TÉCNICO LISBOA

DS 2023

Data Science Project

Team nr: 38	**Student 1 :** Francisco Guilherme	**IST nr:** 99069
	Student 2 : Henrique Anjos	**IST nr:** 99081
	Student 3 : Luis Molina	**IST nr:** 105438
	Student 4 : Tomé Agostinho	**IST nr:** 96948

CLASSIFICATION

1 DATA PROFILING

Dataset 1 (D1) is in the Health Domain and the pos-covid situation (target = CovidPos). Dataset 2 (D2) is in the Services Domain with credit data collected by a finance company (target = Credit_Score).

Data Dimensionality

D1 has 40 variables (10 symbolic, 24 binary, 6 numeric) and 380932 records. Variables like “BMI” or “TetanusLast10Tdap” have high missing values, impacting dimensionality. The Curse of Dimensionality risks computational complexity and overfitting. D2 has 28 variables (12 symbolic, 1 binary, 15 numeric) and 100000 records. Variables like "Credit Mix" show significant missing values, requiring aware preprocessing. Both highlight the importance of managing missing data and dimensionality effects.

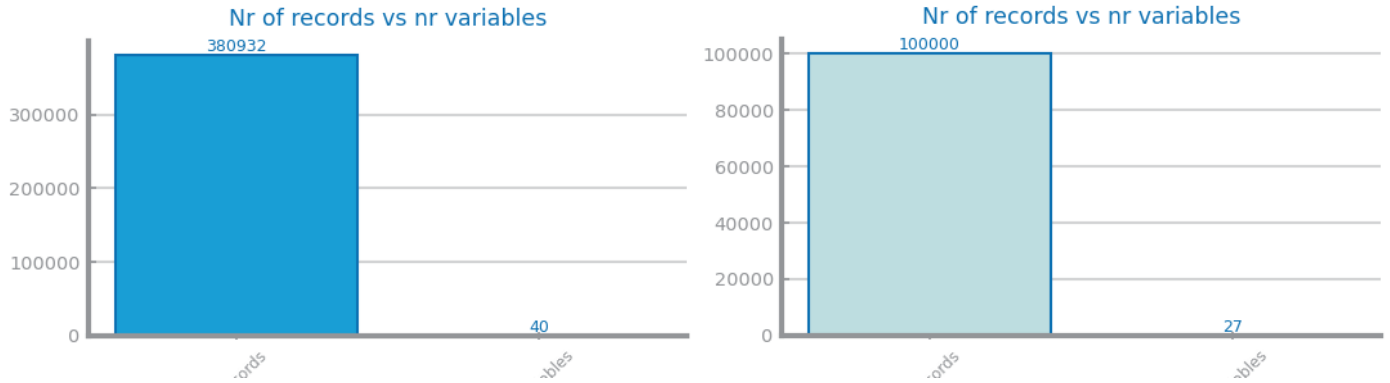


Figure 1 Nr Records x Nr variables for dataset 1 (left) and dataset 2 (right)

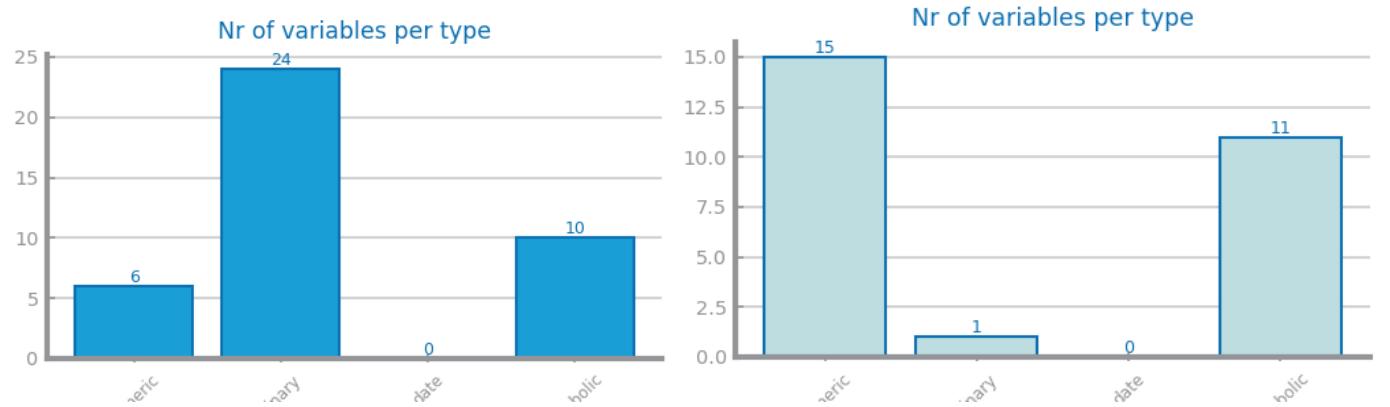
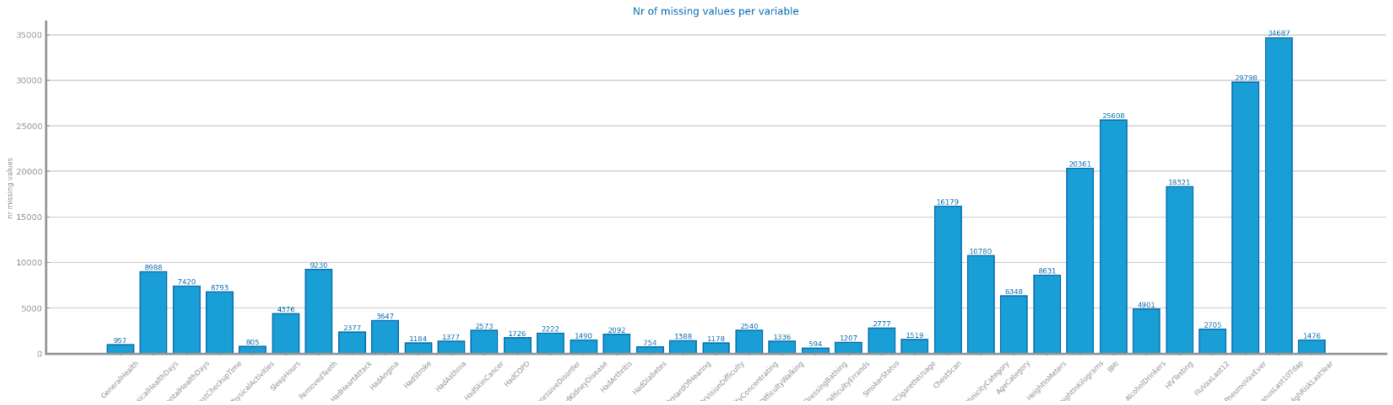


Figure 2 Nr variables per type for dataset 1 (left) and dataset 2 (right)



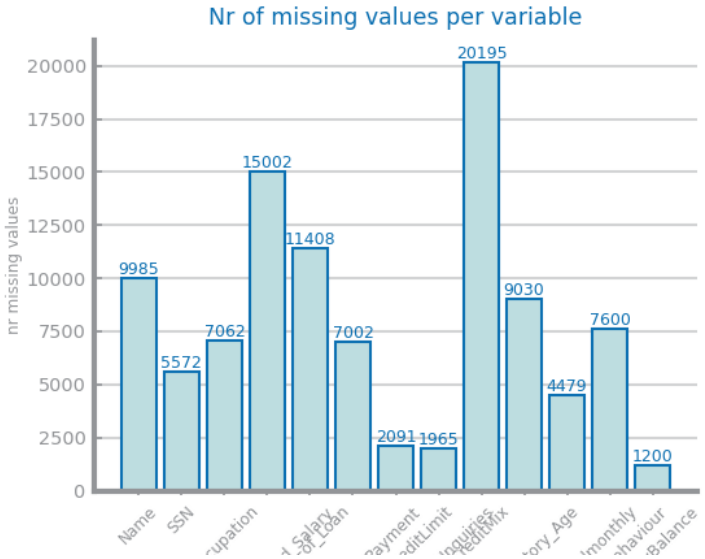


Figure 3 Nr missing values for dataset 1 (up) and dataset 2 (bottom)

Data Distribution

D1 shows varied scales and distributions across variables. Binary and symbolic variables require scale transformations. Outlier analysis, particularly for 'PhysicalHealthDays' and 'MentalHealthDays', reveals varying sensitivities based on threshold adjustments. D2 also shows differing scales, with symbolic variables needing transformation. Outliers like 'Annual_Income' and 'OutstandingDebt' vary in counts across detection methods, emphasizing threshold sensitivity and data preprocessing needs.

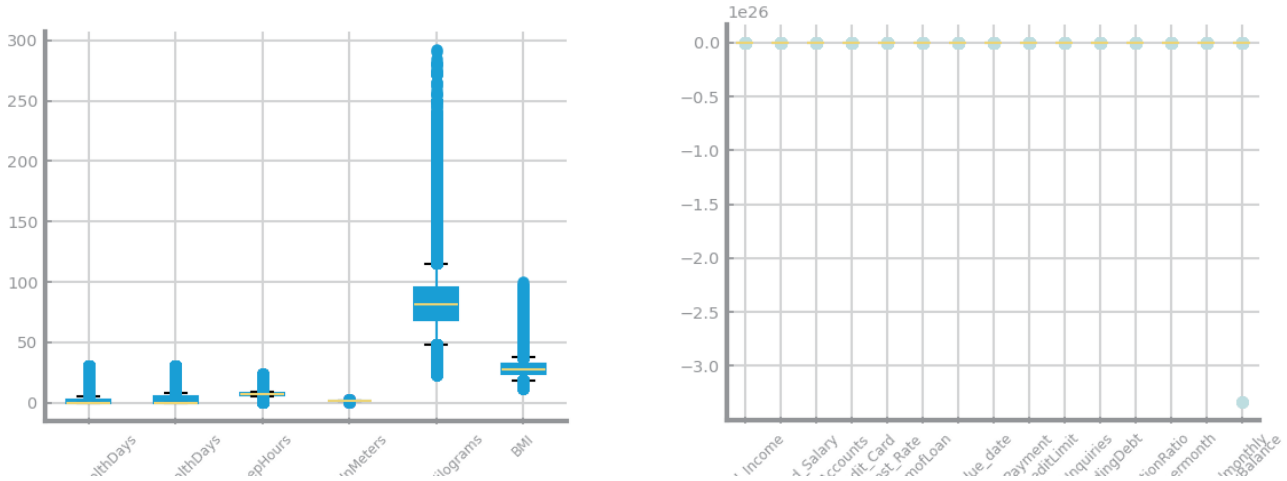


Figure 4 Global boxplots dataset 1 (left) and dataset 2 (right)

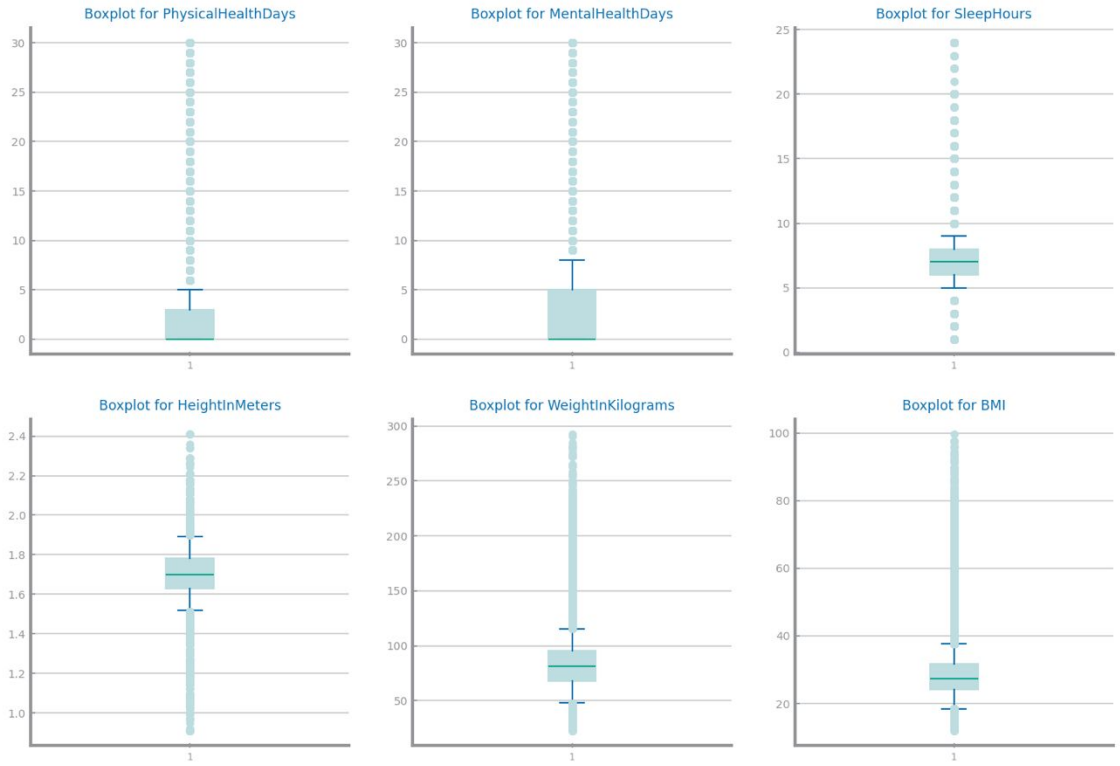


Figure 5 Single variable boxplots for dataset 1

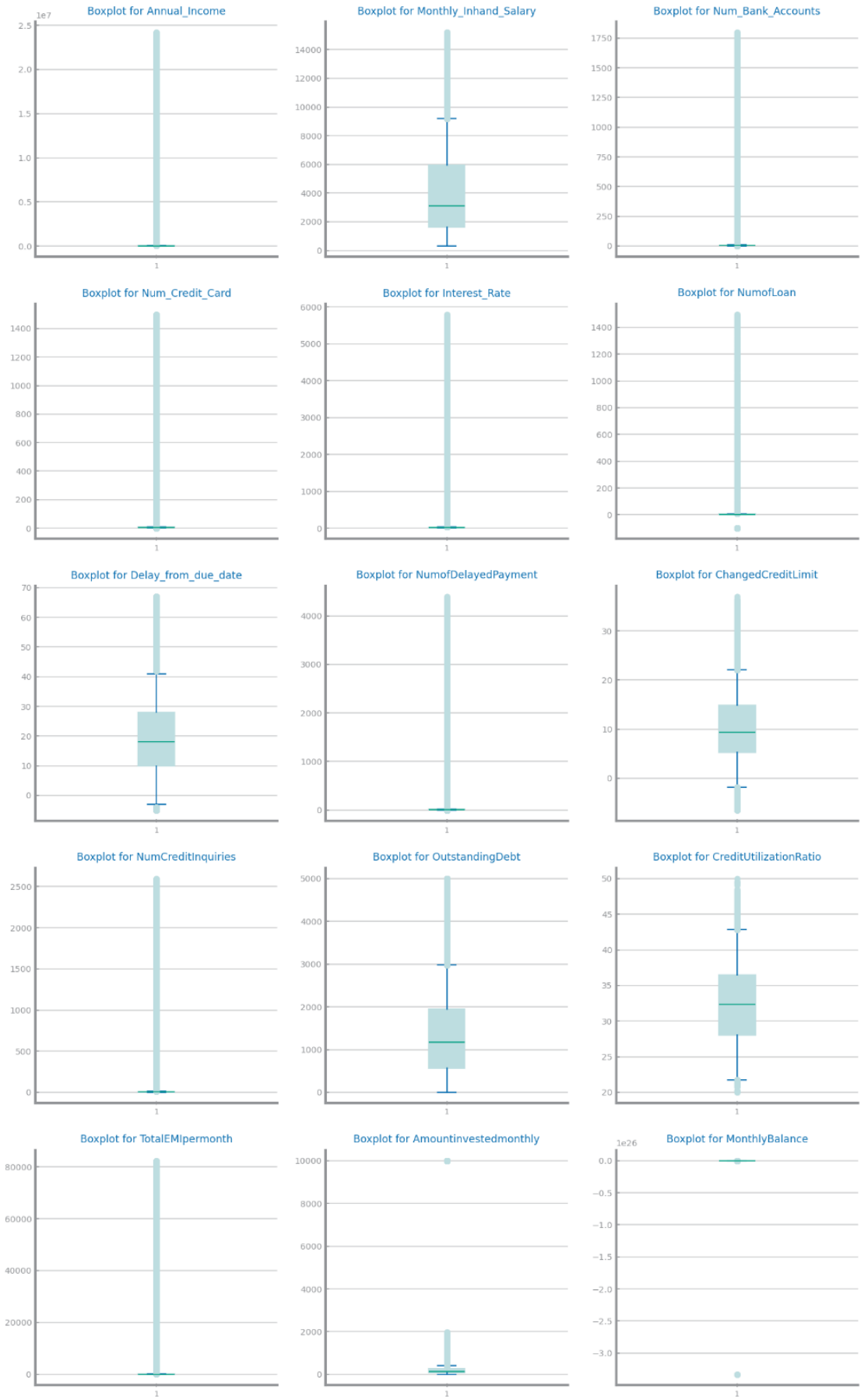


Figure 6 Single variable boxplots s for dataset 2

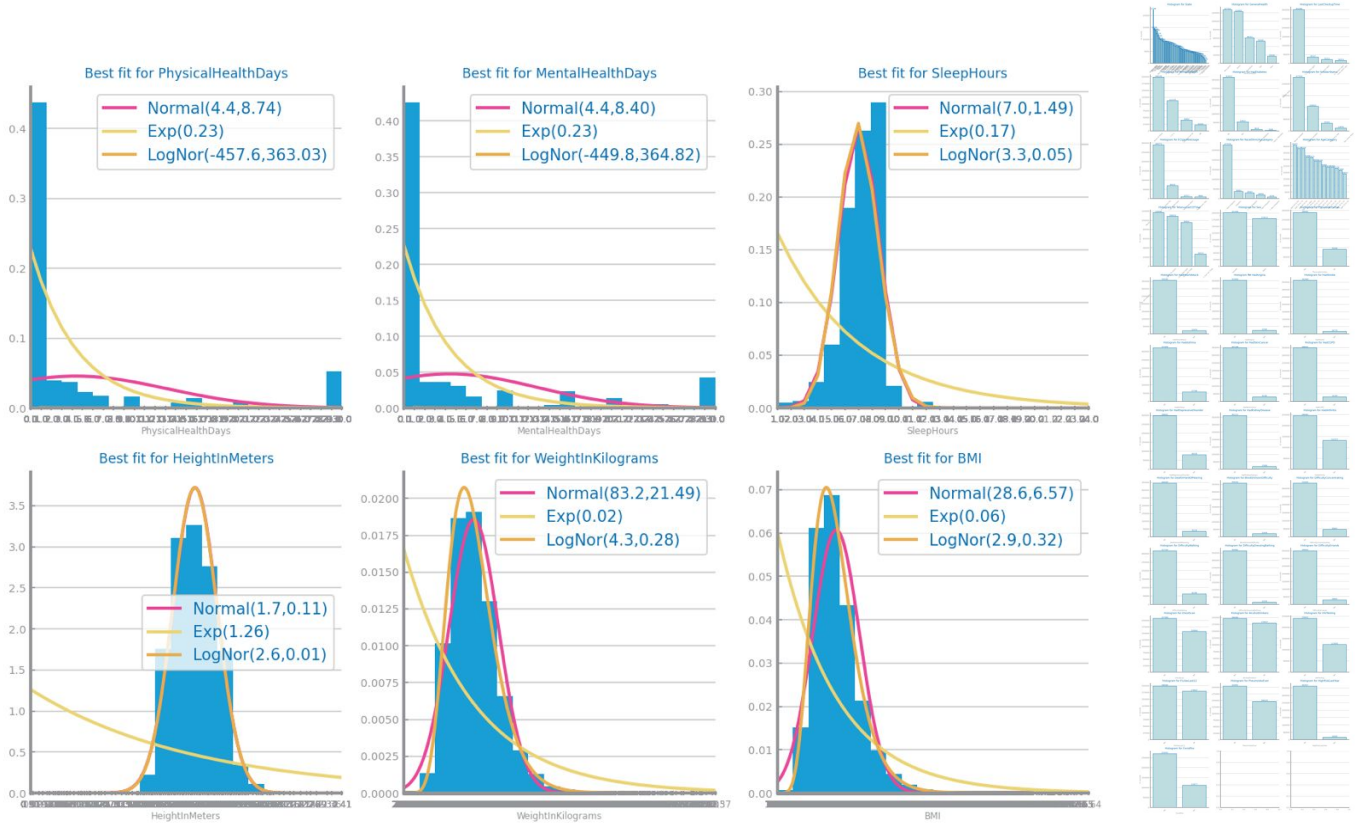


Figure 7 Histograms for dataset 1 (with distributions is enough)

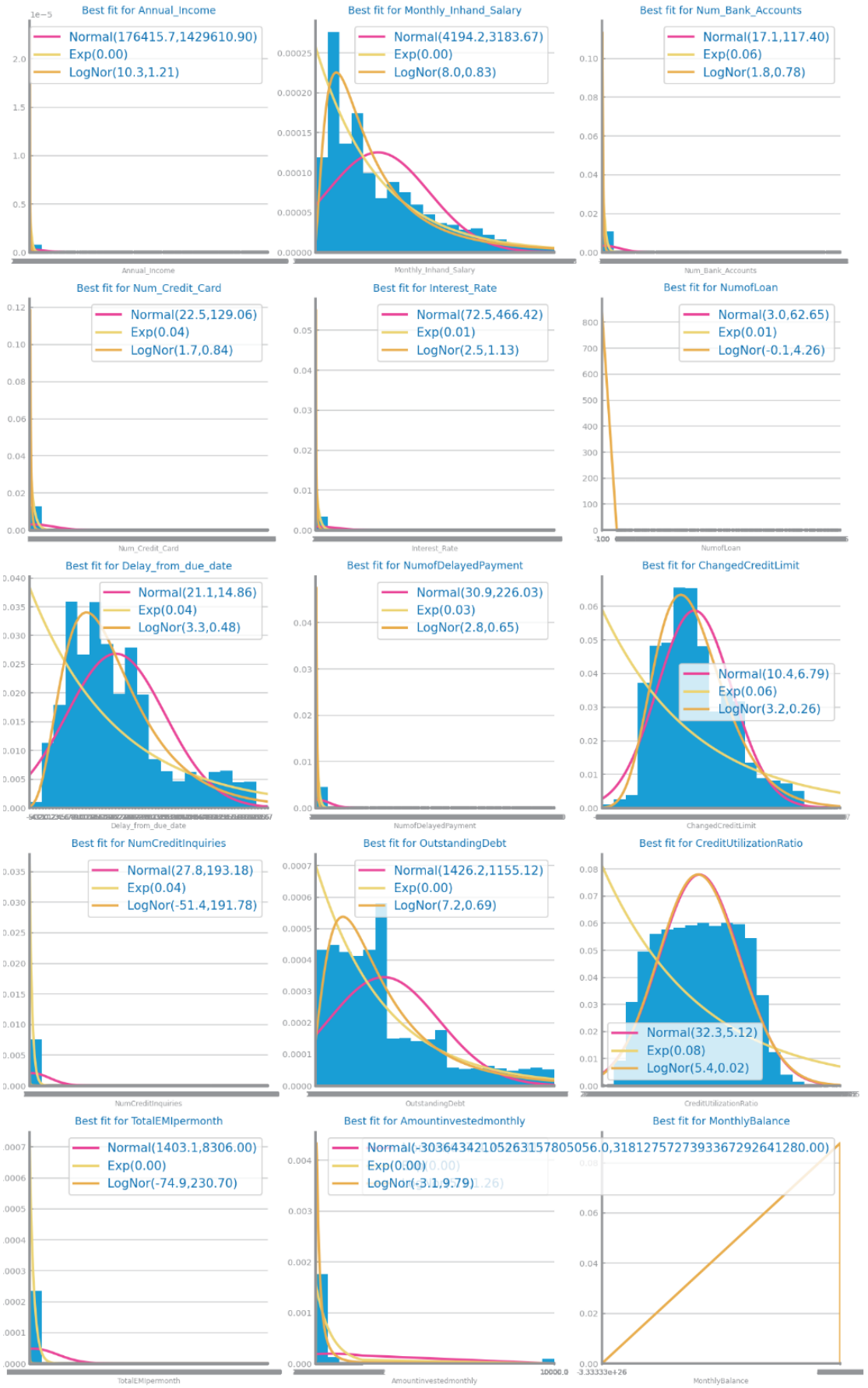


Figure 8 Histograms for dataset 2 (with distributions is enough)

