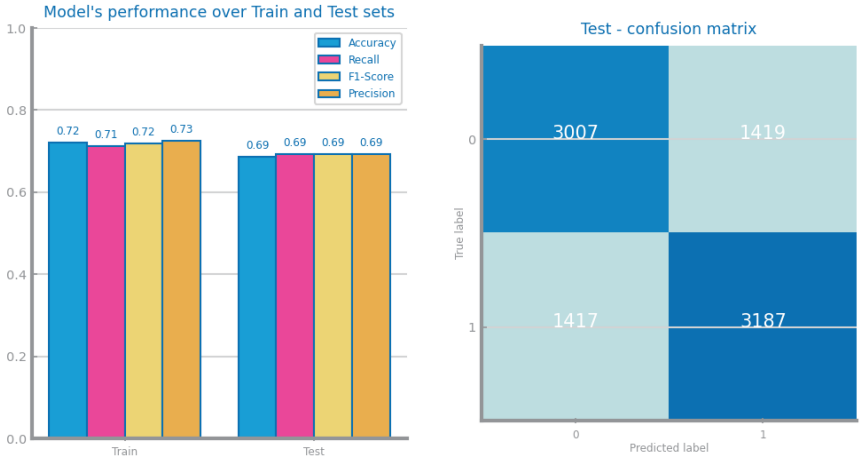


Figure 85: KNN best model results for dataset 2

Decision Trees

In set 1, we're always overfitting as trends are always opposite (max_depth => more overfitting); the performance was bad. There's a slight improvement in accuracy and recall of class 2 but recall of class 1 got worse (comparing to KNN). In set 2, after max_depth=20 the model loses generalization ability. The results were much better than KNN, as it is shown.

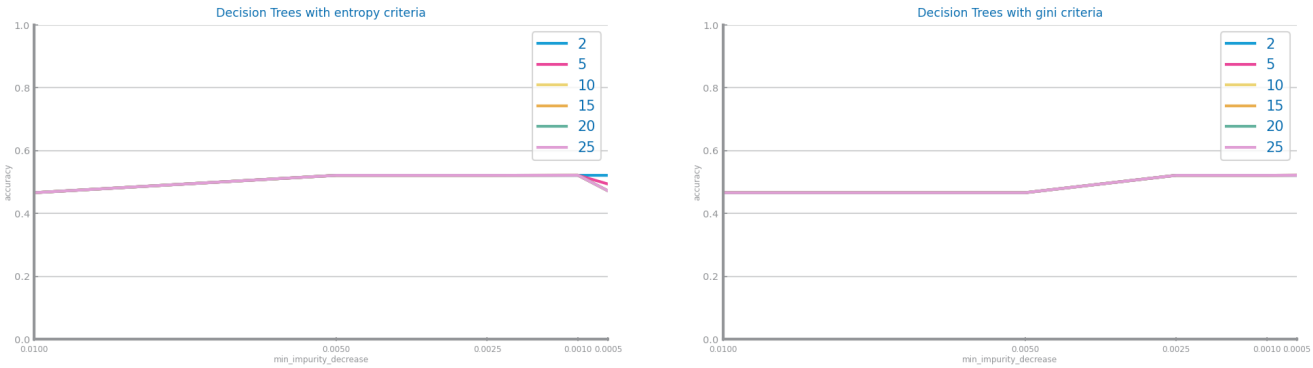


Figure 86: Decision Trees different parameterizations comparison for dataset 1

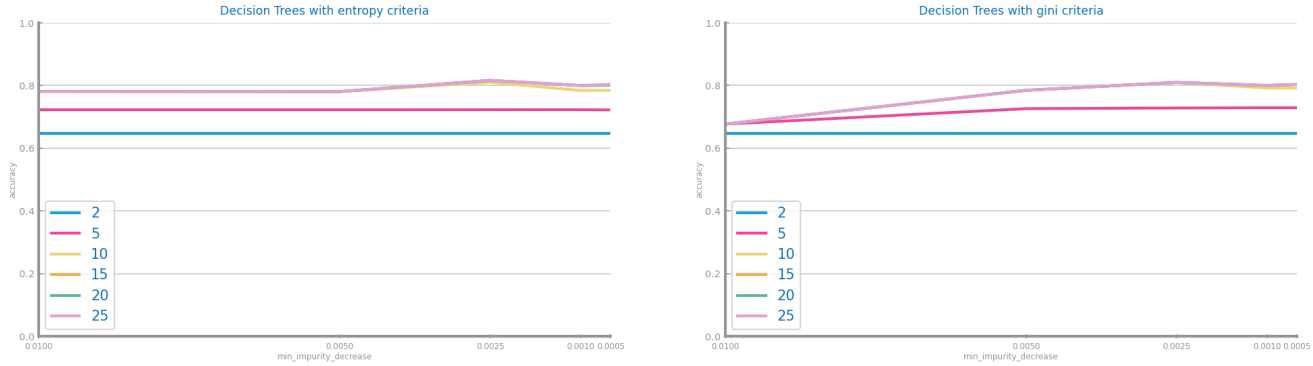


Figure 87: Decision Trees different parameterizations comparison for dataset 2

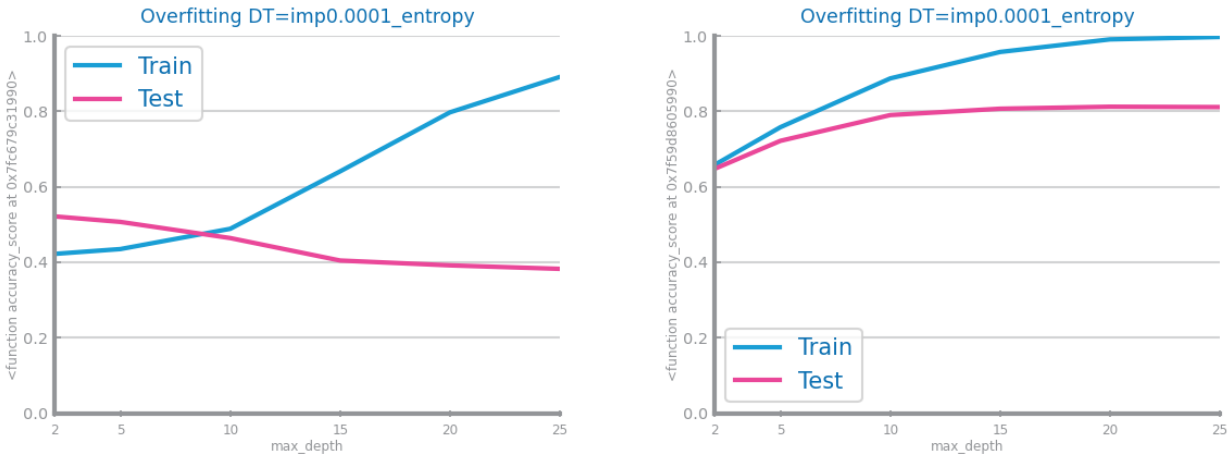


Figure 88: Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

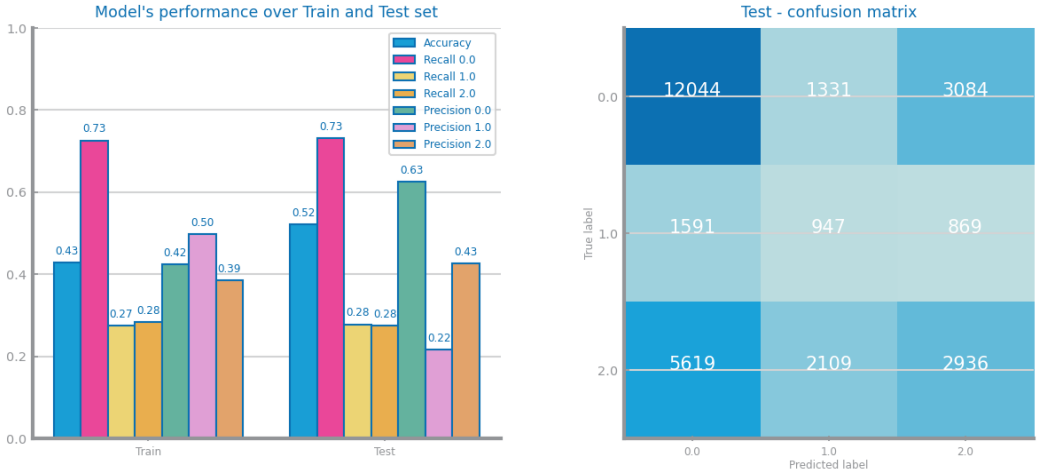


Figure 89: Decision trees best model results for dataset 1

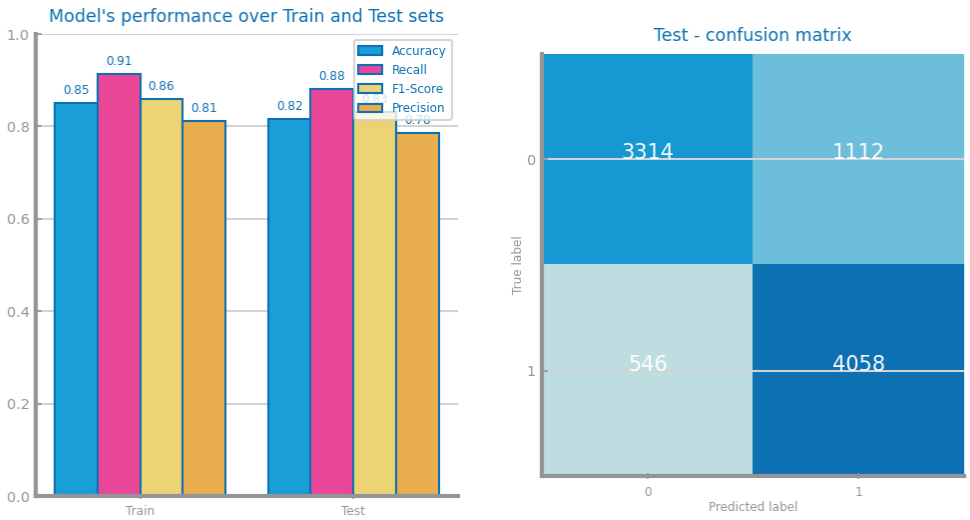


Figure 90: Decision trees best model results for dataset 2

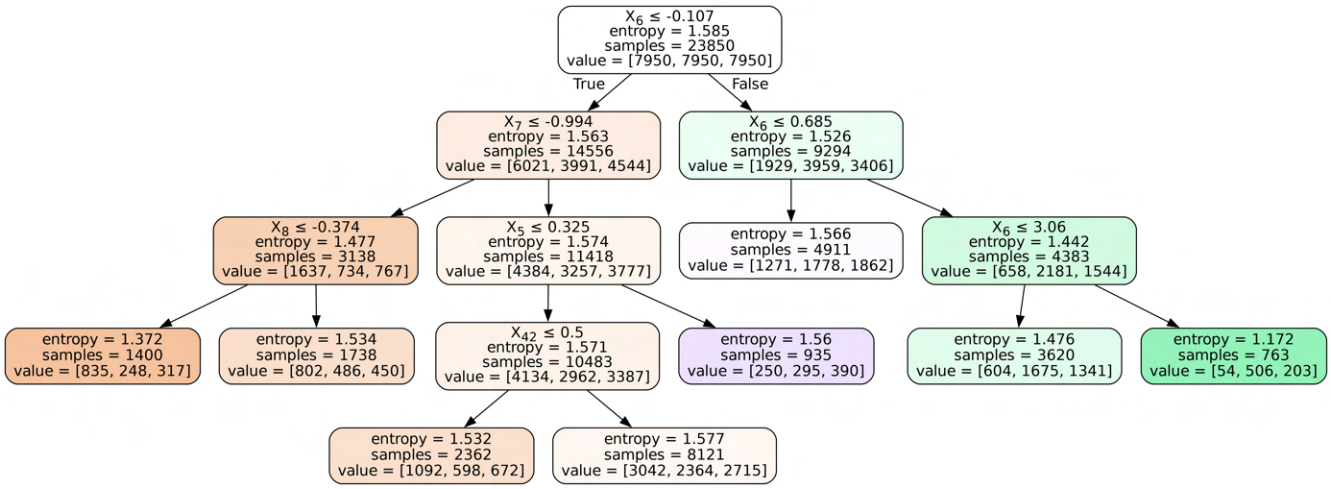


Figure 91: Best tree for dataset 1

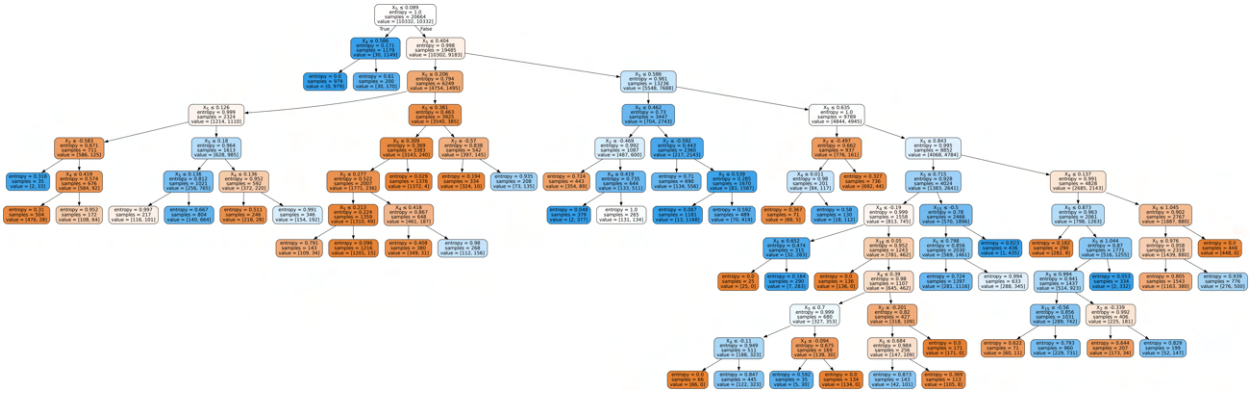


Figure 92: Best tree for dataset 2

Random Forests

In set 1, max_depths of 5 and 10 were slightly better. In set 2, 25 was better. Because ensembles are better than individual models (as they combine their strengths), it improved overall results in both datasets. Exhibits no overfitting, which isn't a surprise since combining diverse models allows for better generalization.

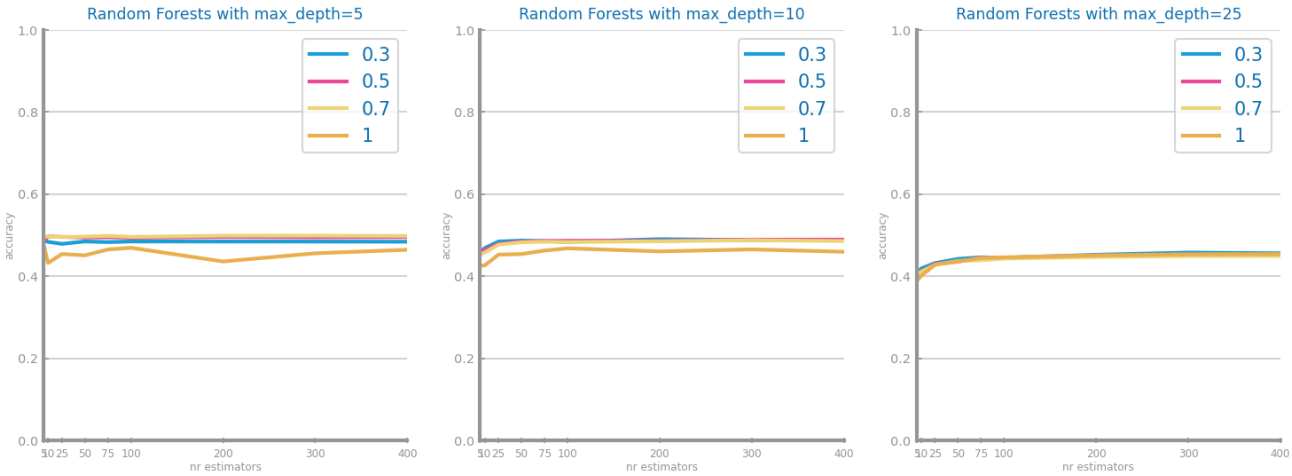


Figure 93: Random Forests different parameterizations comparison for dataset 1

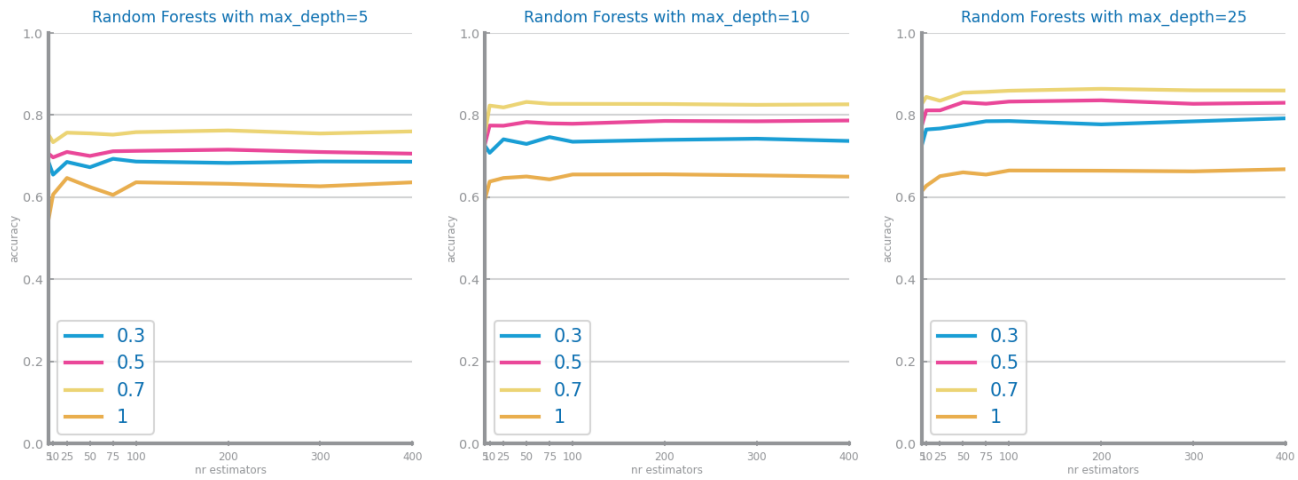


Figure 94: Random Forests different parameterizations comparison for dataset 2

Figure 95: Random Forests overfitting analysis for dataset 1 (left) and dataset 2 (right)

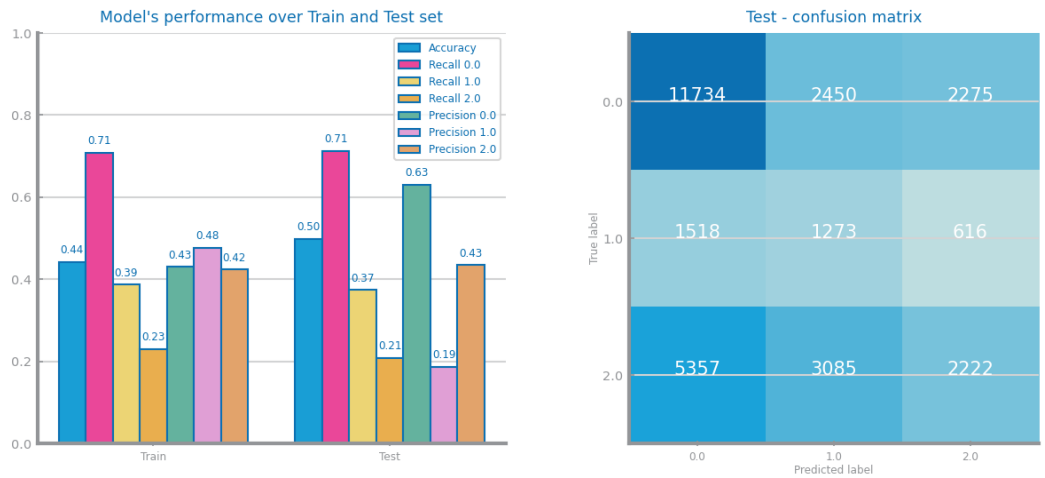


Figure 96: Random Forests best model results for dataset 1

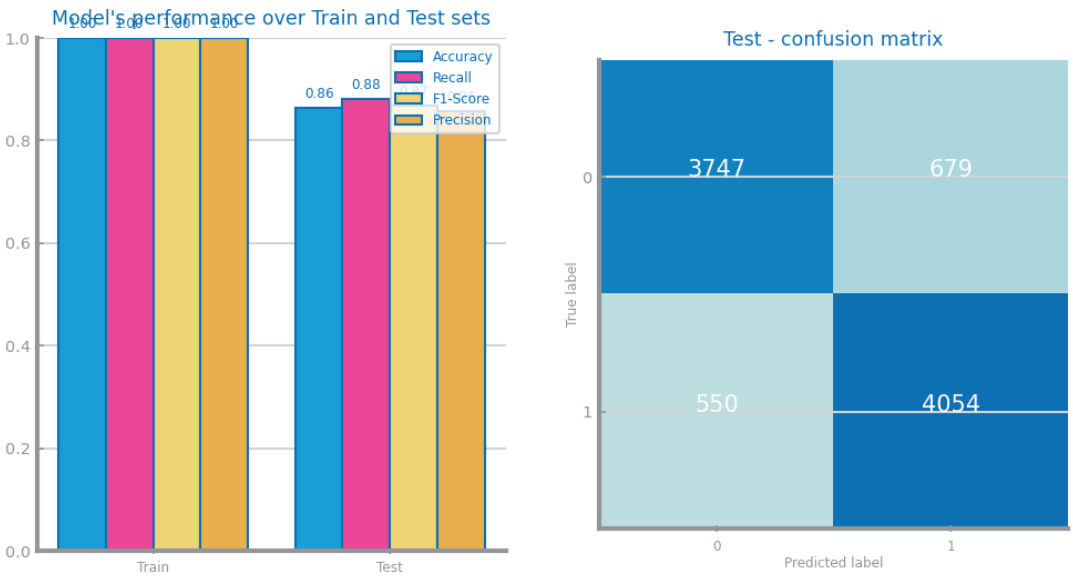


Figure 97: Random Forests best model results for dataset 2

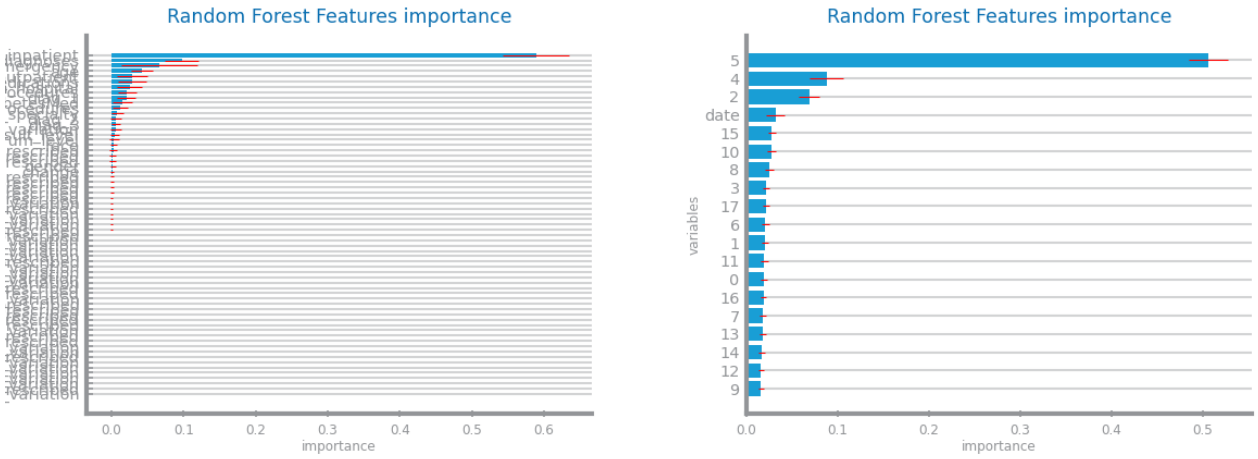


Figure 98: Random Forests variables importance for dataset 1 (left) and dataset 2 (right)

Gradient Boosting

In set 1, max_depth=5 has a slightly better accuracy for a few number of estimators however, with the increase of the number of estimators, the accuracies are similiar. In set 2, max_depth=10 had better results. In the overfitting plots, it is shown that the accuracy slightly oscilates with the number of estimators. It has improved overall results in both datasets, in comparission to individual models.