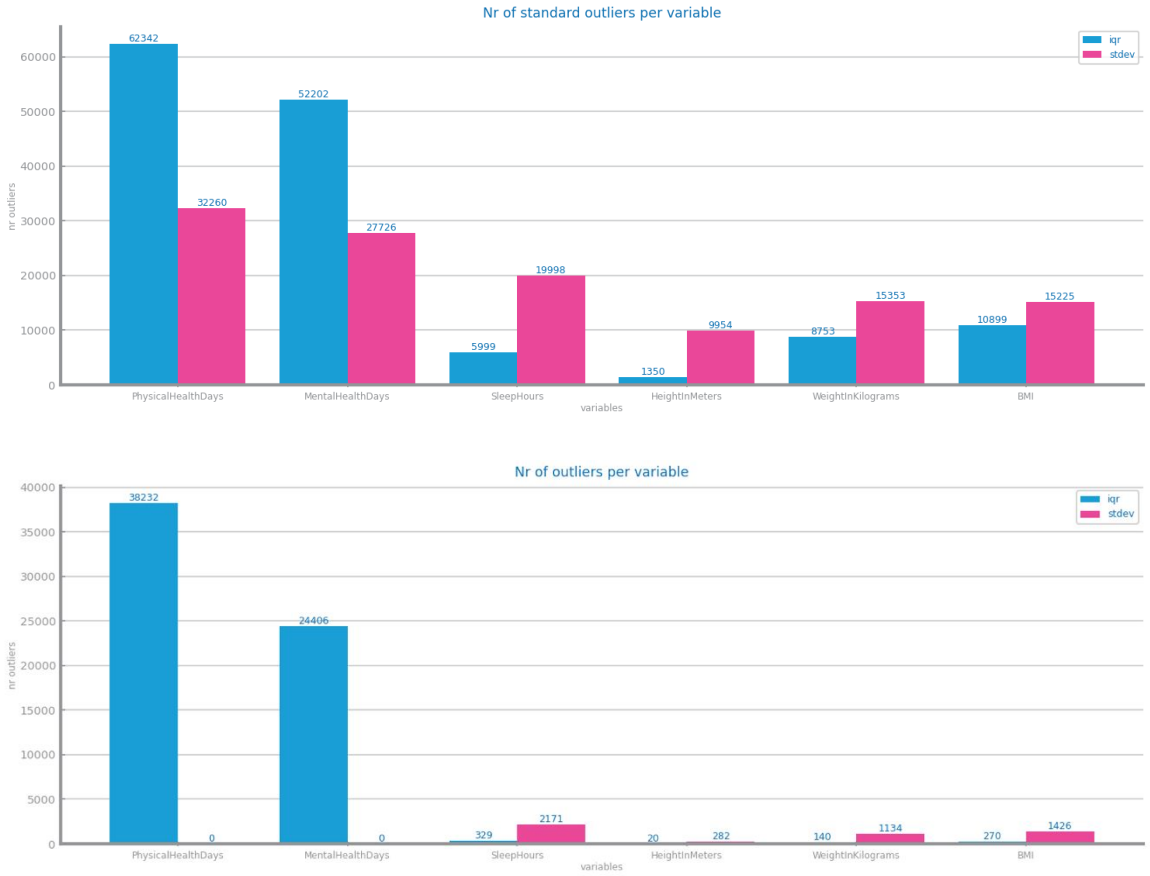


Figure 9 Outliers study dataset 1

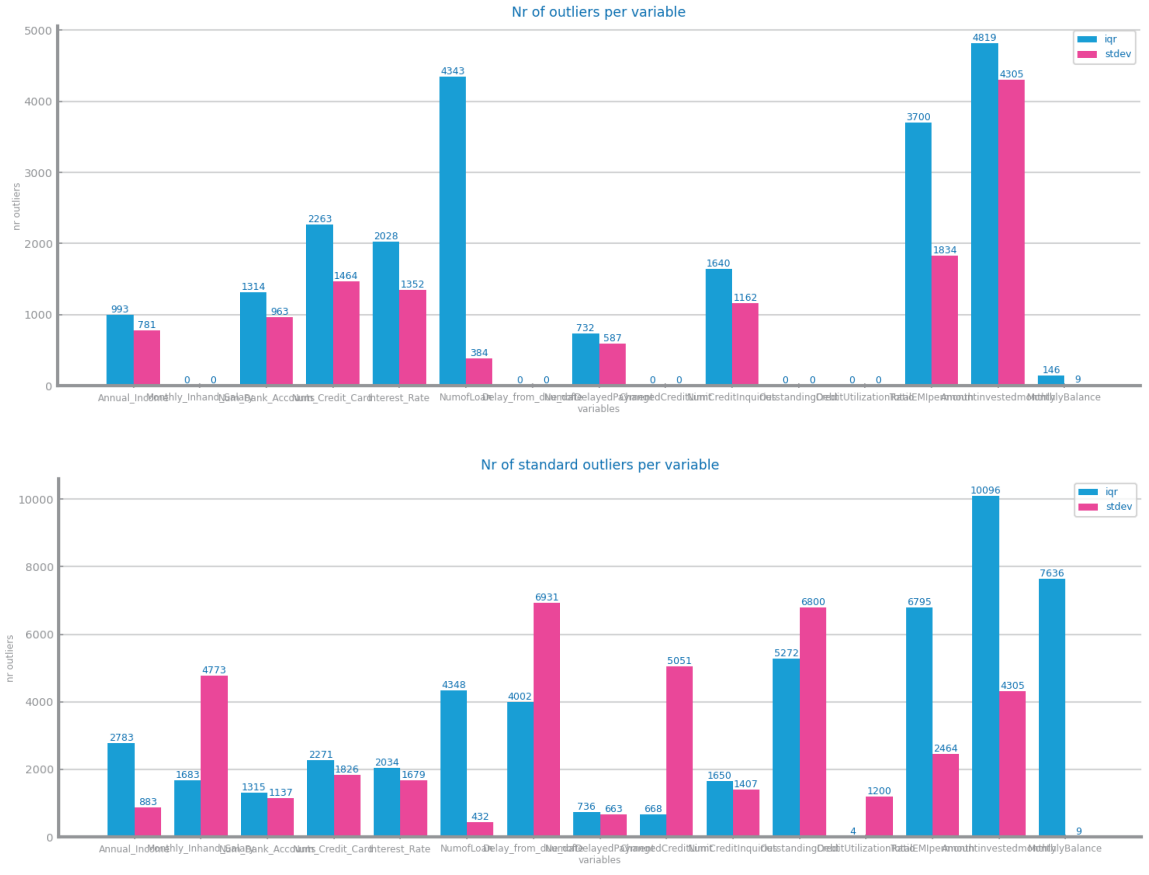


Figure 10 Outliers study for dataset 2

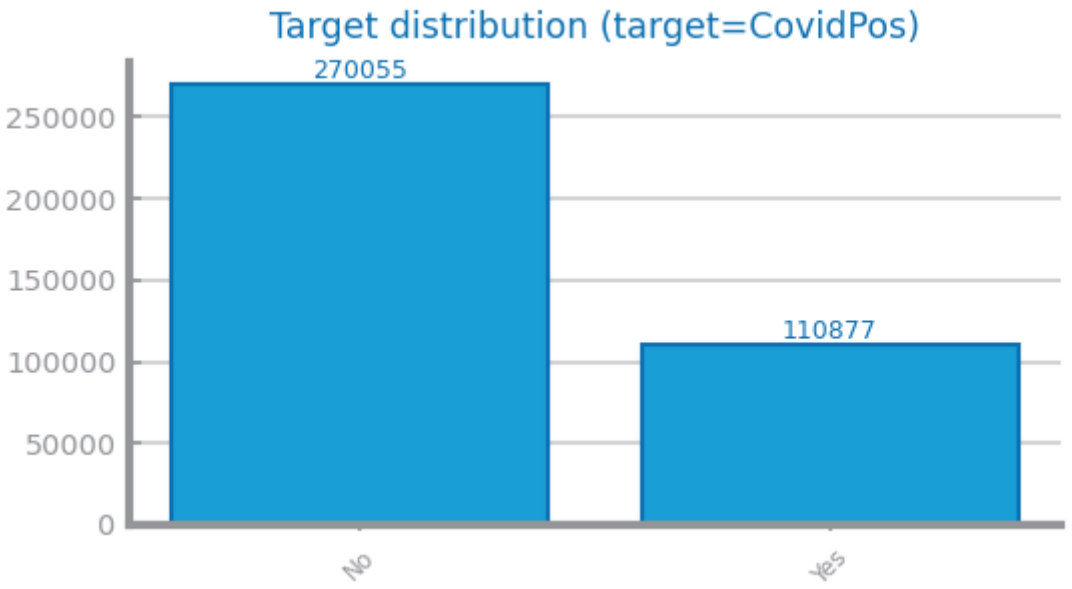


Figure 11 Class distribution for dataset 1

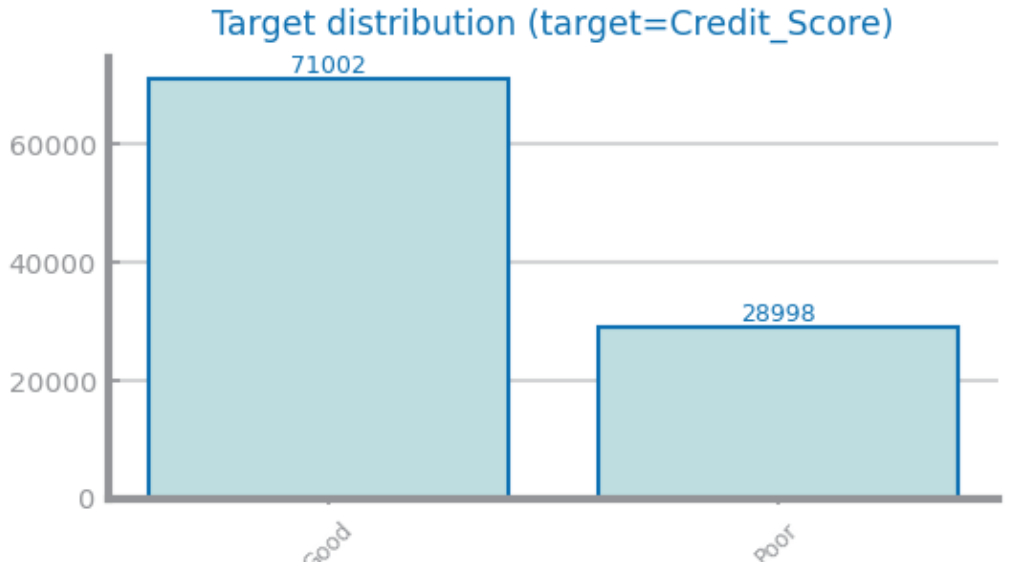


Figure 12 Class distribution for dataset 2

Data Granularity

Dataset 1 and Dataset 2 exhibit structured patterns in categorical variables, revealing hierarchies in demographics, health, and lifestyle attributes. Geographic and temporal trends are evident, aiding interpretation and decision-making. However, anomalies like negative or alphanumeric age values in Dataset 2 highlight data quality concerns. Identifying and addressing these anomalies is crucial for accurate analysis and meaningful insights.

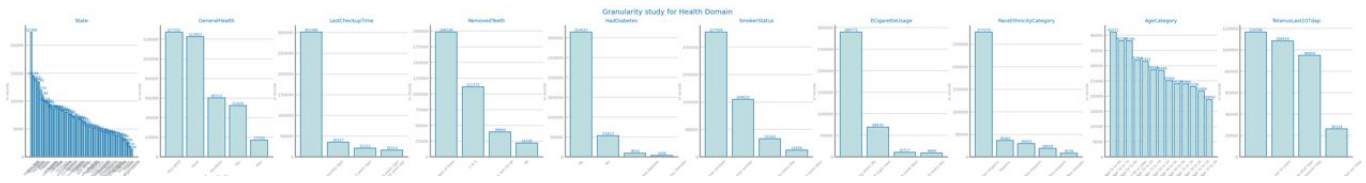


Figure 13 Granularity analysis for dataset 1

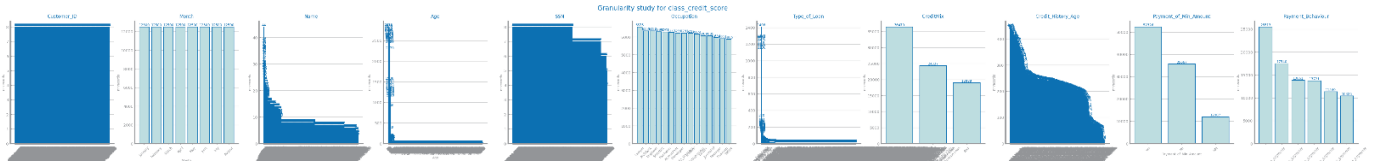


Figure 14 Granularity analysis for dataset 2

Data Sparsity

Dataset 1 shows strong intra-health variable correlation; BMI notably correlated with health. Dataset 2 shows strong correlations between financial variables, with significant relationships between debt, due date, and credit factors. Regarding the variable sparsity, Dataset 2 exhibits greater density in its correlations, suggesting that its variables may hold more significant insights compared to those in Dataset 1.

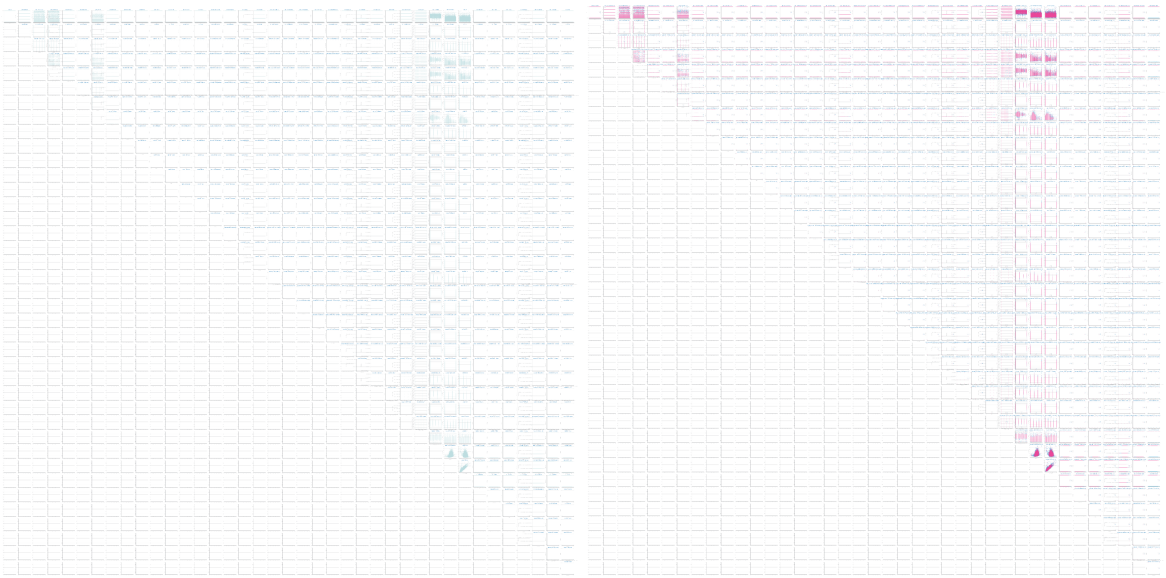


Figure 15 Sparsity analysis for dataset 1

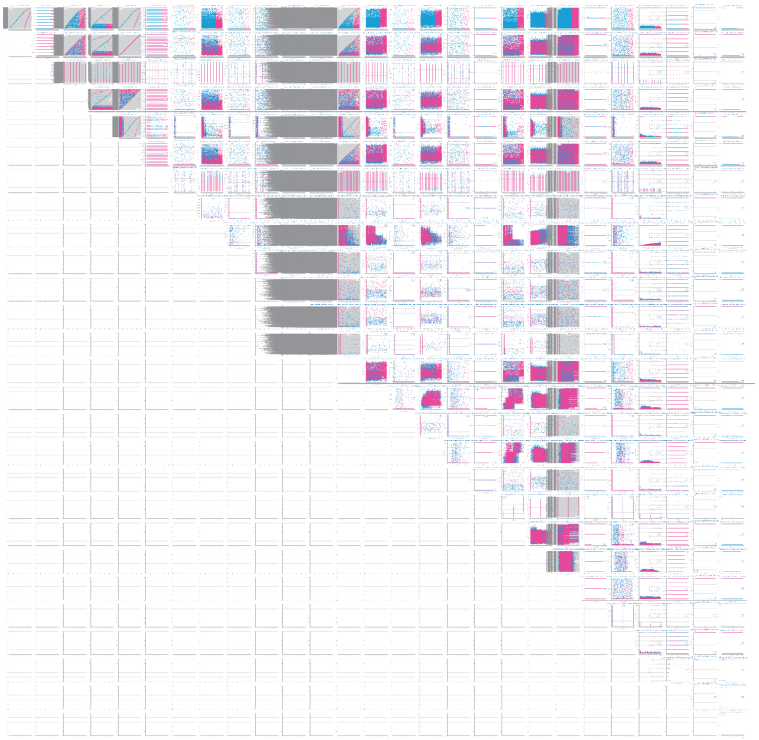


Figure 16 Sparsity analysis for dataset 2

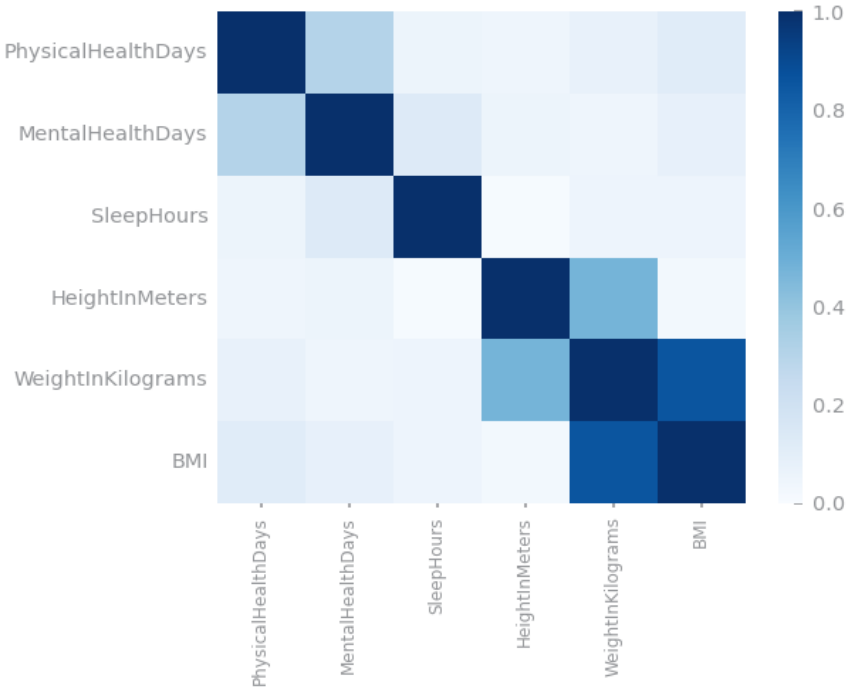


Figure 17 Correlation analysis for dataset 1

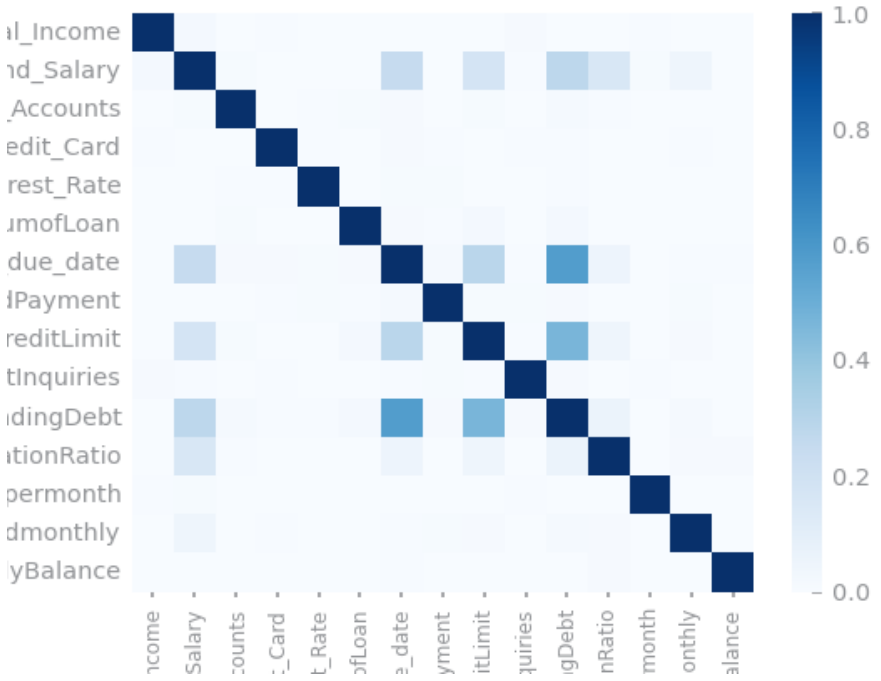


Figure 18 Correlation analysis for dataset 2

2 Data Preparation

Variables Encoding

Dataset 1:
6 variables were left as they were; 32 variables were ordinal linear encoded; 1 variable was ordinal encoded based on taxonomy; variable “State” was transformed into 3 new variables after cross_checking information on the dataset.

Dataset 1	AS-IS	ORDINAL LINEAR ENCODING	ORDINAL ENCODING BASED ON TAXONOMY	OTHER TRANSFORMATION	NEW VARIABLES
State				Dataset crossing to create 3 numeric variables	HealthCareQuality HealthCareAccess PublicHealth
Sex		x			
GeneralHealth		x			
PhysicalHealthDays	x				
MentalHealthDays	x				
LastCheckupTime		x			
PhysicalActivities		x			
SleepHours	x				
RemovedTeeth		x			
HadHeartAttack		x			
HadAngina		x			
HadStroke		x			
HadAsthma		x			
HadSkinCancer		x			
HadCOPD		x			
HadDepressiveDisorder		x			
HadKidneyDisease		x			
HadArthritis		x			
HadDiabetes		x			
DeafOrHardOfHearing		x			
BlindOrVisionDifficulty		x			
DifficultyConcentrating		x			
DifficultyWalking		x			
DifficultyDressingBathing		x			
DifficultyErrands		x			
SmokerStatus		x			
ECigaretteUsage		x			
ChestScan		x			
RaceEthnicityCategory			x		
AgeCategory		x			
HeightInMeters	x				
WeightInKilograms	x				
BMI	x				
AlcoholDrinkers		x			
HIVTesting		x			
FluVaxLast12		x			
PneumoVaxEver		x			
TetanusLast10Tdap		x			
HighRiskLastYear		x			
CovidPos		x			

Dataset 2:

16 variables were left as they were; “Name” and “SSN” were dropped because they were irrelevant to the target variable (due to the existence of Costumer_ID); 3 variables were ordinal linear encoded; “Month” was cyclic encoded; "Costumer_ID" and "Age" required cleaning (irrelevant characters); variable "Costumer_ID" was converted to decimal and "Credit_History_Age" to months; 2 variables were split into dummy variables for each category; "MonthlyBalance" was split into 2 numeric variables.

Dataset 2	AS-IS	DROP	ORDINAL LINEAR ENCODING	CYCLIC ENCODING	OTHER TRANSFORMATION	NEW VARIABLES
ID	x					
Customer_ID					Clean Data and convert from hexadecimal to decimal	
Month				x		
Name		x				
Age					Clean Data	
SSN		x				
Occupation			x			
Annual_Income	x					
Monthly_Inhand_Salary	x					
Num_Bank_Accounts	x					
Num_Credit_Card	x					
Interest_Rate	x					
NumofLoan	x					
Type_of_Loan					Create a dummy variable for each type of loan	Not Specified; Mortgage Loan; Auto Loan; Pesonal Loan; Debt Consolidation Loan; Student Loan; Home Equity Loan; Payday Loan; Credit-Builder Loan
Delay_from_due_date	x					
NumofDelayedPayment	x					
ChangedCreditLimit	x					
NumCreditInquiries	x					
CreditMix			x			
OutstandingDebt	x					
CreditUtilizationRatio	x					
Credit_History_Age					Convert to months and create numeric variable	Credit_History_Age_Numeric
Payment_of_Min_Amount	x					
TotalEMIpermonth	x					
Amountinvestedmonthly	x					
Payment_Behaviour					Create a dummy variable for each Payment Behaviour	Low_spent_Small_value_payments High_spent_Medium_value_payments Low_spent_Medium_value_payments High_spent_Large_value_payments High_spent_Small_value_payments Low_spent_Large_value_payments
MonthlyBalance					Create 1 non-negative variable and 1 non-positive variable	Positive_Monthly_Balance Negative_Monthly_Balance
Credit_Score			x			

Missing Value Imputation

Two approaches were used for missing value handling in datasets 1 and 2. App. 1 removed variables with <10% missing values, preserving data integrity. App. 2 kept records with <1.5% missing values, imputing with the "frequent" method. For D1, both approaches yielded similar model results. However, for D2, App. 2 slightly enhanced recall

without compromising accuracy. Chosen for D1 due to slightly better outcomes, App. 2 balanced data integrity with model performance, underscoring its relevance.

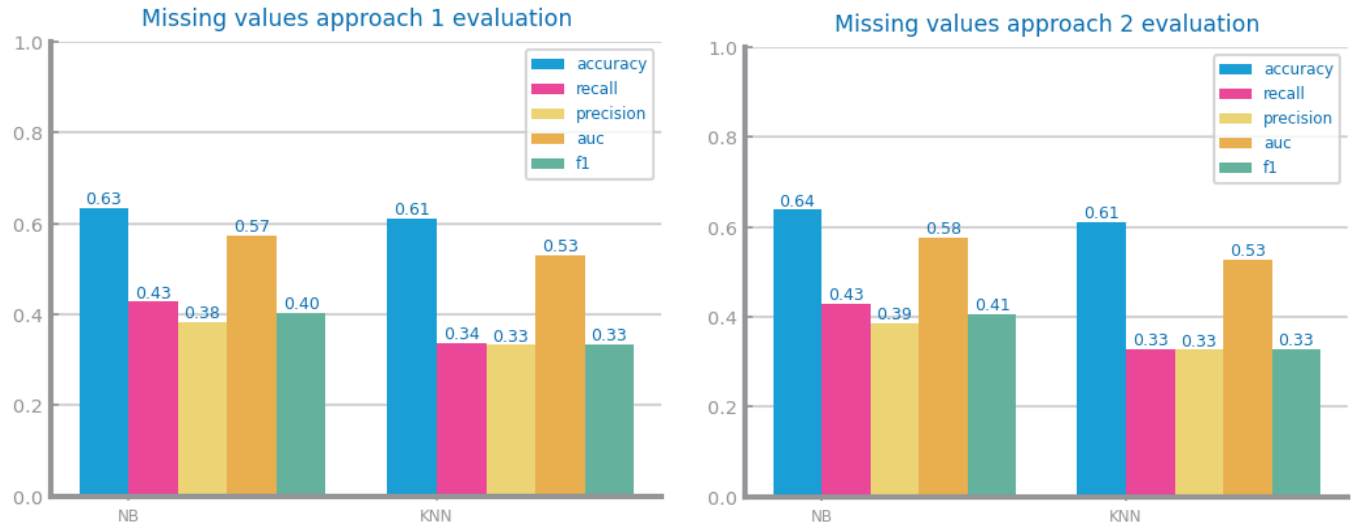


Figure 19 Missing values imputation results with different approaches for dataset 1

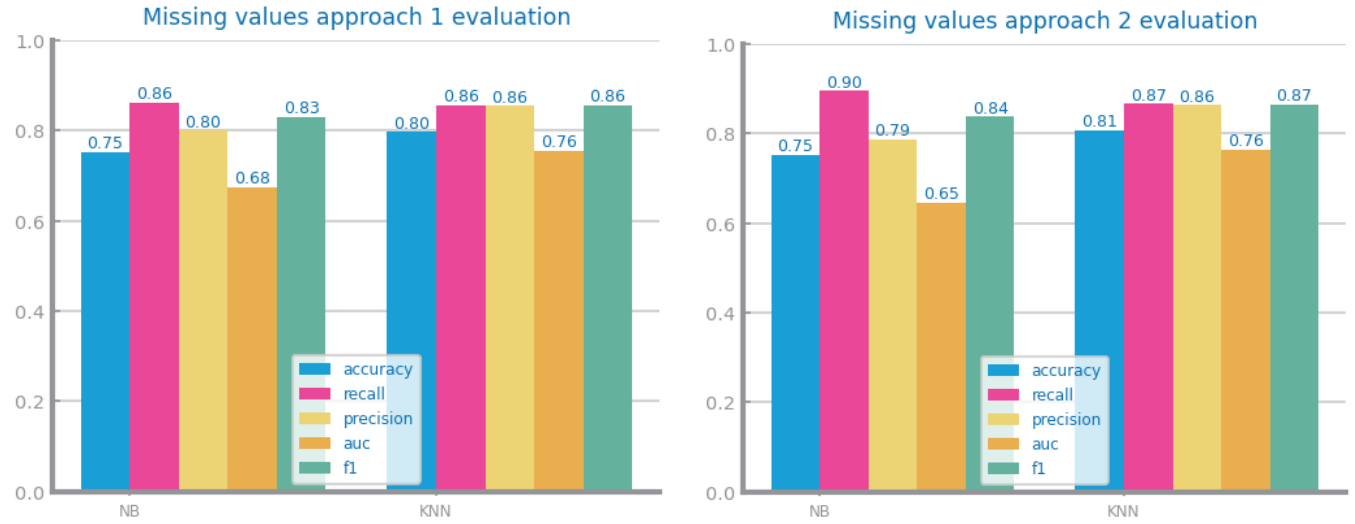


Figure 20 Missing values imputation results with different approaches for dataset 2

Outliers Treatment

Two outlier handling approaches were used for datasets 1 and 2: App 1 removed outliers, while App 2 replaced them with fixed threshold values. For D1, both approaches yielded similar performance, but App 1 was chosen due to its simplicity and data integrity maintenance. For D2, App 2 was preferred as it slightly outperformed App 1 and retained potentially valuable outliers, emphasizing data completeness and domain relevance.

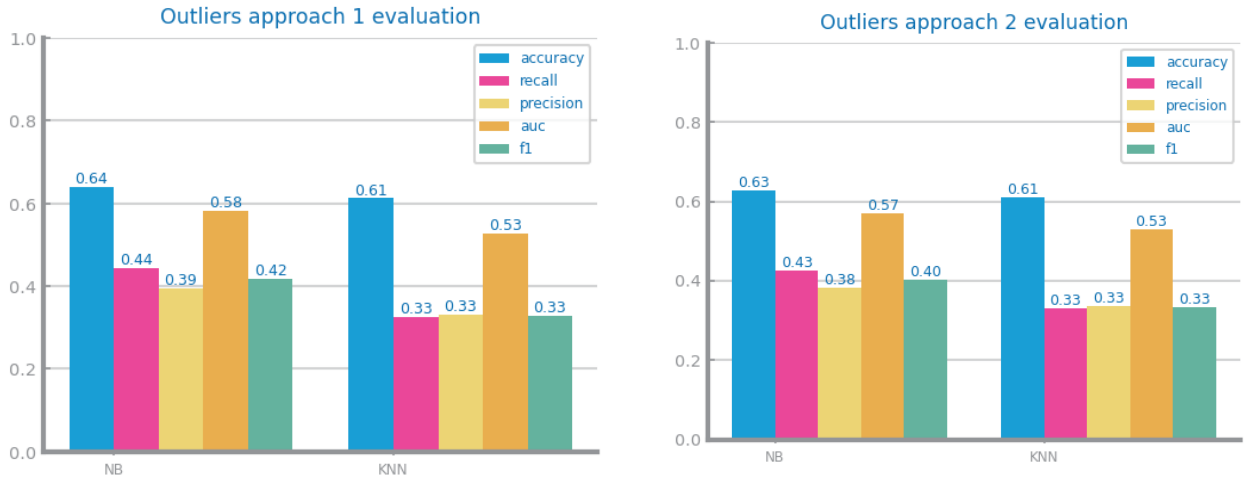


Figure 21 Outliers imputation results with different approaches for dataset 1

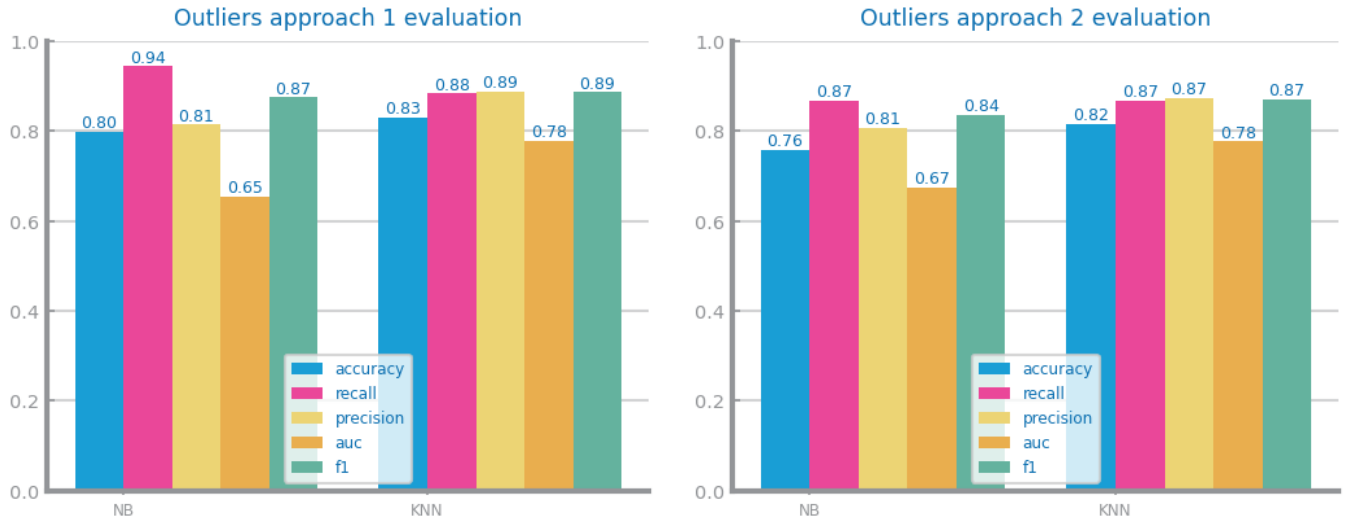
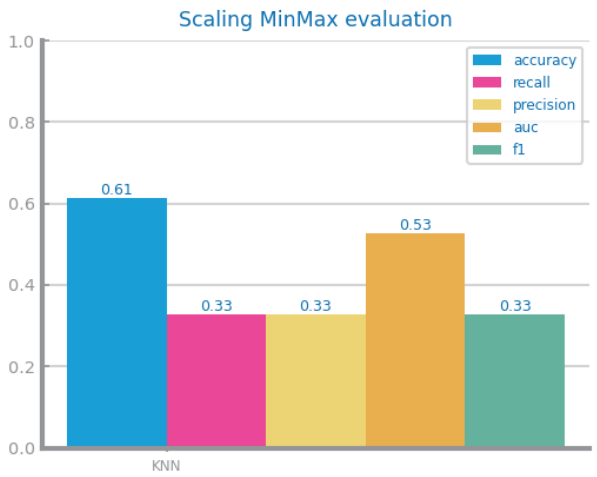


Figure 22 Outliers imputation results with different approaches for dataset 2

Scaling

For D1, Min-Max and Z-score scaling were tested for KNN but skipped as they didn't improve results. For D2 Min-Max was chosen, with results on par with Z-score, showing flexibility in scaling methods.



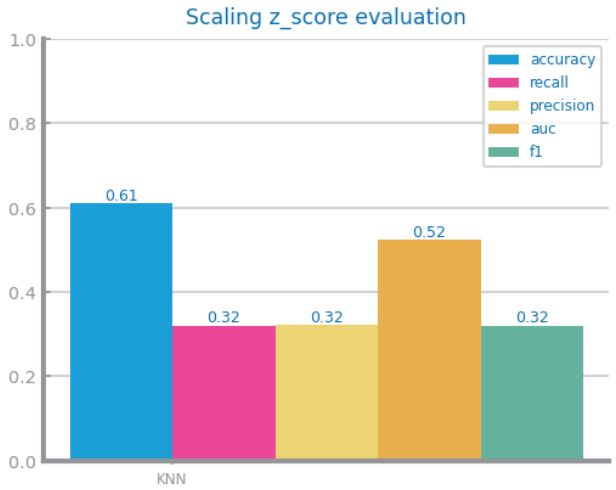
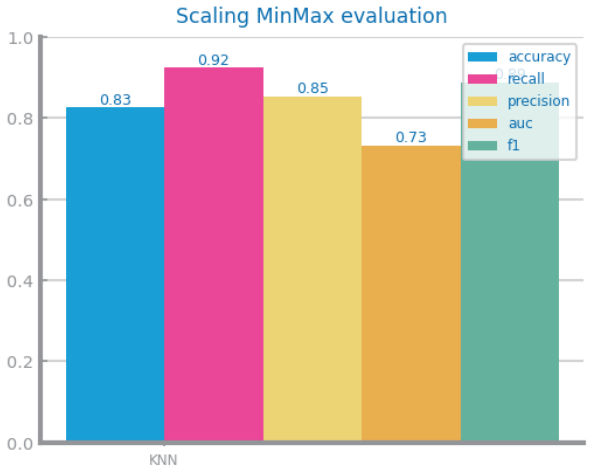


Figure 23 Scaling results with different approaches for dataset 1



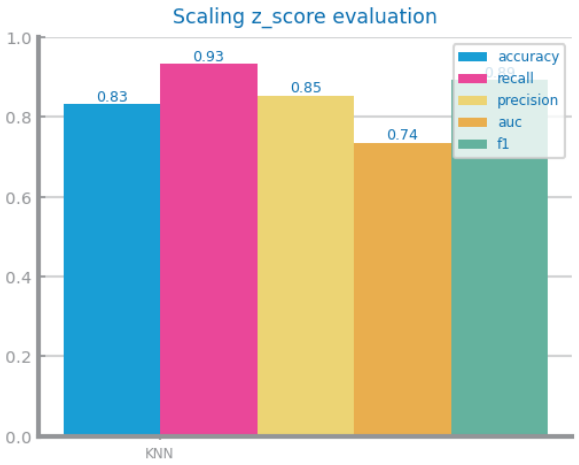


Figure 24 Scaling results with different approaches for dataset 2

Balancing

For D1, aiming to optimize F1, SMOTE outperforms over and under-sampling, notably improving KNN's F1 score. NB also benefits from SMOTE, making it the selected approach. For D2, with accuracy as the target, SMOTE again leads, boosting KNN's accuracy the most. NB maintains good accuracy across methods. Thus, SMOTE is chosen for both datasets, given its consistent enhancement of the desired metrics for both classifiers.

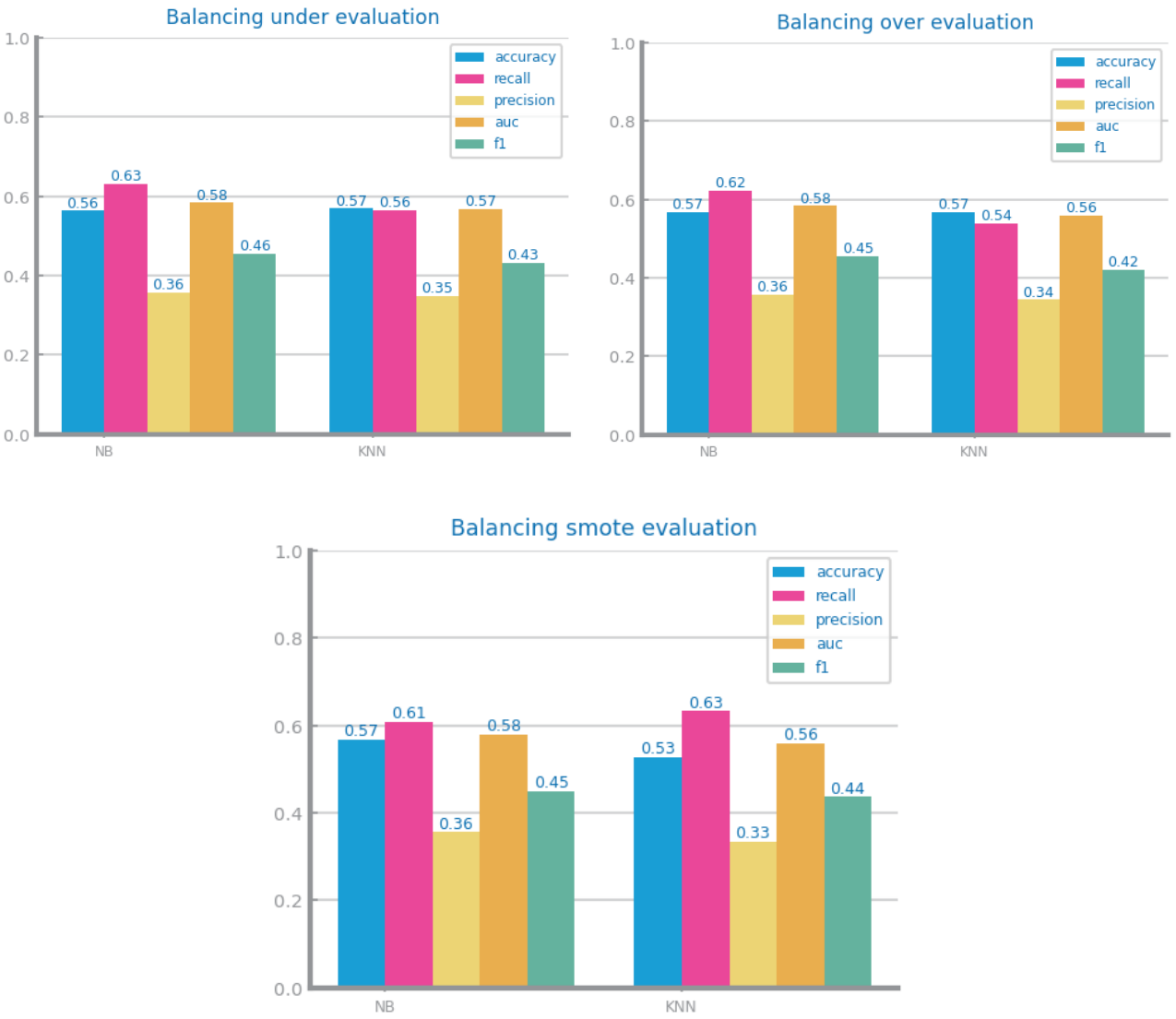
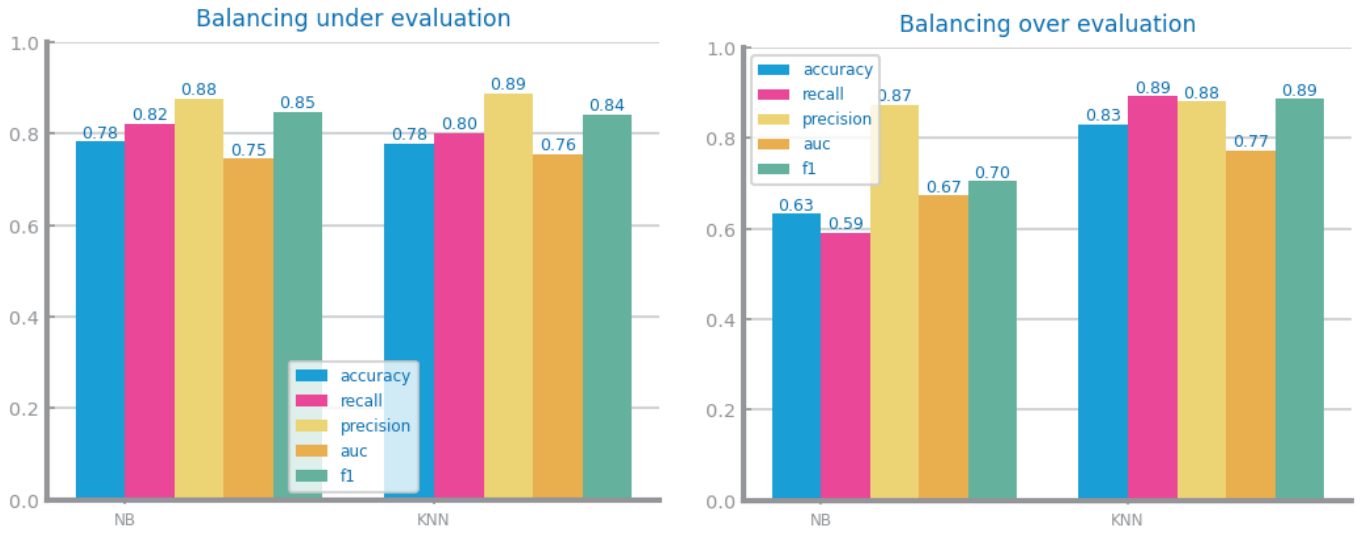


Figure 25 Balancing results with different approaches for dataset 1



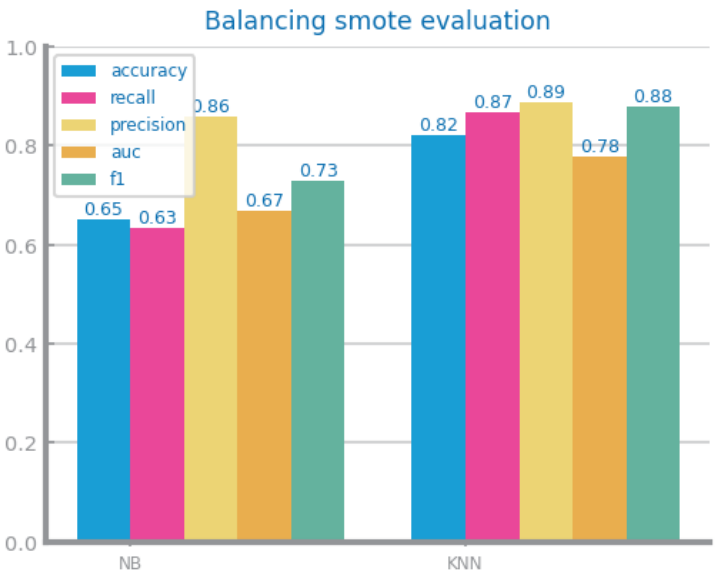
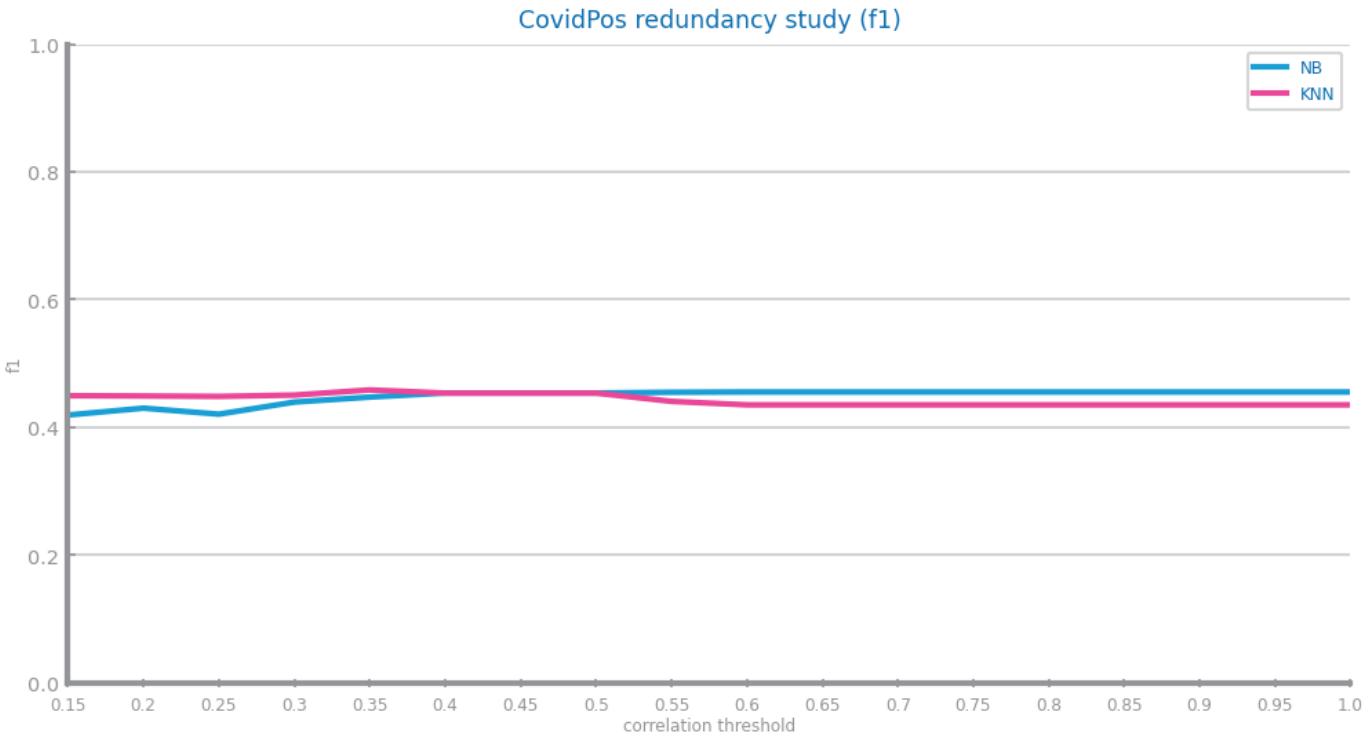
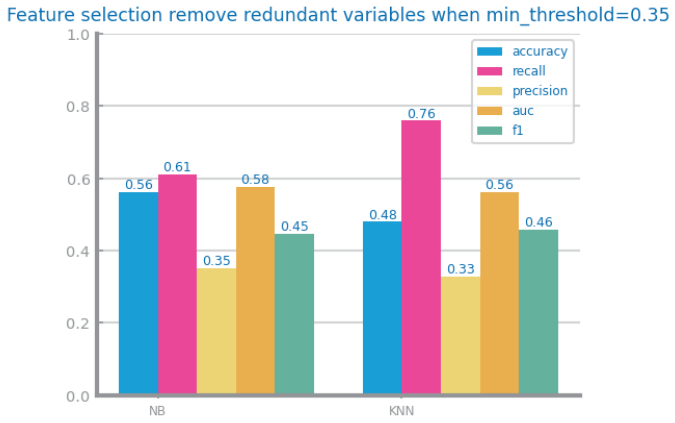


Figure 26 Balancing results with different approaches for dataset 2

Feature Selection

In D1, feature selection showed negligible impact on F1 scores, implying the inherent model robustness to feature redundancy or low variance. Consequently, no features were removed to avoid potential loss of informative signals. For D2, setting a variance threshold of 0.15 for feature selection notably enhanced model accuracy, particularly for NB. The optimal for D2 is thus at variance = 0.15, balancing dimensionality and maximizing accuracy, fitting the targeted model improvement.





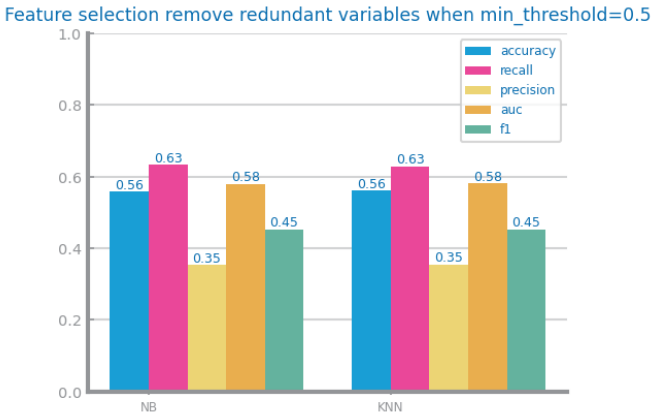
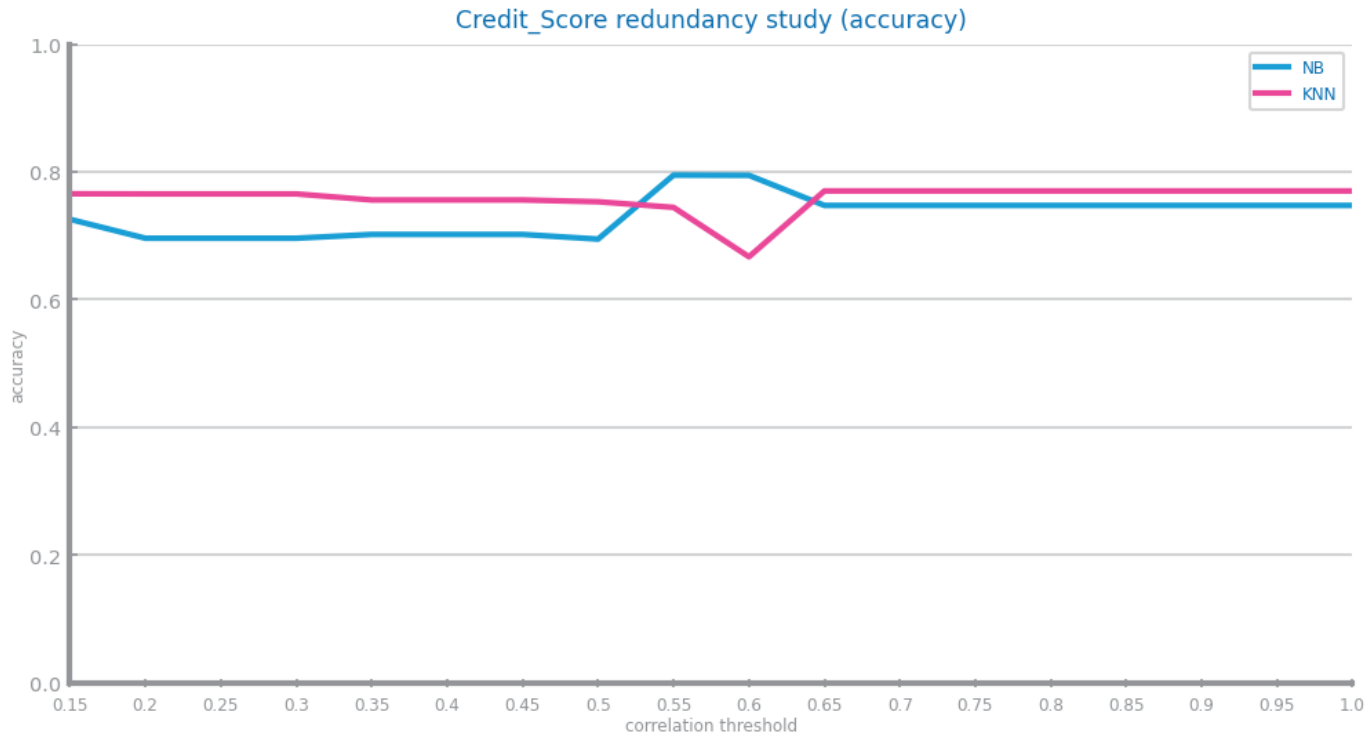


Figure 27 Feature selection of redundant variables results with different parameters for dataset 1



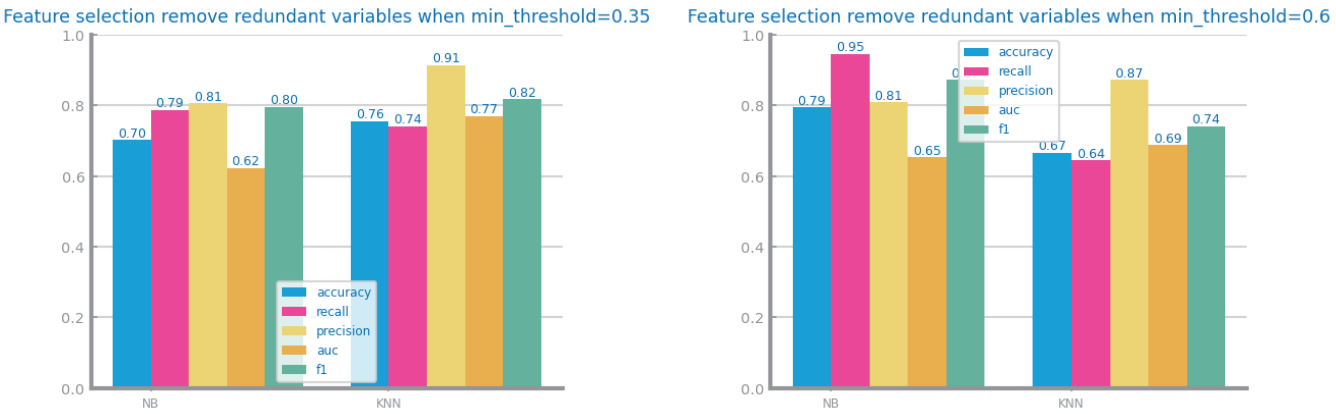
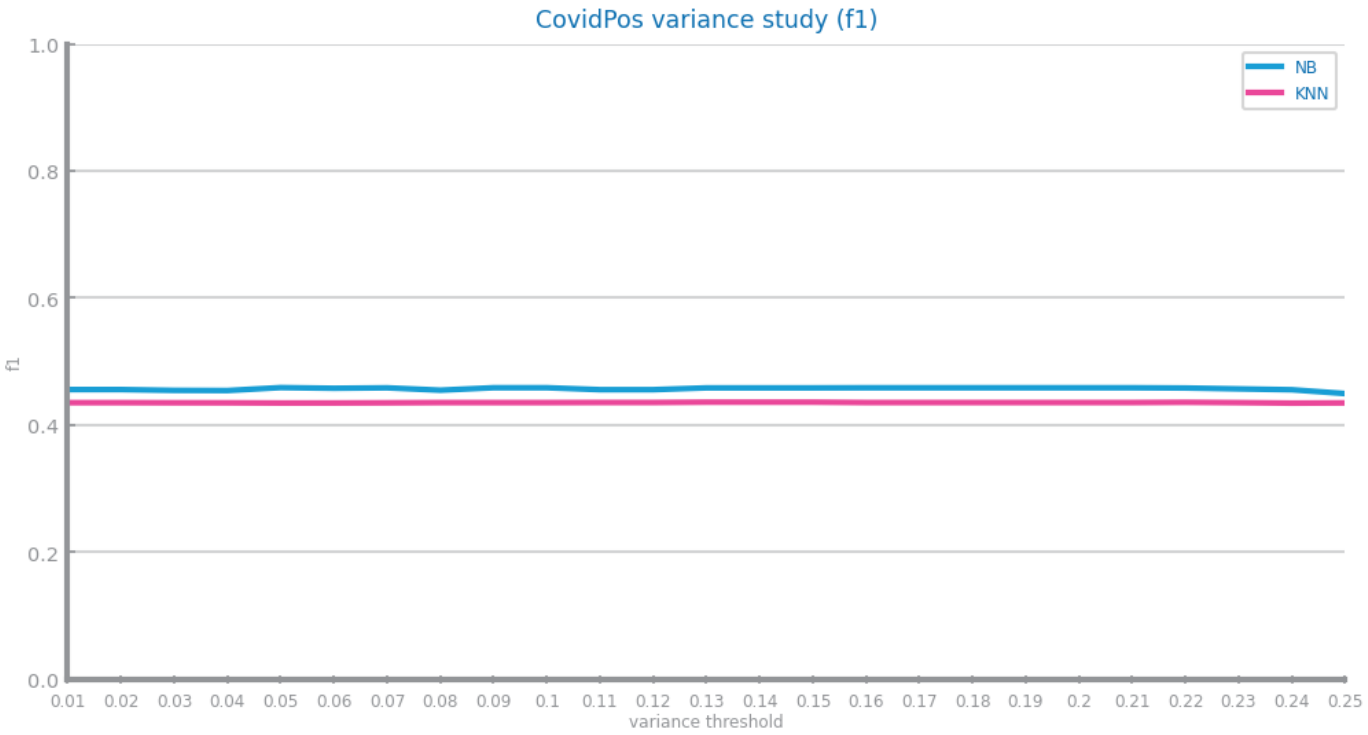


Figure 28 Feature selection of redundant variables results with different parameters for dataset 2



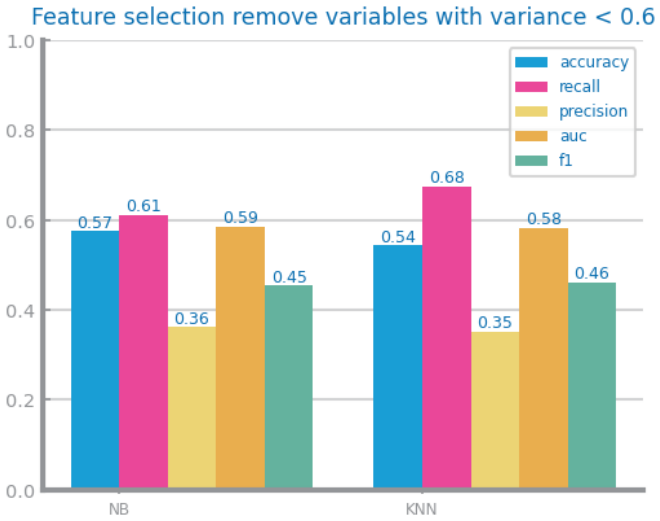


Figure 29 Feature selection of relevant variables results with different parameters for dataset 1 (variance study)

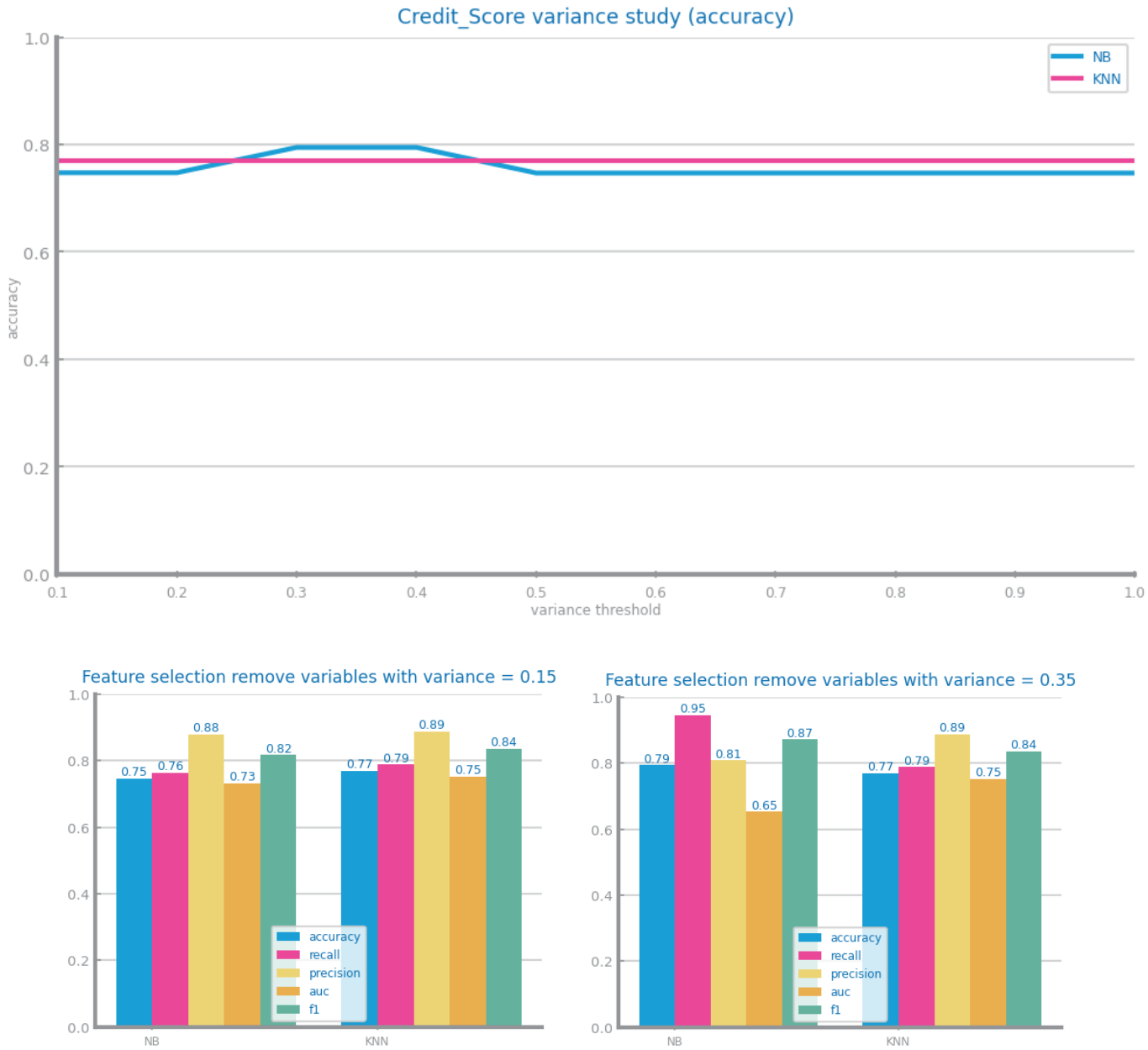
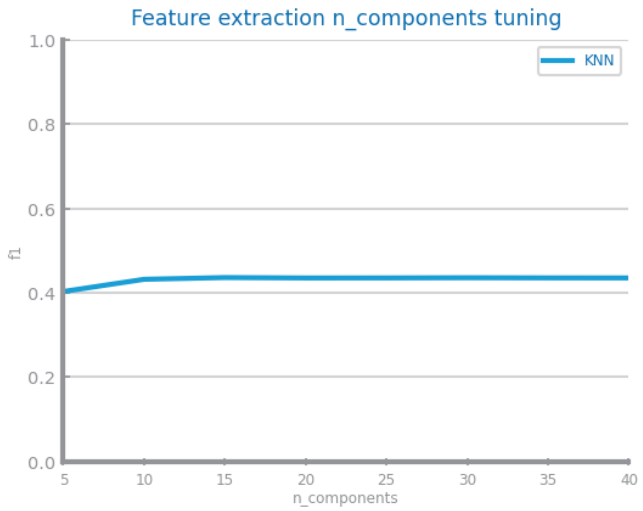
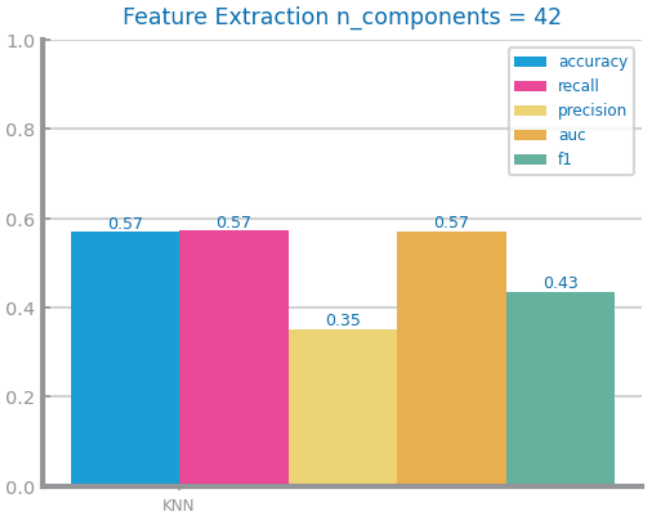