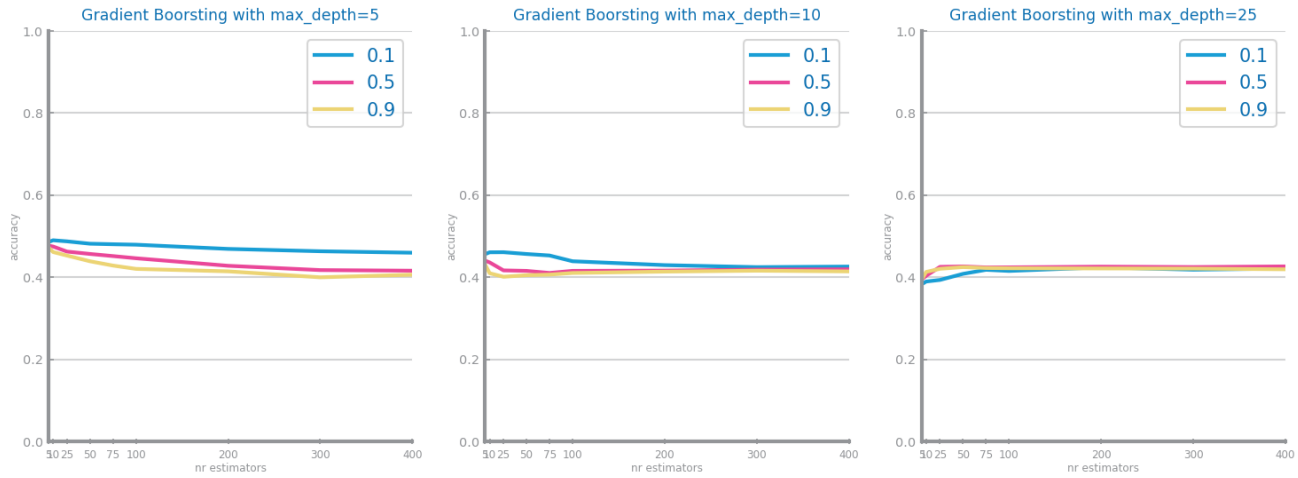


Figure 99: Gradient boosting different parameterizations comparison for dataset 1

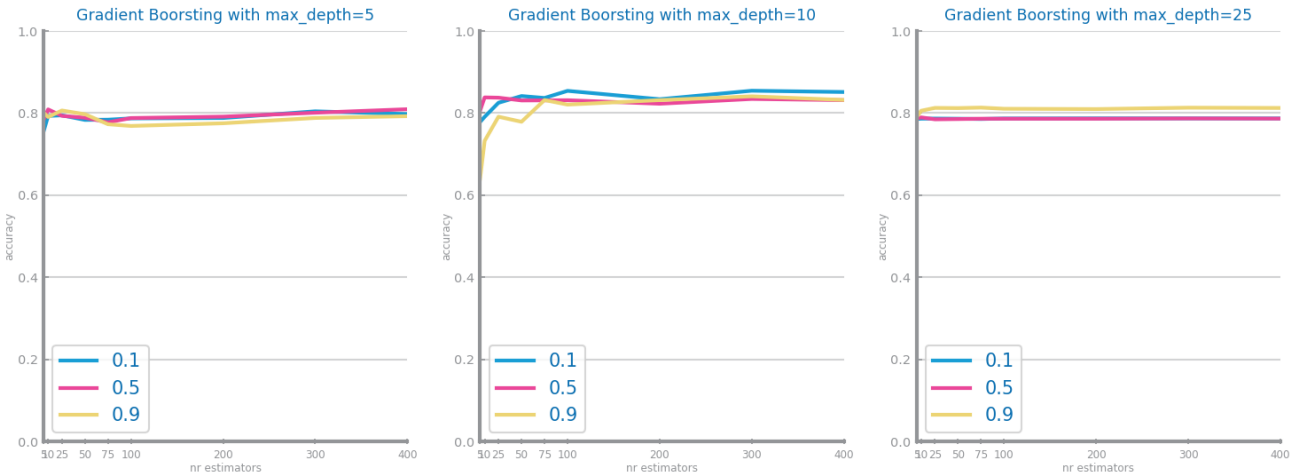


Figure 100: Gradient boosting different parameterizations comparison for dataset 2

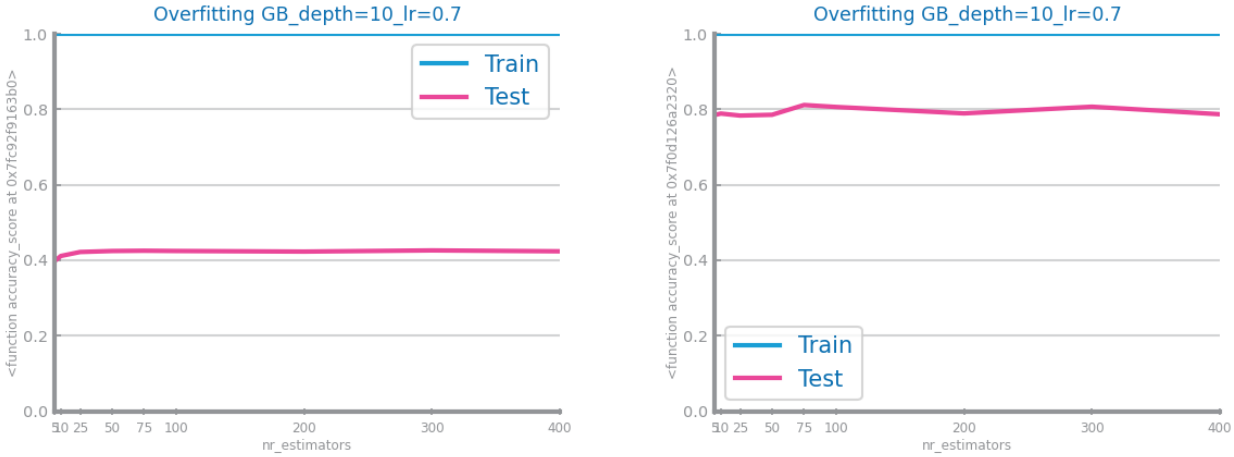


Figure 101: Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

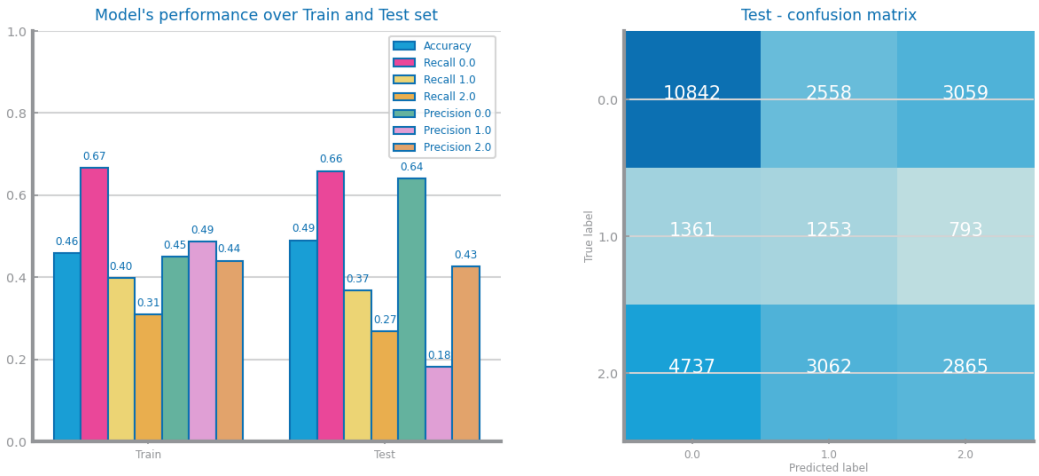


Figure 102: Gradient boosting best model results for dataset 1

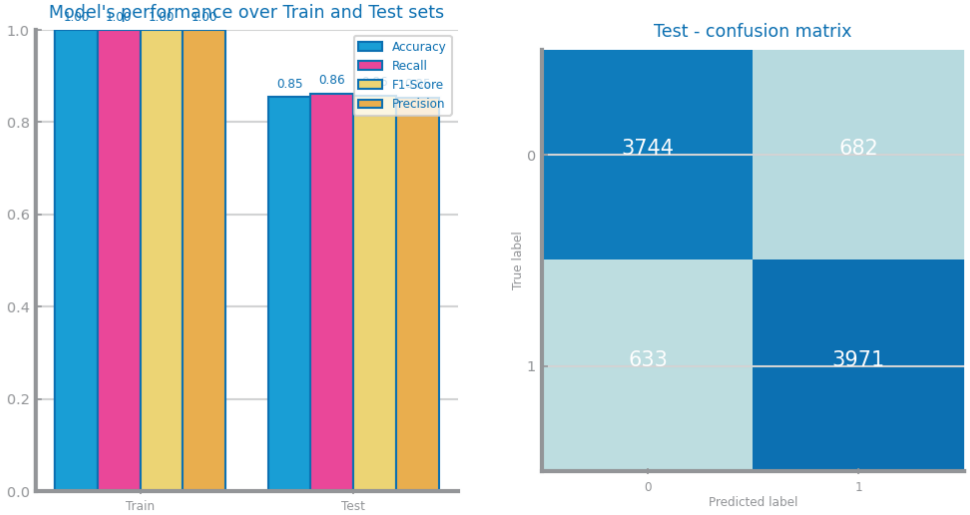


Figure 103: Gradient boosting best model results for dataset 2

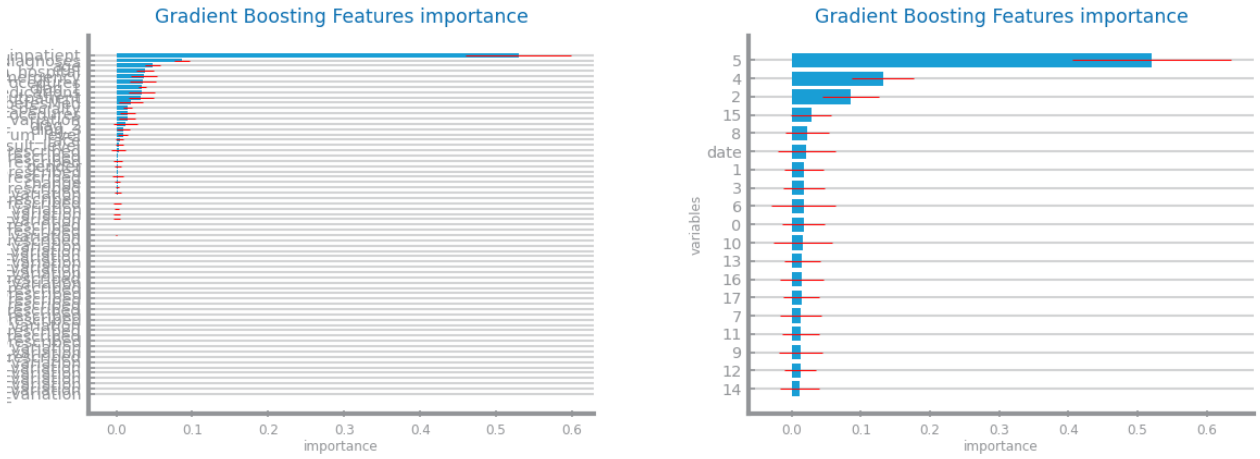
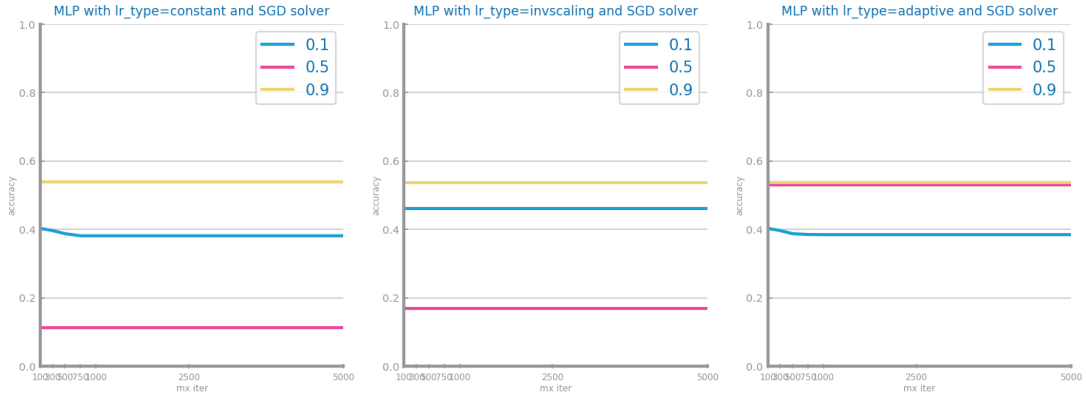


Figure 104: Gradient boosting variables importance for dataset 1 (left) and dataset 2 (right)

Multi-Layer Perceptrons

We used different solvers for each dataset because Adam is better than SGD for large datasets (which is the case of set 1). We combined Adam with warm start. Results were constant, probably due to having constant (and large) learning rates: the model could have been jumping from local minima to local minima without finding a better parametrization. In set 1, the model stagnated (loss became constant) and couldn't learn any more. In 2, it kept learning, although in an unregular way.



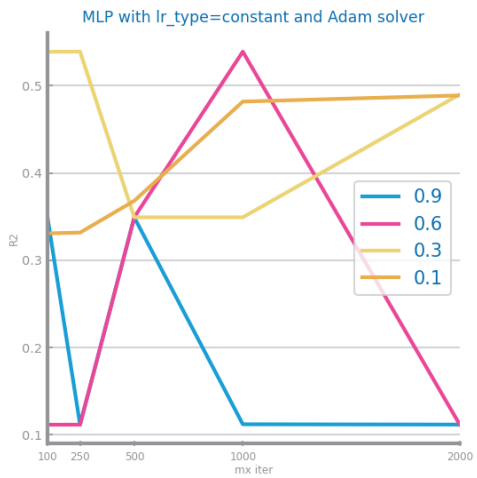
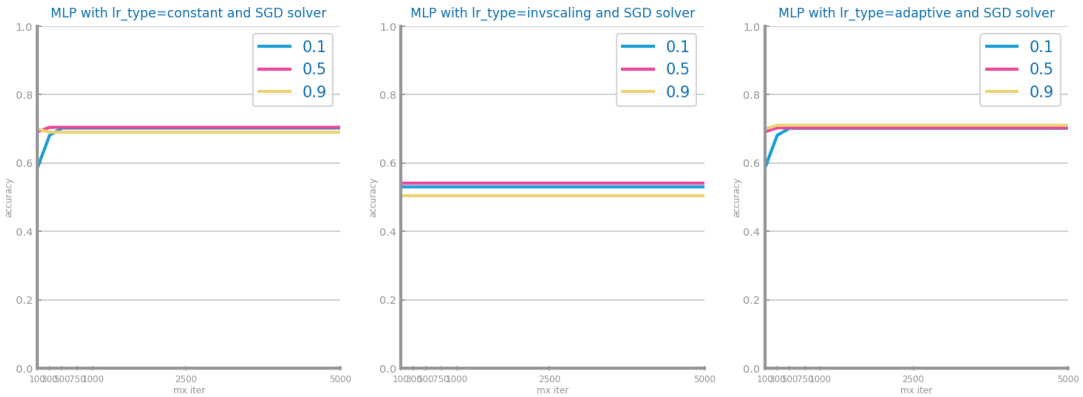


Figure 105: MLP different parameterizations comparison for dataset 1



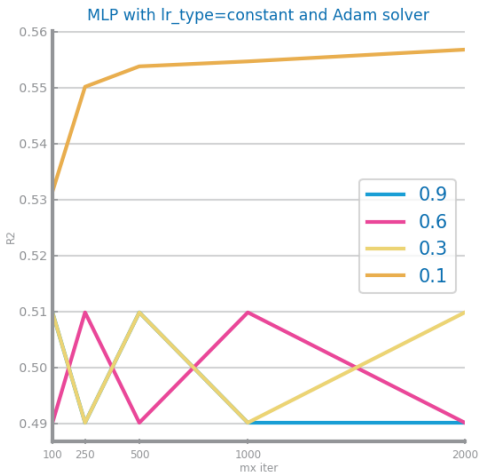


Figure 106: MLP different parameterizations comparison for dataset 2

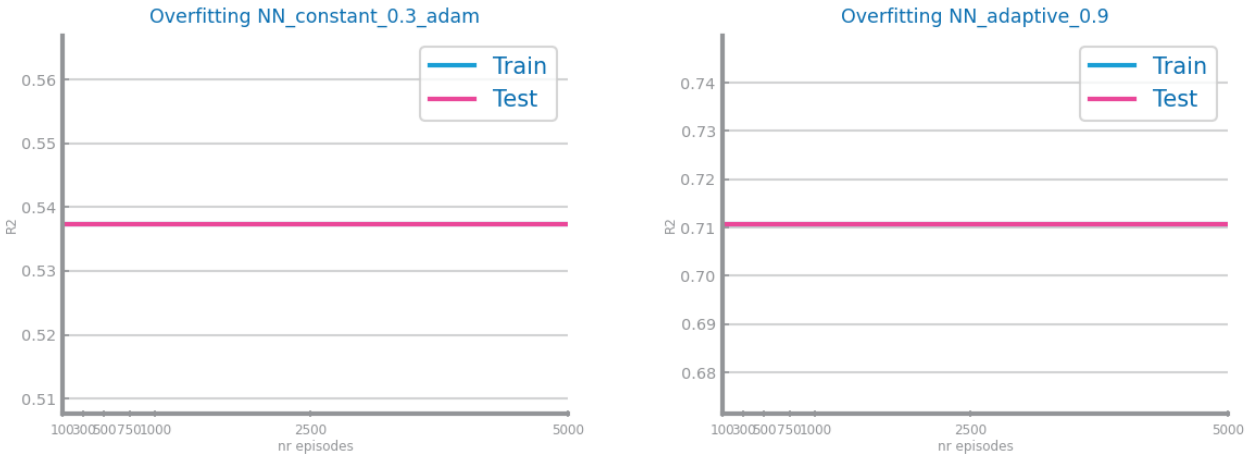


Figure 107: MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

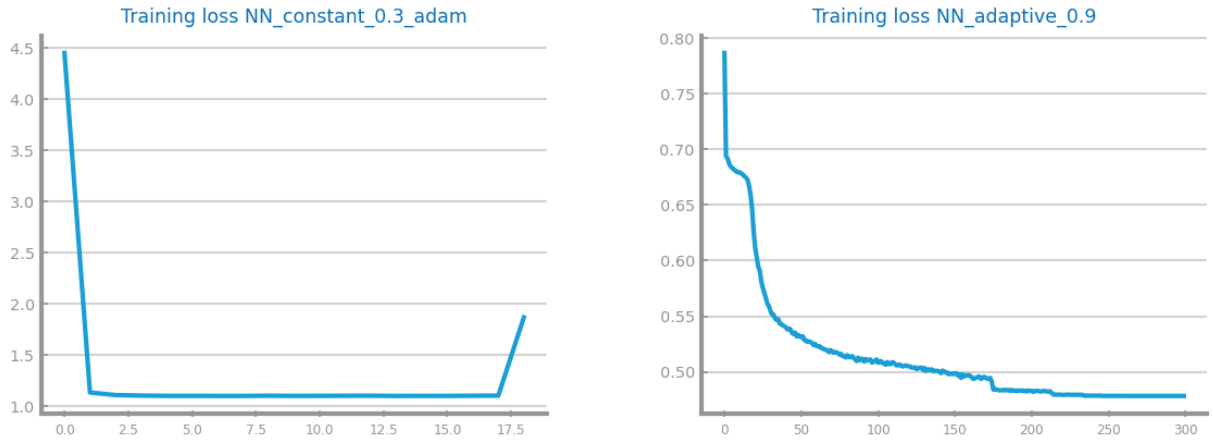


Figure 108: Loss curves analysis for dataset 1 (left) and dataset 2 (right)

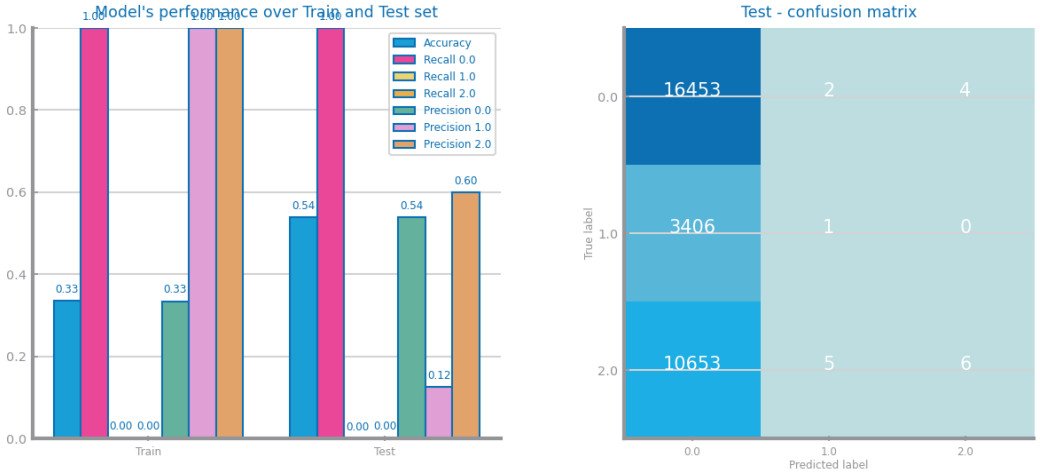


Figure 109: MLP best model results for dataset 1

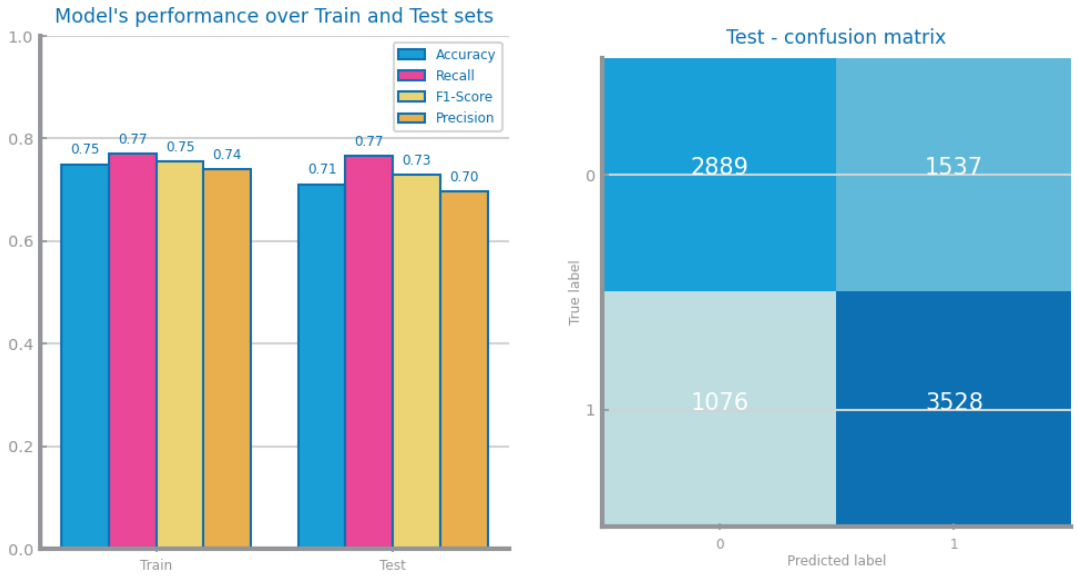


Figure 110: MLP best model results for dataset 2

4 CRITICAL ANALYSIS

The encoding was fundamental to improve the predictions over the datasets. In dataset 1, the paper's results were important to encode the variables in a way that expressed their relation to the target variable. In dataset 2, the transformation of *date* into cyclical data was essential to obtain better results since *date* had a great impact on the classification.

Throught the Data Preparation phase, we kept getting improvements on the data which led to better classification with the different models.

Dataset 1's results were worse. Likely due to class imbalance as well as in the variables: many records are similar, however, have different classes. If two records are similar but differ only in the class, it is hard for the models to learn how to differentiate between them.

For both sets, KNN had better results than NB. This is because in both there's an order inherent to the variables (and, e.g. in set 1, we used domain knowledge for it) which aids prediction. The analysis of overfitting for KNN is opposite to the one of DT (in the 1st smaller is worse and in the 2nd it's the opposite).

MLPs were not a good choice for set 2, since it has a high number of correlated variables which makes it hard for the model to learn useful representations of the data. Moreover, require a lot of data to learn. Our datasets aren't very big and, especially in set 1 (where we did undersampling), the sample is unsufficient. They also struggle to extract useful features from the data and to find complex relationships between the features.