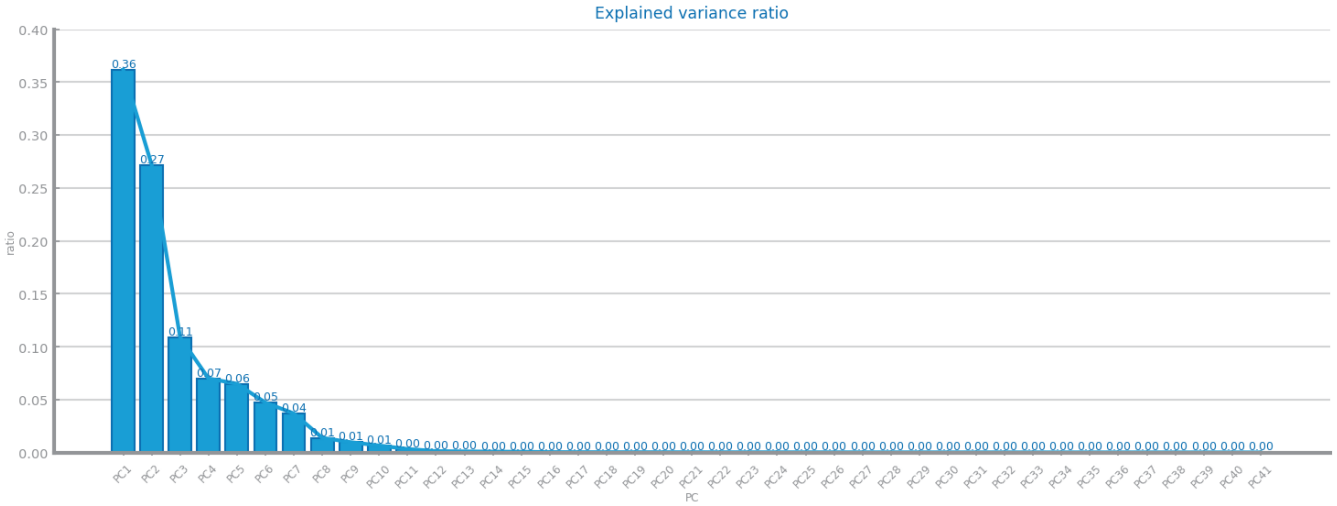
Figure 30 Feature selection of relevant variables results with different parameters for dataset 2 (variance study)

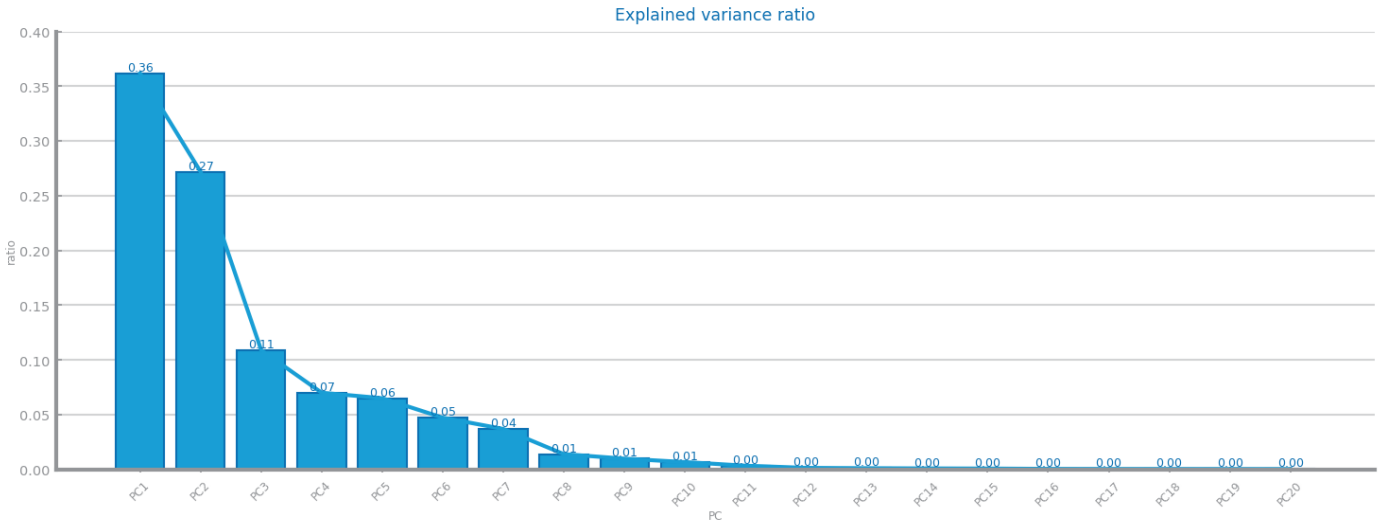
Feature Extraction (optional)

PCA analysis shows no F1 score improvement beyond 0.43 for KNN with >10 components. Non-PCA yields higher F1 (0.46). Thus, not using PCA is optimal.









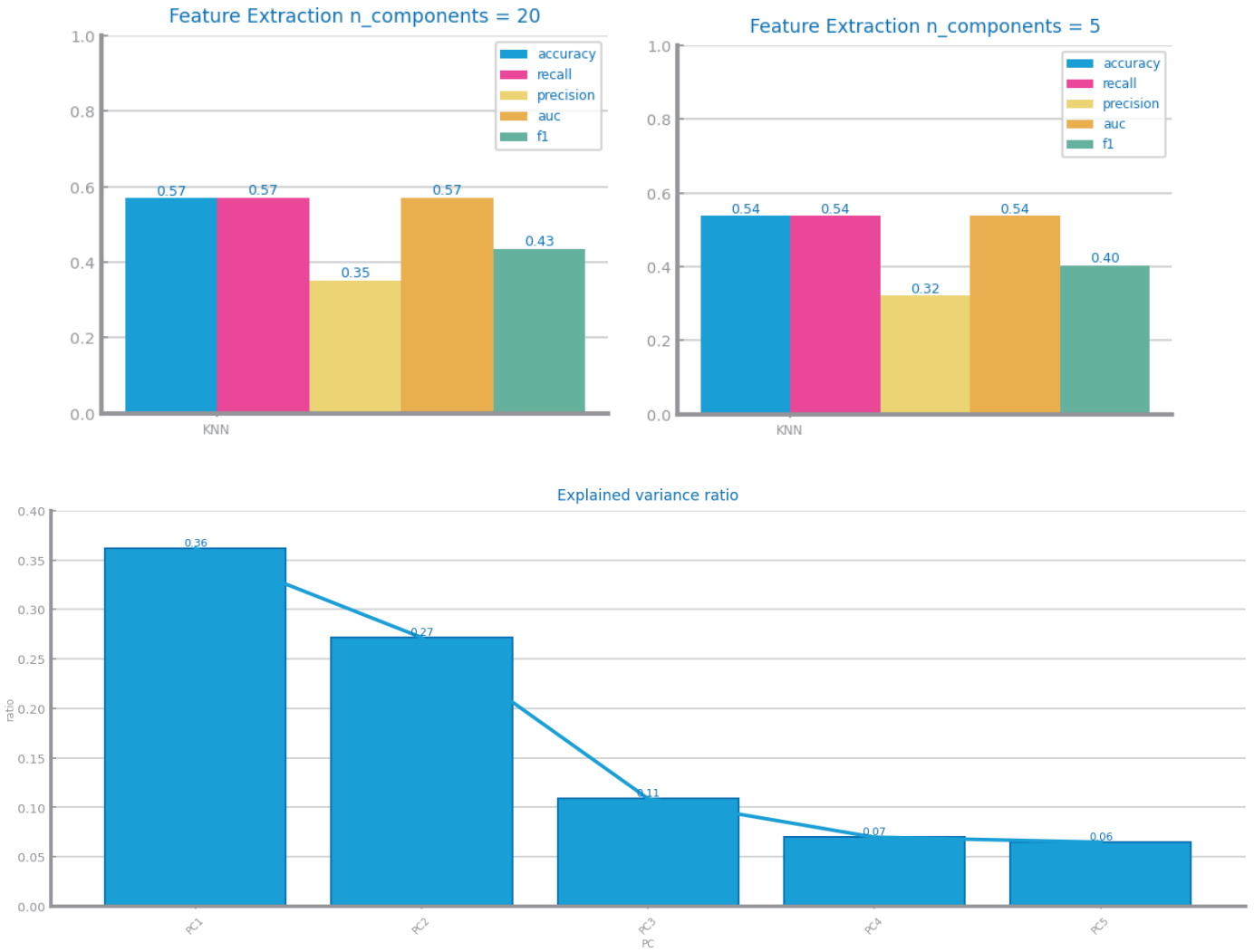


Figure 31 Principal components analysis and feature extraction results for dataset 1

3 Models' Evaluation

Accuracy, recall, AUC (Area Under the Curve), F1 Score and precision were the evaluation measures used to assess the performance of classification for both datasets.

For Dataset 1, the goal was to maximize the F1 score as there is a need for balance between accuracy and recall. For Dataset 2, the goal was to maximize accuracy.

Naïve Bayes

In D1, GaussianNB showed 0.70 recall but struggled with overfitting while scoring the best F1 (0.46); BernoulliNB was not utilized. In D2, BernoulliNB excelled with 0.73 recall, outperforming GaussianNB. Dataset specifics influenced model outcomes, emphasizing adaptability for optimal results.

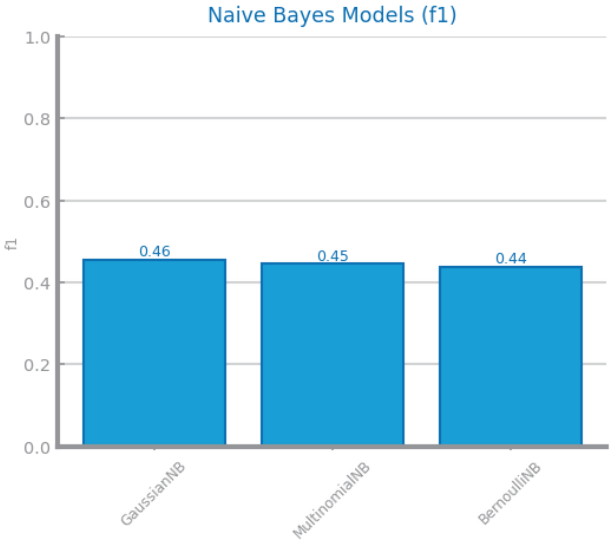


Figure 35 Naïve Bayes alternatives comparison for dataset 1

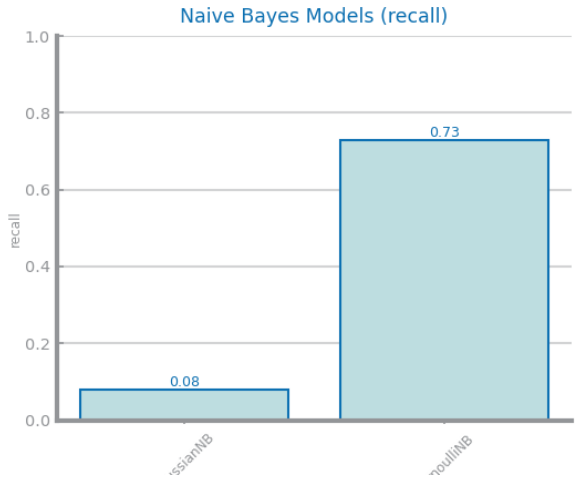


Figure 36 Naïve Bayes alternative comparison for dataset 2

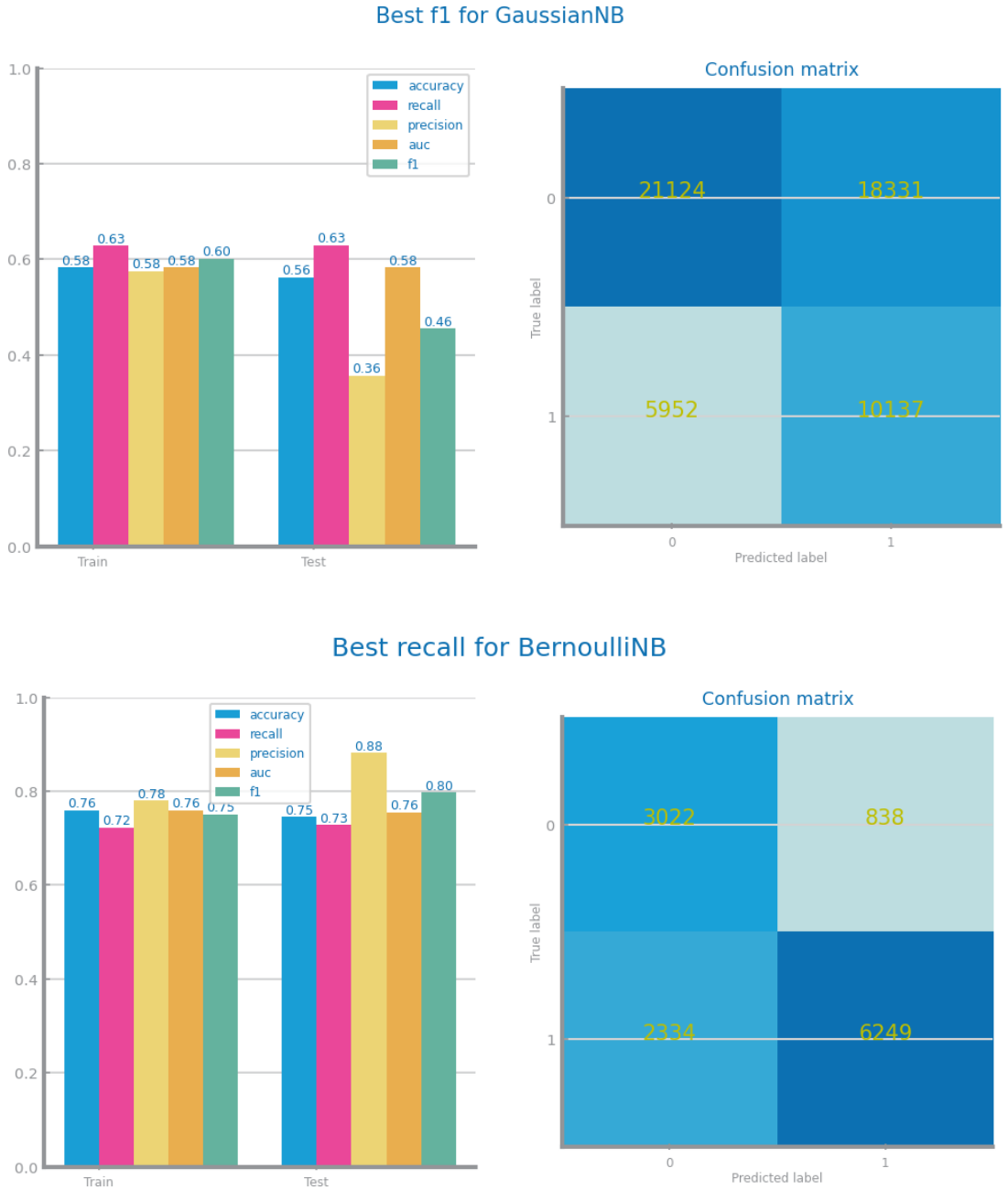


Figure 37 Naïve Bayes best model results for dataset 1 (up) and dataset 2 (bottom)

KNN

In D1 (Fig. 36), Manhattan and Euclidean distances excelled for smaller k values, with Manhattan showing stability. D2 (Fig. 37) also favored Manhattan initially. Overfitting was evident in both datasets (Fig. 38), more pronounced in D1. D1's KNN model showed training accuracy of ~0.6 but only ~0.45 on the test set, highlighting generalization challenges. D2's model performed better, underscoring the need for tailored parameterization to enhance KNN's generalizability, especially in D1.

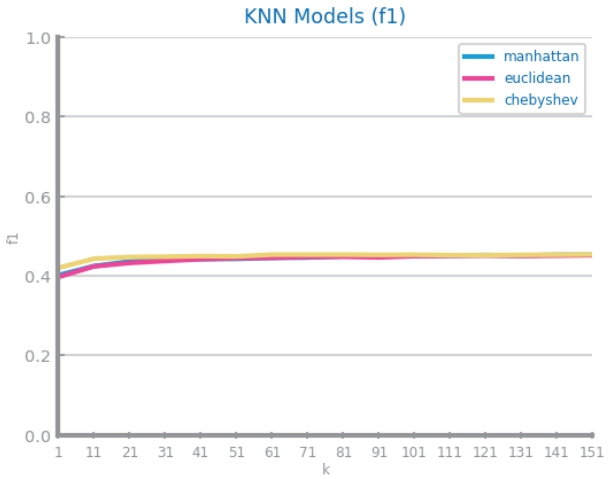


Figure 38 KNN different parameterisations comparison for dataset 1

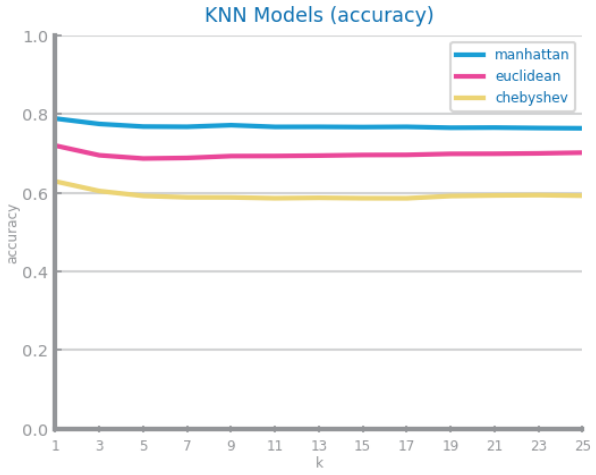


Figure 39 KNN different parameterisations comparison for dataset 2

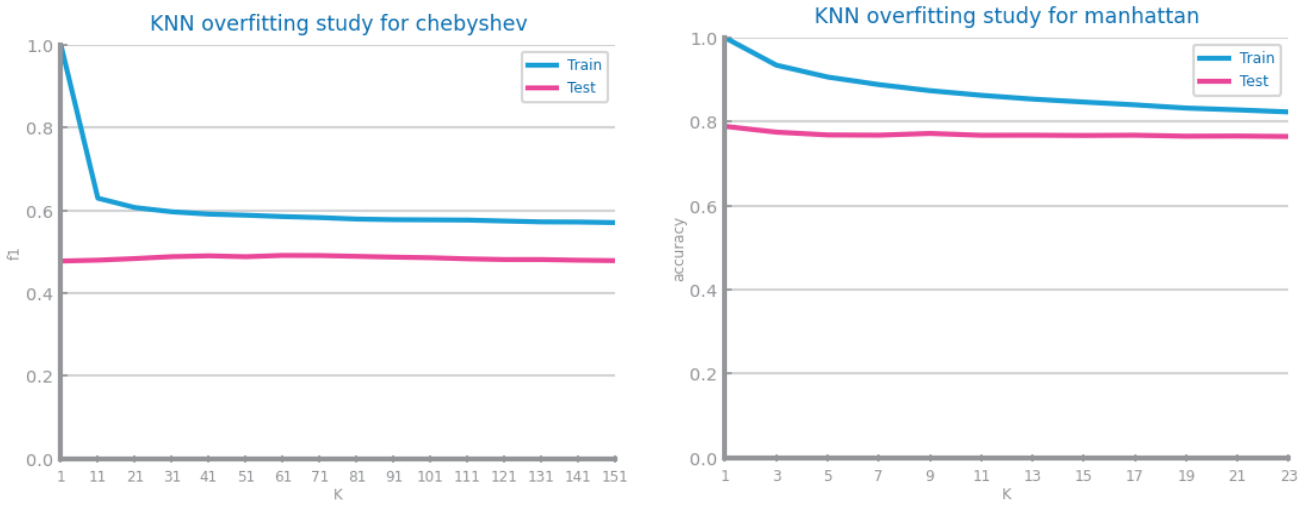


Figure 40 KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 41 KNN best model results for dataset 1 (up) and dataset 2 (bottom)

Decision Trees

In D1 (Fig. 42), both Gini and Entropy rose in accuracy until depth 6, declining thereafter; Gini was chosen. D2 (Fig. 43) showed peak accuracy at depth 10 for both, with Gini slightly outperforming. Overfitting was evident in both datasets (Fig. 44), with D1's test f1 dropping to 40% at depth 24, while D2 remained stable around 80%. D1's best tree (Fig. 44) emphasized age and health indicators; D2 (Fig. 45) highlighted financial attributes, particularly debt-related metrics.

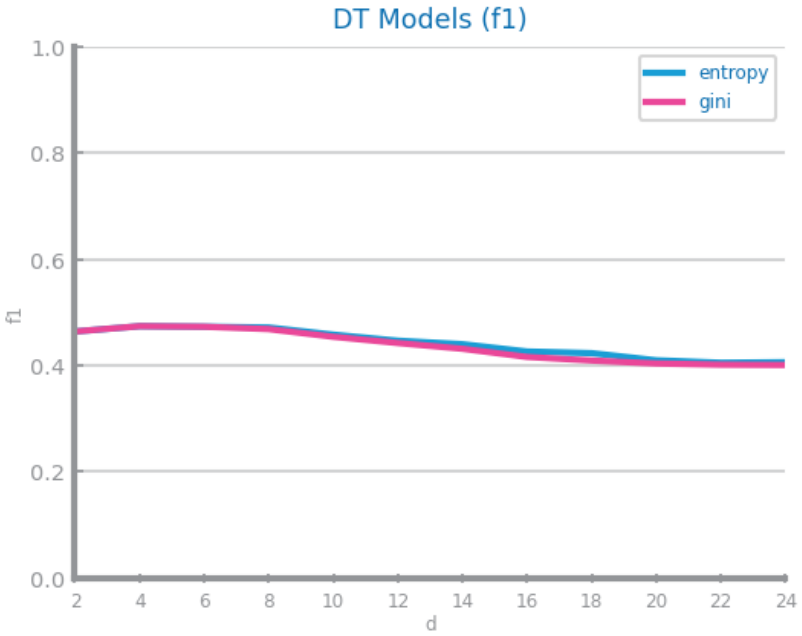


Figure 42 Decision Trees different parameterisations comparison for dataset 1

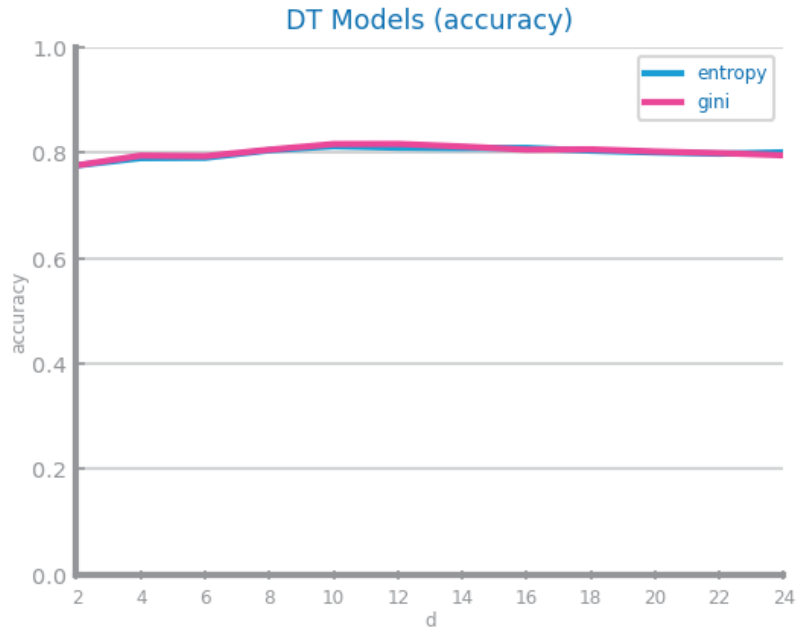


Figure 43 Decision Trees different parameterisations comparison for dataset 2

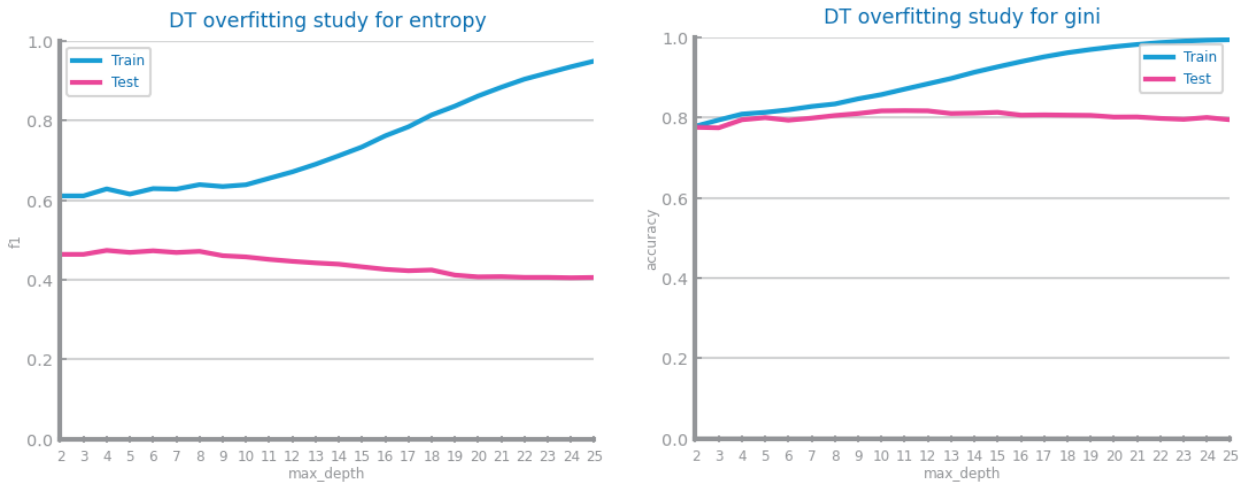


Figure 44 Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

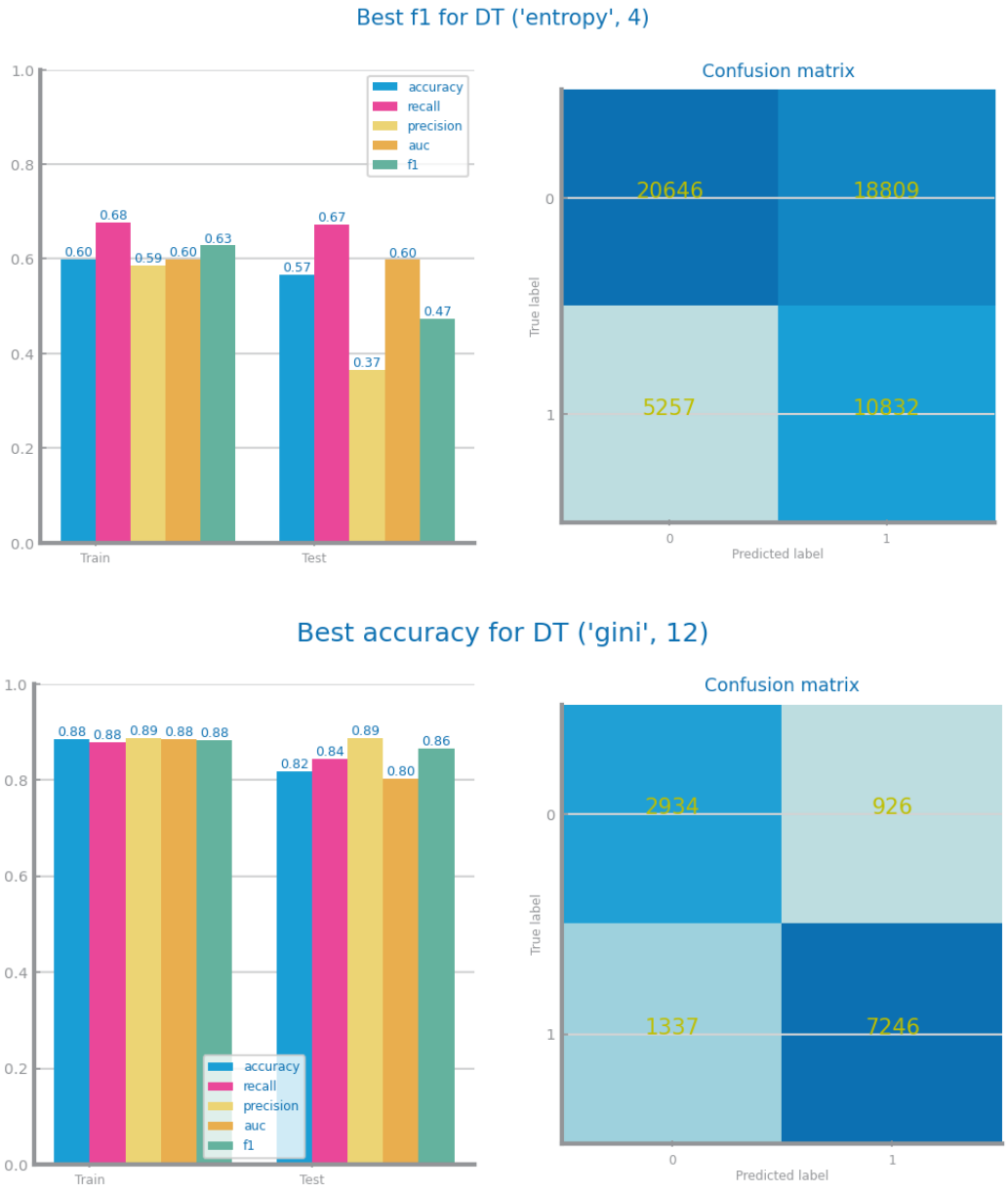


Figure 45 Decision trees best model results for dataset 1 (up) and dataset 2 (bottom)