TIME SERIES ANALYSIS

5 DATA PROFILING

Data Granularity

We plotted the data for different granularities, looking for repetitions. In set 1, the most atomic granularity where regularities were encountered was hourly, while in set 2 was daily. The two other granularities considered were daily and weekly for set 1, and weekly and monthly for set 2.

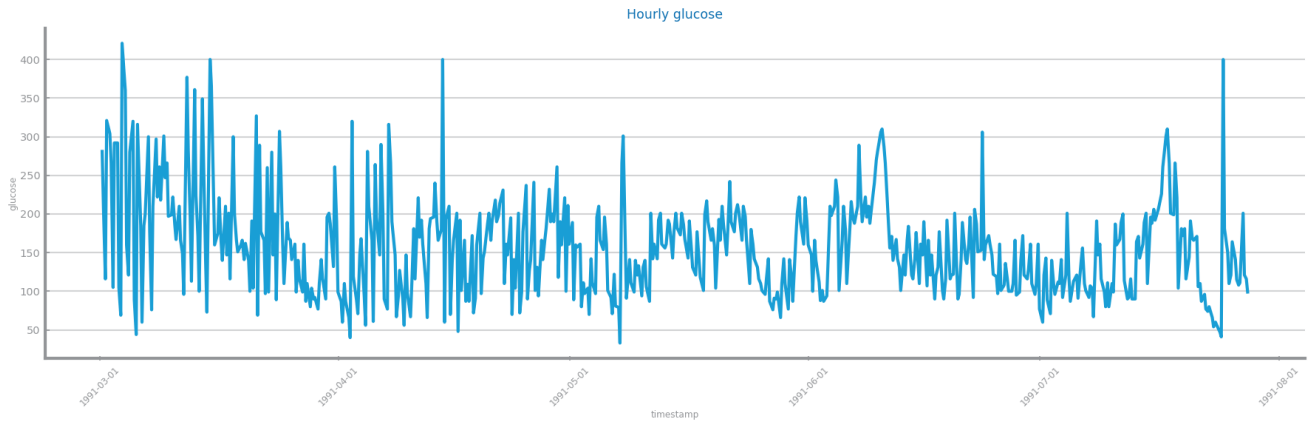


Figure 111: Time series 1 at the most granular detail

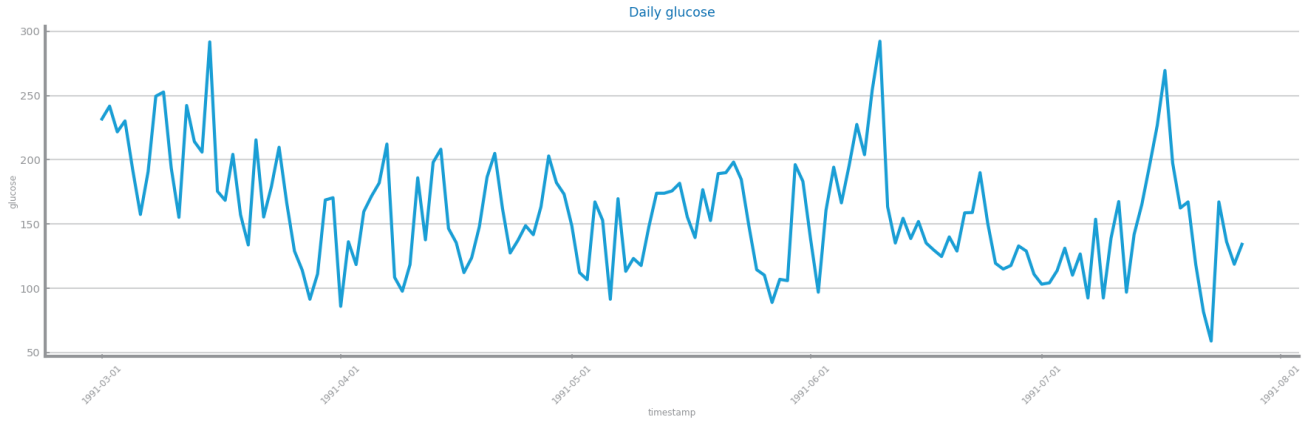


Figure 112: Time series 1 at the second chosen granularity

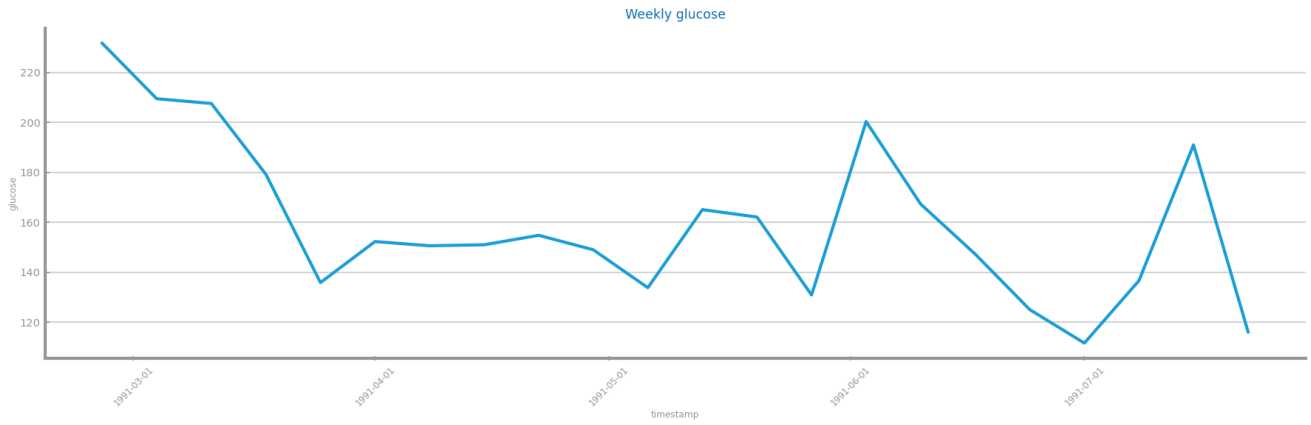


Figure 113: Time series 1 at the third chosen granularity

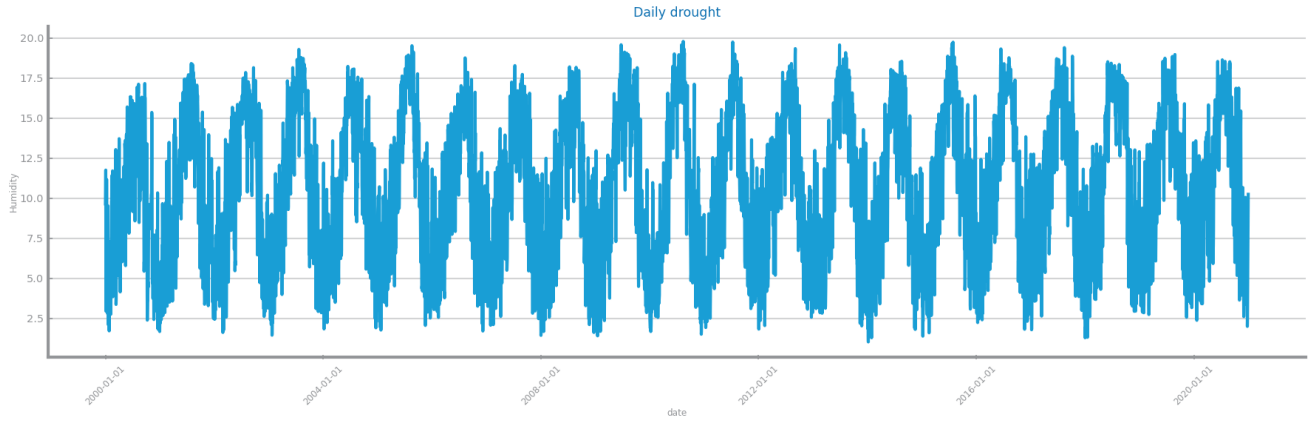


Figure 114: Time series 2 at the most granular detail

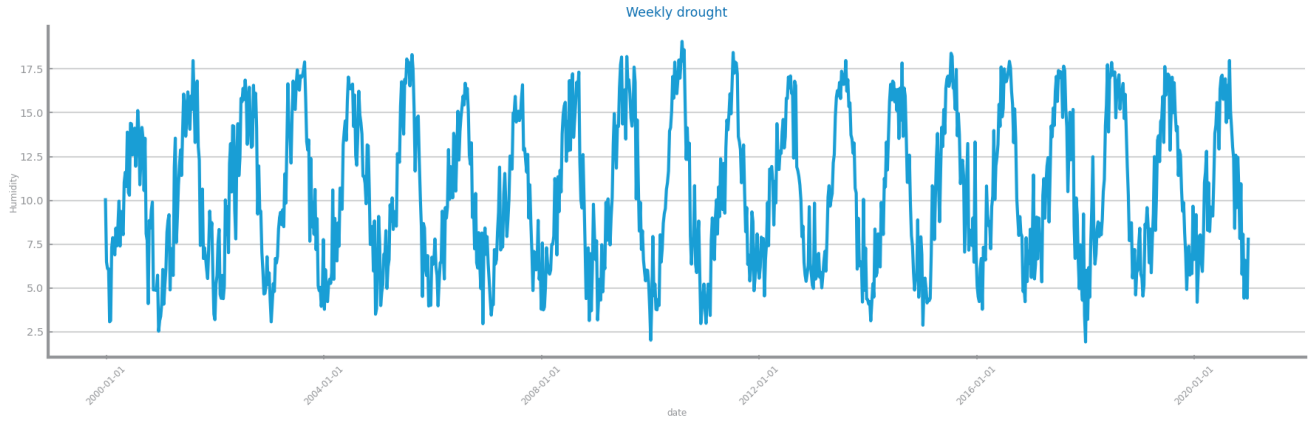


Figure 115: Time series 2 at the second chosen granularity

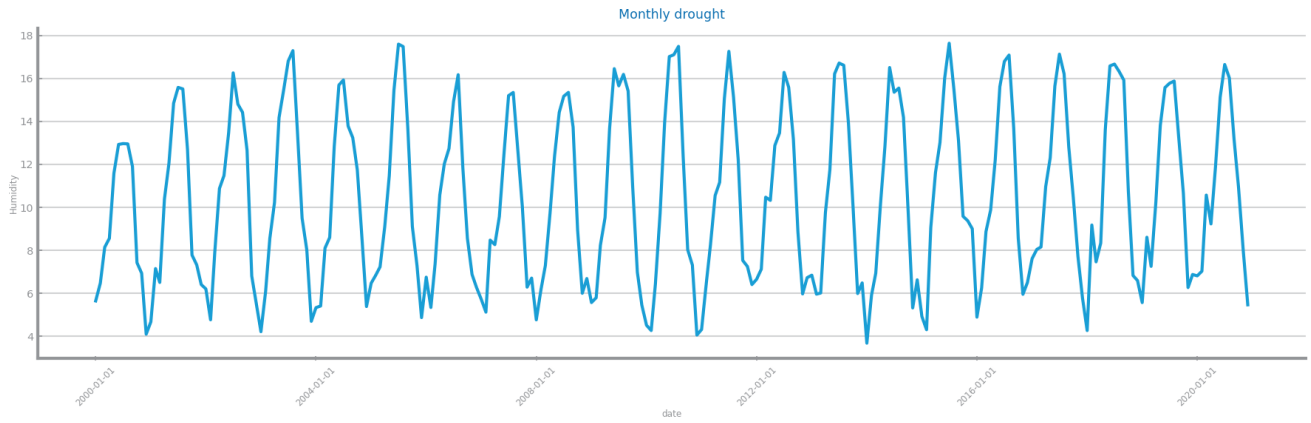


Figure 116: Time series 2 at the third chosen granularity

Data Distribution and Stationarity

We plotted the data for the 3 different granularities even though we could've only considered the most atomic. The boxplots seem to follow a normal distribution, however the histograms prove otherwise. In terms of stationarity, neither show a big mean variance, although 1 oscillates a bit more.

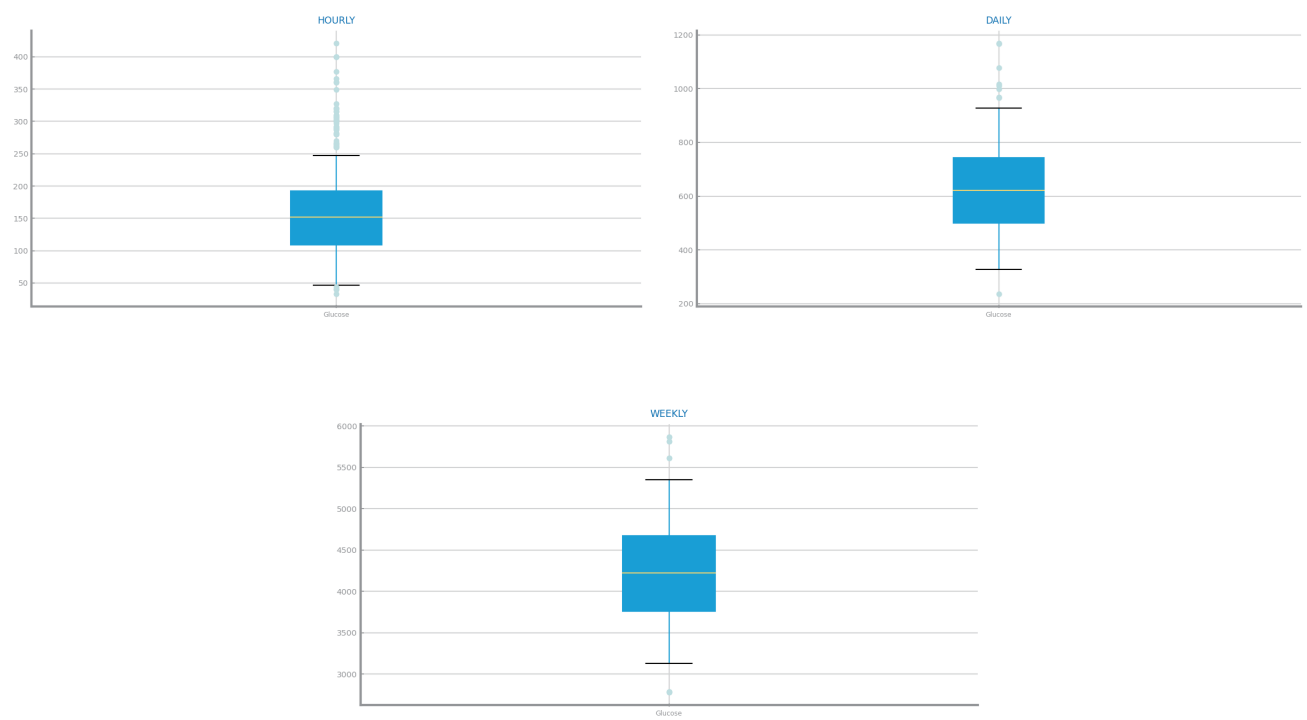


Figure 118: Boxplot(s) for time series 1

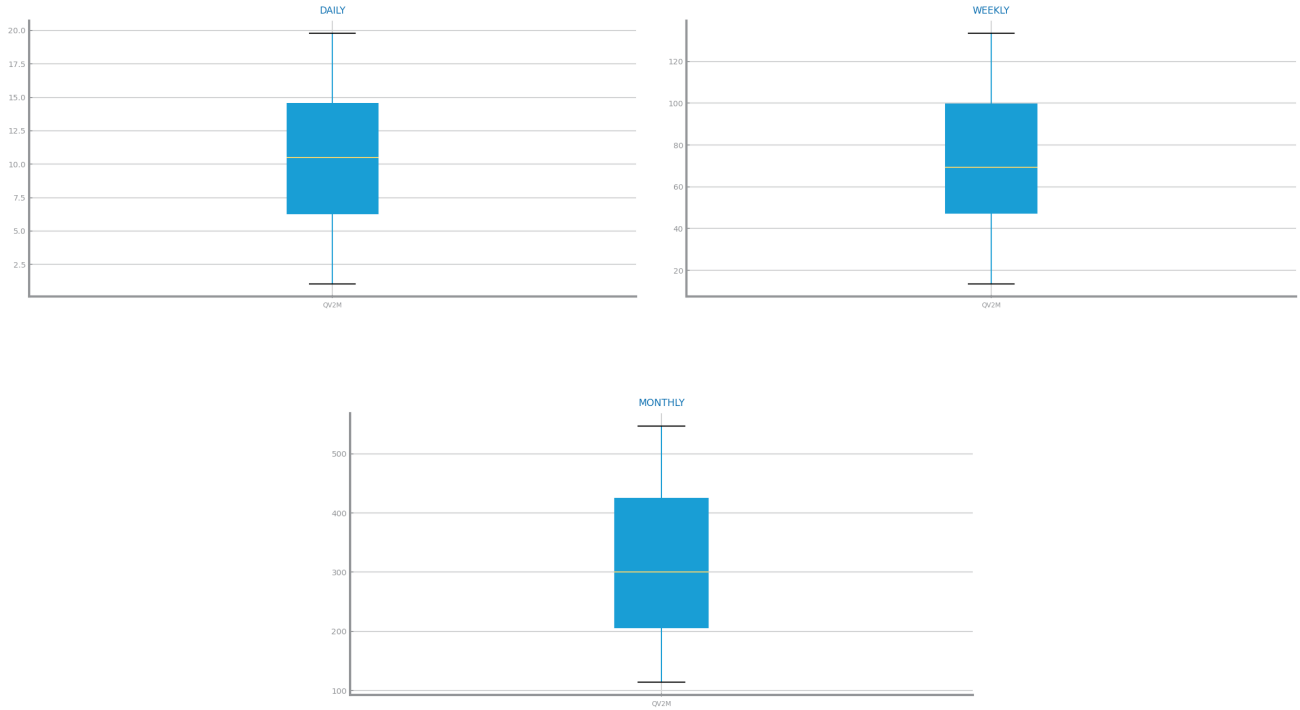
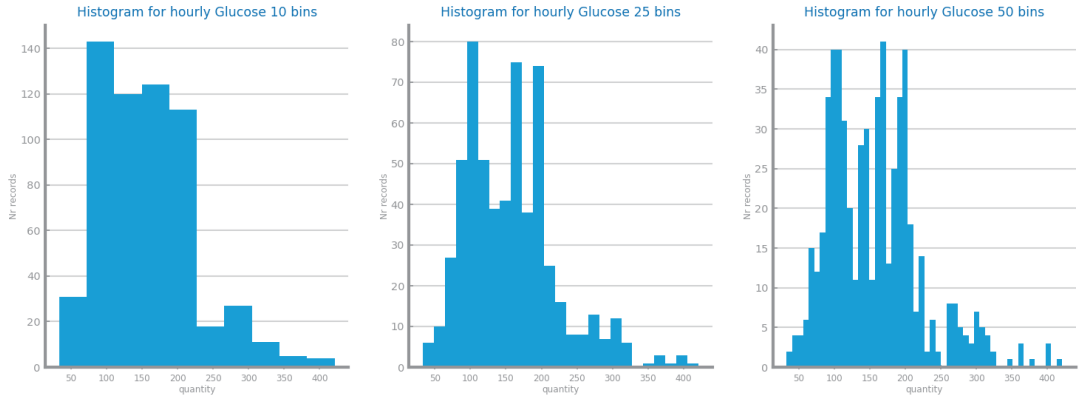


Figure 120: Boxplot(s) for time series 2



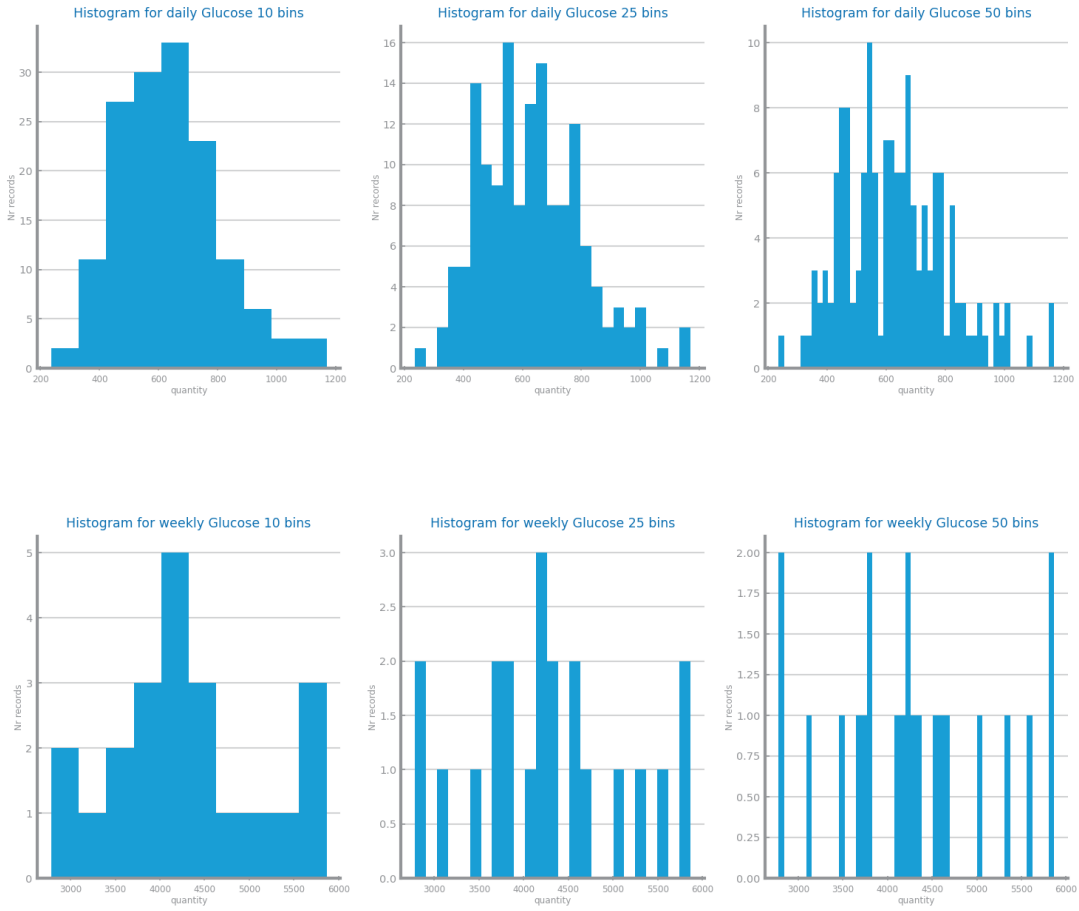
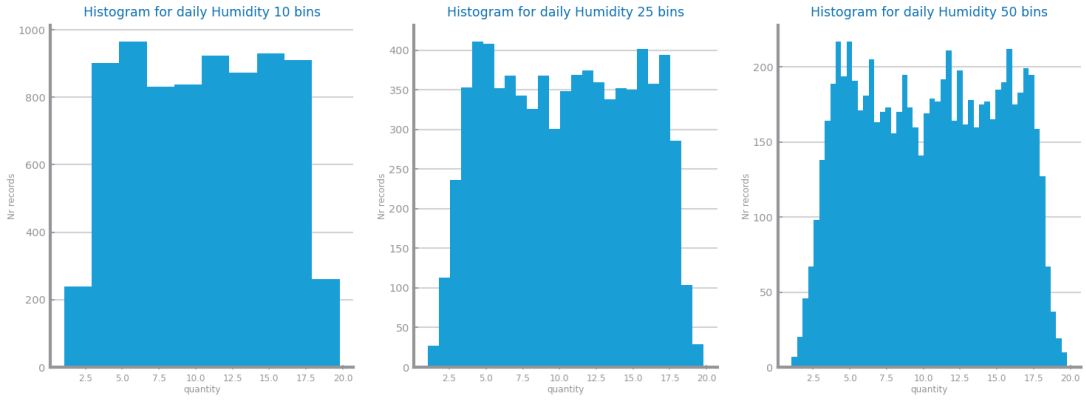