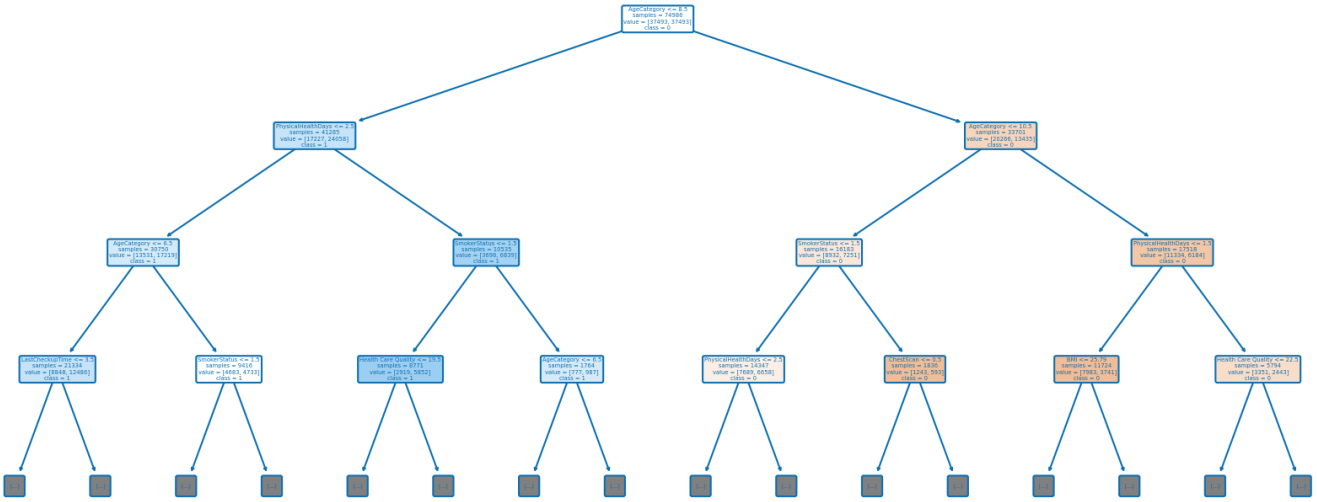


Figure 46 Best tree for dataset 1

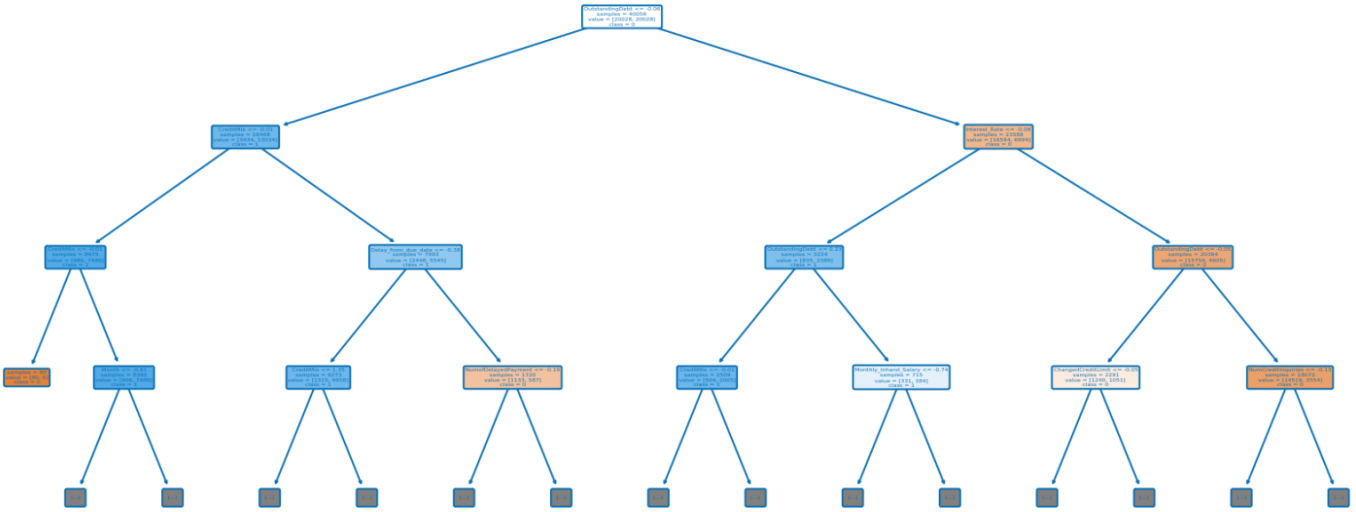


Figure 47 Best trees for dataset 2

Random Forests

For D1, RF models with max_depth=7 and f=0.7, avoiding overfitting, gave the best F1. High feature importance was on age, BMI, and physical act. For D2, the best accuracy was at max_depth=7, f=0.7, and estimators=100, with a balanced confusion matrix. Important features were outstanding debt and interest rate. Both datasets show stable performance across various parameterizations, with no overfitting as train and test scores converge, indicating robust models tuned to each d's characteristics.

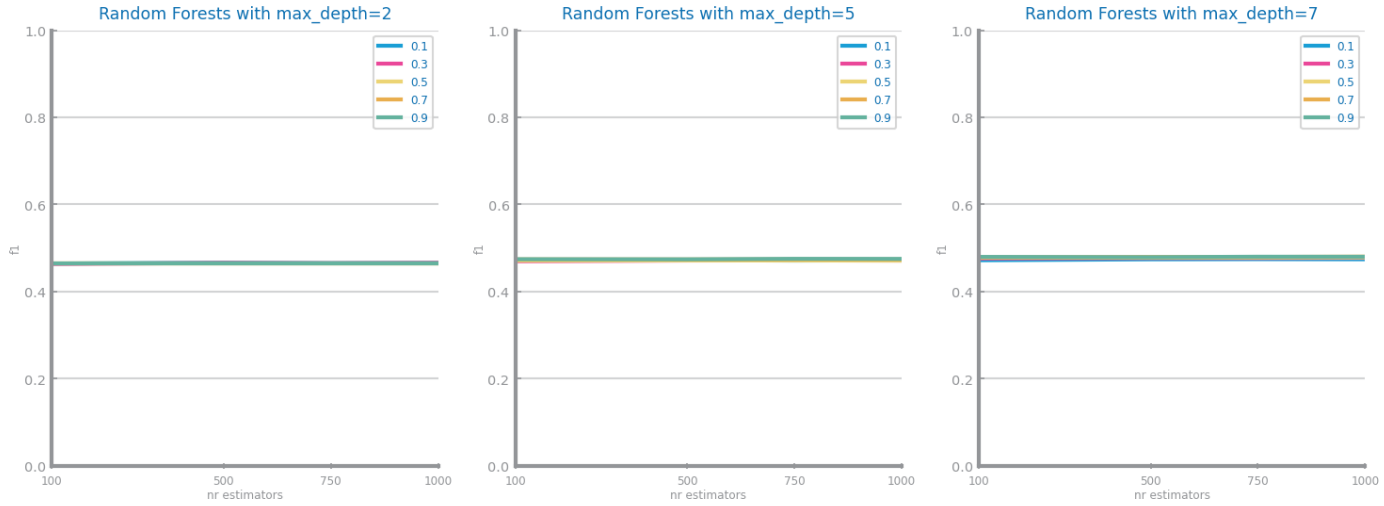


Figure 48 Random Forests different parameterisations comparison for dataset 1

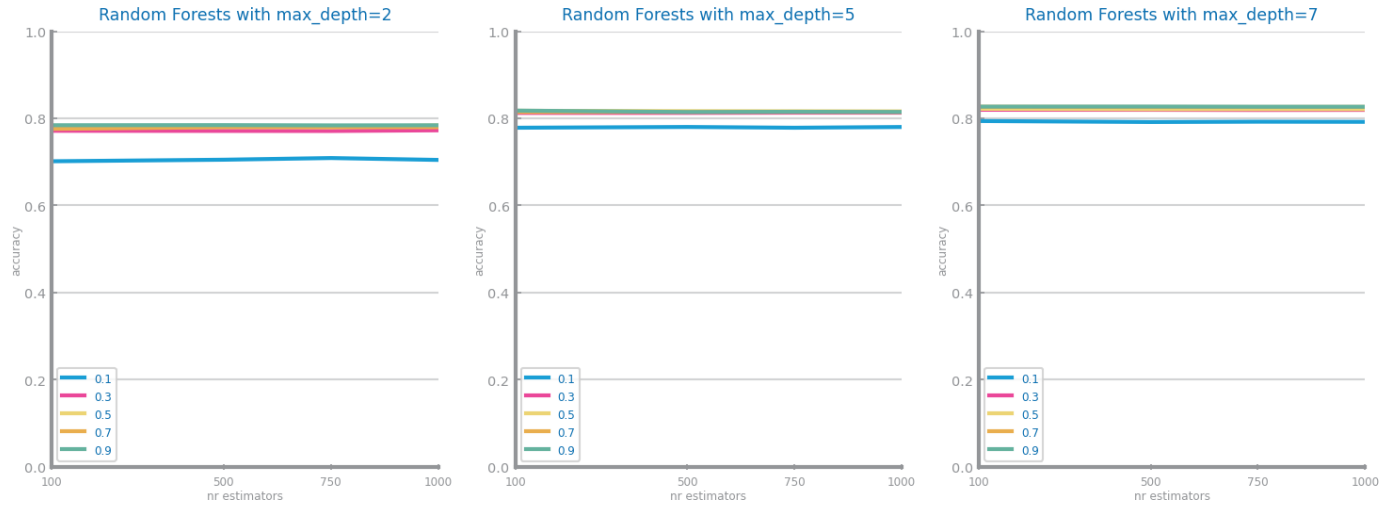
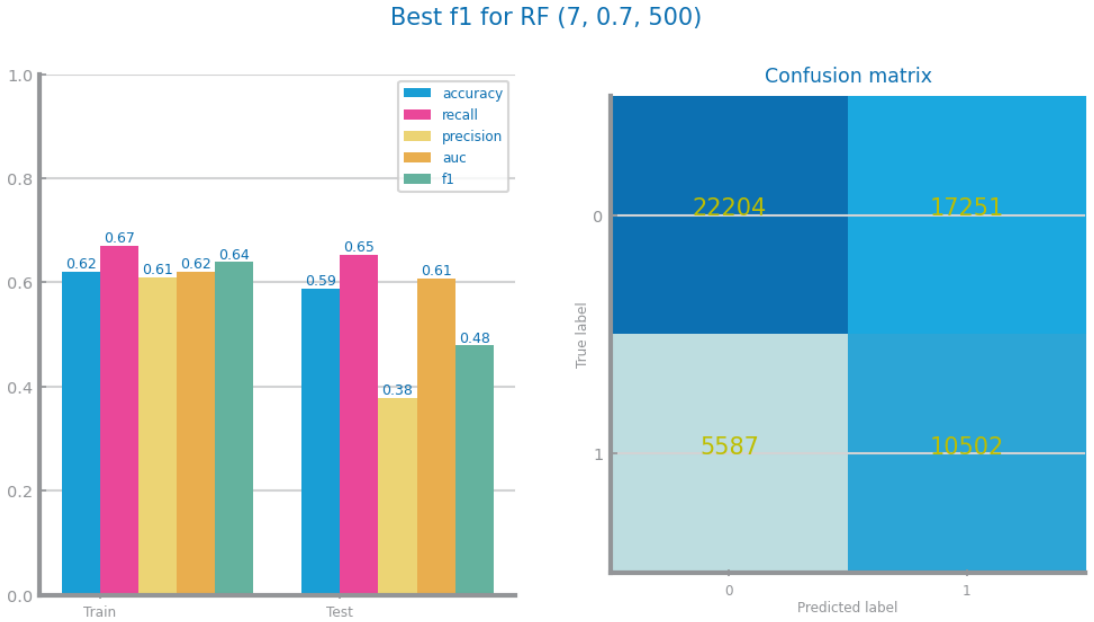


Figure 49 Random Forests different parameterisations comparison for dataset 2

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



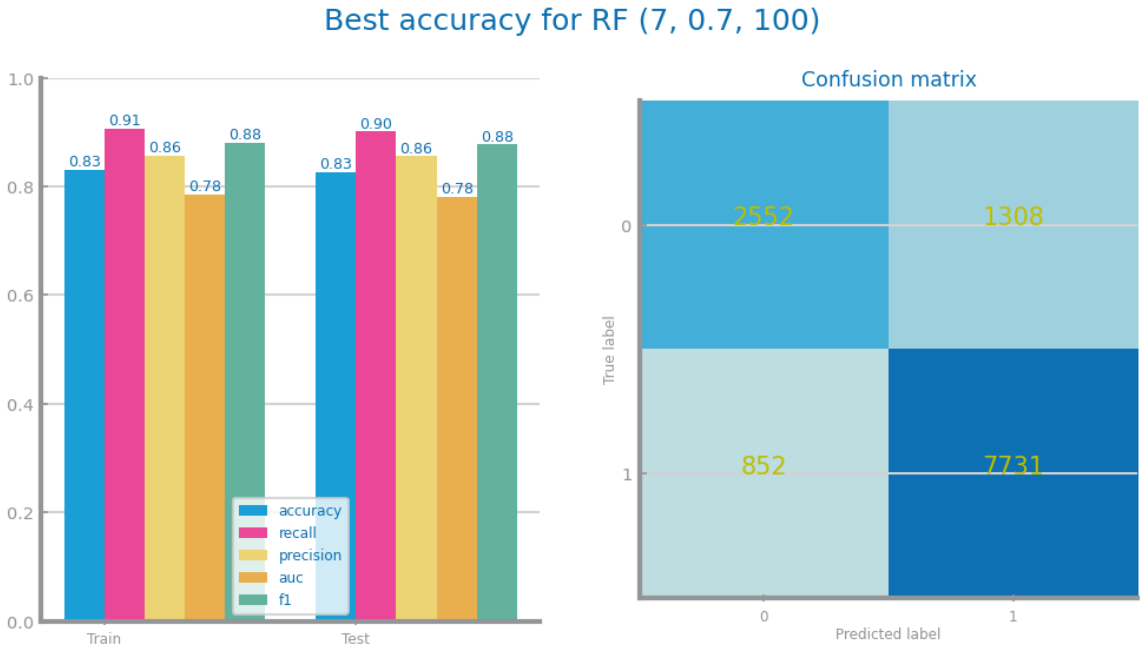


Figure 51 Random Forests best model results for dataset 1 (up) and dataset 2 (bottom)

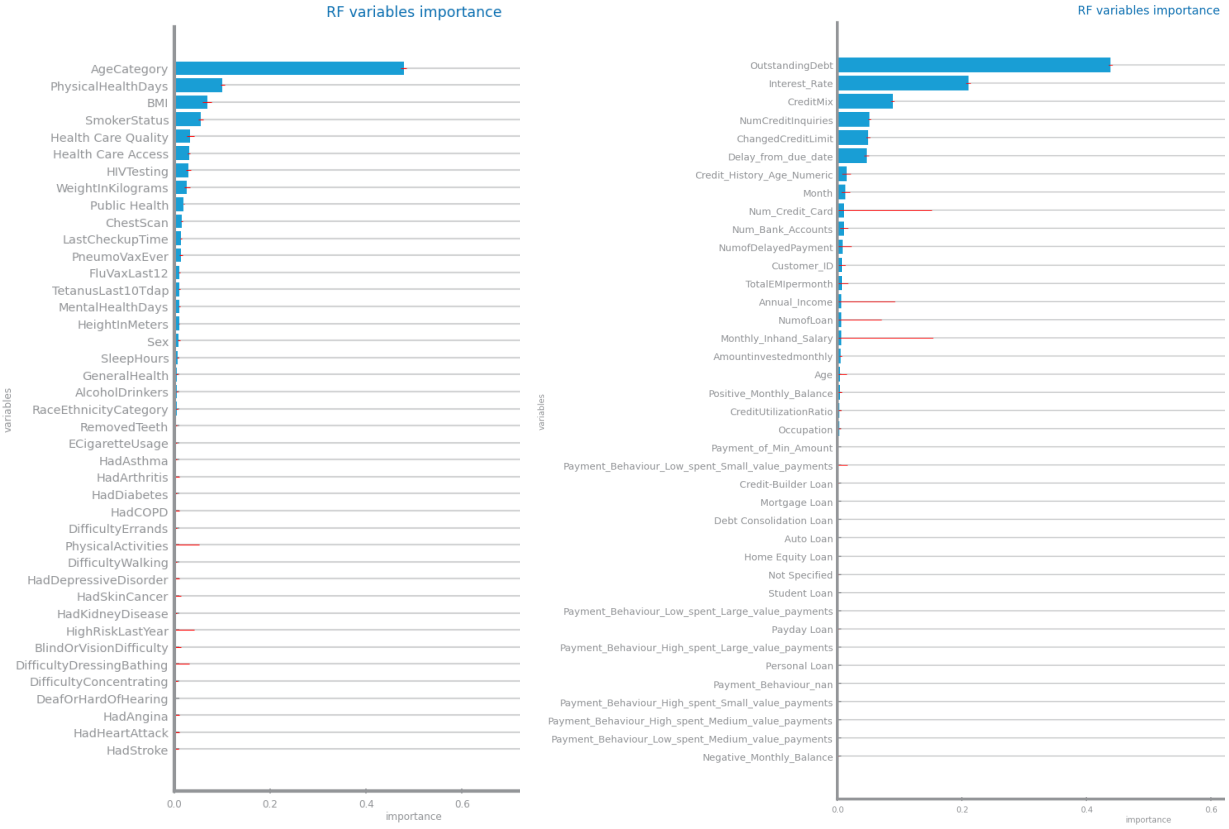


Figure 52 Random Forests variables importance for dataset 1 (left) and dataset 2 (right)

Gradient Boosting

D1's GB model with an f1 ~0.63 shows mild overfitting; 'AgeCategory' is notably significant. D2's GB model has high accuracy (~0.91), but stark overfitting is a concern; 'OutstandingDebt' is highly influential. D2 exceeds in accuracy, but D1's consistency suggests better generalization. False negatives in D1's confusion matrix indicate prediction caution, while D2's balanced matrix shows strength in classification despite potential overfitting risks.

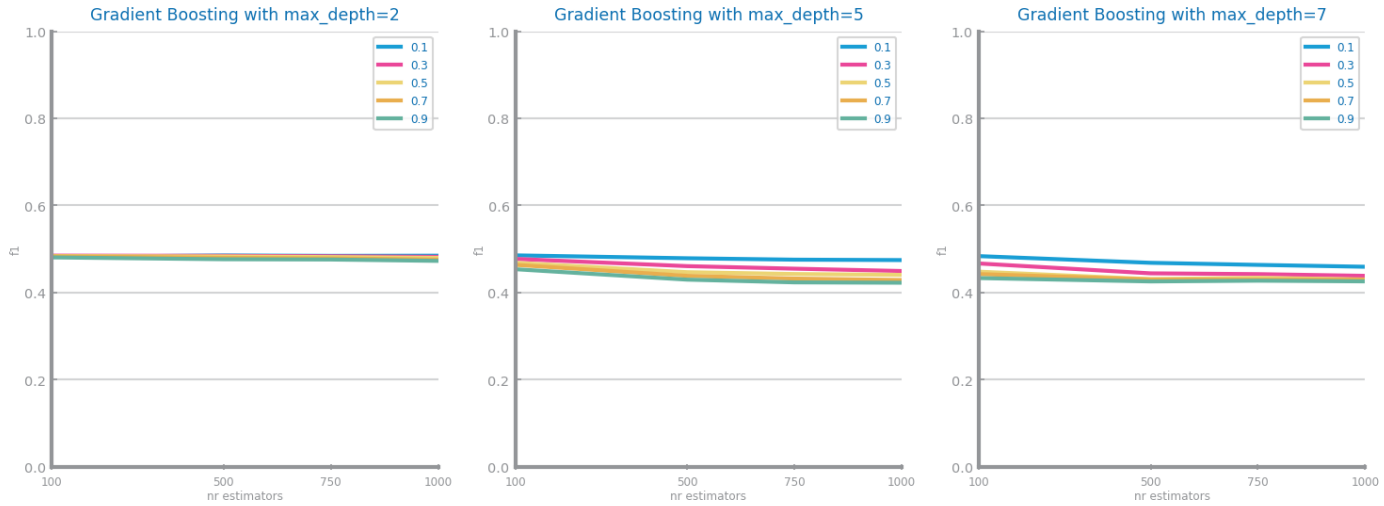


Figure 53 Gradient boosting different parameterisations comparison for dataset 1

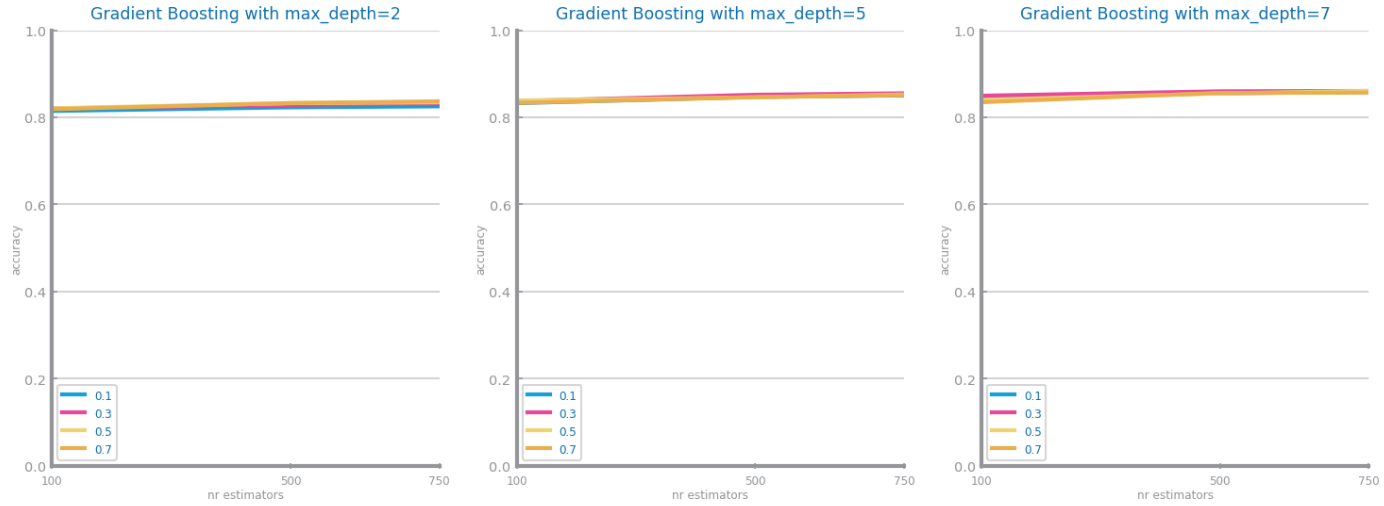


Figure 54 Gradient boosting different parameterisations comparison for dataset 2

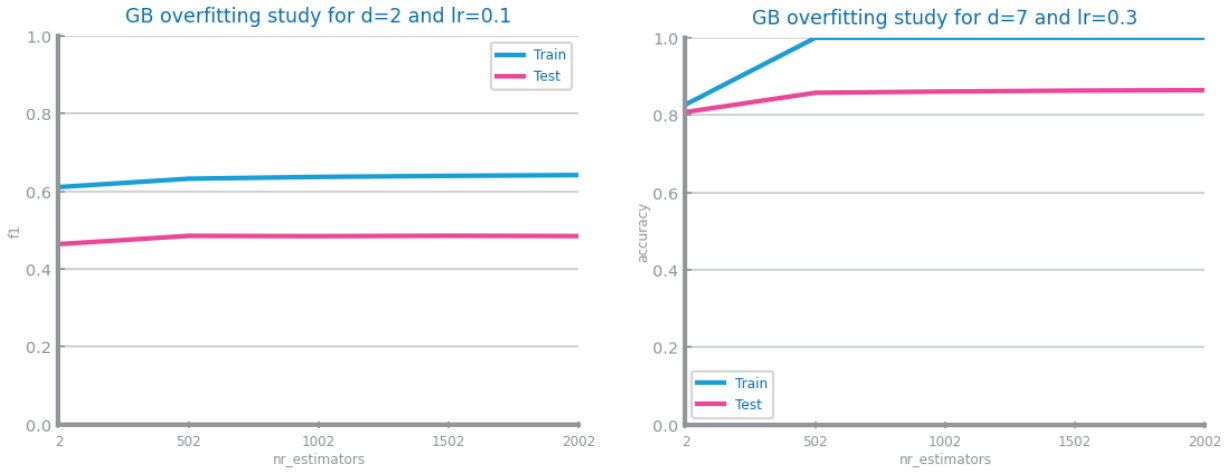


Figure 55 Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

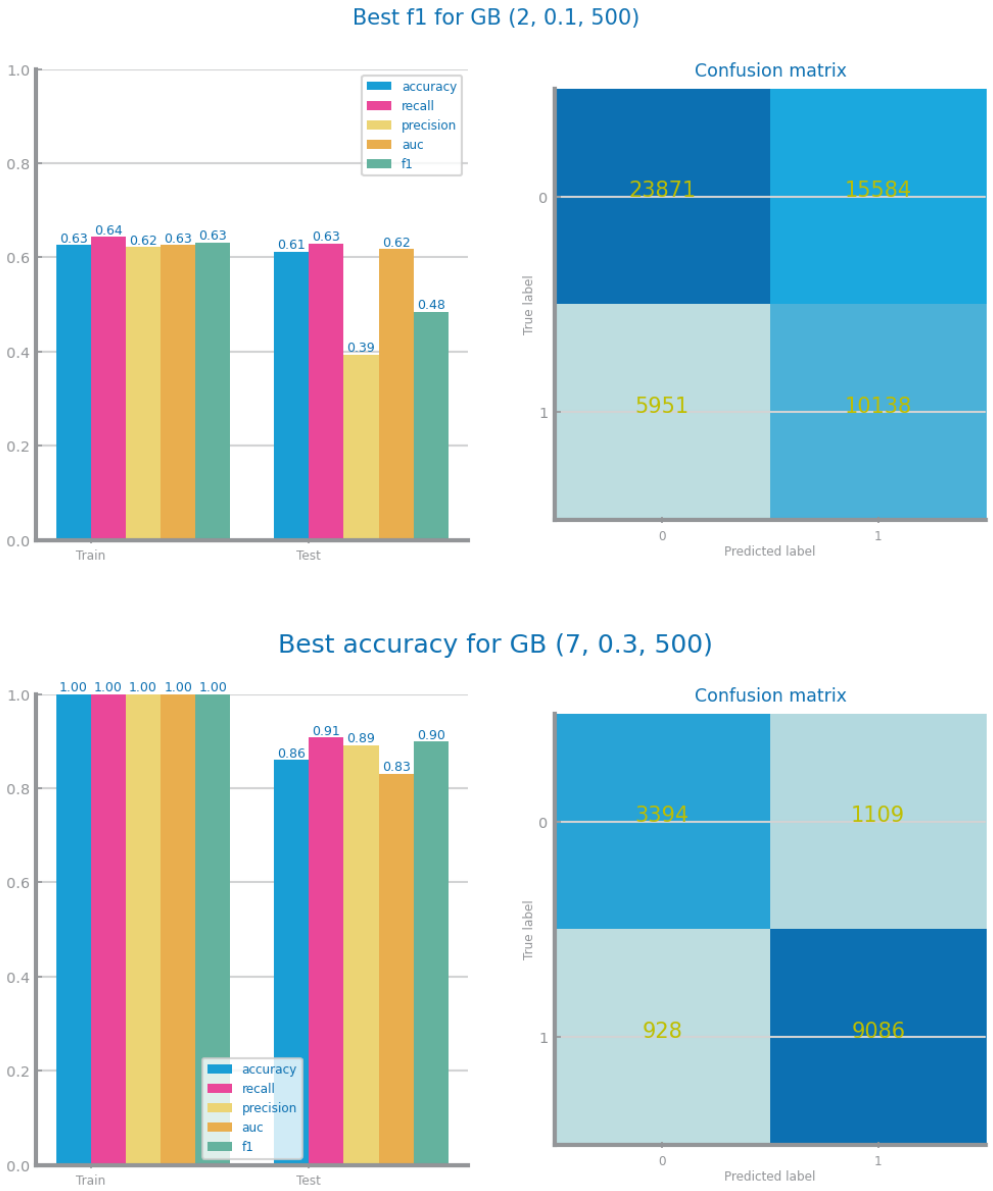


Figure 56 Gradient boosting best model results for dataset 1 (up) and dataset 2 (bottom)