Figure 121: Histogram(s) for time series 1



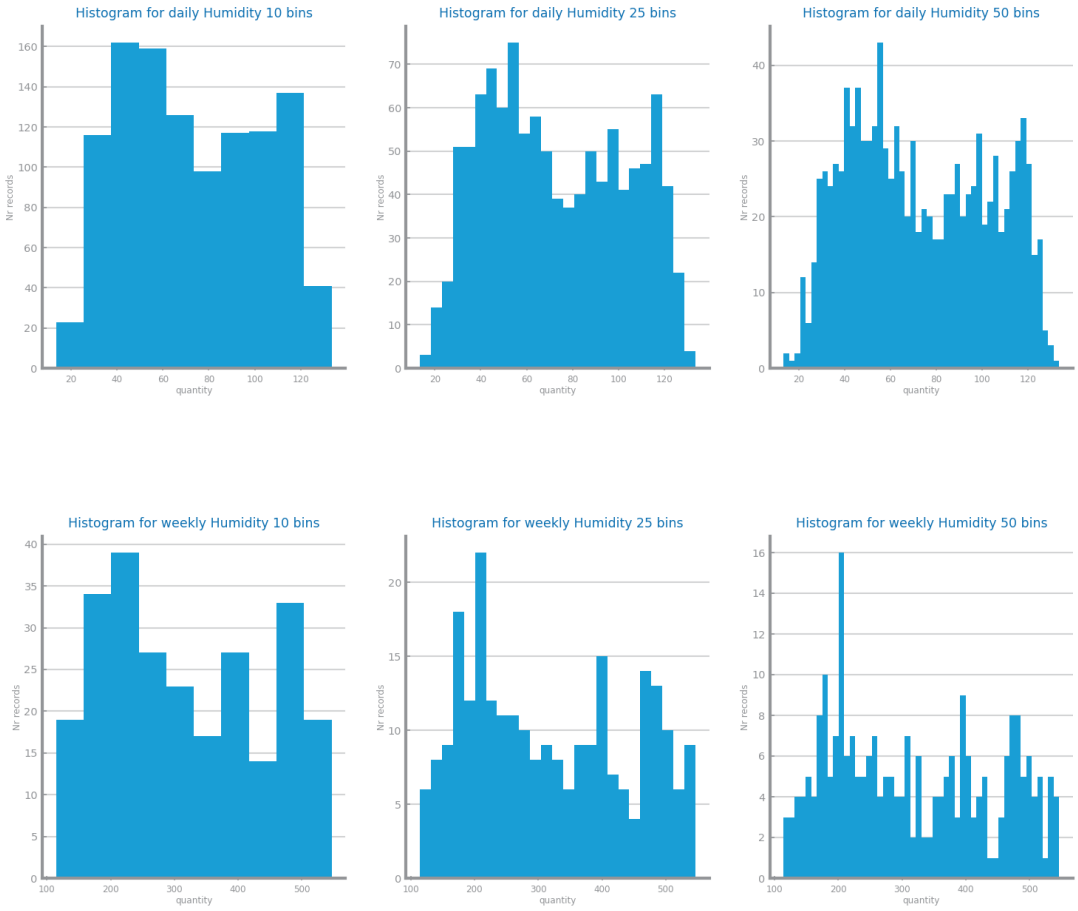
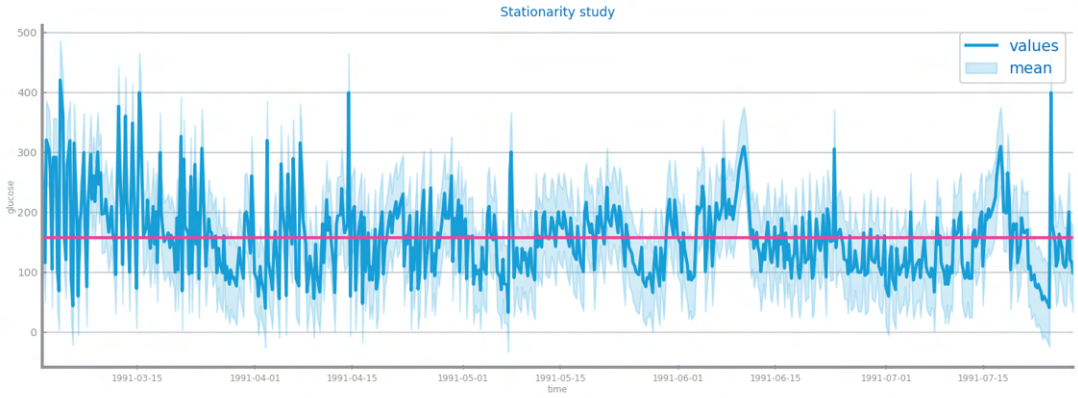


Figure 122: Histogram(s) for time series 2



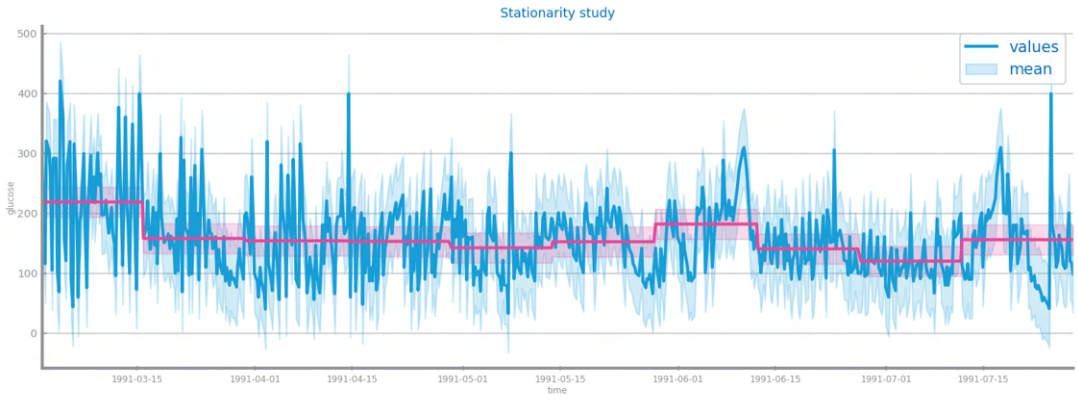
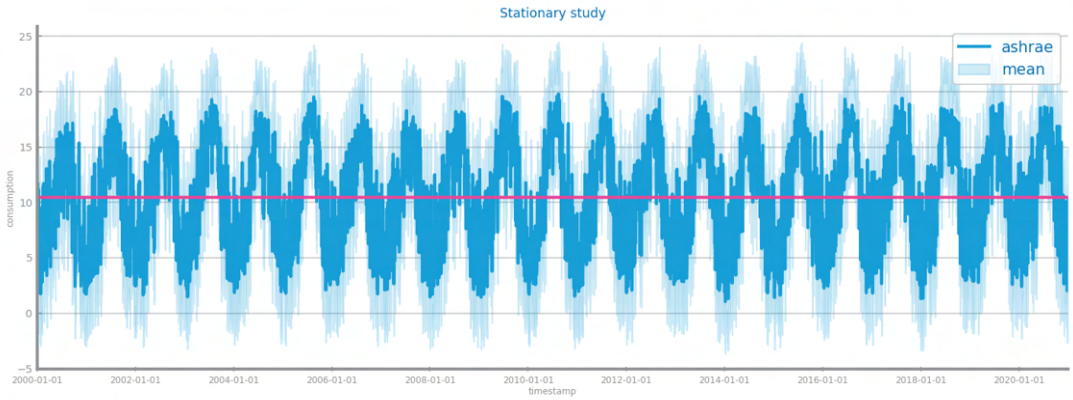


Figure 123: Stationarity study for time series 1



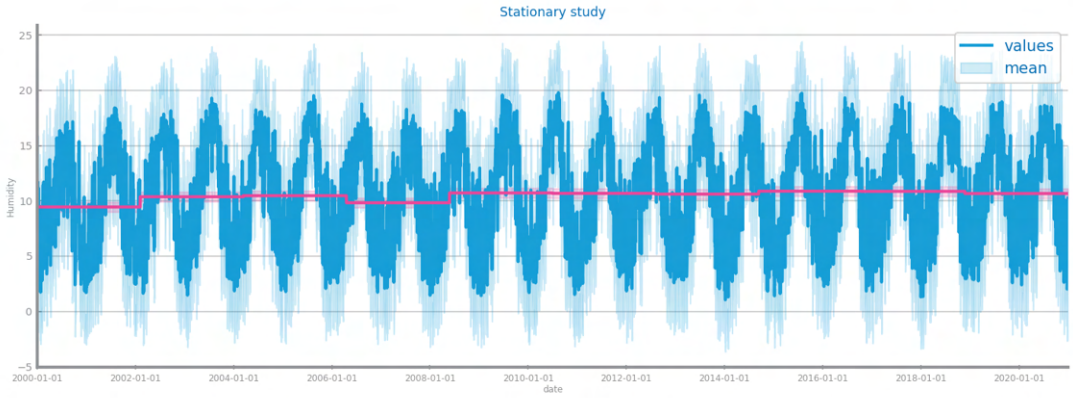


Figure 124: Stationarity study for time series 2

6 DATA TRANSFORMATION

Aggregation

We aggregated the data given the three most atomic periods. Set 1, hours, days and weeks. Set 2, days, weeks and months. For both, we decided to chose the most atomic one given the similar results.

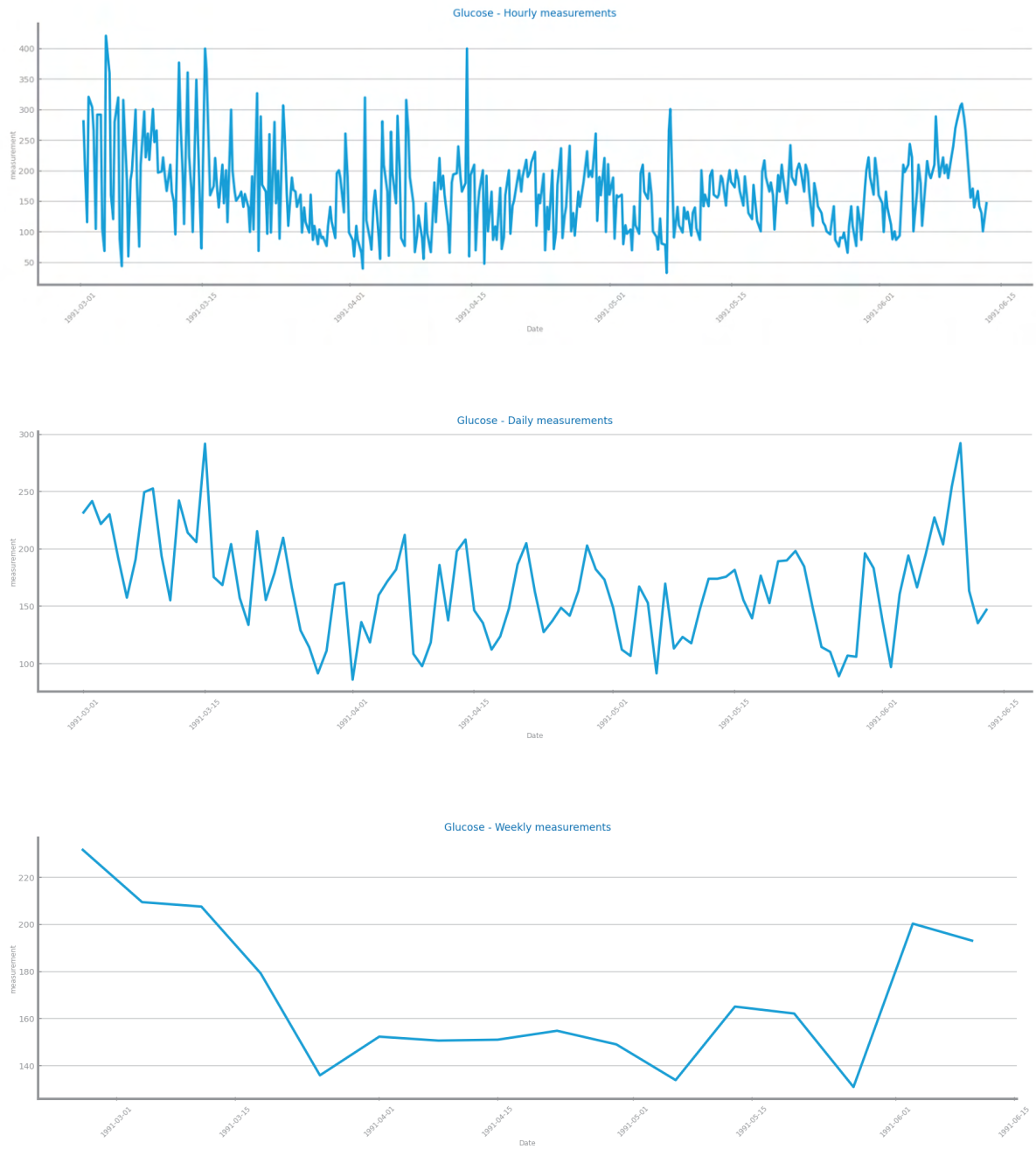


Figure 125: Time series plots after different aggregations on time series 1

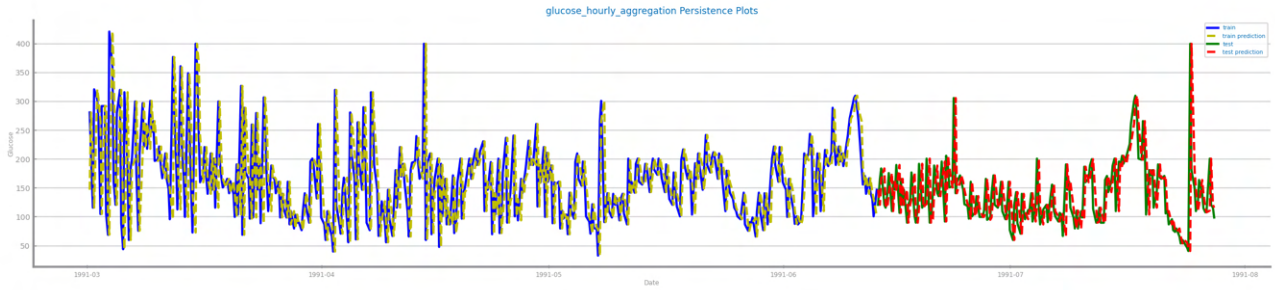


Figure 126: Forecasting plots after hourly aggregation on time series 1

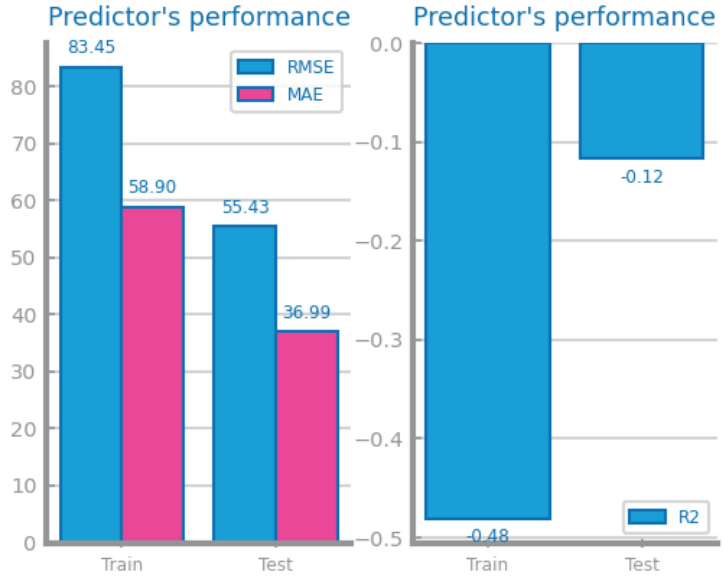


Figure 127: Forecasting results after hourly aggregation on time series 1

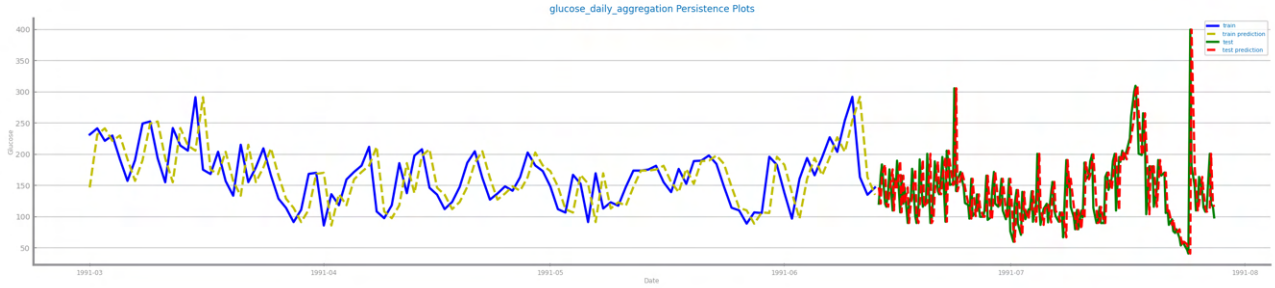


Figure 128: Forecasting plots after daily aggregation on time series 1

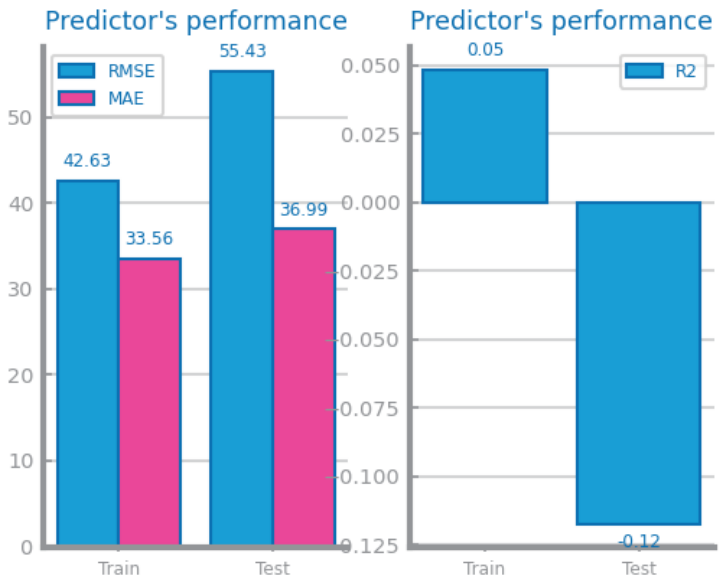


Figure 129: Forecasting results after daily aggregation on time series 1

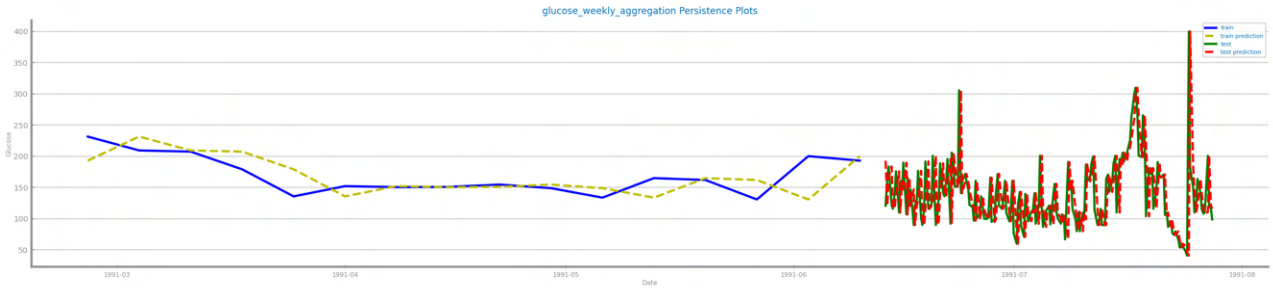


Figure 130: Forecasting plots after weekly aggregation on time series 1

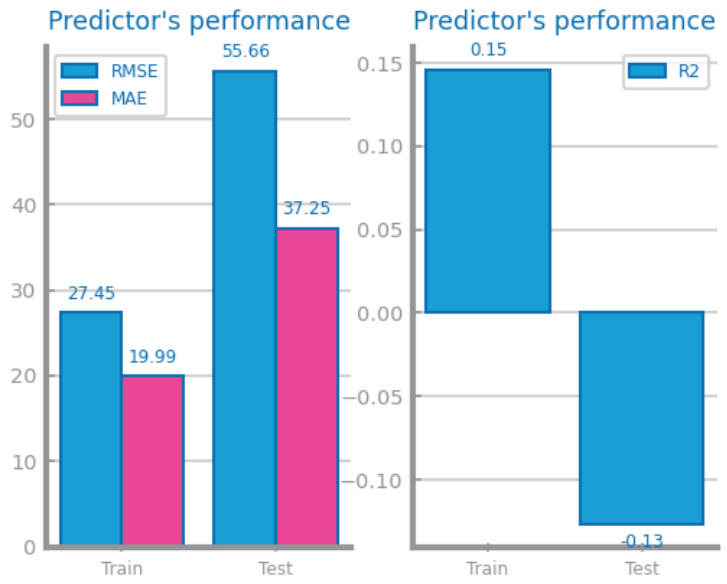
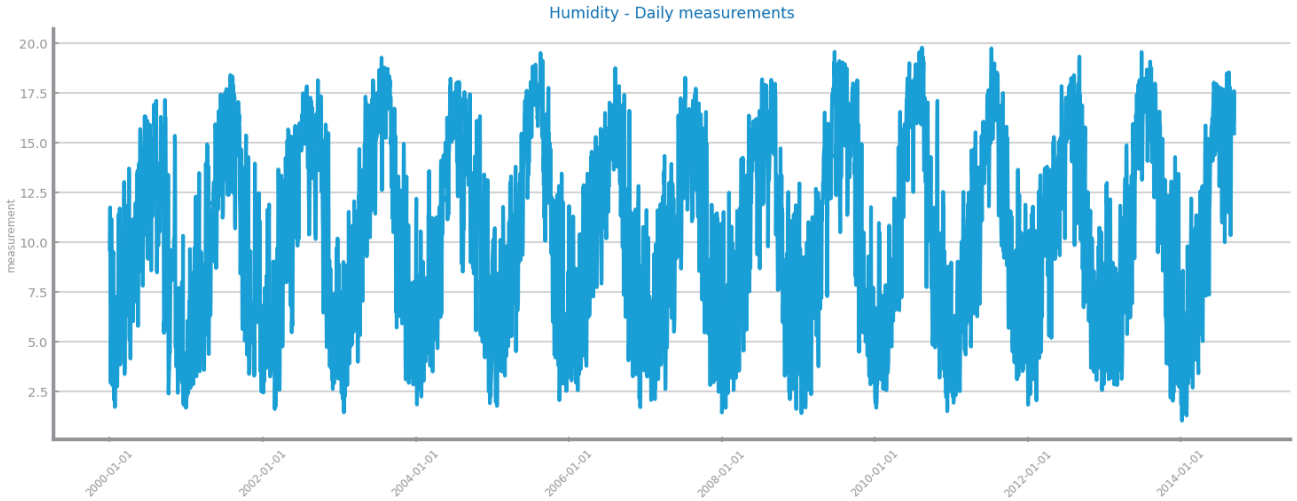


Figure 131: Forecasting results after weekly aggregation on time series 1



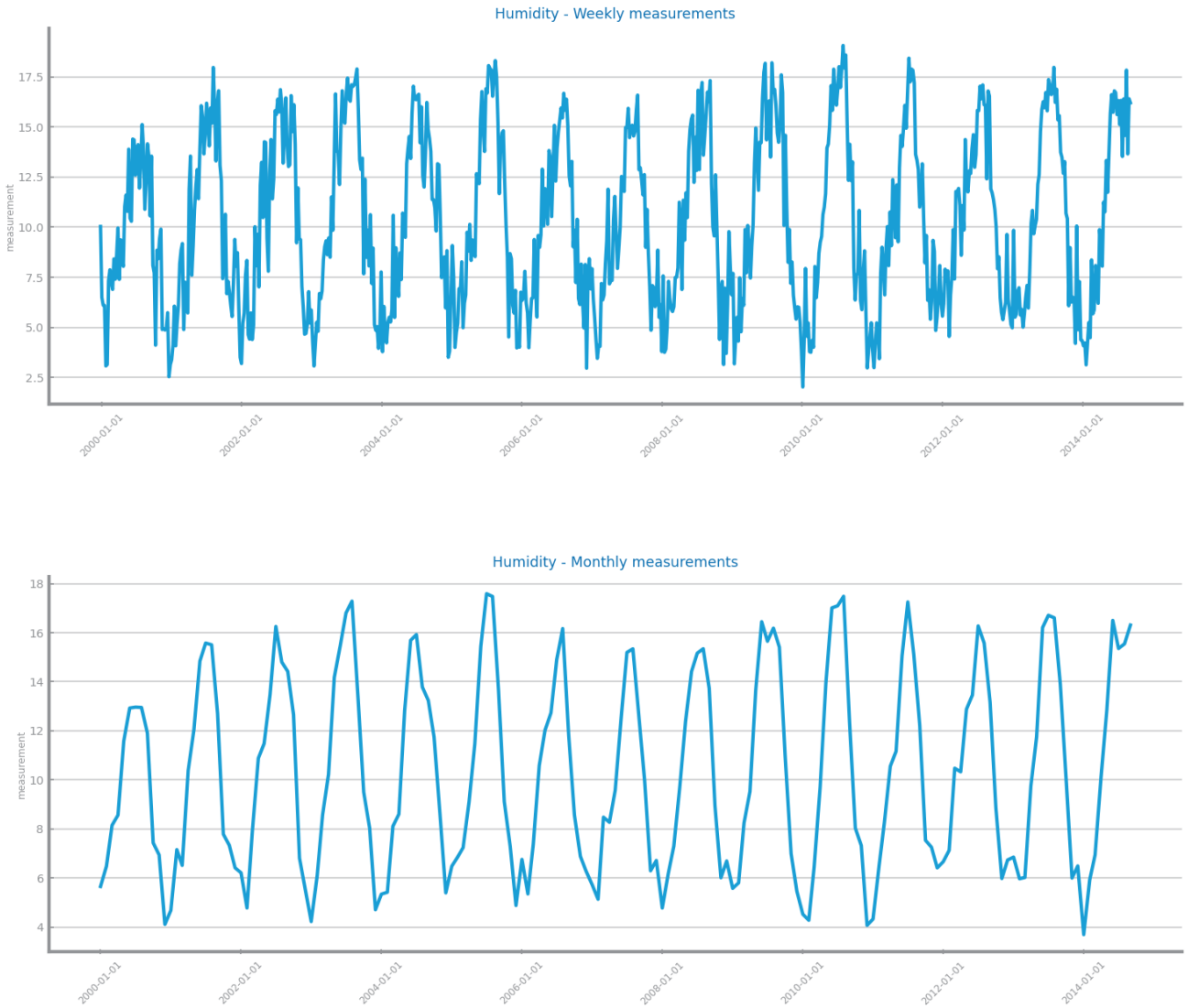


Figure 132: Time series plots after different aggregations on time series 2

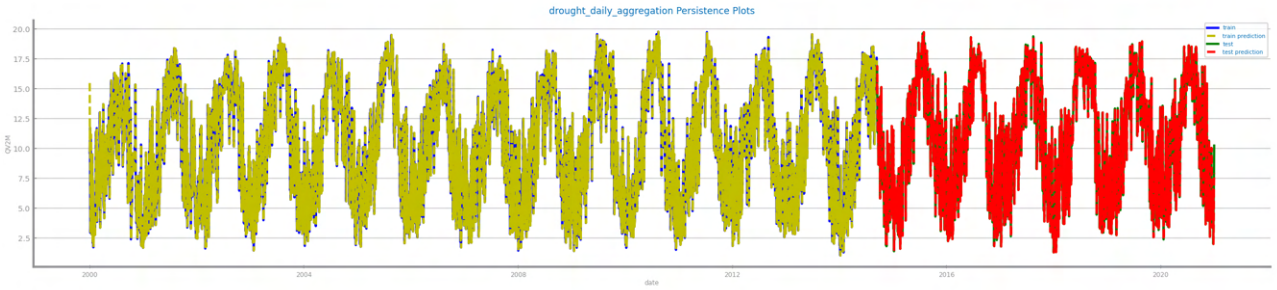


Figure 133: Forecasting plots after daily aggregation on time series 2

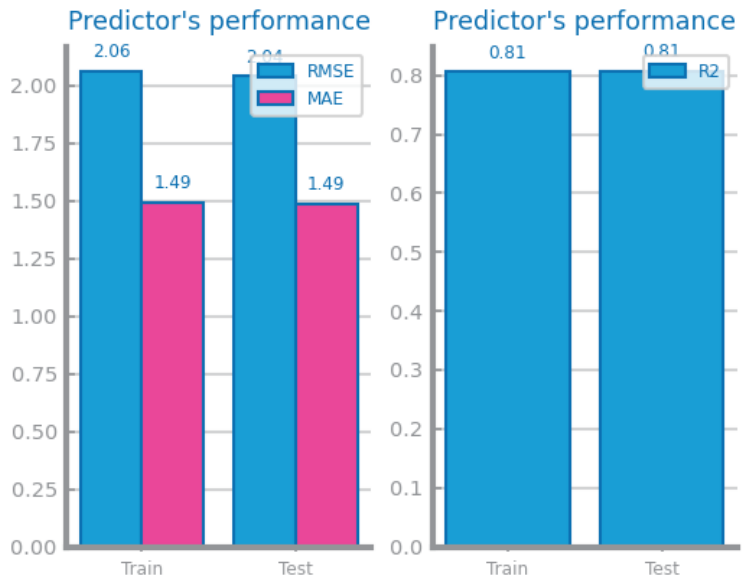


Figure 134: Forecasting results after daily aggregation on time series 2

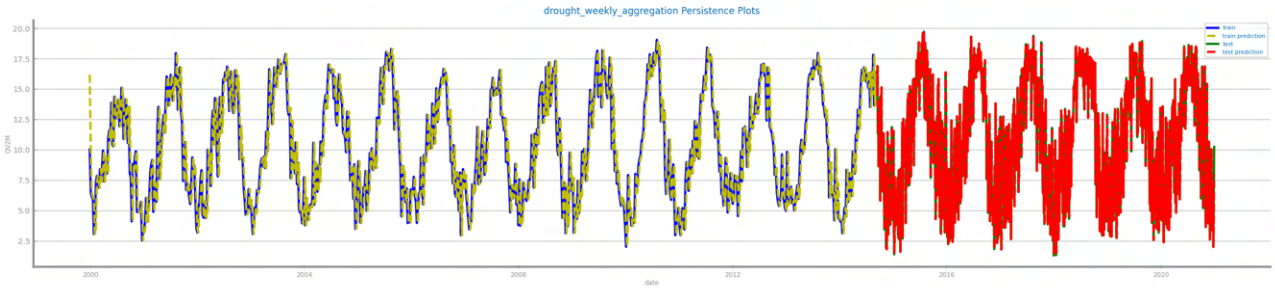


Figure 135: Forecasting plots after weekly aggregation on time series 1

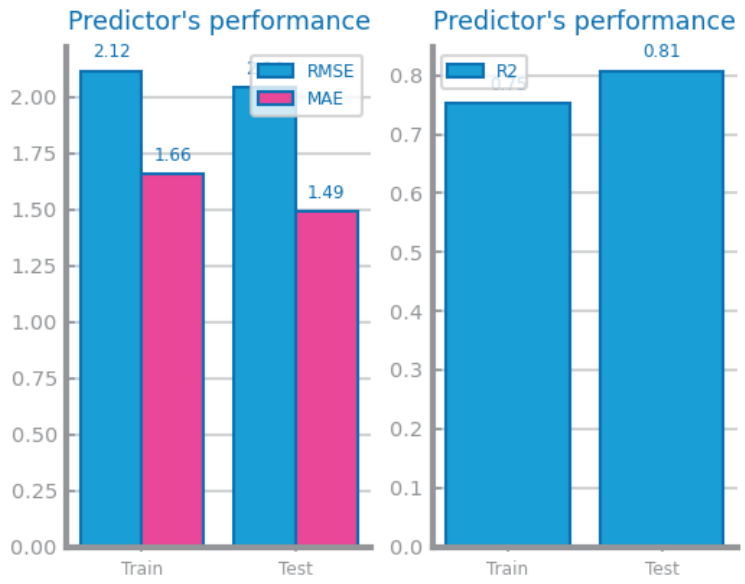


Figure 136: Forecasting results after weekly aggregation on time series 1

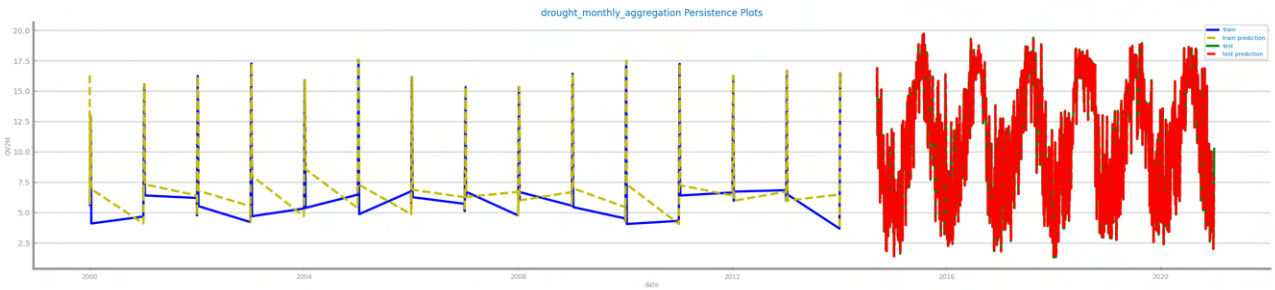


Figure 137: Forecasting plots after monthly aggregation on time series 1

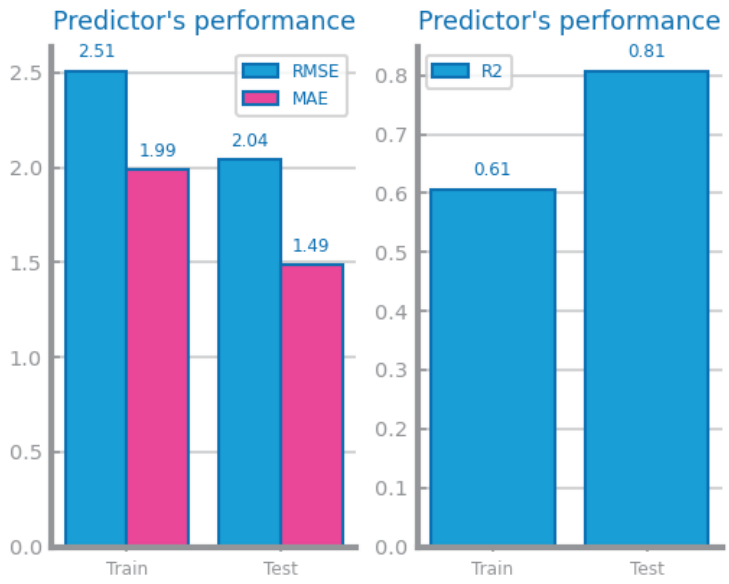
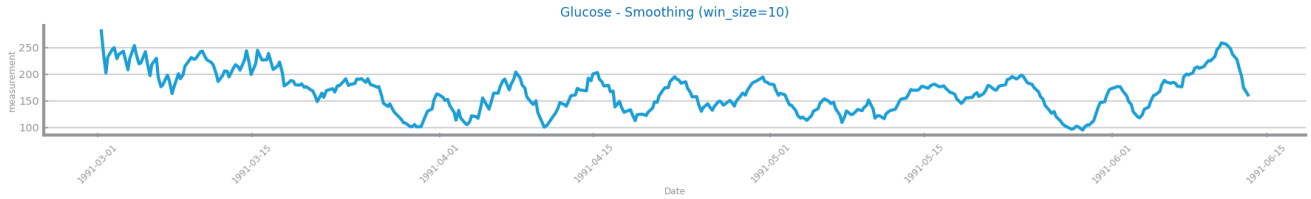
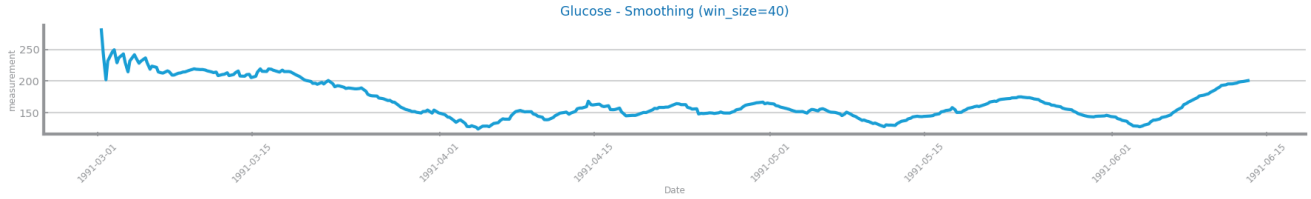


Figure 138: Forecasting results after monthly aggregation on time series 1

Smoothing

For set 1, win sizes of 10, 40, 100. Best R2 and error: 10. For set 2, win sizes of 10, 30, 100, 200. Chosen: 30 - the results were similar and we did not want to lose too much sharpness on the data.





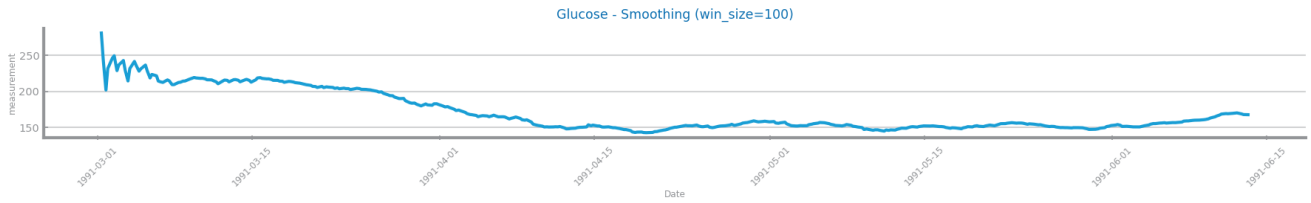


Figure 139: Time series plots after different smoothing parameterizations on time series 1

Figure 140: Forecasting plots after smoothing with window of size 10 on time series 1

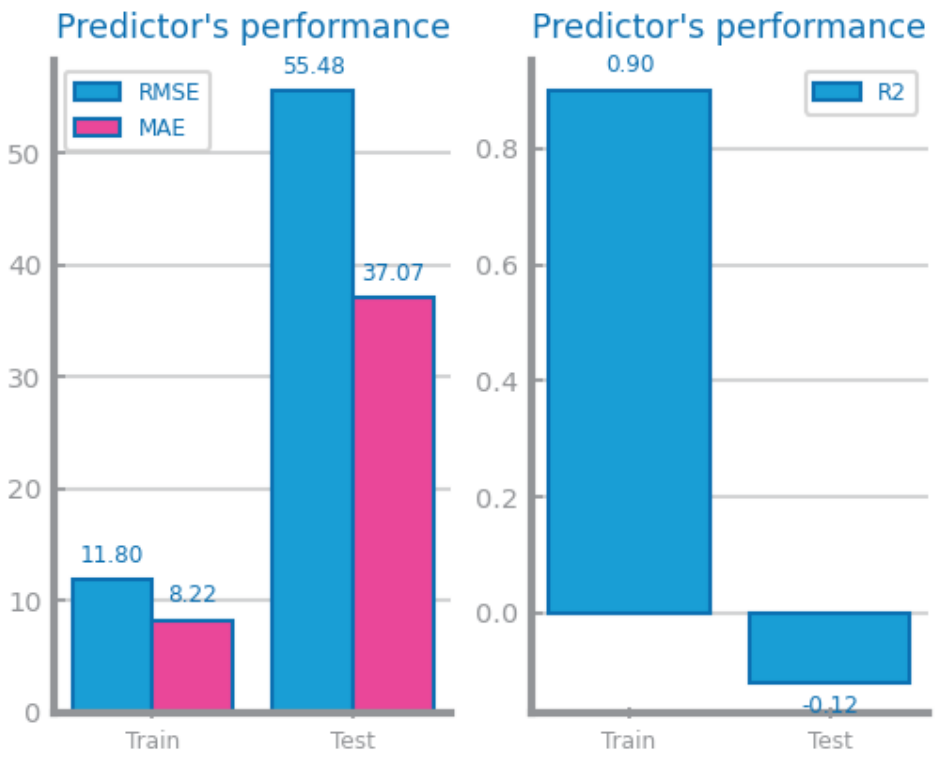


Figure 141: Forecasting results after smoothing with window of size 10 on time series 1

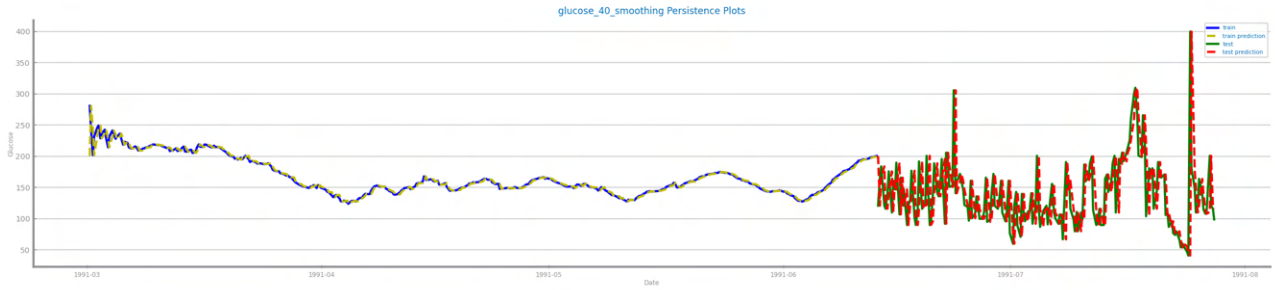


Figure 142: Forecasting plots after smoothing with window of size 40 on time series 1

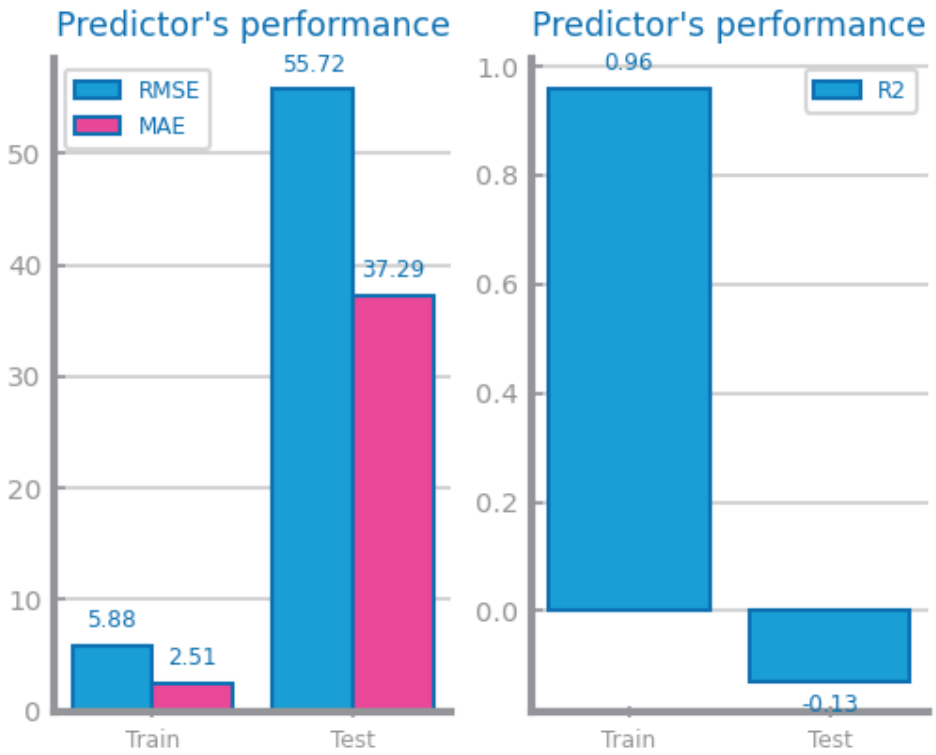


Figure 143: Forecasting results after smoothing with window of size 40 on time series 1

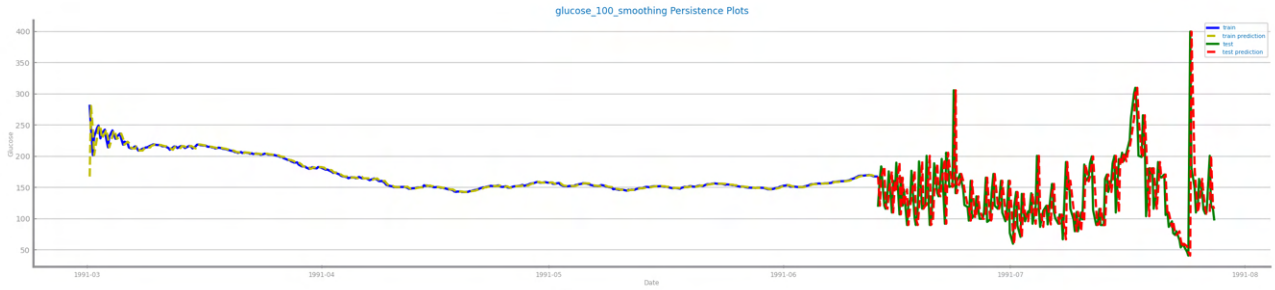


Figure 144: Forecasting plots after smoothing with window of size 100 on time series 1

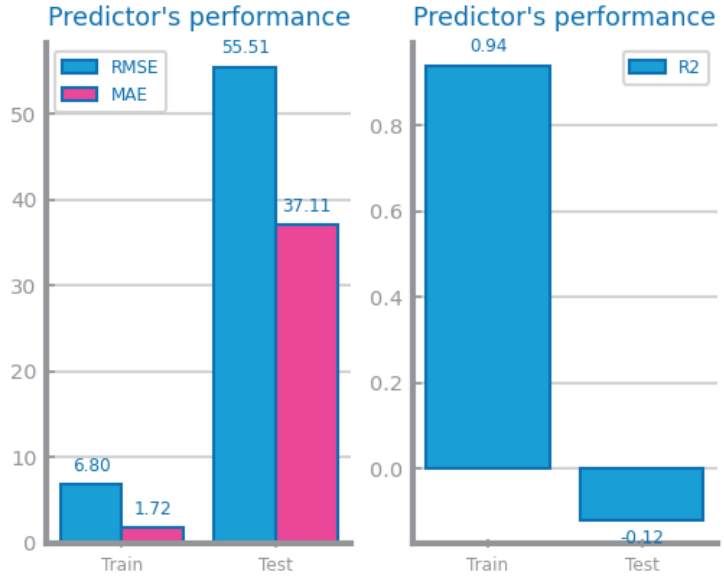
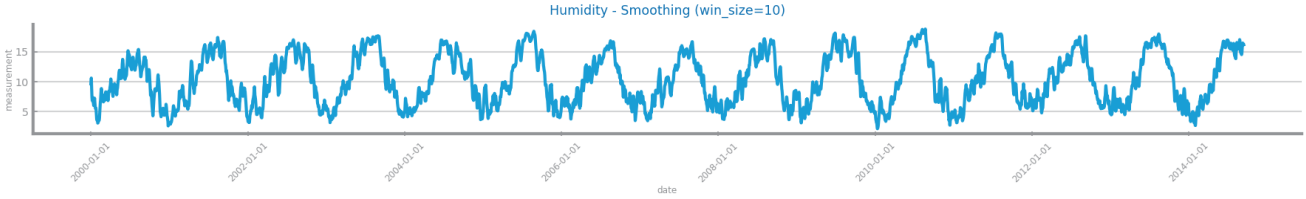
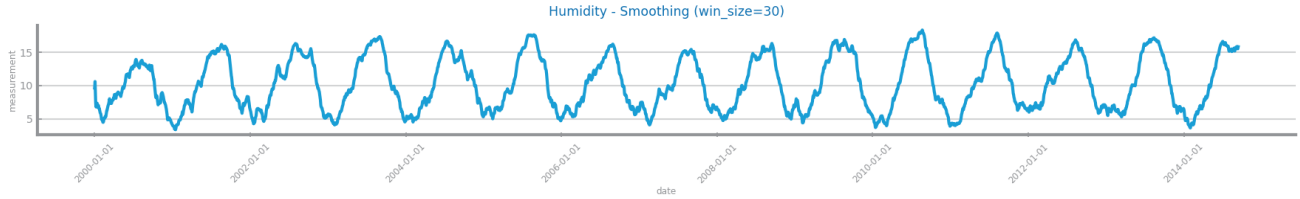


Figure 145: Forecasting results after smoothing with window of size 100 on time series 1





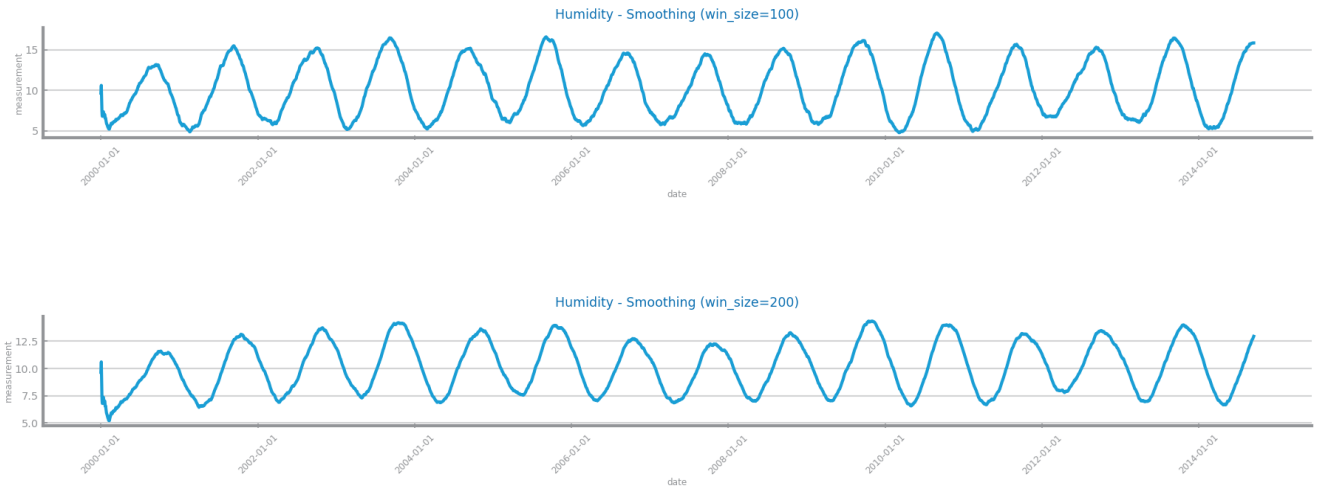


Figure 146: Time series plots after different smoothing parameterizations on time series 2