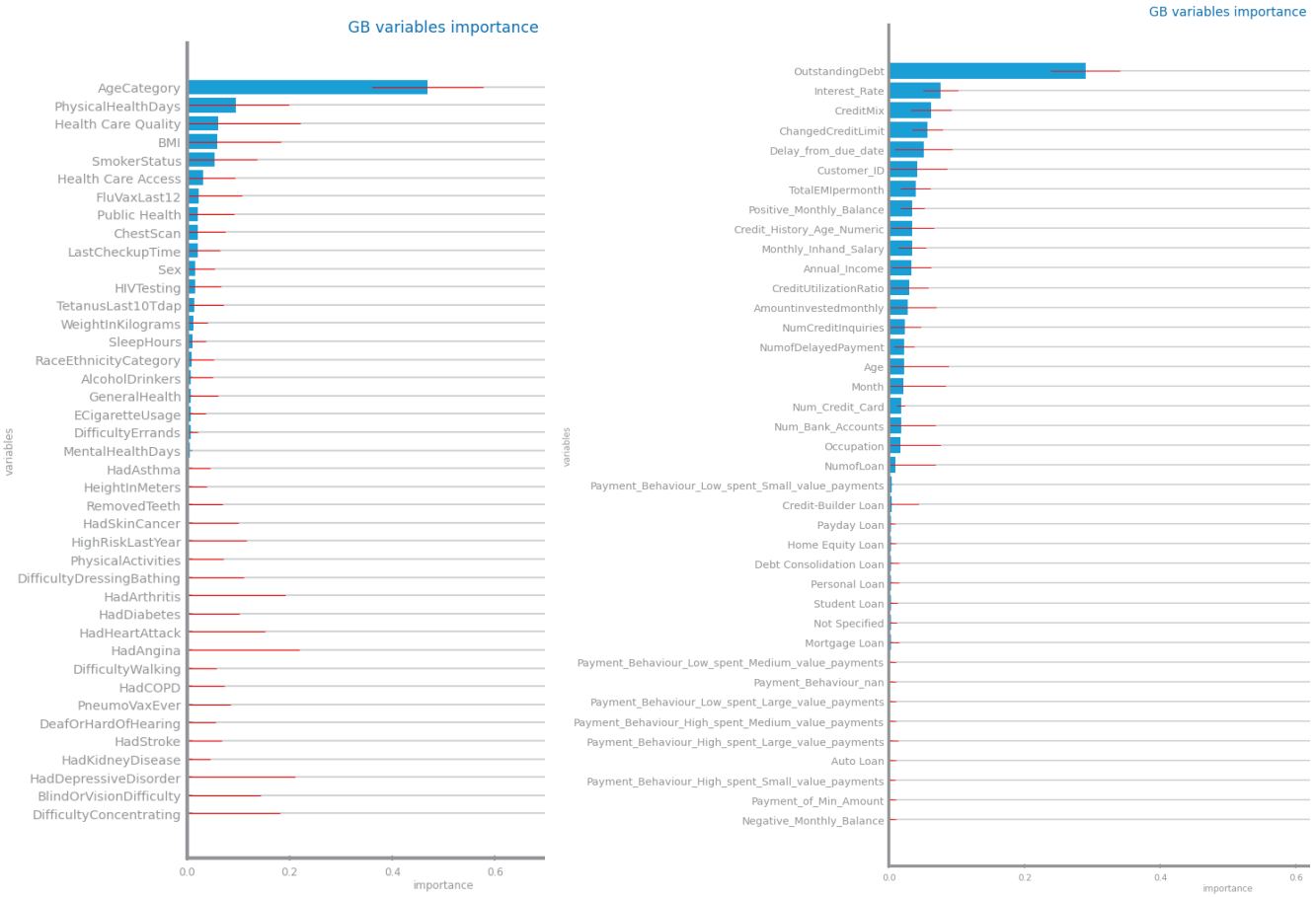


Figure 57 Gradient boosting variables importance for dataset 1 (left) and dataset 2 (right)

Multi-Layer Perceptrons

D1's MLP achieves an f1 of ~0.45, indicating modest performance with significant overfitting, evidenced by a larger drop in test metrics and a high false positive rate. D2's MLP performs better with an accuracy of ~0.75 and displays a more balanced confusion matrix with fewer false predictions. Lower learning rates appear to mitigate overfitting across datasets, yet D2 exhibits enhanced stability and maintains higher performance even as the learning rate increases, suggesting a more generalizable model.

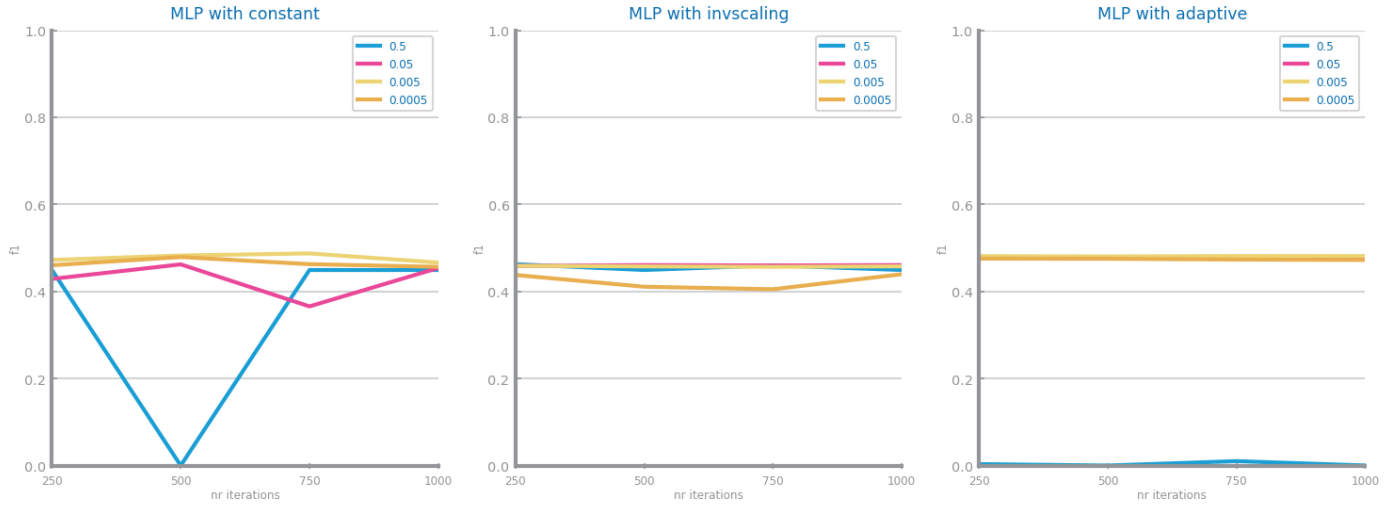


Figure 58 MLP different parameterisations comparison for dataset 1

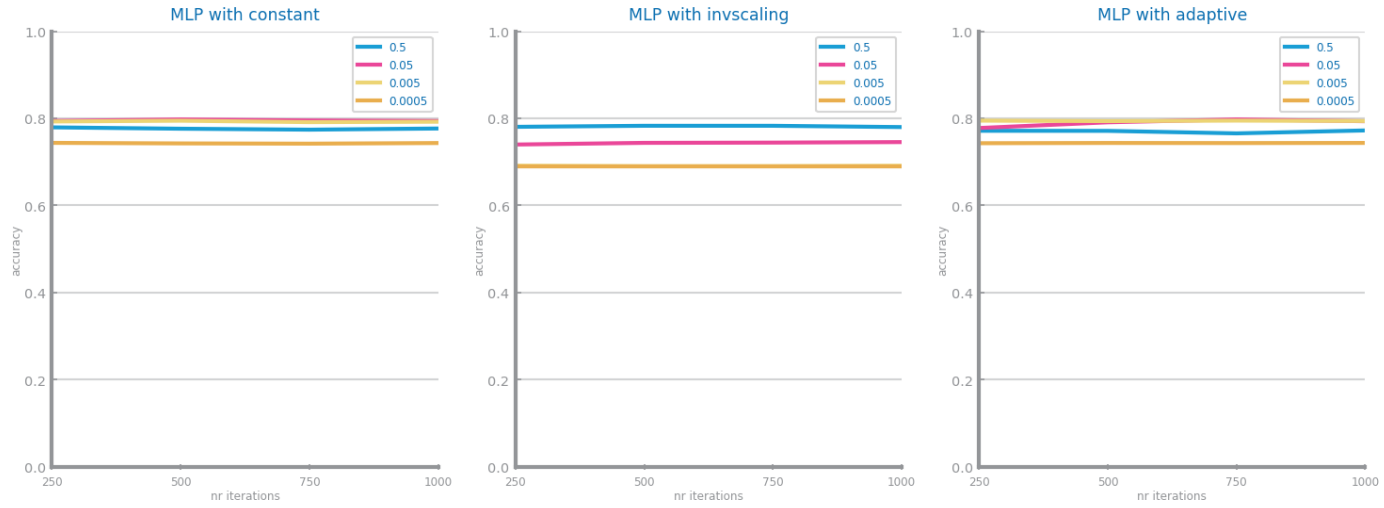


Figure 59 MLP different parameterisations comparison for dataset 2

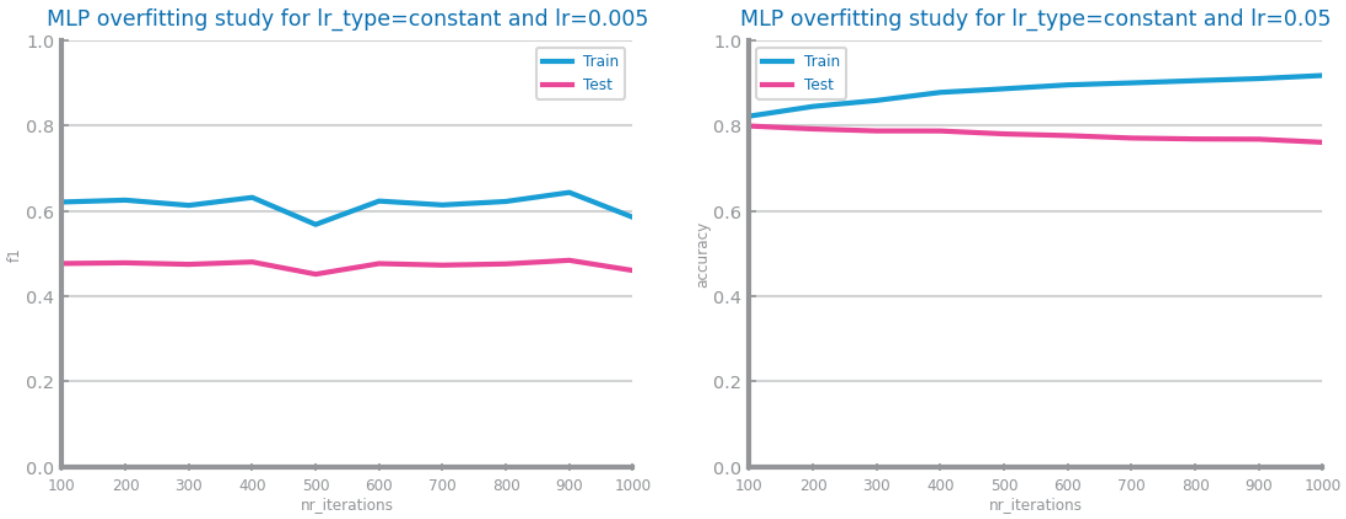


Figure 60 MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

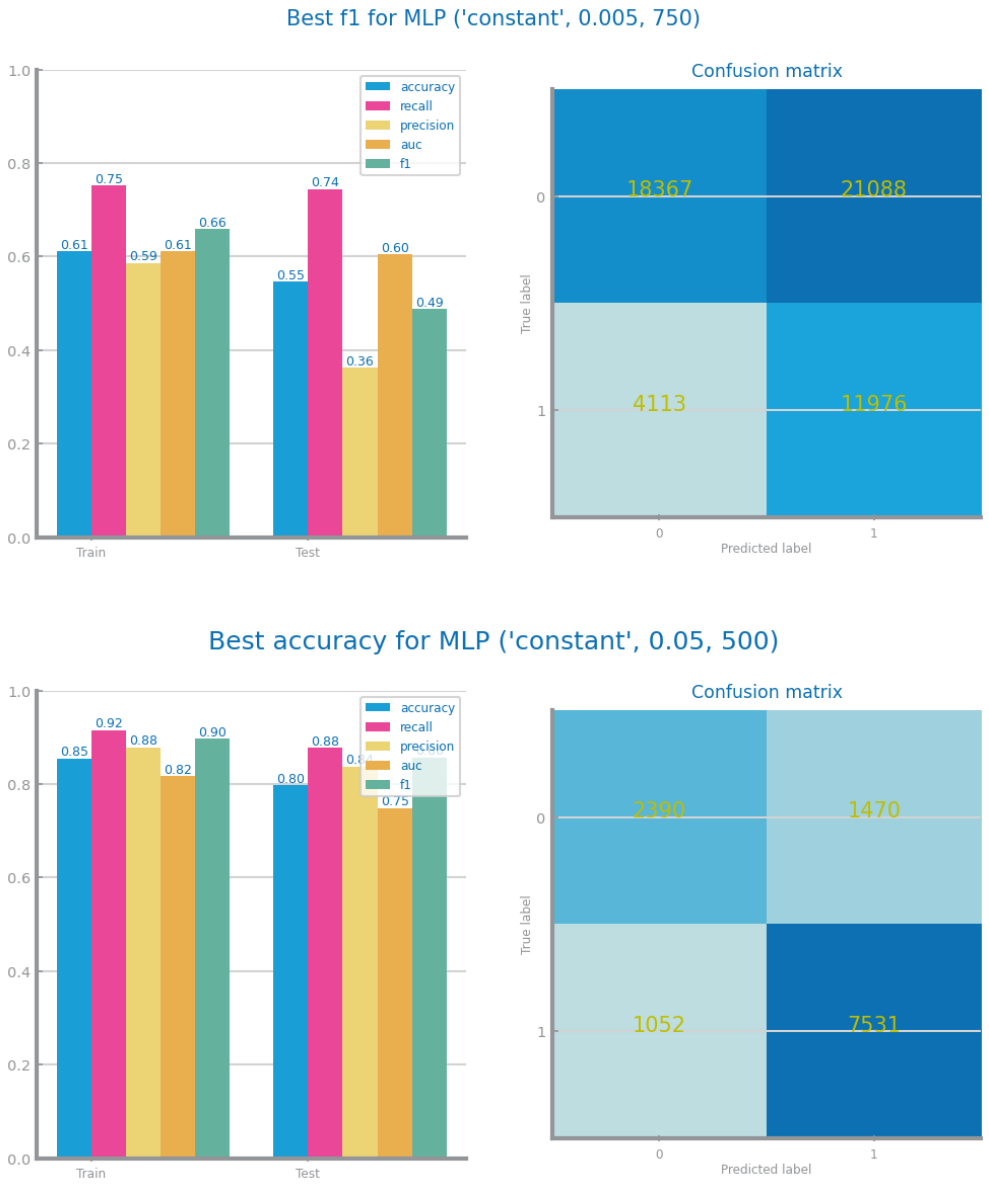


Figure 61 MLP best model results for dataset 1 (up) and dataset 2 (bottom)

4 Critical Analysis

D1 in the health domain and D2 in the finance domain, each presenting unique challenges and objectives. D1, targeting post-COVID scenarios, included 380,932 records with a diverse array of 40 variables. The primary issue encountered was the high rate of MV in key variables like “BMI,” which led to challenges such as the Curse of Dimensionality, increasing the risk of overfitting and complicating data analysis. The focus for D1 was on optimizing the F1 score, aiming to balance precision and recall, a critical aspect for accuracy in health-related predictive modeling

D2's analysis centered on credit data, comprising 100,000 records with 28 variables. Here, the significant challenge was handling the substantial MV in variables like "Credit Mix," necessitating meticulous preprocessing to maintain the integrity of the predictive models. The focus placed on maximizing accuracy, aligning with high stakes of predictive accuracy in financial decision-making processes

For D1, GB and RF were integral, effectively managing the imbalanced data, a crucial factor for optimizing the F1 score. The strategic application of FS notably enhanced their efficacy. GB stood out for its capability to handle missing data efficiently and prevent overfitting, a common issue in extensive health datasets

In contrast, D2 saw the dominance of RF and MLP, which demonstrated superior performance, largely benefiting from appropriate data scaling and balancing. While Decision Trees and KNN showed potential, they were surpassed by the more complex models, underlining the necessity for advanced analytical techniques in the financial sector

A cross-model comparative analysis revealed that variables such as age and income were consistently influential across both datasets, underscoring their pivotal role in predictive modeling. This project exemplified the importance of selecting specific models and data preparation techniques that are tailored to the distinct characteristics and requirements of each dataset

5 DATA PROFILING

Data Dimensionality and Granularity

S1's granularities: hourly (H) reveals high fluctuation, weekly (W) smooths variations, and monthly (M) shows clearest trends. S2: daily (D) captures most variance, weekly (W) offers a balance, and yearly (Y) presents the broadest trends. Finest granularities (H for S1, D for S2) provide detailed data, whereas coarser granularities (W, M/Y) are better for trend analysis, sacrificing detail for clarity and reducing noise in the time series.

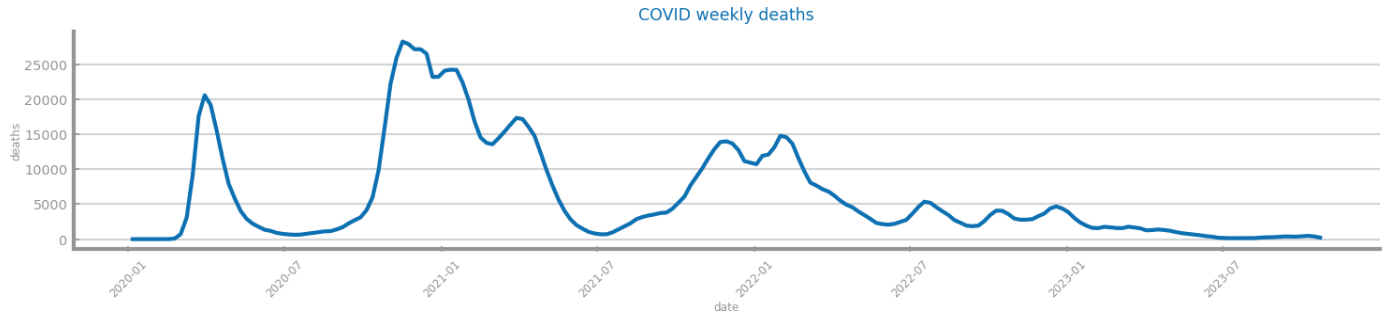


Figure 62 Original time series 1 (the most atomic detail)

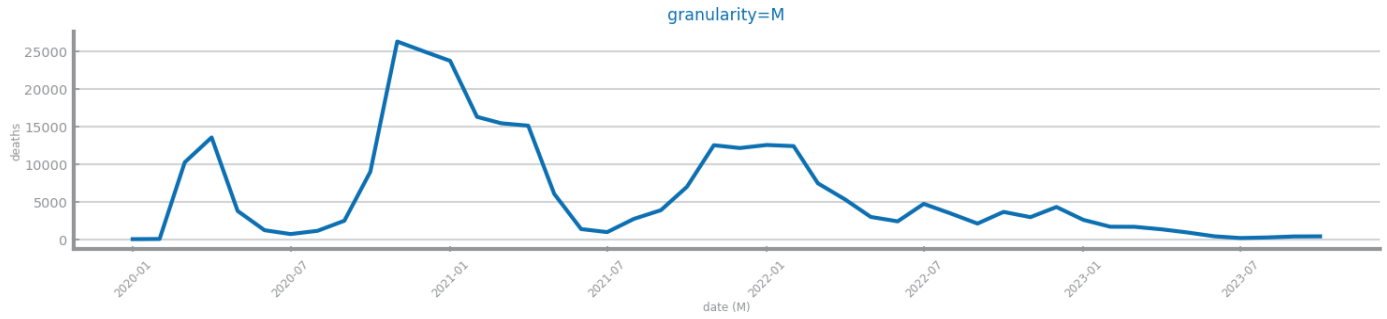


Figure 63 Time series 1 at the second chosen granularity

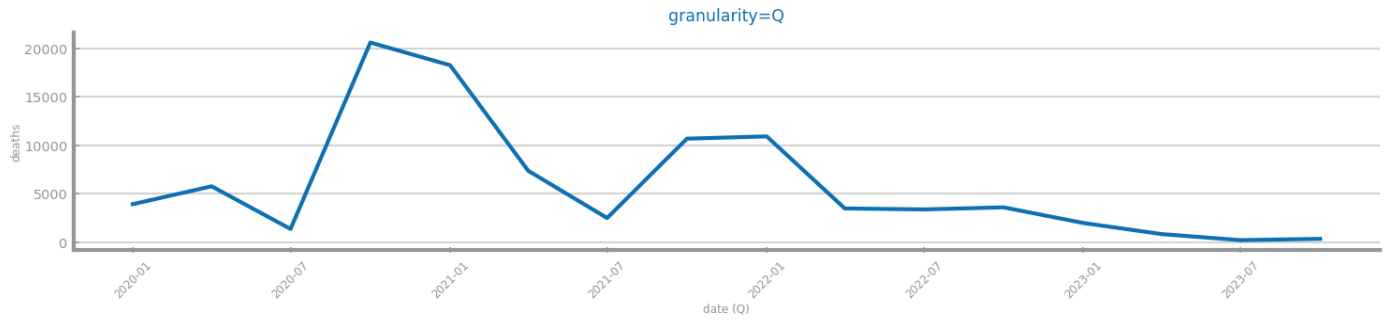


Figure 64 Time series 1 at the third chosen granularity

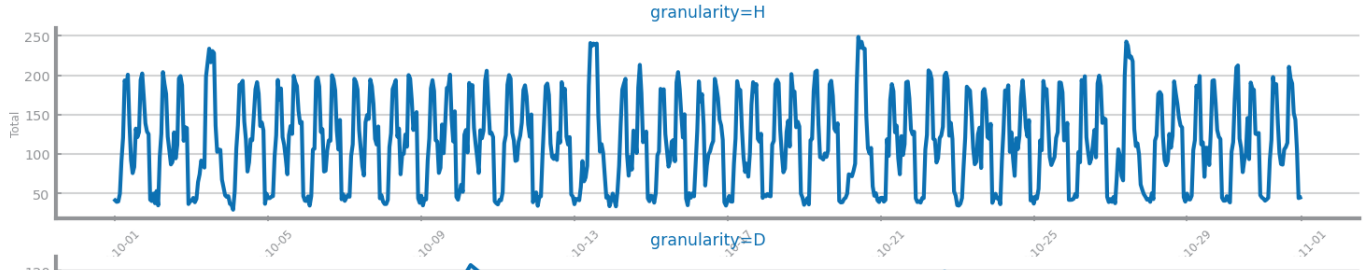


Figure 65 Original time series 2 (the most atomic detail)

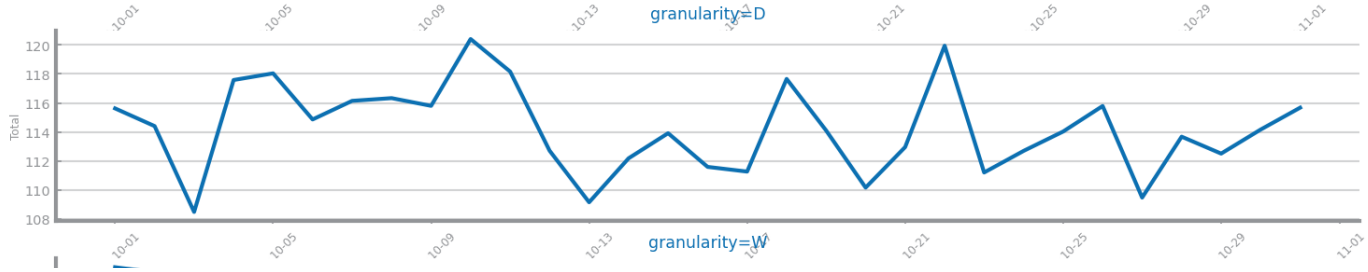


Figure 66 Time series 2 at the second chosen granularity

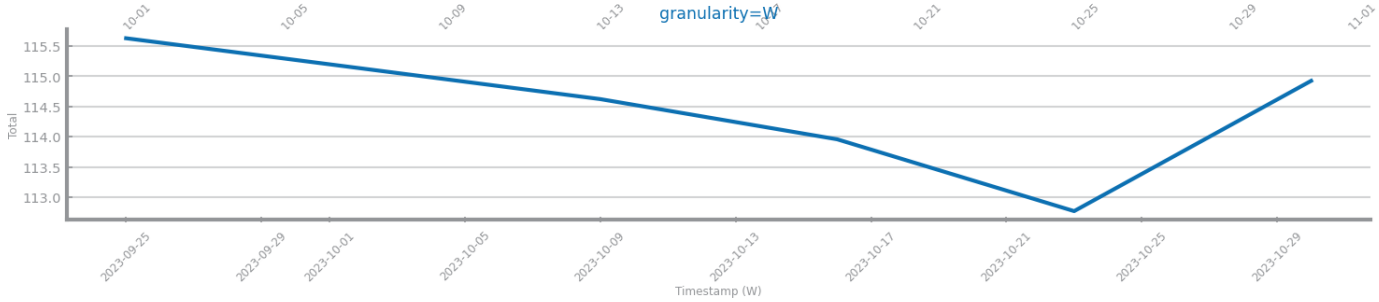
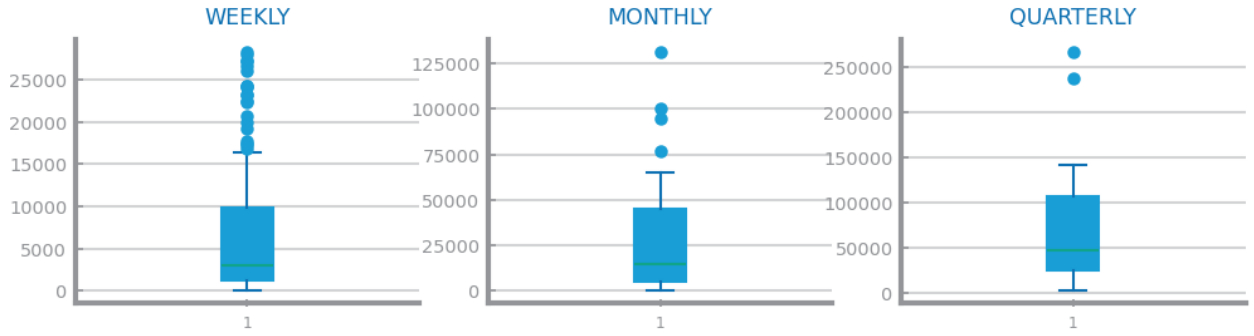


Figure 67 Time series 2 at the third chosen granularity

Data Distribution

S1 displays significant autocorrelation; weekly aggregation captures COVID death trends with lower variance and a pronounced right skew. S2 shows consistent hourly autocorrelation, indicating the importance of finer granularity to model traffic patterns accurately. Both choose granularities—weekly for S1 and hourly for S2—are appropriate for the distinct natures of the datasets, emphasizing the need for models that can account for strong temporal correlations and varying volatility.



count	199.000000
mean	6224.371859
std	7140.064648
min	0.000000
25%	1189.000000
50%	3088.000000
75%	9886.000000
max	28296.000000

Name: deaths, dtype: float64

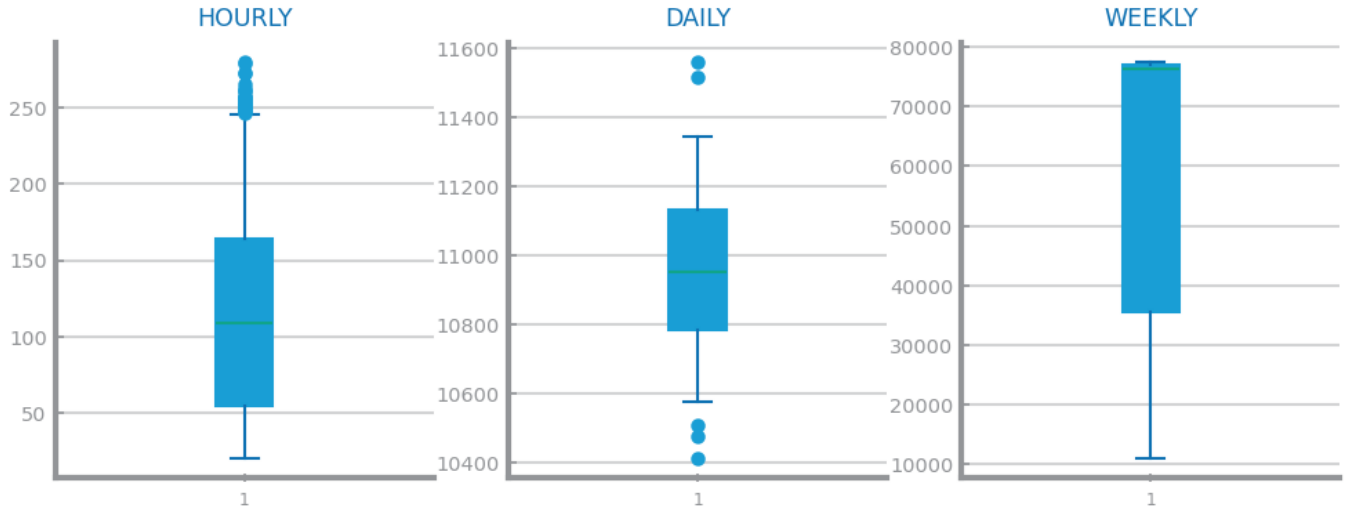
count	46.000000
mean	26927.173913
std	30714.588492
min	1.000000
25%	5418.250000
50%	14911.000000
75%	45427.000000
max	131687.000000