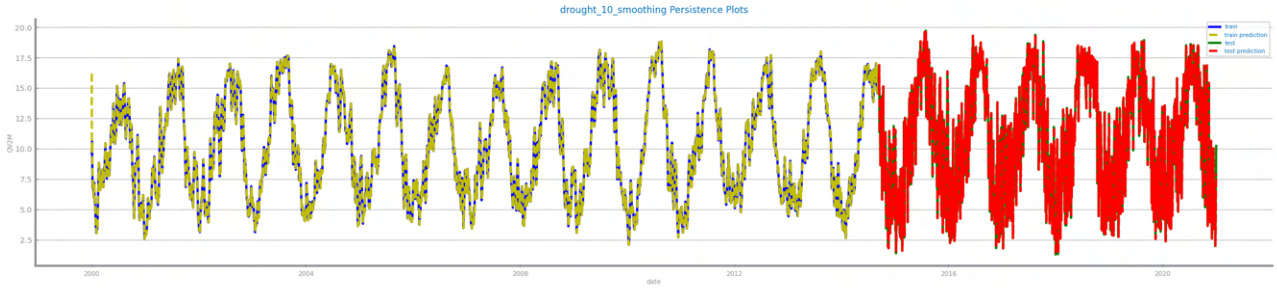


Figure 147: Forecasting plots after smoothing with window of size 10 on time series 2

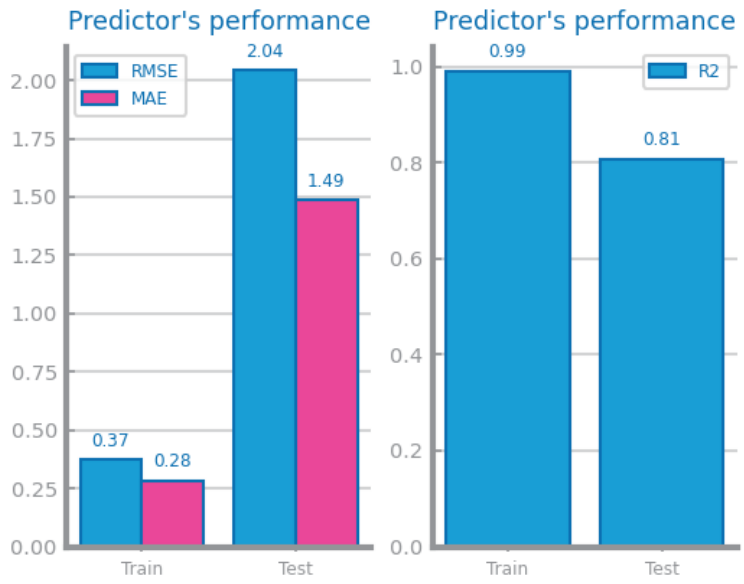


Figure 148: Forecasting results after smoothing with window of size 10 on time series 2

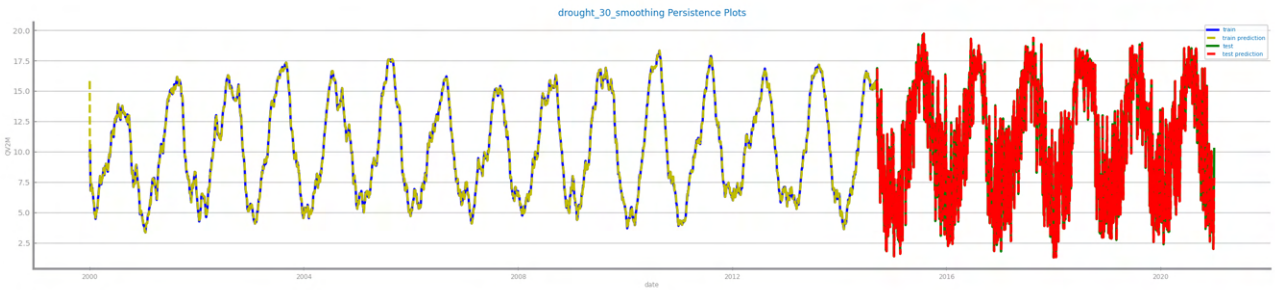


Figure 149: Forecasting plots after smoothing with window of size 30 on time series 2

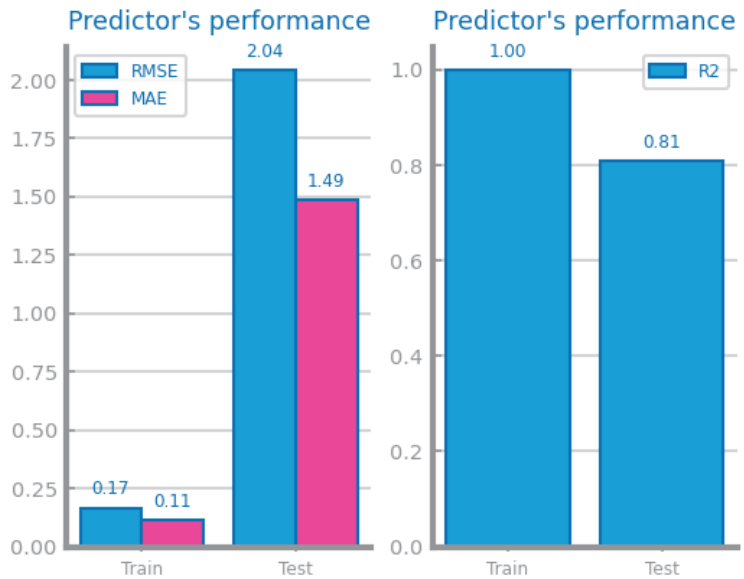


Figure 150: Forecasting results after smoothing with window of size 30 on time series 2

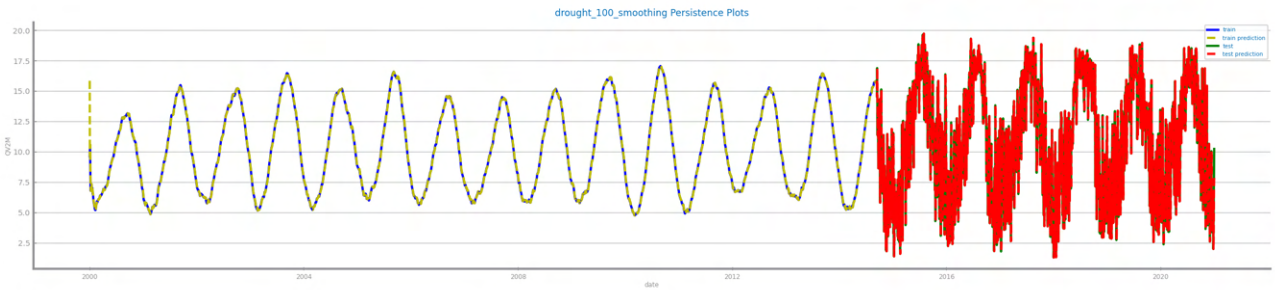


Figure 151: Forecasting plots after smoothing with window of size 100 on time series 2

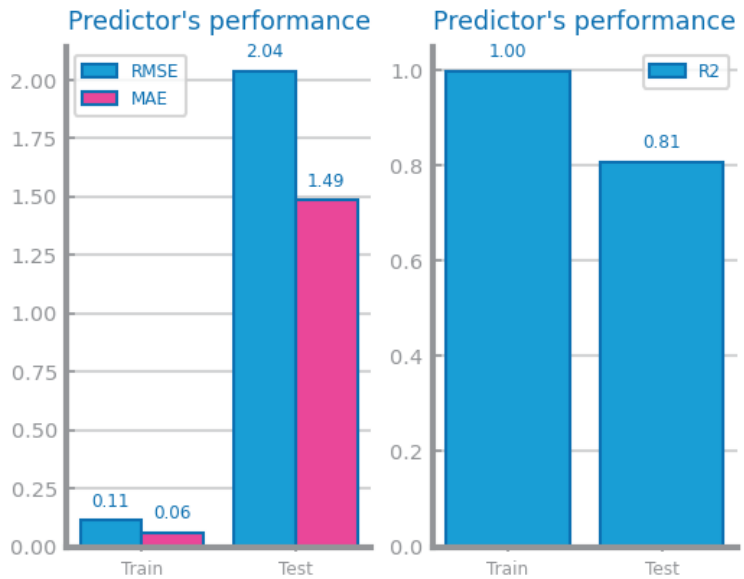


Figure 152: Forecasting results after smoothing with window of size 100 on time series 2

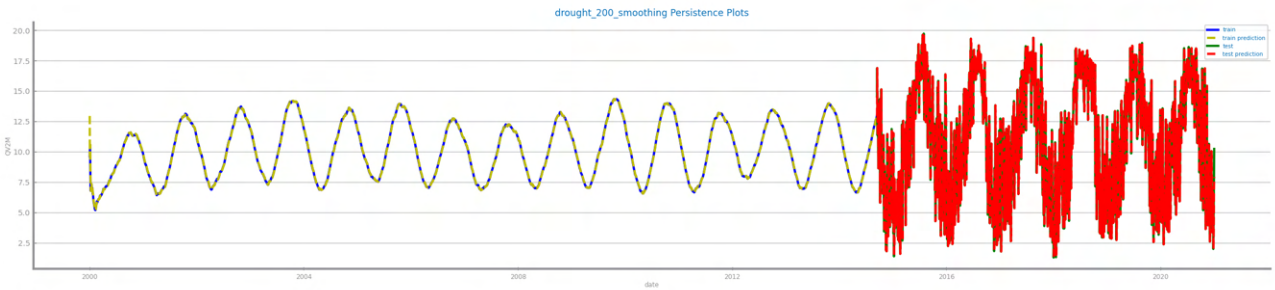


Figure 153: Forecasting plots after smoothing with window of size 200 on time series 2

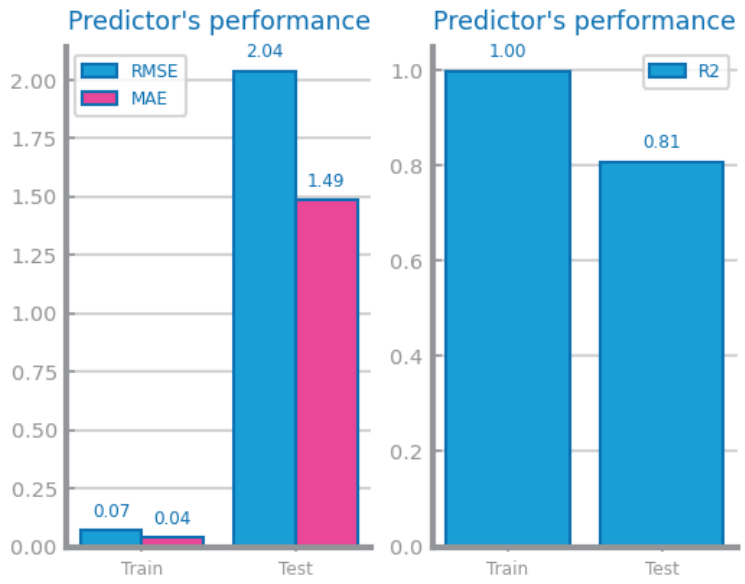
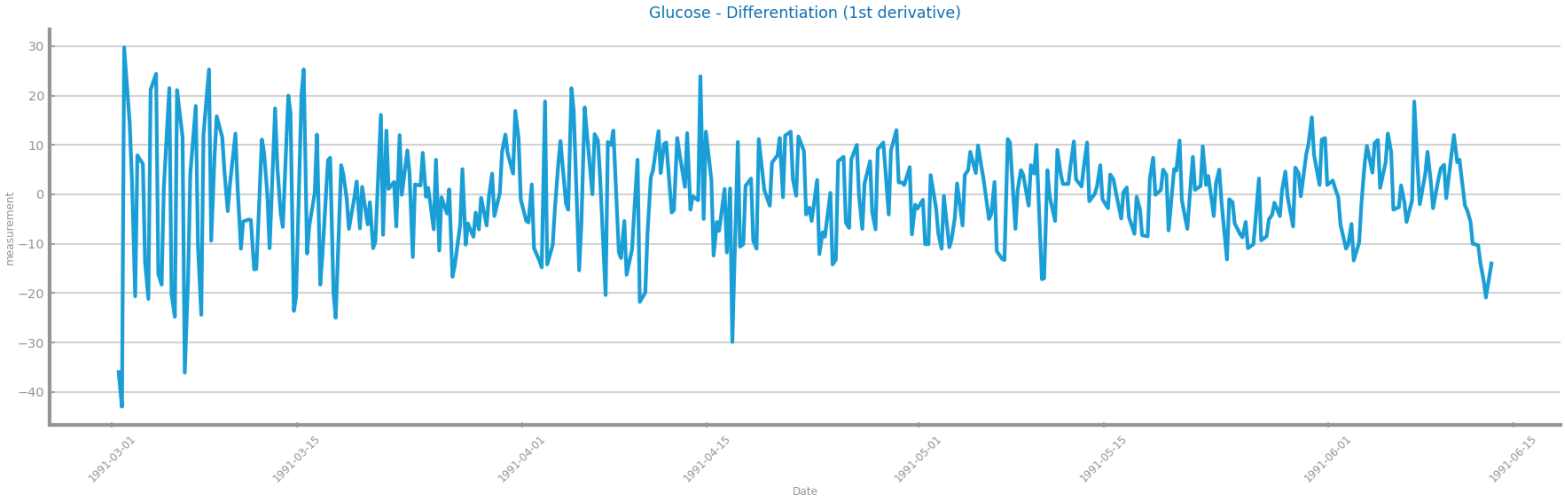


Figure 154: Forecasting results after smoothing with window of size 200 on time series 2

Differentiation

For this task, we applied the transformation to the entire series, since we are changing the data space. Differentiation did not improve neither of the results so we kept the previous series.



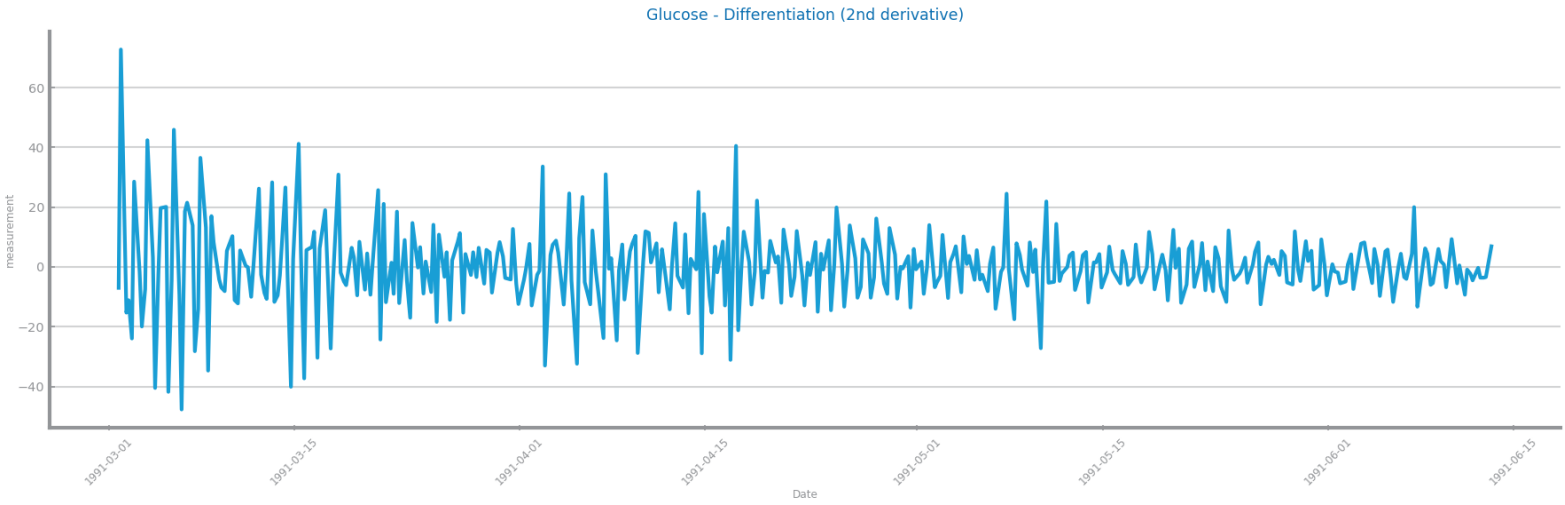
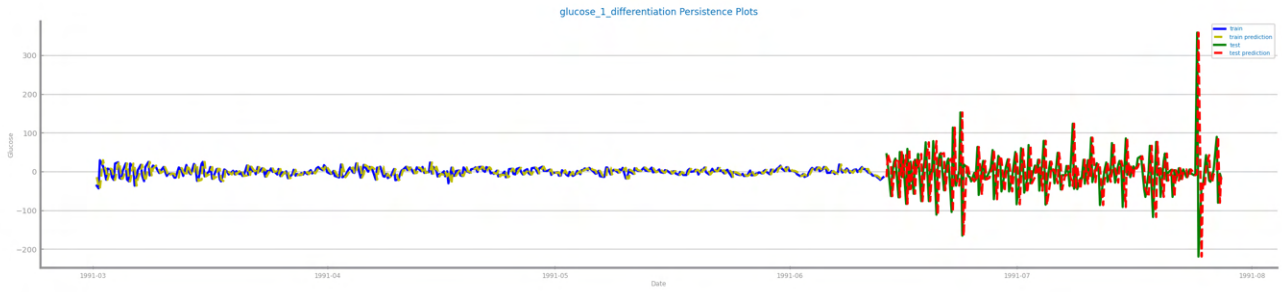


Figure 155: Time series plots after first and second differentiation of time series 1



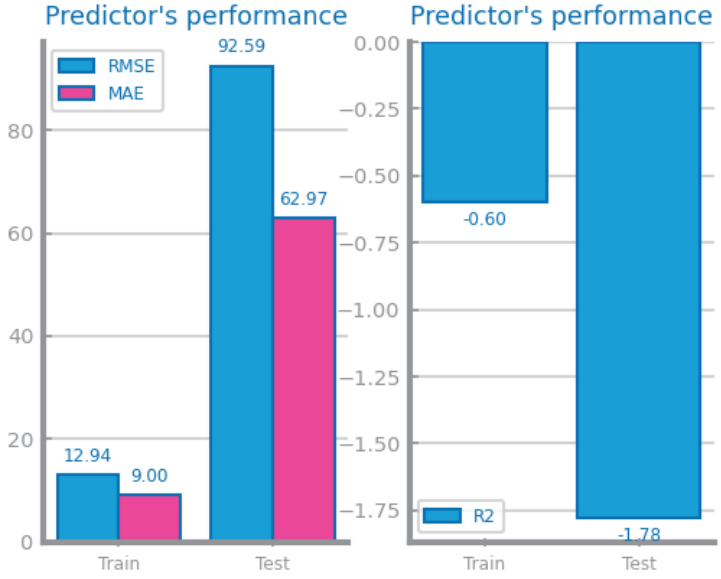
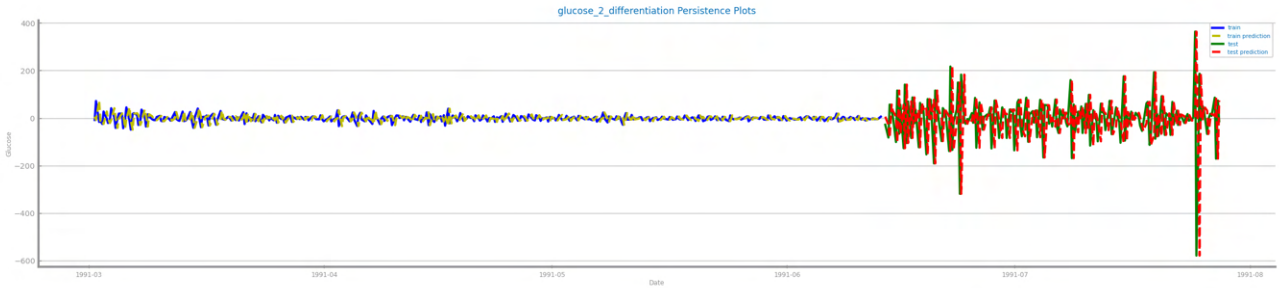


Figure 156: Forecasting results after first differentiation of time series 1



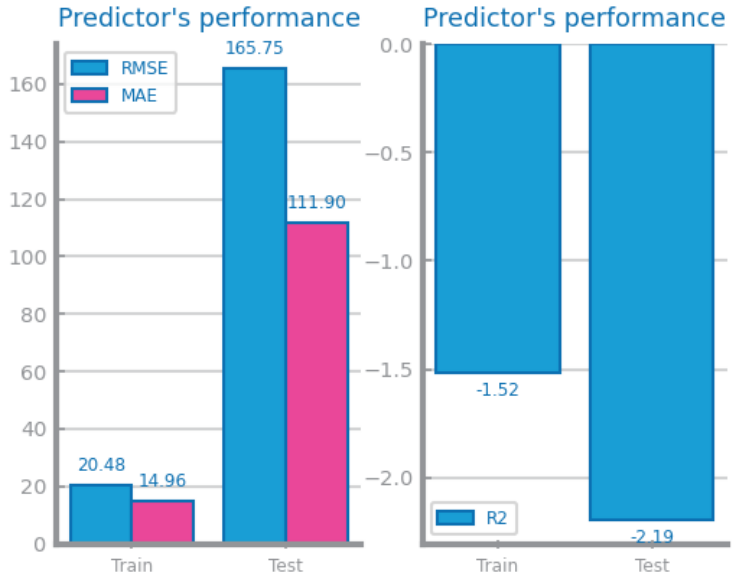
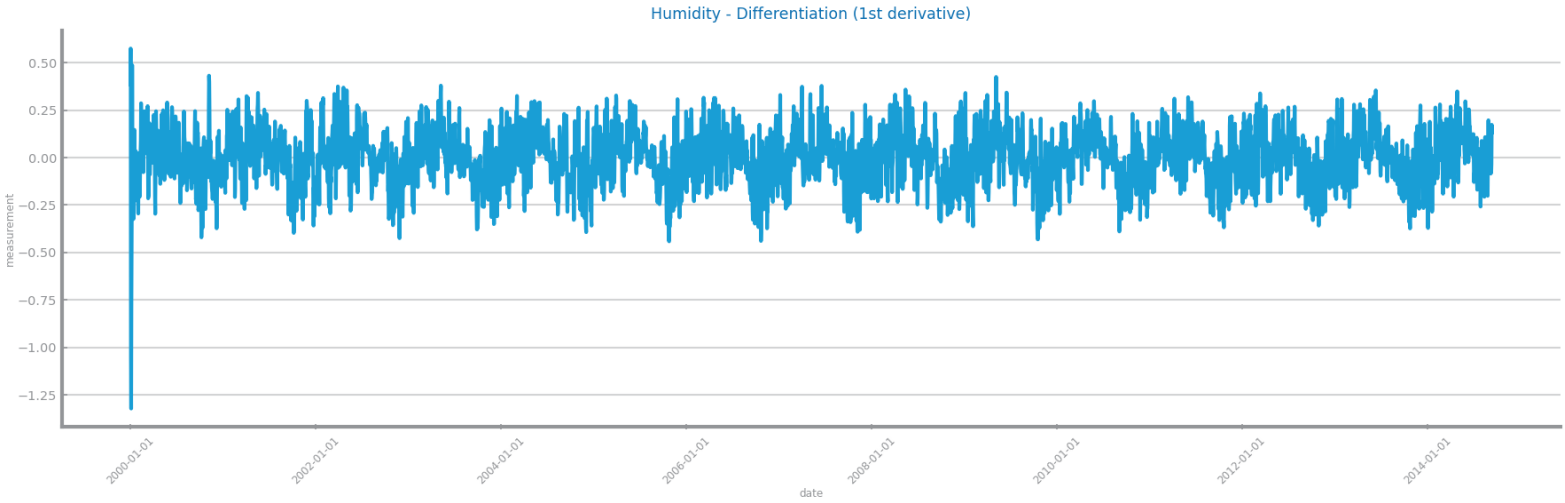