Name: deaths, dtype: float64

count	16.000000
mean	77415.625000
std	80730.901127
min	1490.000000
25%	24073.500000
50%	46294.500000
75%	106797.500000
max	267869.000000

Name: deaths, dtype: float64

Figure 68 Boxplots for time series 1 at different granularities



count	2976.000000
mean	114.218414
std	60.190627
min	21.000000
25%	55.000000
50%	109.000000
75%	164.000000
max	279.000000

Name: Total, dtype: float64

count	31.000000
mean	10964.967742
std	292.135868
min	10415.000000
25%	10785.500000
50%	10954.000000
75%	11133.500000
max	11559.000000

Name: Total, dtype: float64

count	6.000000
mean	56652.333333
std	31235.531535
min	11100.000000
25%	35494.250000
50%	76180.500000
75%	76913.500000
max	77363.000000

Name: Total, dtype: float64

Figure 69 Boxplots for time series 2 at different granularities

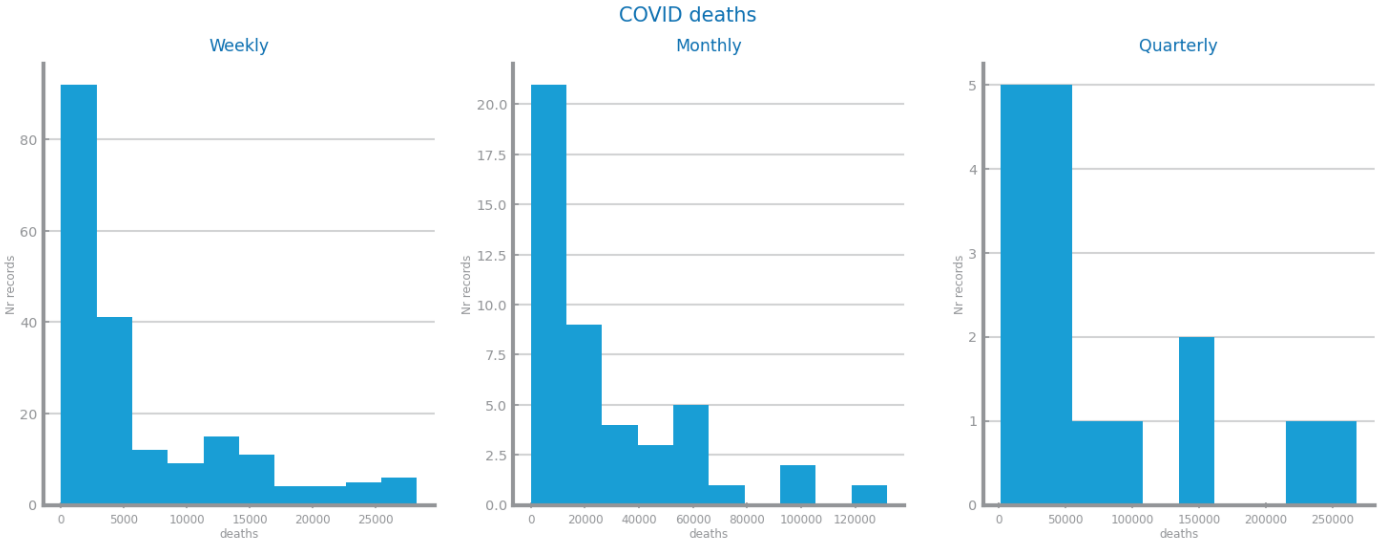


Figure 70 Histograms for time series 1 at different granularities

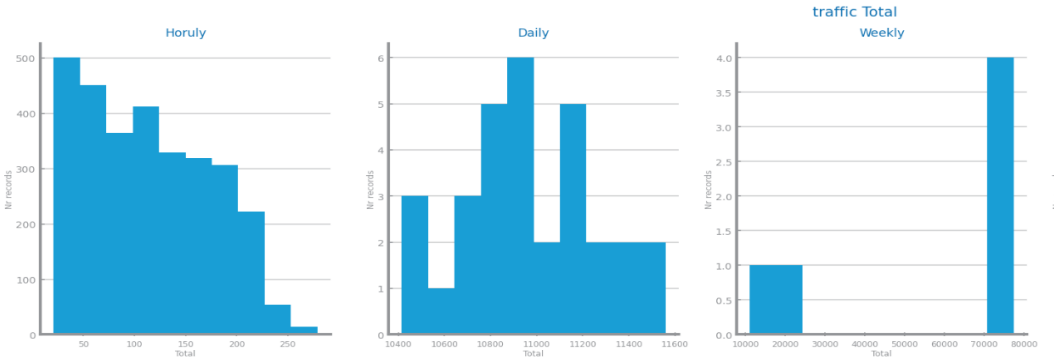


Figure 71 Histograms for time series 2 at different granularities

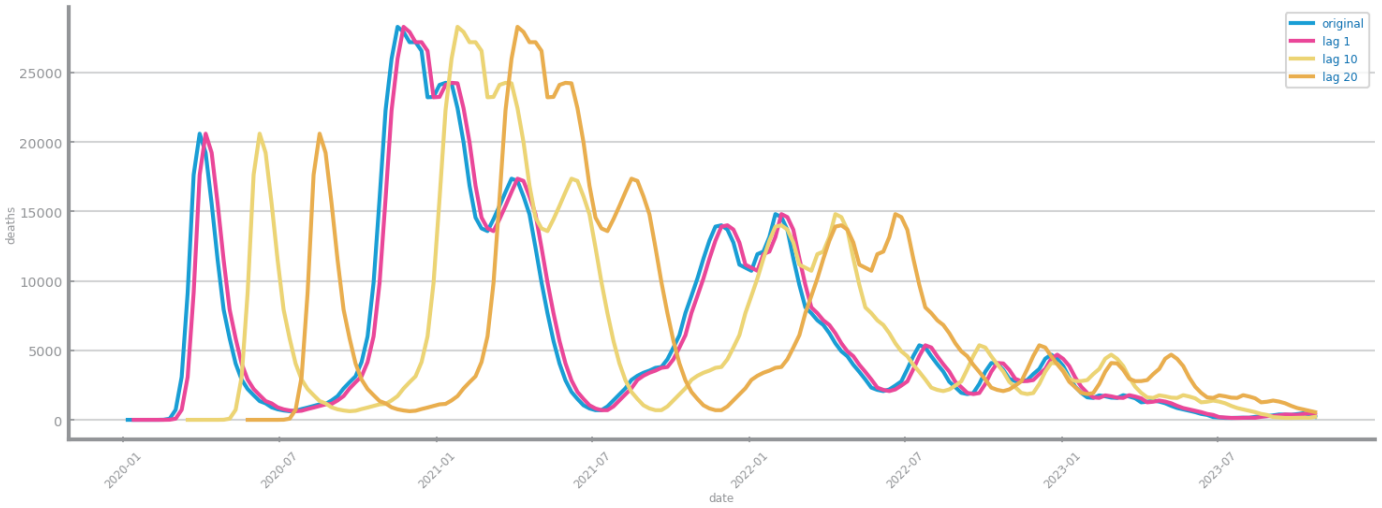


Figure 72 Autocorrelation lag-plots for original time series 1

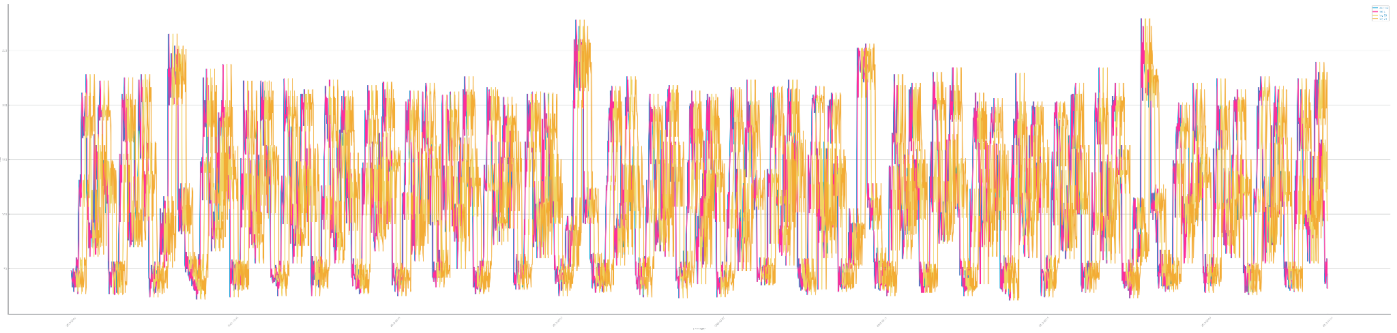


Figure 73 Autocorrelation lag-plots for original time series 2

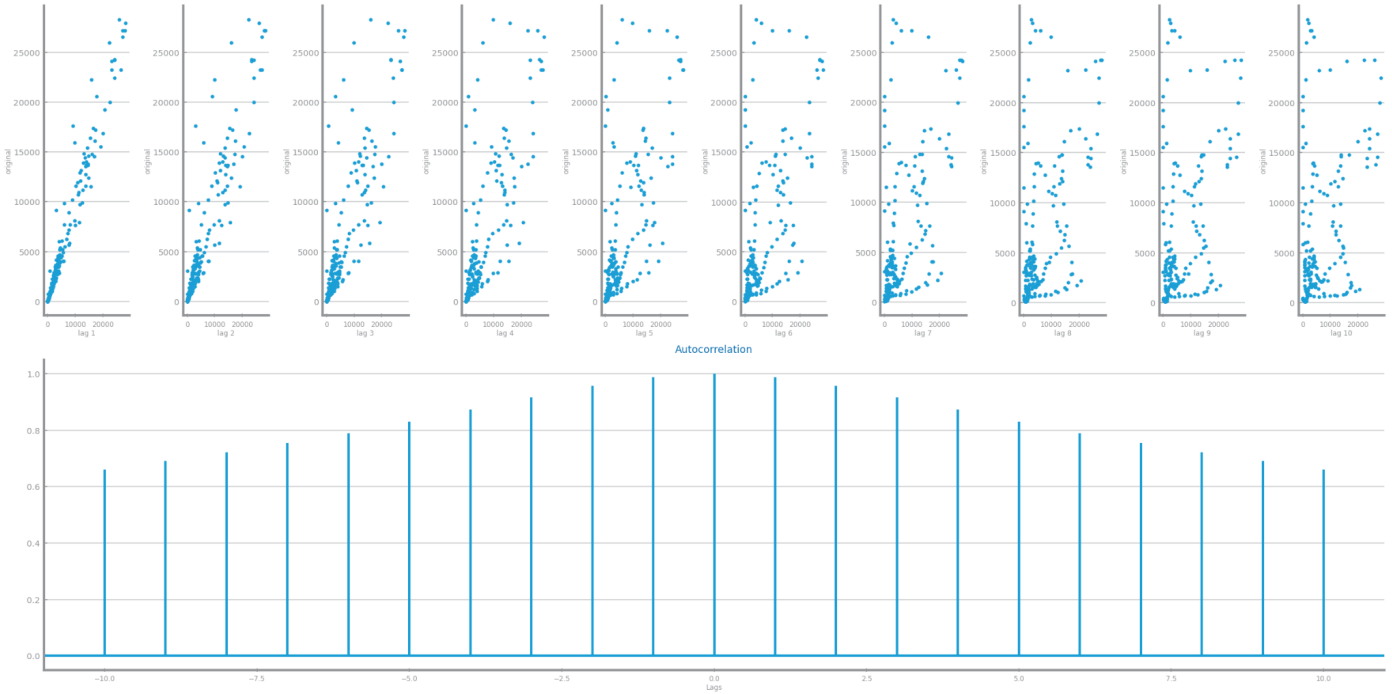


Figure 74 Autocorrelation correlogram for original time series 1

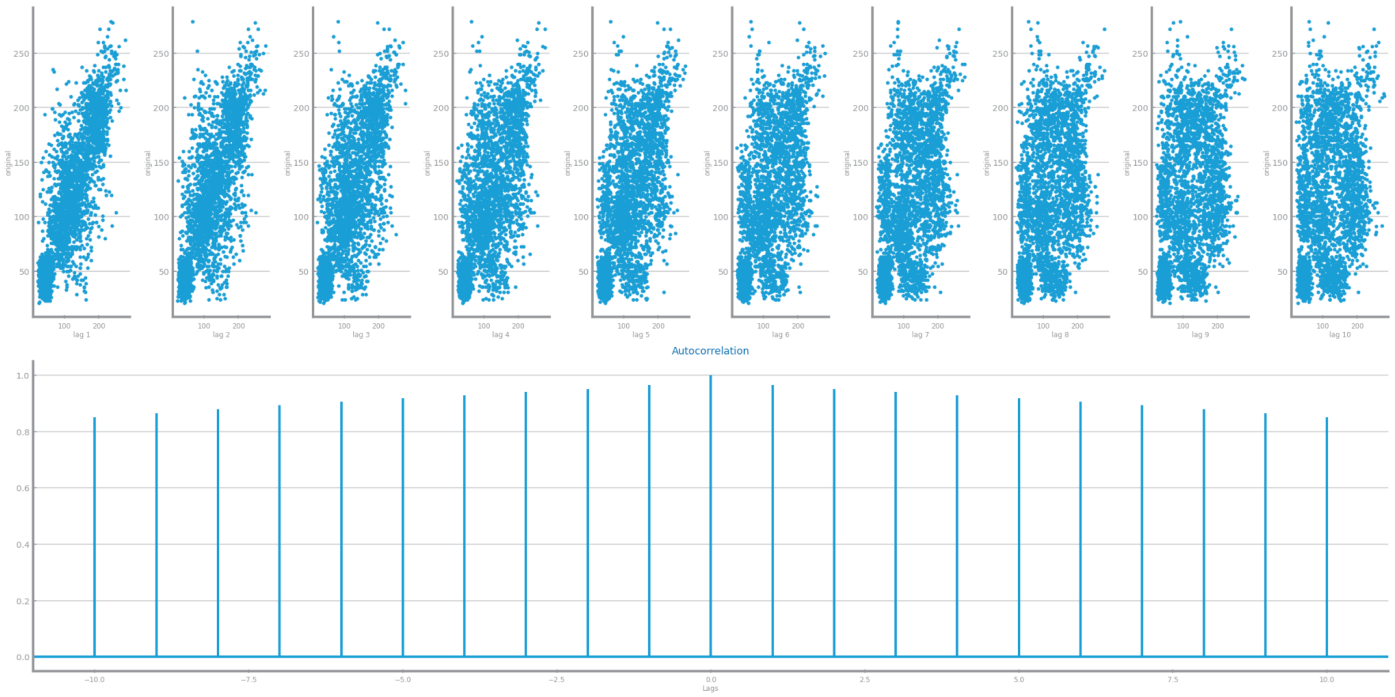


Figure 75 Autocorrelation correlogram for original time series 2

Data Stationarity

S1, with observed seasonality and trend, is non-stationary (ADF: -2.608, p > 0.05). Graphs show mean and variance shifts, corroborating ADF results. S2, despite pronounced seasonality, is stationary (ADF: -9.927, p ≈ 0). Rolling stats likely remain within confidence intervals, aligning with ADF findings. S1 requires differencing; S2 is model-ready. Analyzing at different granularities ensures robust forecasting, as seen with S2's daily cycles and S1's broader trends.

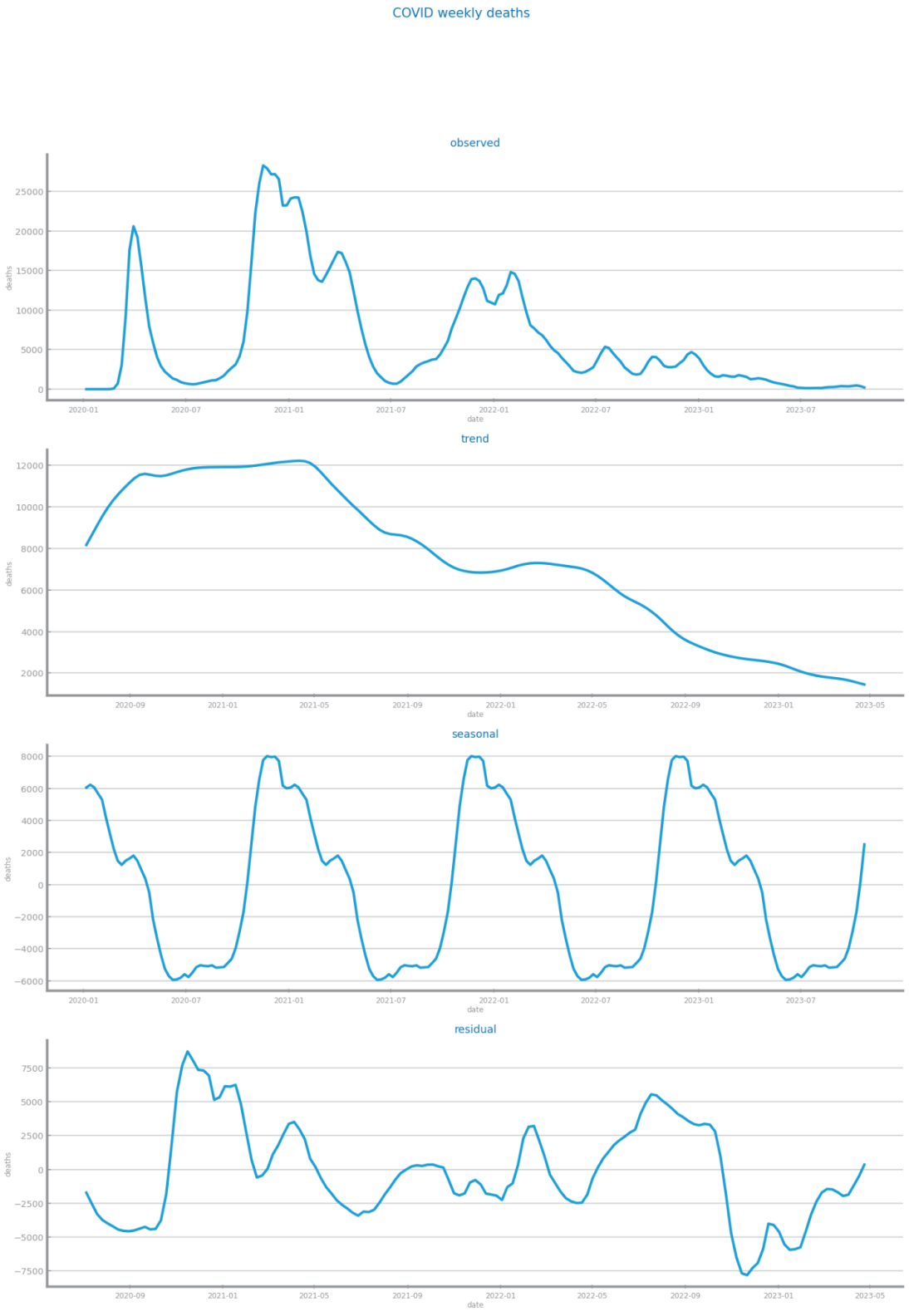


Figure 76 Components study for time series 1

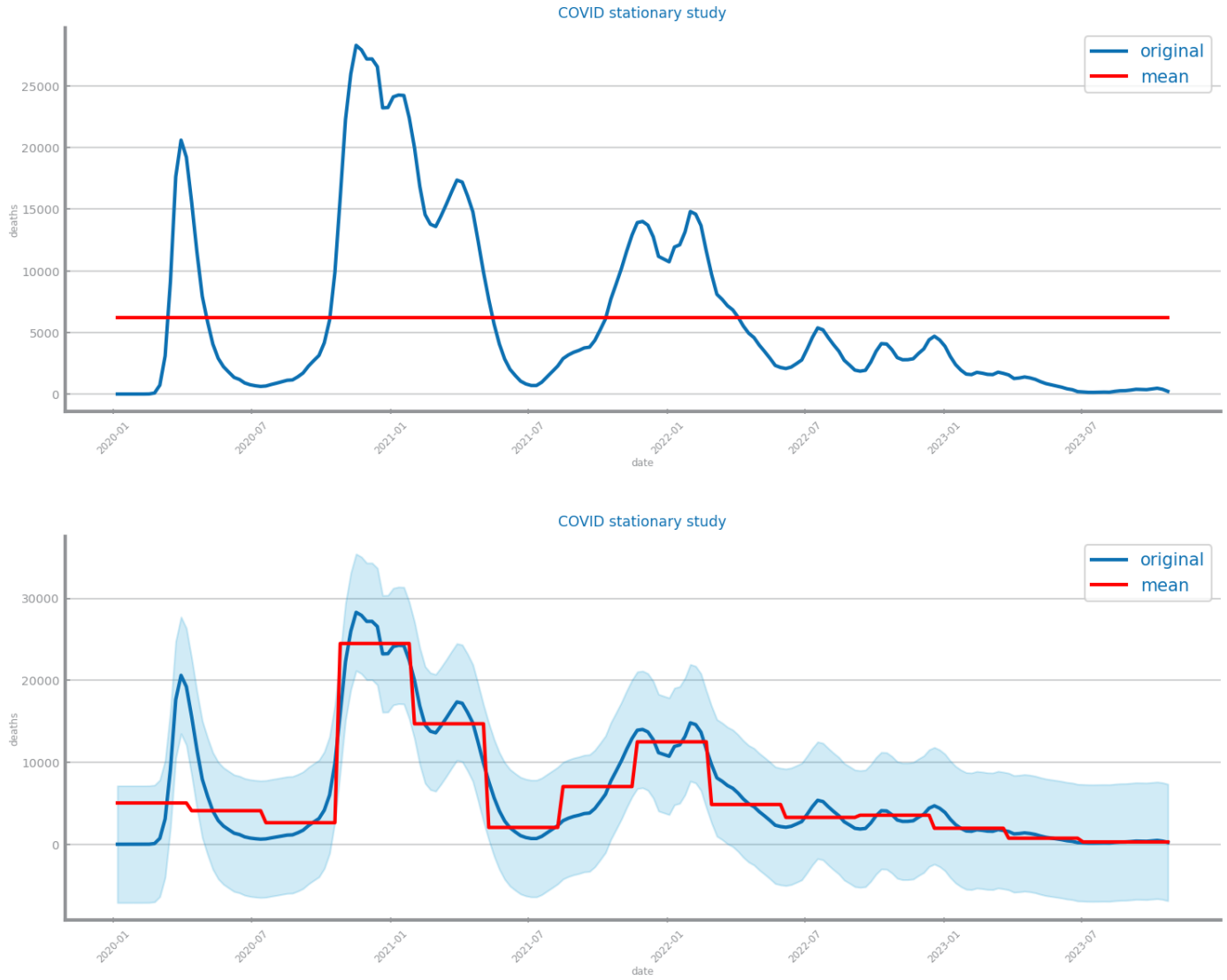


Figure 77 Stationarity study for time series 1

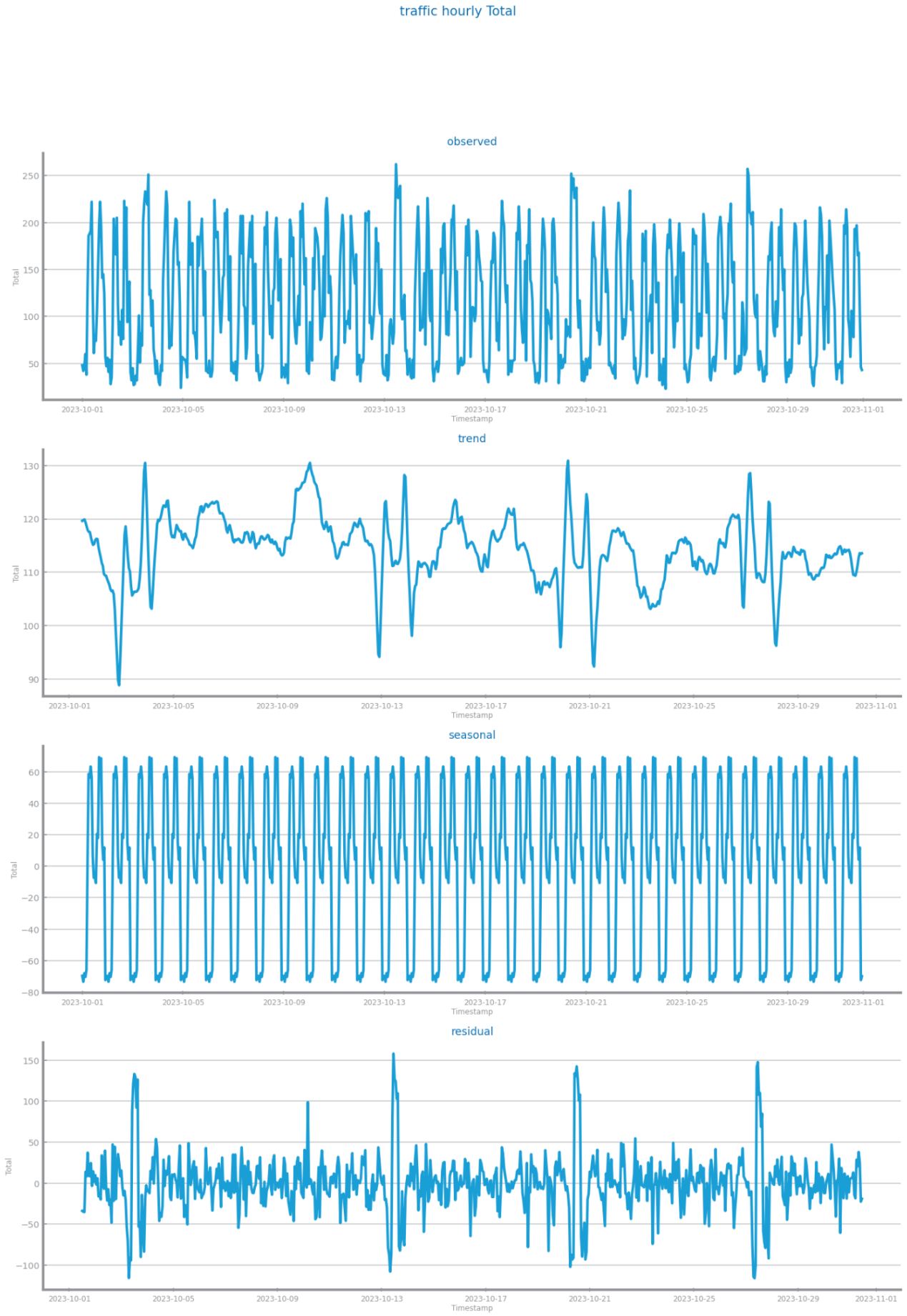


Figure 78 Components study for time series 2

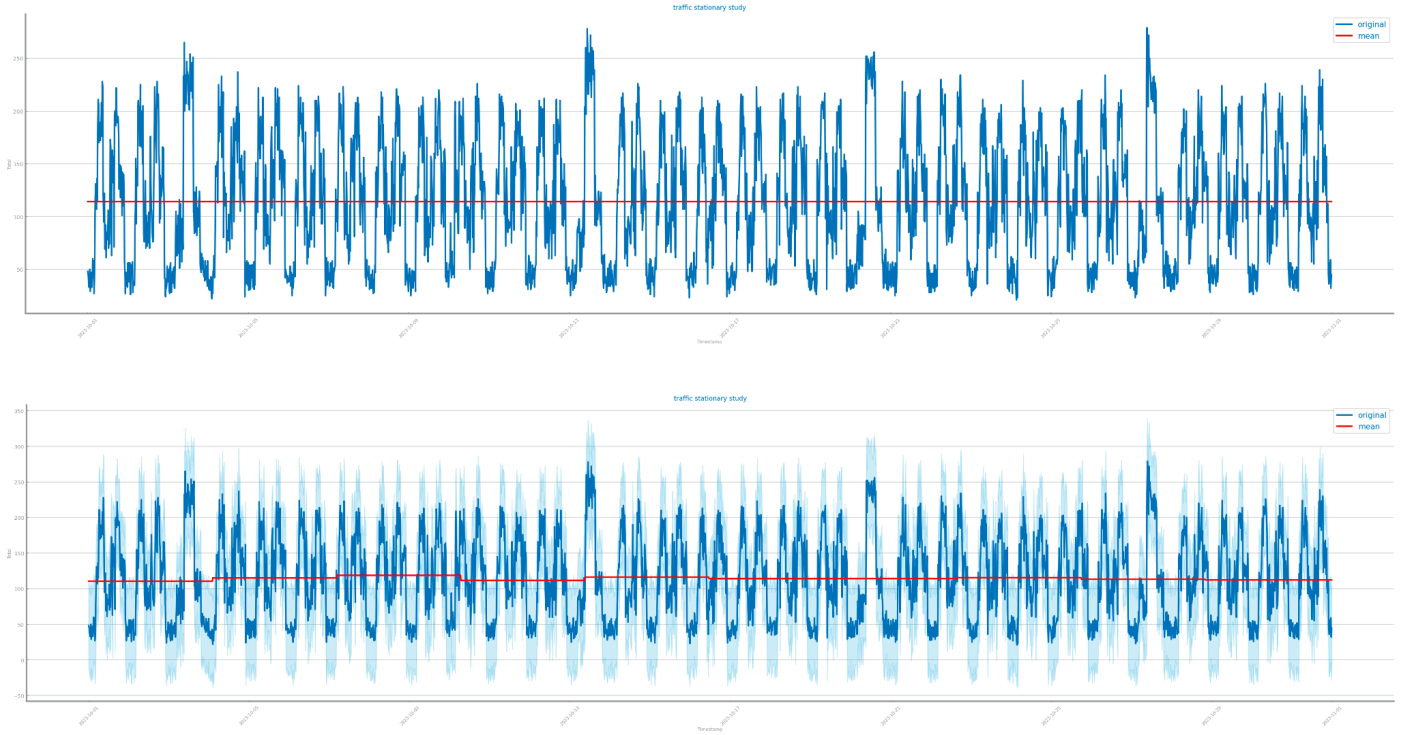


Figure 79 Stationarity study for time series 2

6 Data Transformation

Aggregation

For S1, weekly aggregation balances detail and model accuracy, with moderate RMSE/MAE. For S2, hourly granularity yields the lowest errors and highest R2, optimal for forecasting.