Figure 157: Forecasting results after second differentiation of time series 1



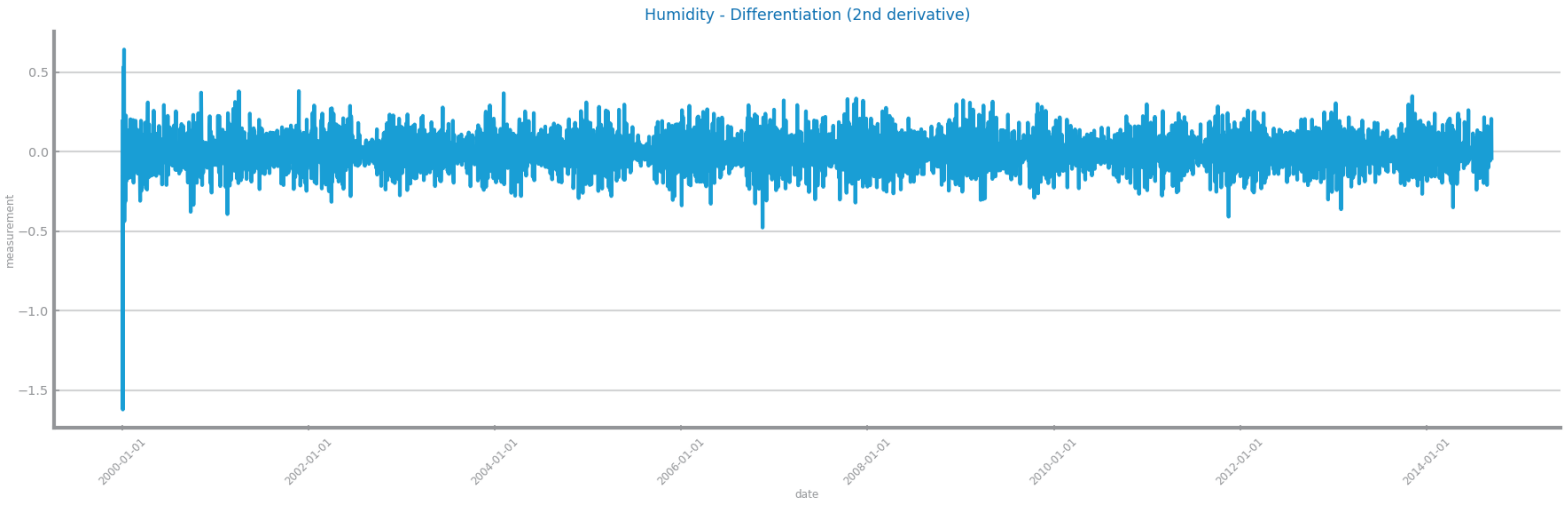


Figure 158: Time series plots after first and second differentiation of time series 2

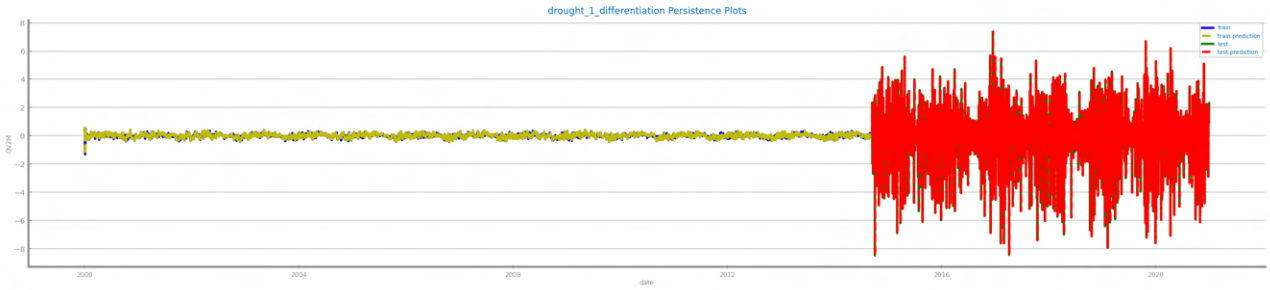


Figure 159: Forecasting plots after first differentiation of time series 2

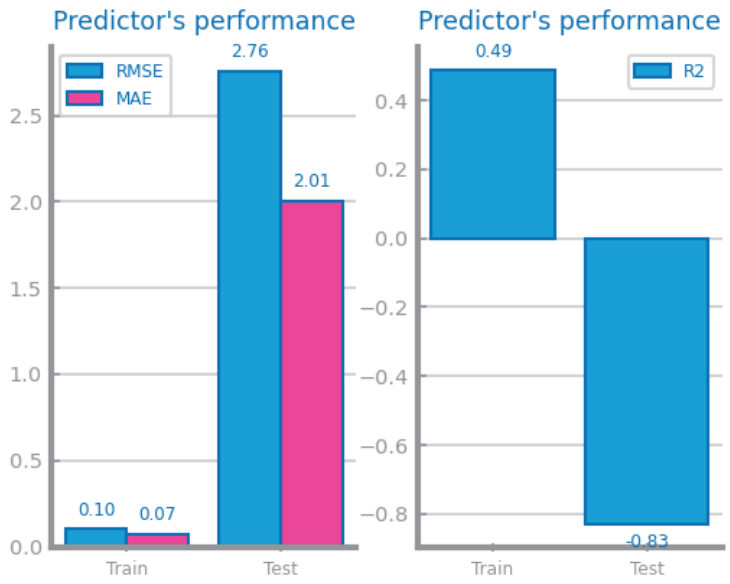


Figure 160: Forecasting results after first differentiation of time series 2

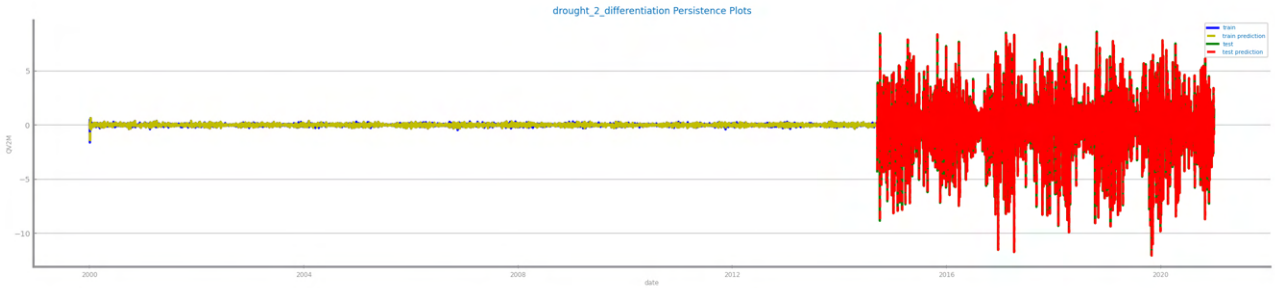


Figure 161: Forecasting plots after second differentiation of time series 2

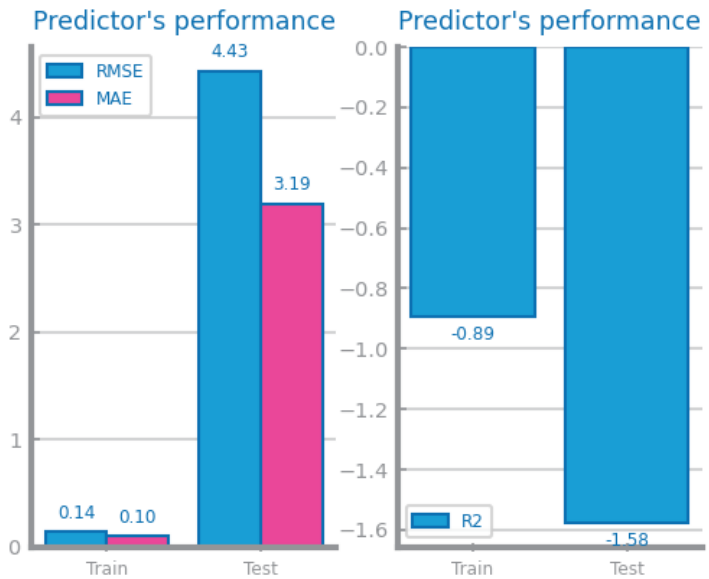


Figure 162: Forecasting results after second differentiation of time series 2

7 MODELS' EVALUATION

We evaluated all models using the original time series, as instructed by the professor. To better evaluate the models and compare results, we decided to split the data always using the same seed.

Simple Average Model

This model makes predictions based on the average of a set of data. It is a good baseline comparison for more complex models. As expected, it does not fit the test data accurately in both sets.

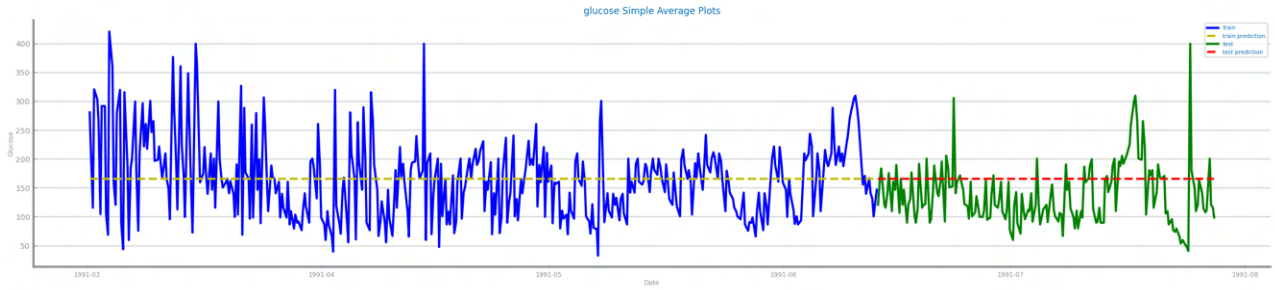


Figure 163: Forecasting plots obtained with Simple Average model over time series 1

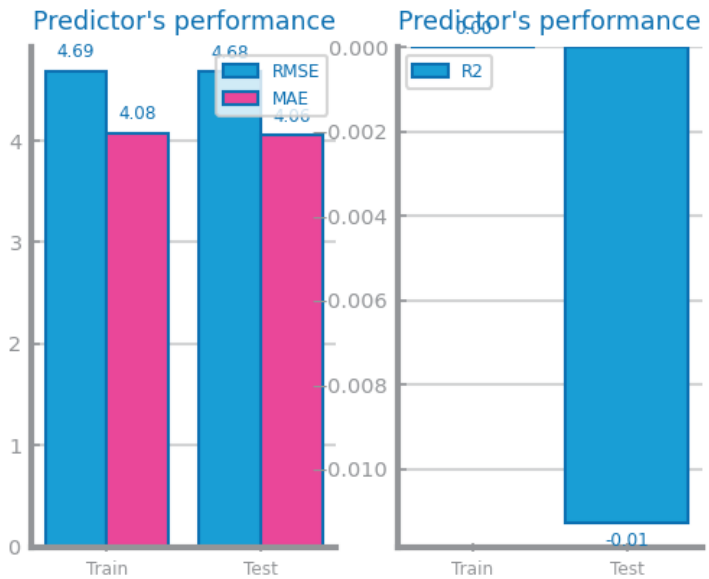


Figure 164: Forecasting results obtained with Simple Average model over time series 1

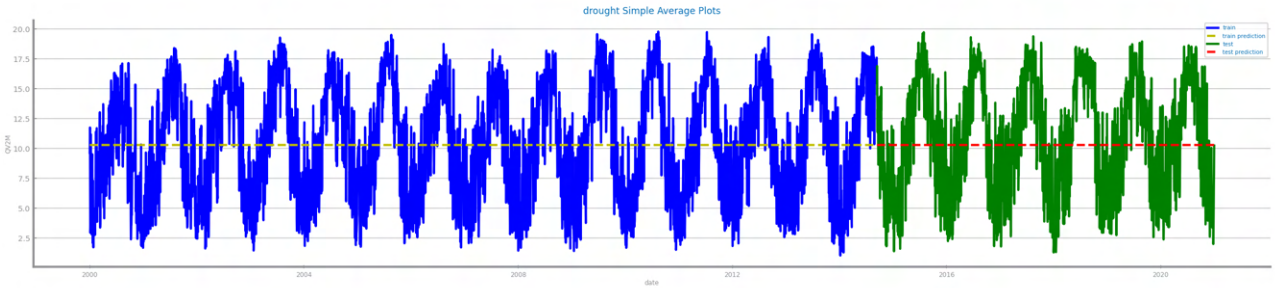


Figure 165: Forecasting plots obtained with Simple Average model over time series 2

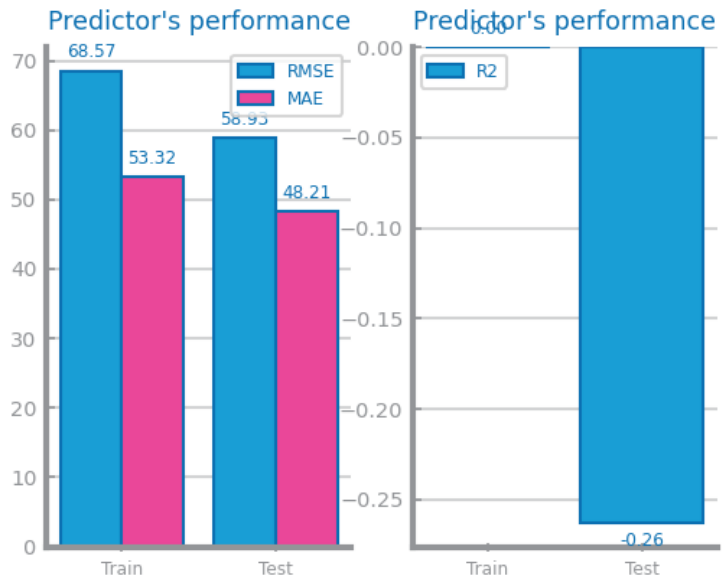


Figure 166: Forecasting results obtained with Simple Average model over time series 2

Persistence Model

This model makes predictions based on the recent past. It performed well because the data isn't very irregular, so the value at t doesn't differ much from the one at t-1.

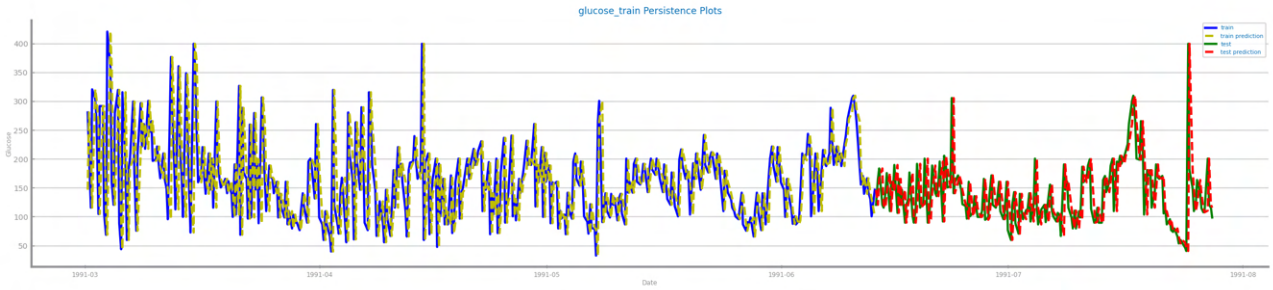


Figure 167: Forecasting plots obtained with Persistence model over time series 1

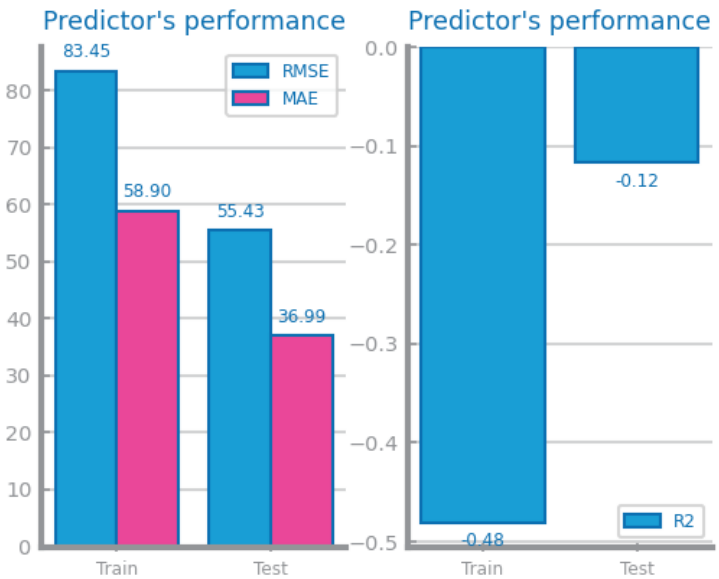


Figure 168: Forecasting results obtained with Persistence model over time series 1

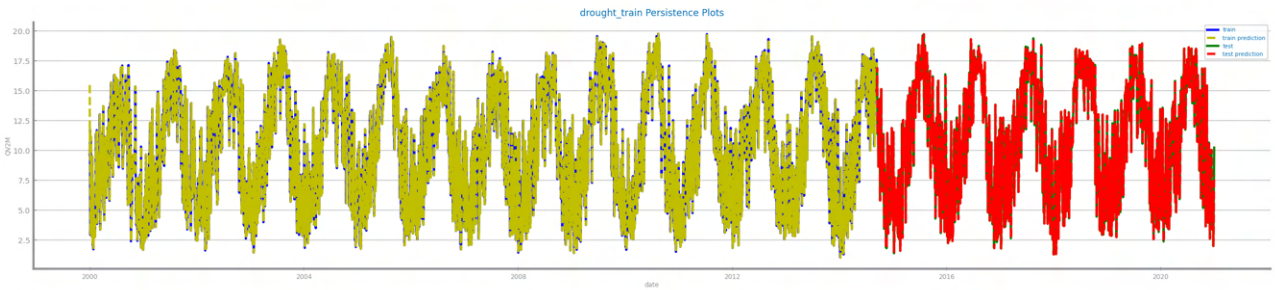


Figure 169: Forecasting plots obtained with Persistence model over time series 2

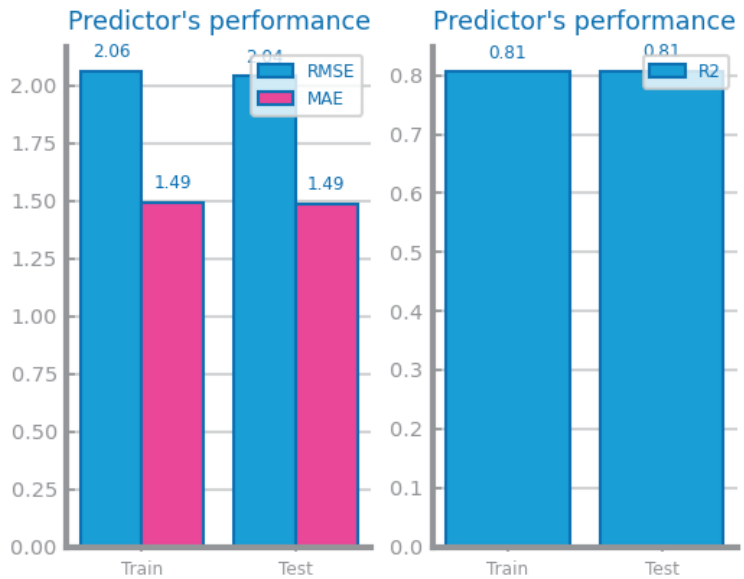


Figure 170: Forecasting results obtained with Persistence model over time series 2

Rolling Mean Model

Results were improved in comparison to the previous regressors for both series as it reduced the influence of fine-grained changes, highlighting the more relevant ones. When the window becomes larger than the periodicity of the data, results become worse - over-smoothing.

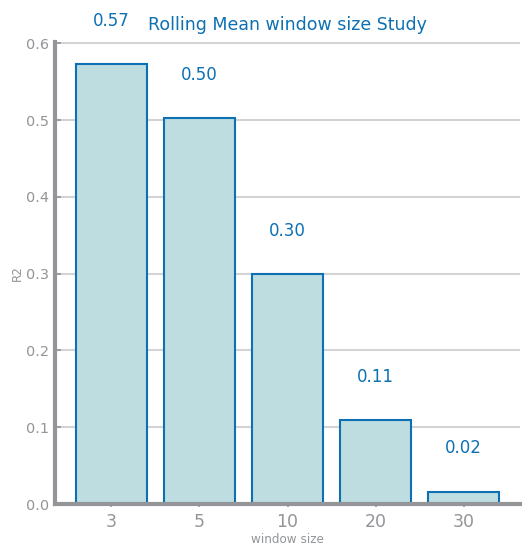


Figure 171: Forecasting study over different parameterizations of the rolling mean algorithm over time series 1

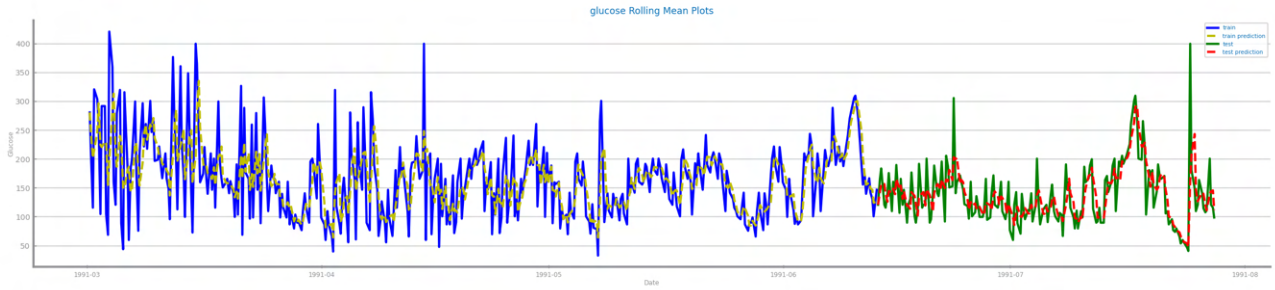


Figure 172: Forecasting plots obtained with the best parameterization of rolling mean algorithm, over time series 1

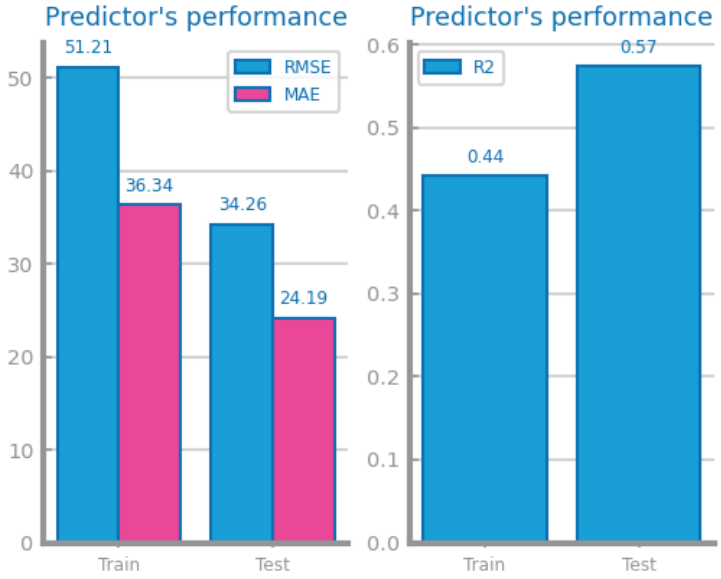


Figure 173: Forecasting results obtained with the best parameterization of rolling mean algorithm, over time series 1

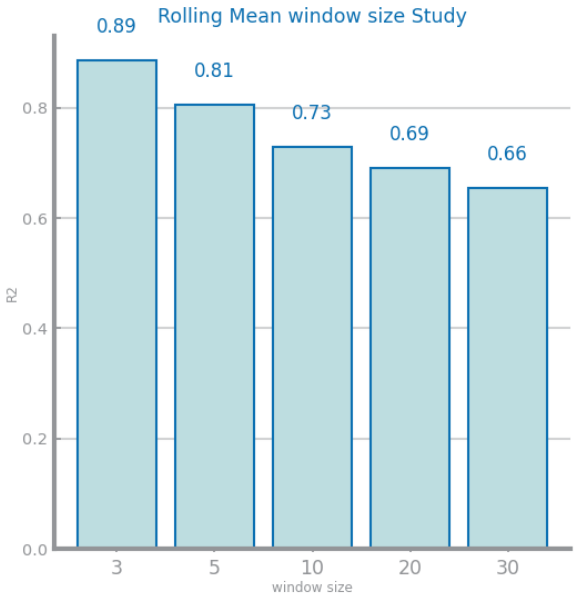


Figure 174: Forecasting study over different parameterizations of the rolling mean algorithm over time series 2

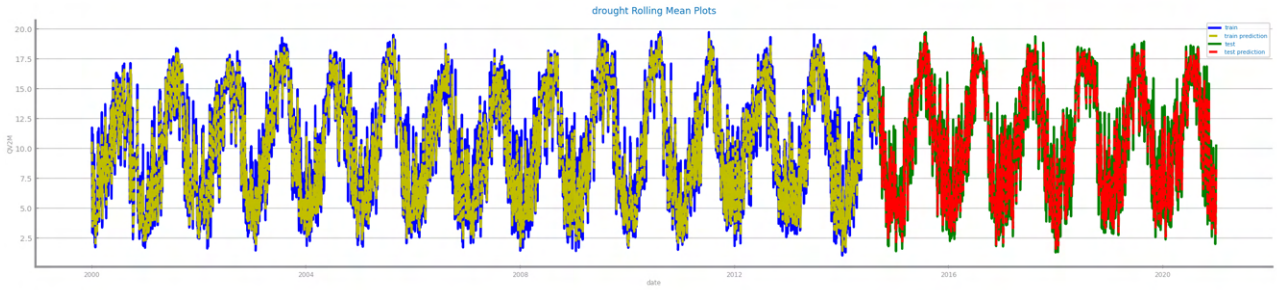


Figure 175: Forecasting plots obtained with the best parameterization of rolling mean algorithm, over time series 2

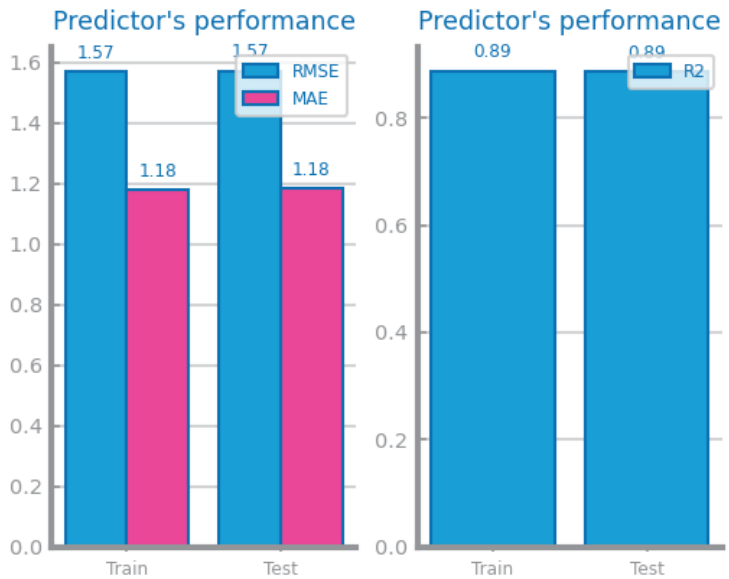


Figure 176: Forecasting results obtained with the best parameterization of rolling mean algorithm, over time series 2

ARIMA Model

ARIMA assumes data is stationary and benefits from data that has temporal dependencies. From the plots we can see the model predicted a stabilized value probably because it is relying on temporal dependent variables, which could be not necessarily true. However, the performance obtained was good.

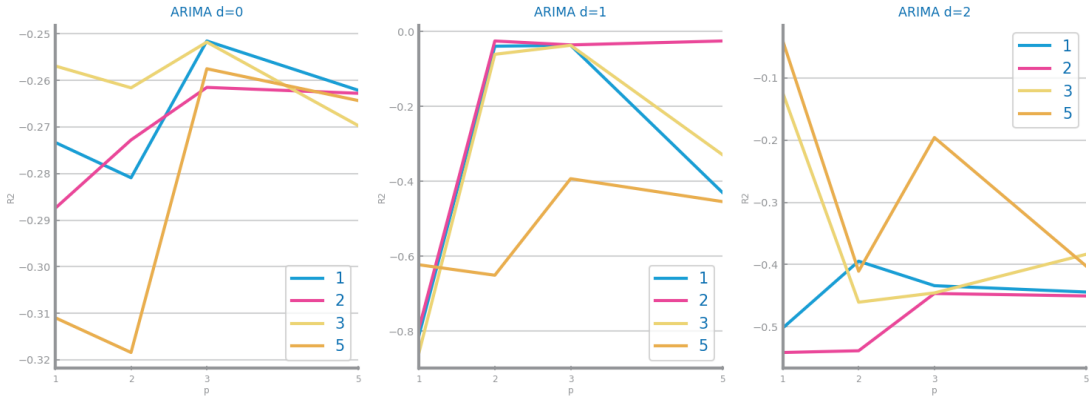


Figure 177: Forecasting study over different parameterizations of the ARIMA algorithm over time series 1

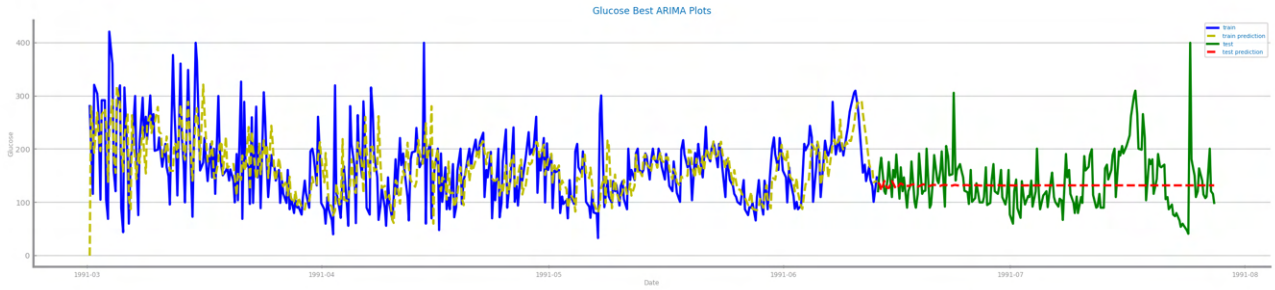


Figure 178: Forecasting plots obtained with the best parameterization of ARIMA algorithm, over time series 1

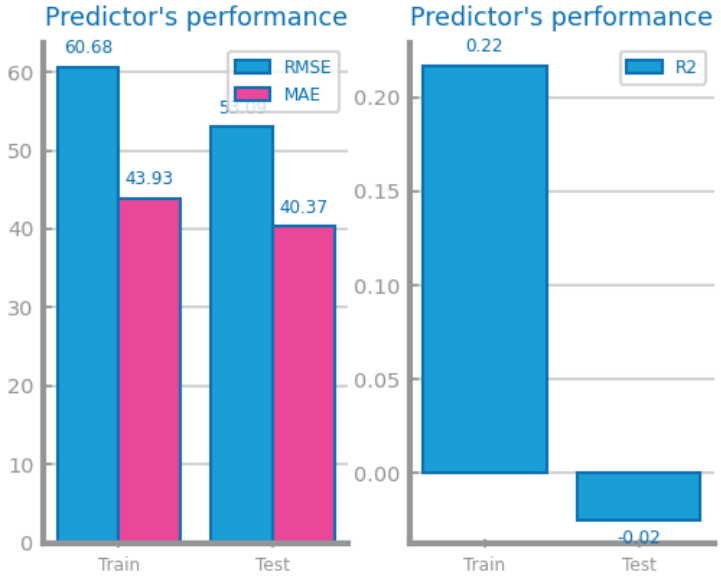


Figure 179: Forecasting results obtained with the best parameterization of ARIMA algorithm, over time series 1

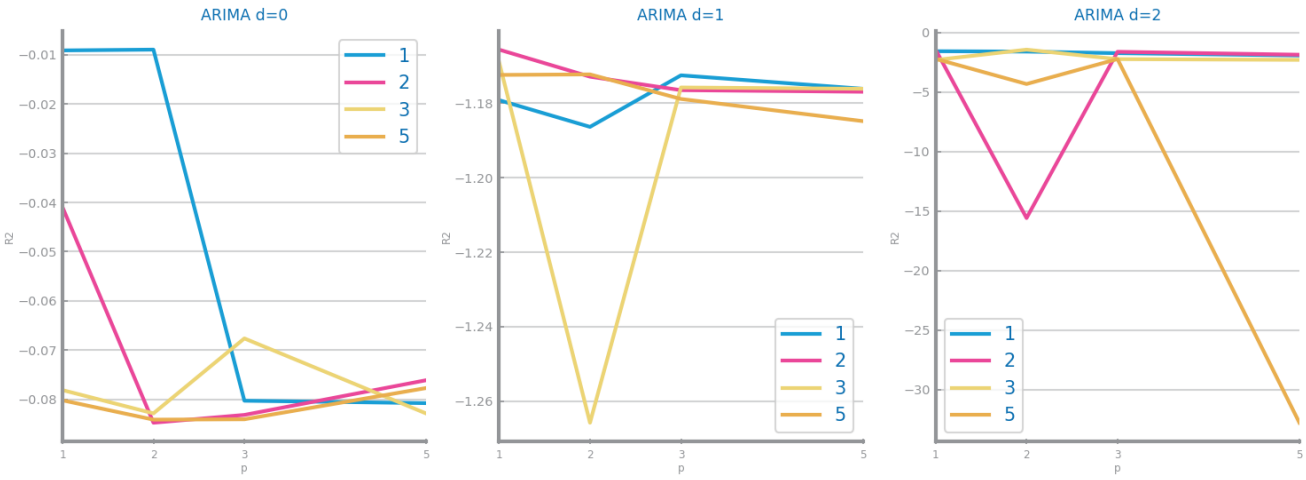


Figure 180: Forecasting study over different parameterizations of the ARIMA algorithm over time series 2

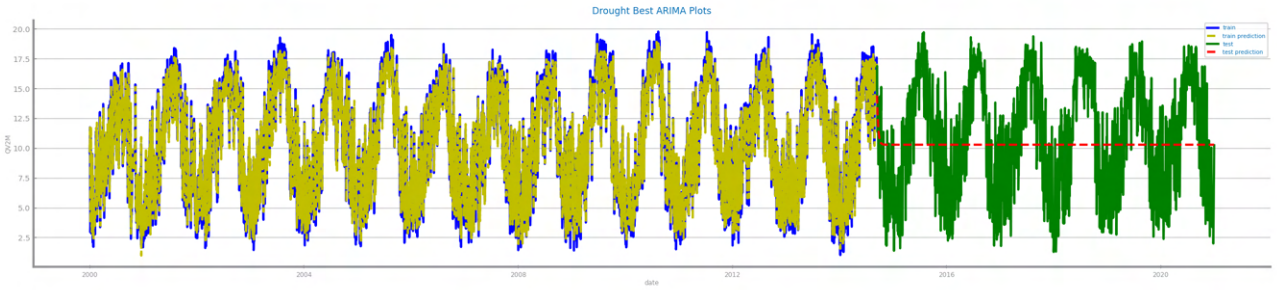


Figure 181: Forecasting plots obtained with the best parameterization of ARIMA algorithm, over time series 2

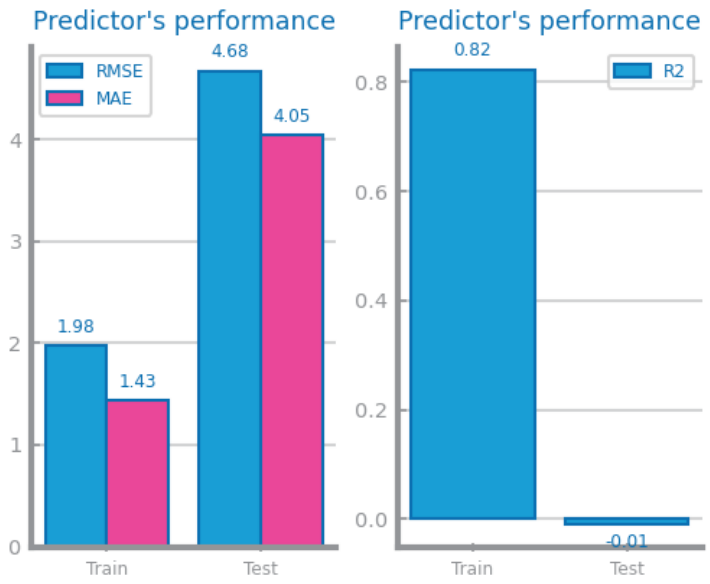


Figure 182: Forecasting results obtained with the best parameterization of ARIMA algorithm, over time series 2

LSTMs Model

Results weren't great in set 1 because it's small (and LSTMs need more data). As for set 2, it accurately predicted the values of the time series, indicating there are dependencies in the sequential data (which is true because drought is a cyclical phenomenon).

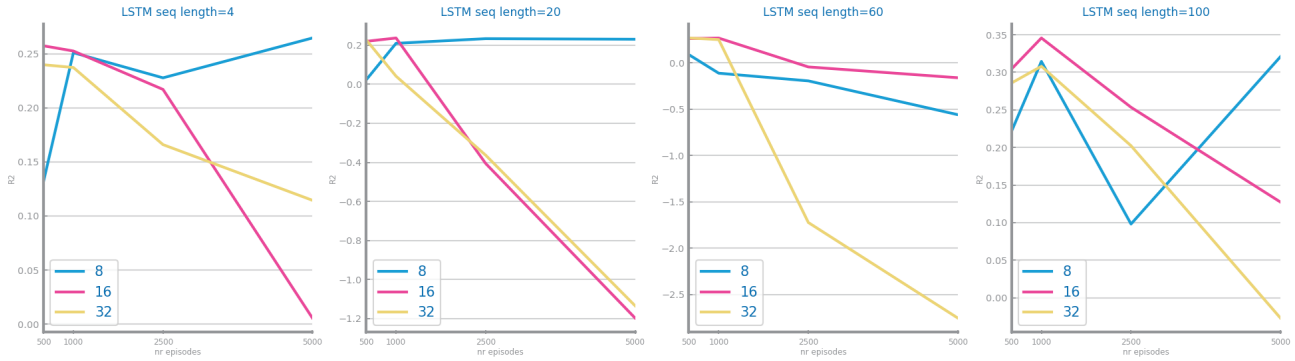


Figure 183: Forecasting study over different parameterizations of LSTMs over time series 1

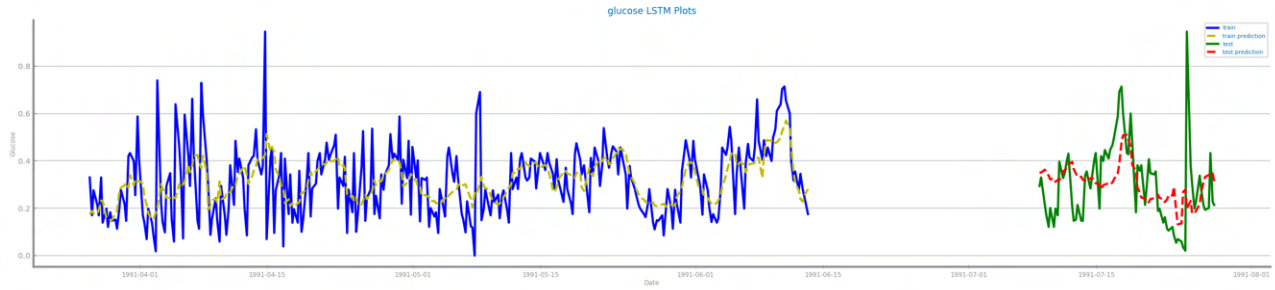


Figure 184: Forecasting plots obtained with the best parameterization of LSTMs, over time series 1

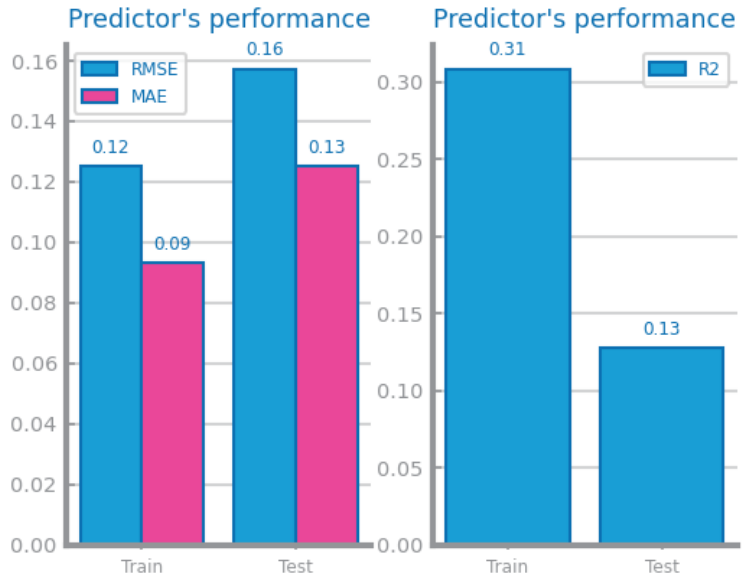


Figure 185: Forecasting results obtained with the best parameterization of LSTMs, over time series 1

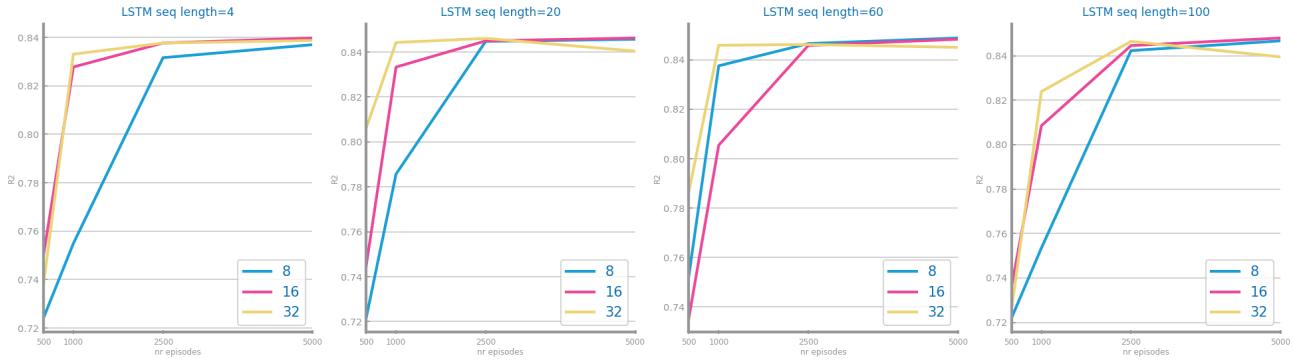


Figure 186: Forecasting study over different parameterizations of the LSTMs over time series 2

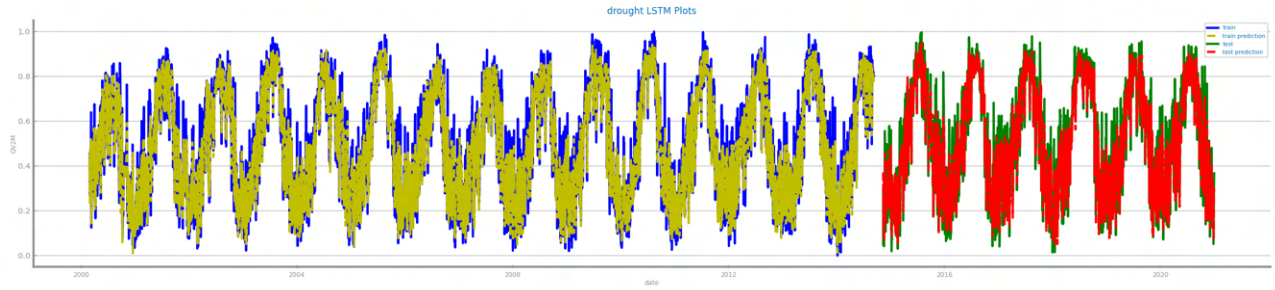


Figure 187: Forecasting plots obtained with the best parameterization of LSTMs, over time series 2

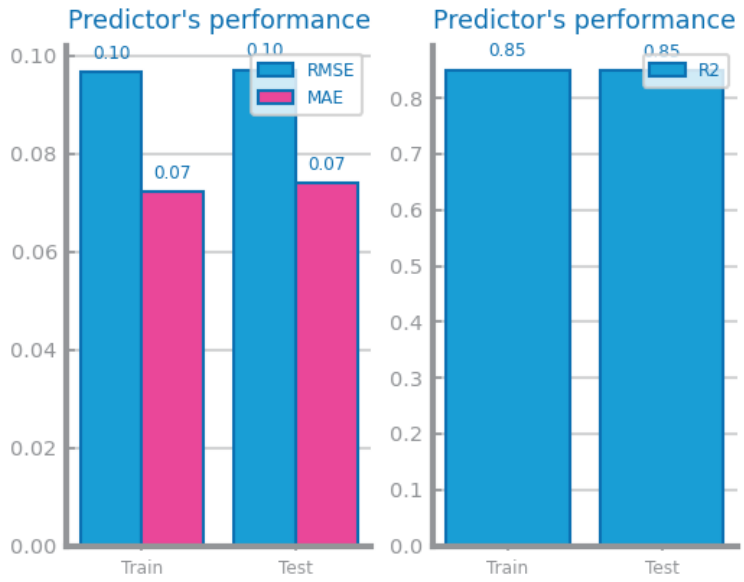


Figure 188: Forecasting results obtained with the best parameterization of LSTMs, over time series 2

8 CRITICAL ANALYSIS

The preparation tasks didn't affect performance because we evaluated all models with the original sets. One improvement we could make is to extract the better transformation and run models on it, i.e., with the most atomic aggregation, smoothing with window size 10 in set 1 and 30 in set 2, and no differentiation.

Given the dimension and periodicity of our data, Rolling Mean proved to be the best model as it is shown on 189. This simple model captured just enough of the similarities between close past samples and extrapolated them to the recent (test) ones.

The simple statistical methods such as Simple Average, Persistence Model and Rolling Mean performed better than the complex ones, contrary to what one may expect. The accuracy of forecasting complex regressors can be improved by using more data, more sophisticated models and fine-tuning the parameters. This could have even a greater impact on dataset 1 as it has very few records.

In almost all models, forecasted values were among what we could expect. However, there was one exception: the ARIMA

model. It was only able to predict a constant value for all testing samples. The LSTM model produced a big gap in the forecasted values. One possible reason for this is the lack of data as it may not be able to learn the underlying patterns in the data, leading to poor predictions and a big gap in the forecasted values.

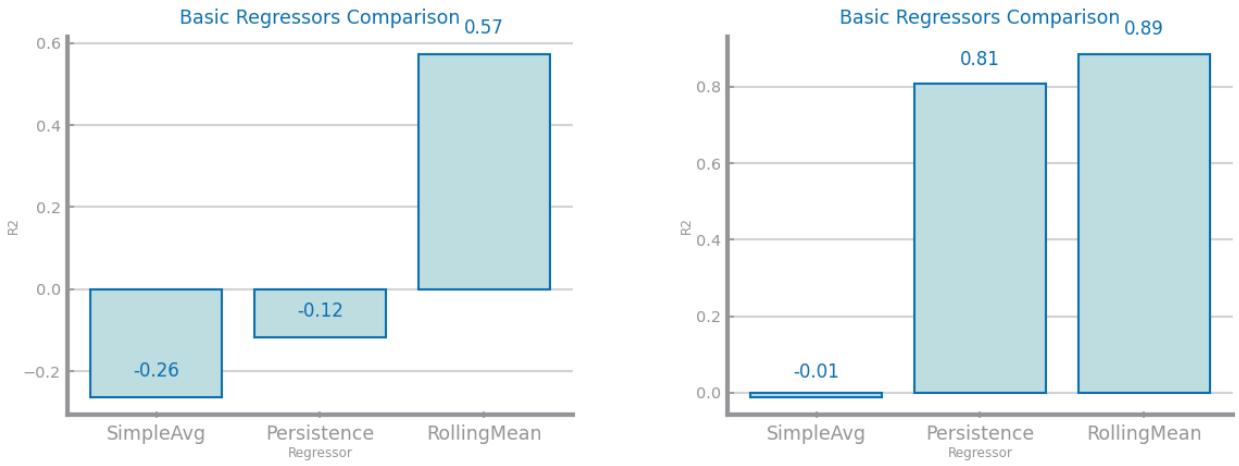


Figure 189: Regressors comparisson on dataset 1 (left) and dataset 2 (right)