Figure 80 Forecasting plots after different aggregations on time series 1

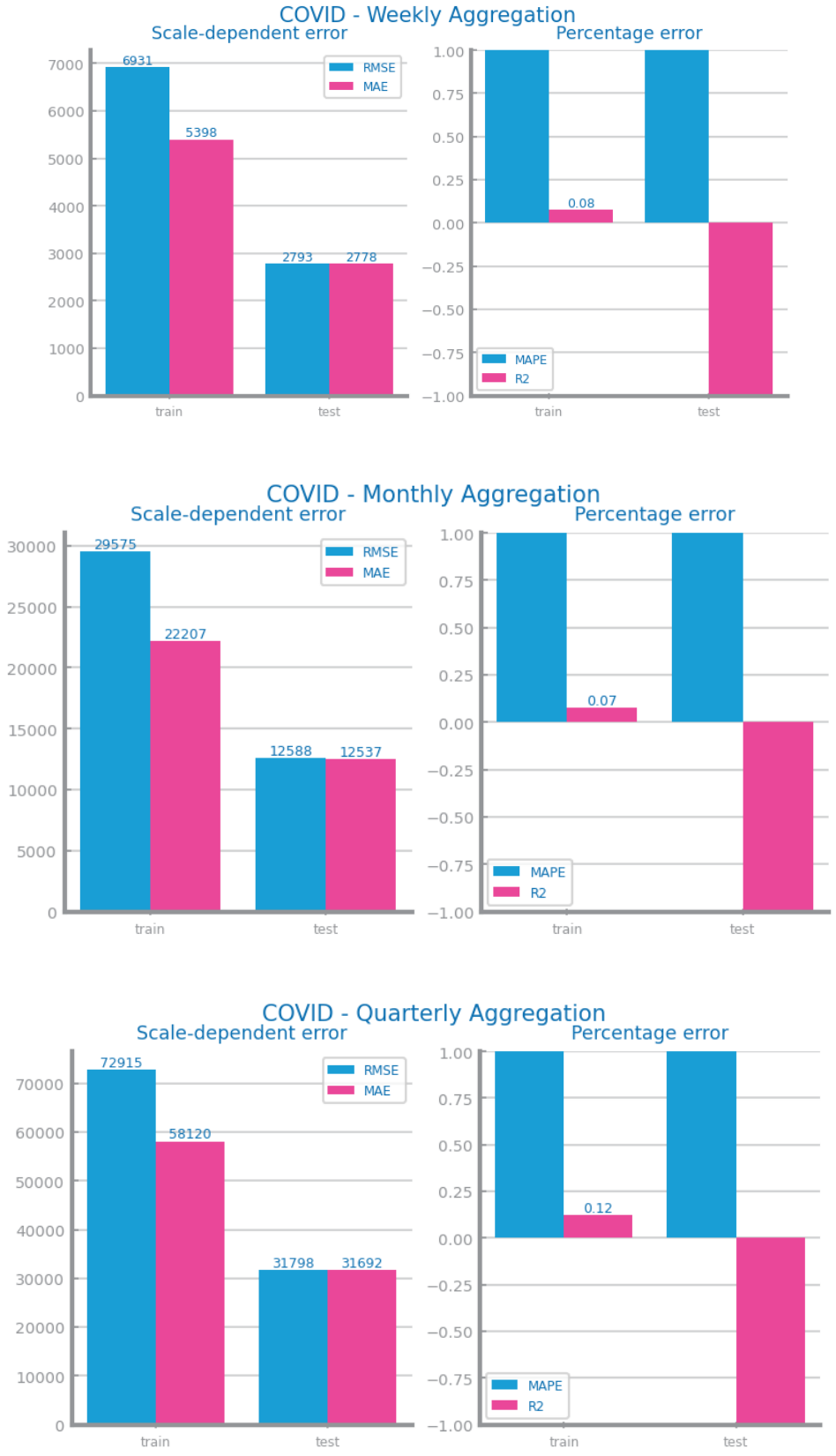


Figure 81 Forecasting results after different aggregations on time series 1

Figure 82 Forecasting plots after different aggregations on time series 2

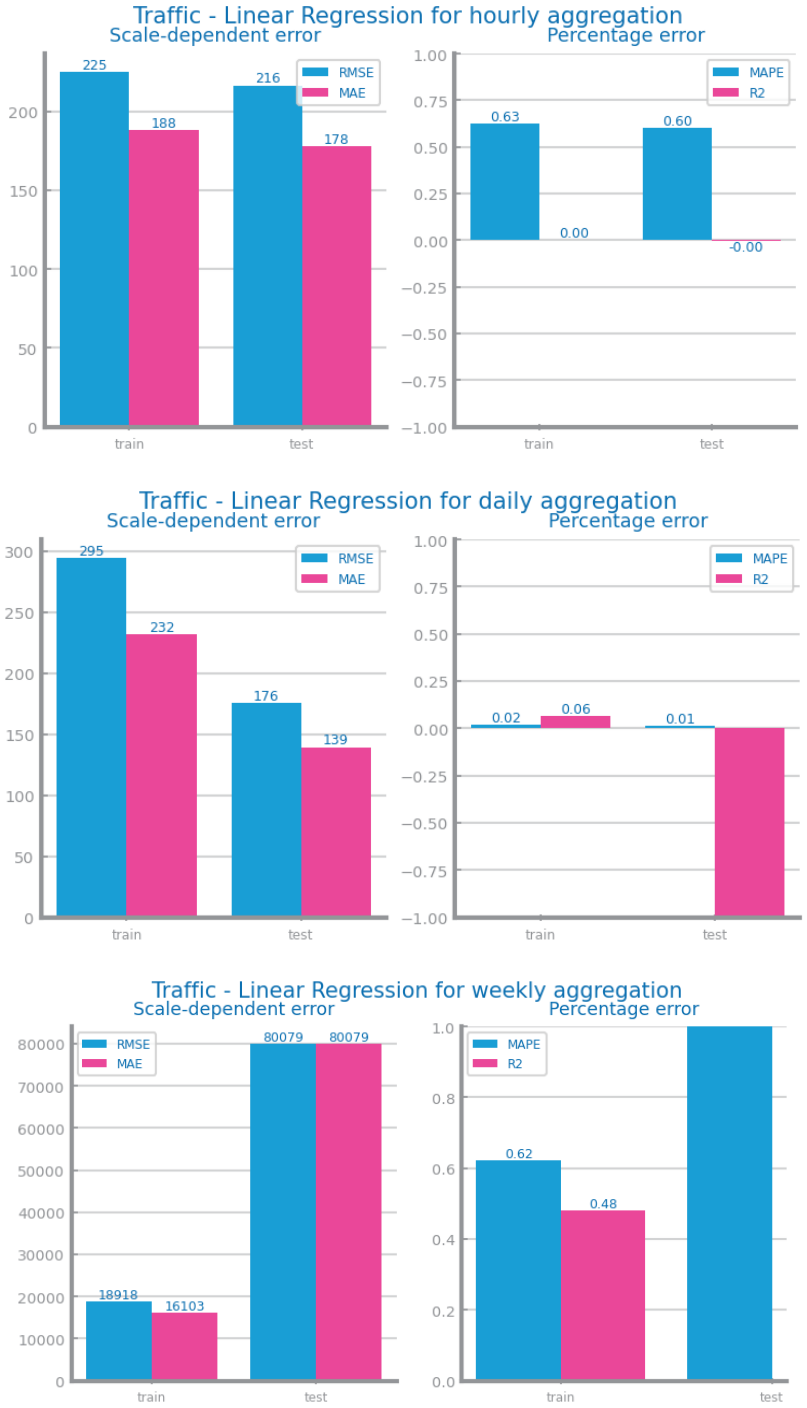


Figure 83 Forecasting results after different aggregations on time series 2

Smoothing

S1 is better the higher the window size, mainly because of the information loss and not because of a better model prediction. So no smoothing was chosen to proceed in S1 due to information loss. In S2 smoothing with window size = 10 was a good balance between information loss and prediction metrics, so it was selected.

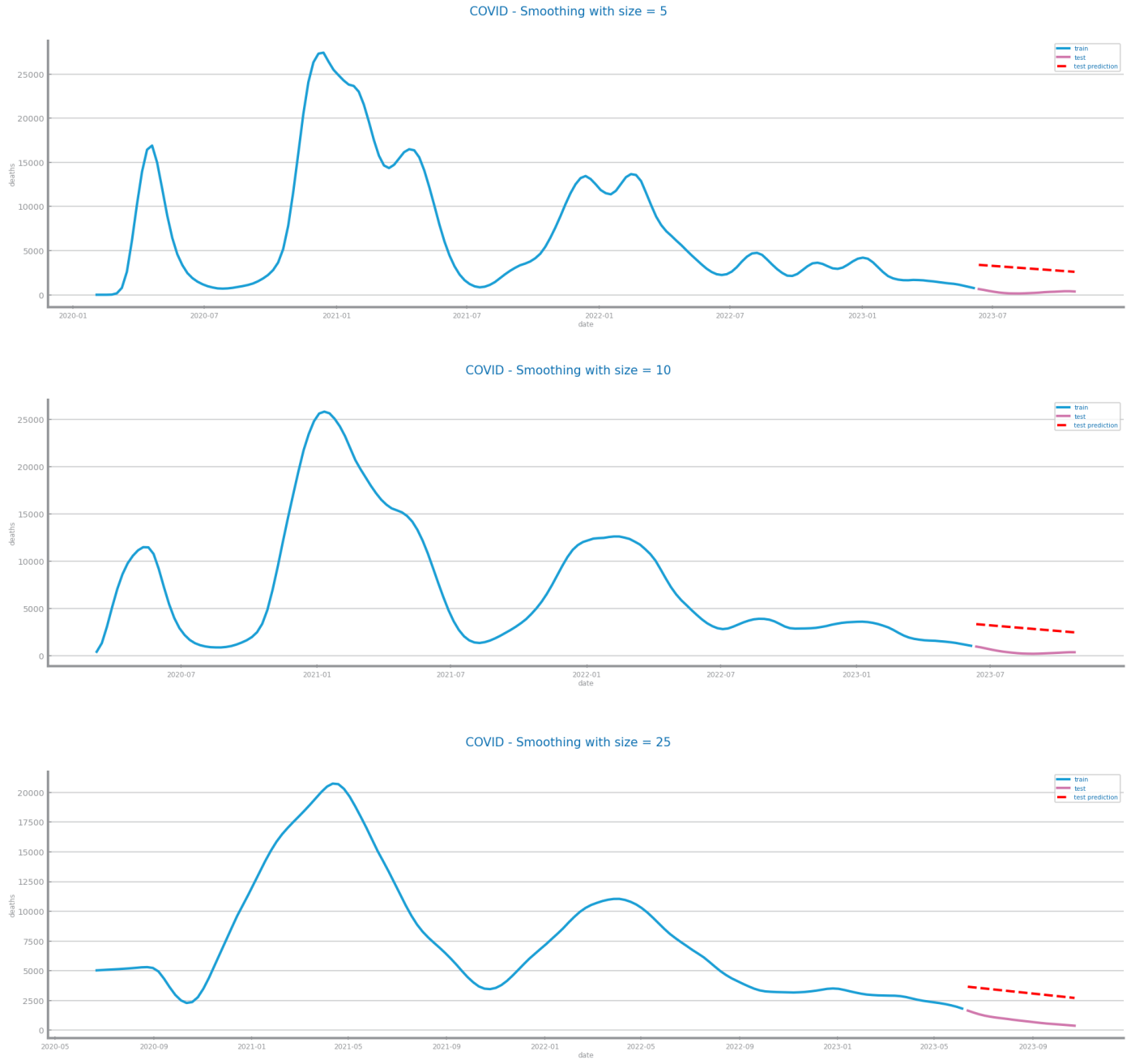


Figure 84 Forecasting plots after different smoothing parameterisations on time series 1

Figure 85 Forecasting results after different smoothing parameterisations on time series 1

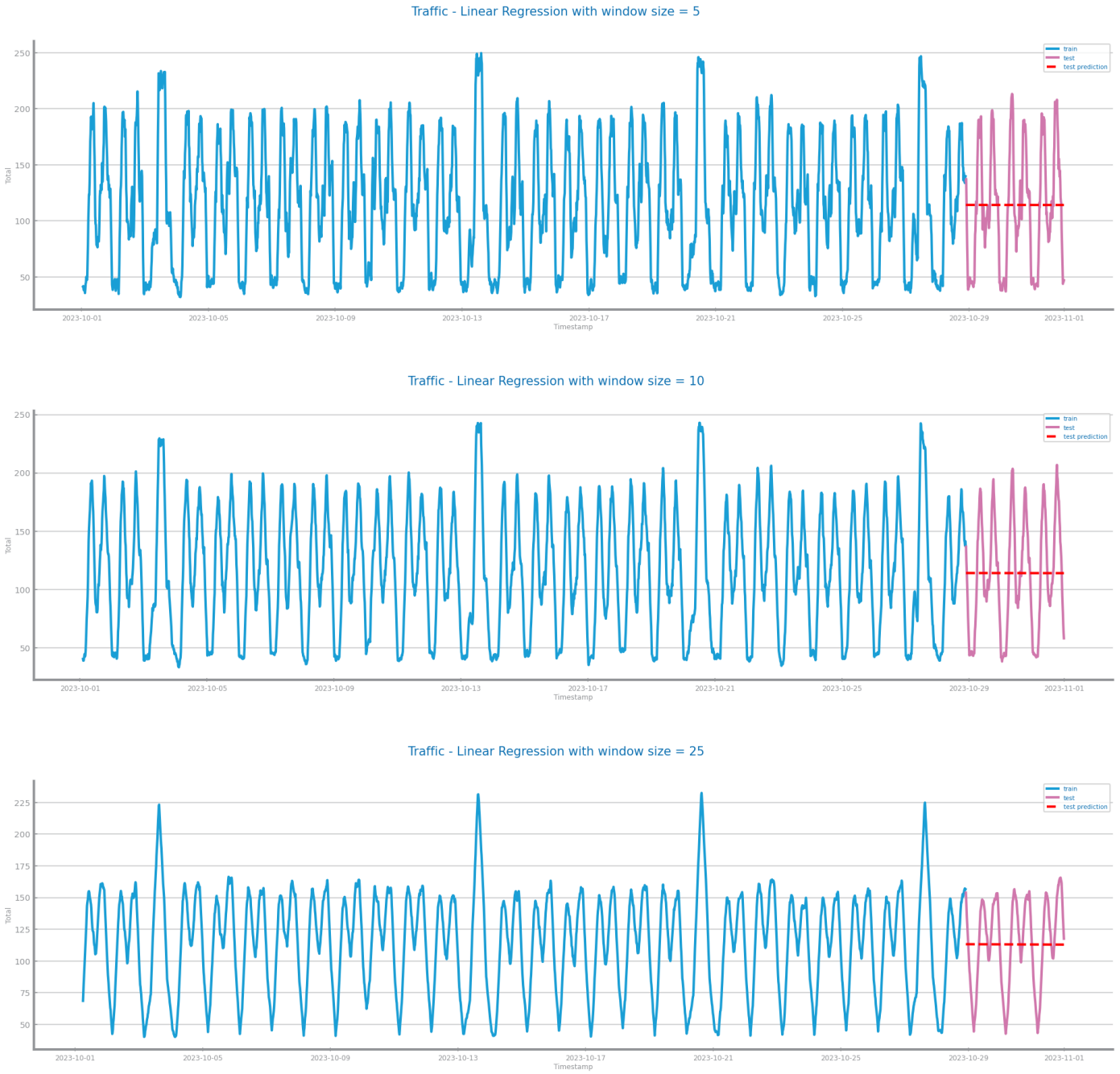


Figure 86 Forecasting plots after different smoothing parameterisations on time series 2

Figure 87 Forecasting results after different smoothing parameterisations on time series 2

Differentiation

1st and 2nd differentiations of S1 show lower RMSE and MAE, with the model predicting fairly well the real values, however with a negative R^2. S2 shows lower RMSE and MAE but zero R^2 in tests, indicating no explanatory power. 1st diff. was chosen to proceed in S1 for its lower error rates and less noise, keeping in mind the identified issues with R^2. No differentiation was applied on S2.

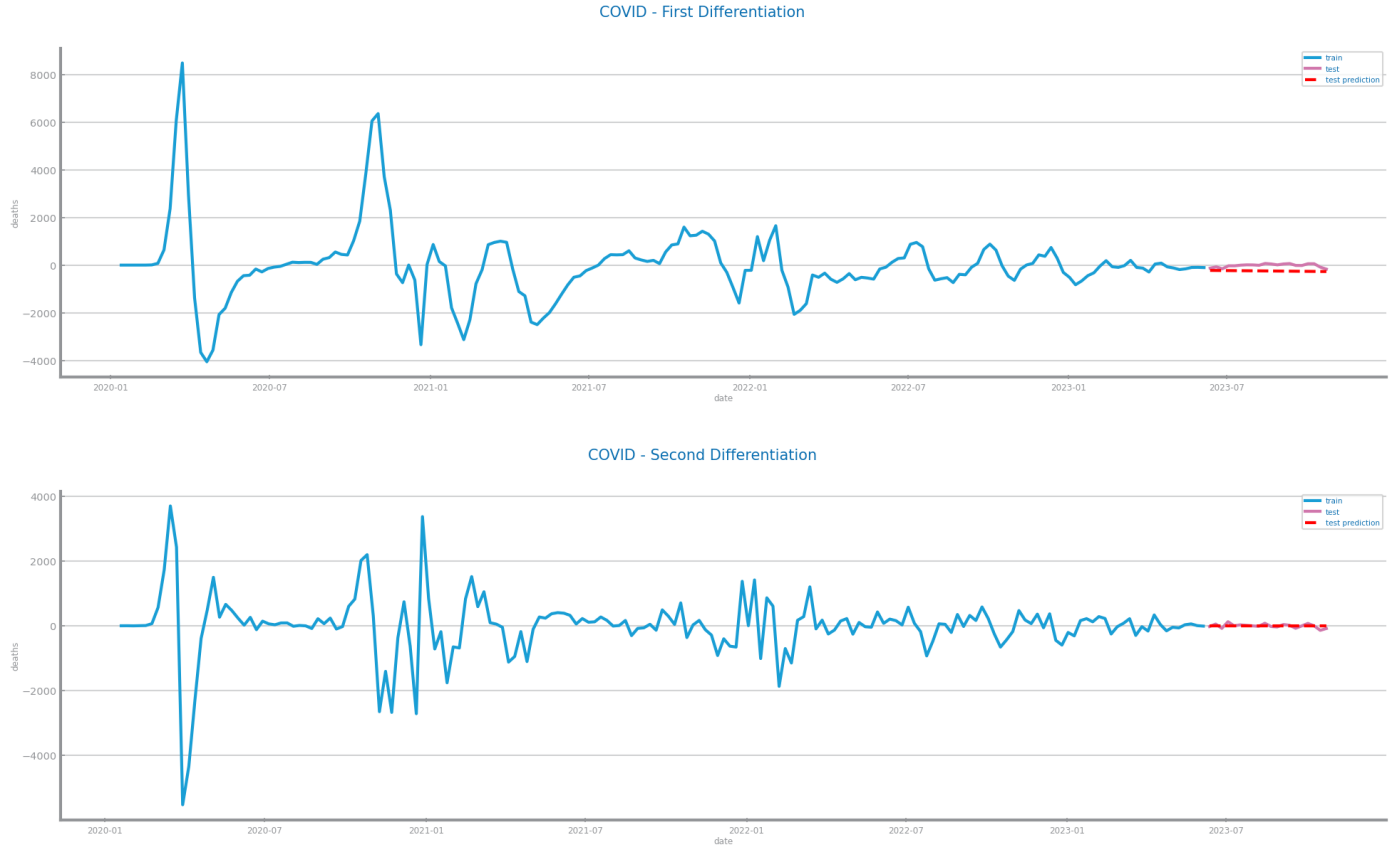
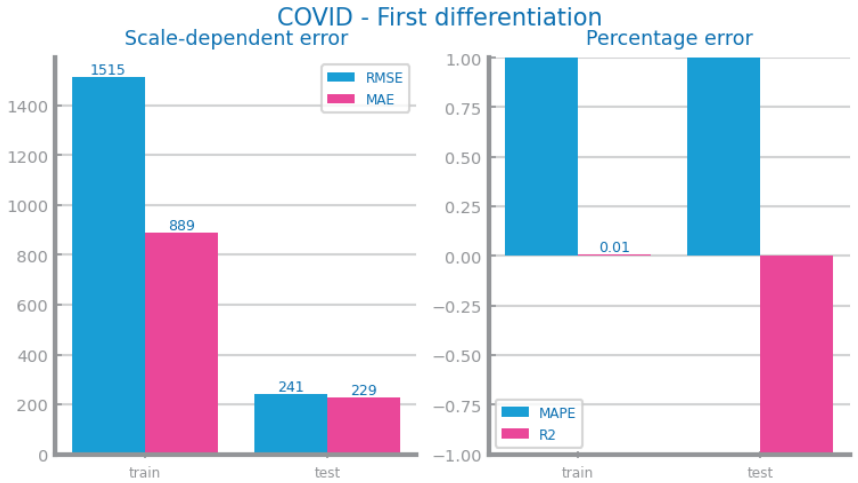


Figure 88 Forecasting plots after first and second differentiation of time series 1



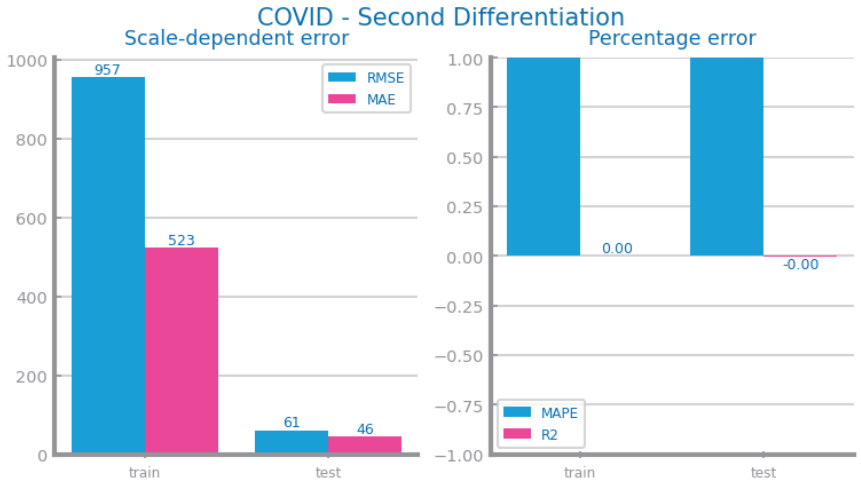


Figure 89 Forecasting results after first and second differentiation of time series 1

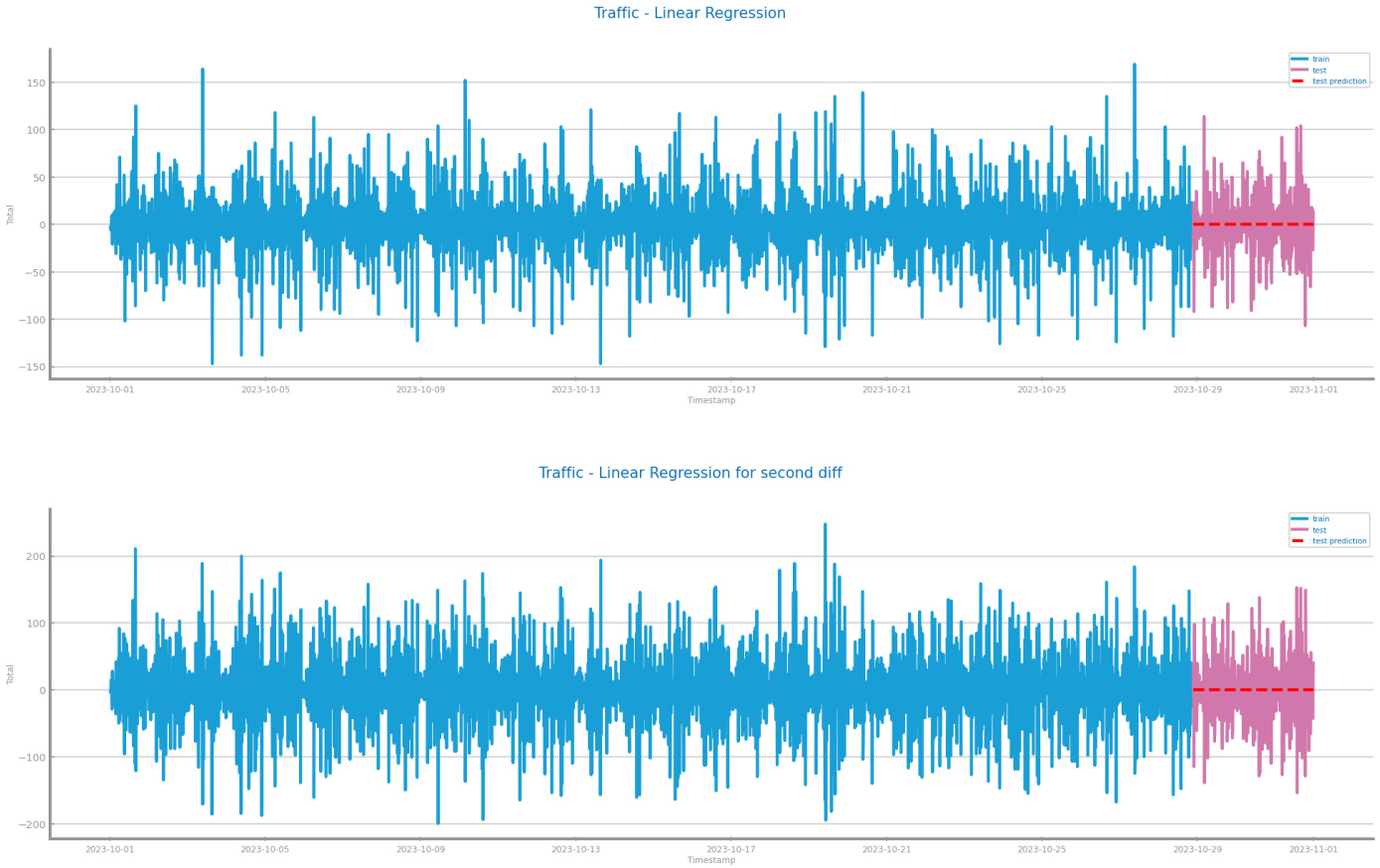


Figure 90 Forecasting plots after first and second differentiation of time series 2

Figure 91 Forecasting results after first and second differentiation of time series 2

7 Models' Evaluation

First differentiation was applied to the COVID time series and smoothing with window size = 10 was applied to the traffic time series before moving to the model's evaluation.

Simple Average Model

S1 shows high errors in training but better in test, with negative R^2. This happens because the train set in rather unpredictable whereas the test set is very stable. For S2, model generalizes well with smaller error in test set but has significant errors. SAModel fits S1, but not S2.

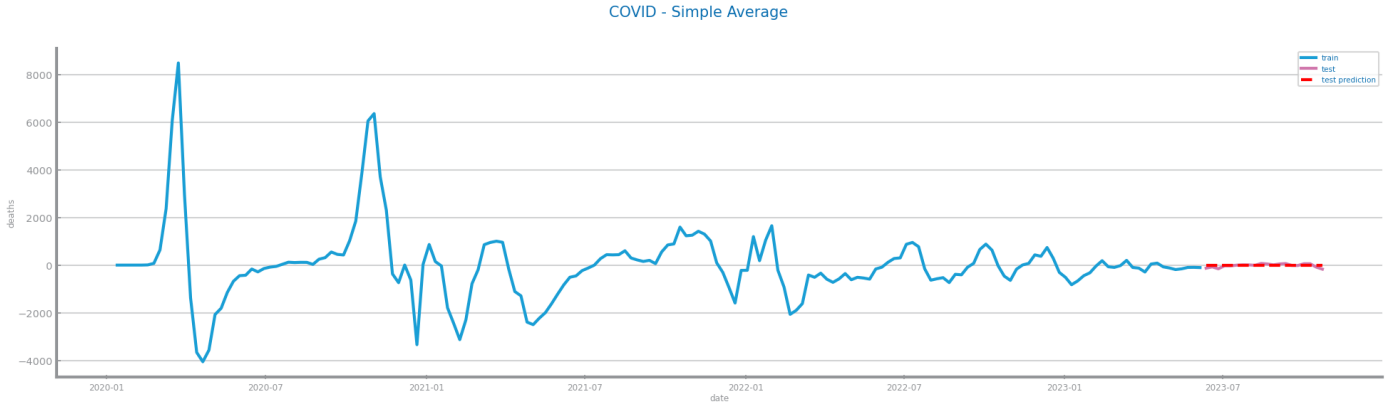


Figure 96 Forecasting plots obtained with Simple Average model over time series 1

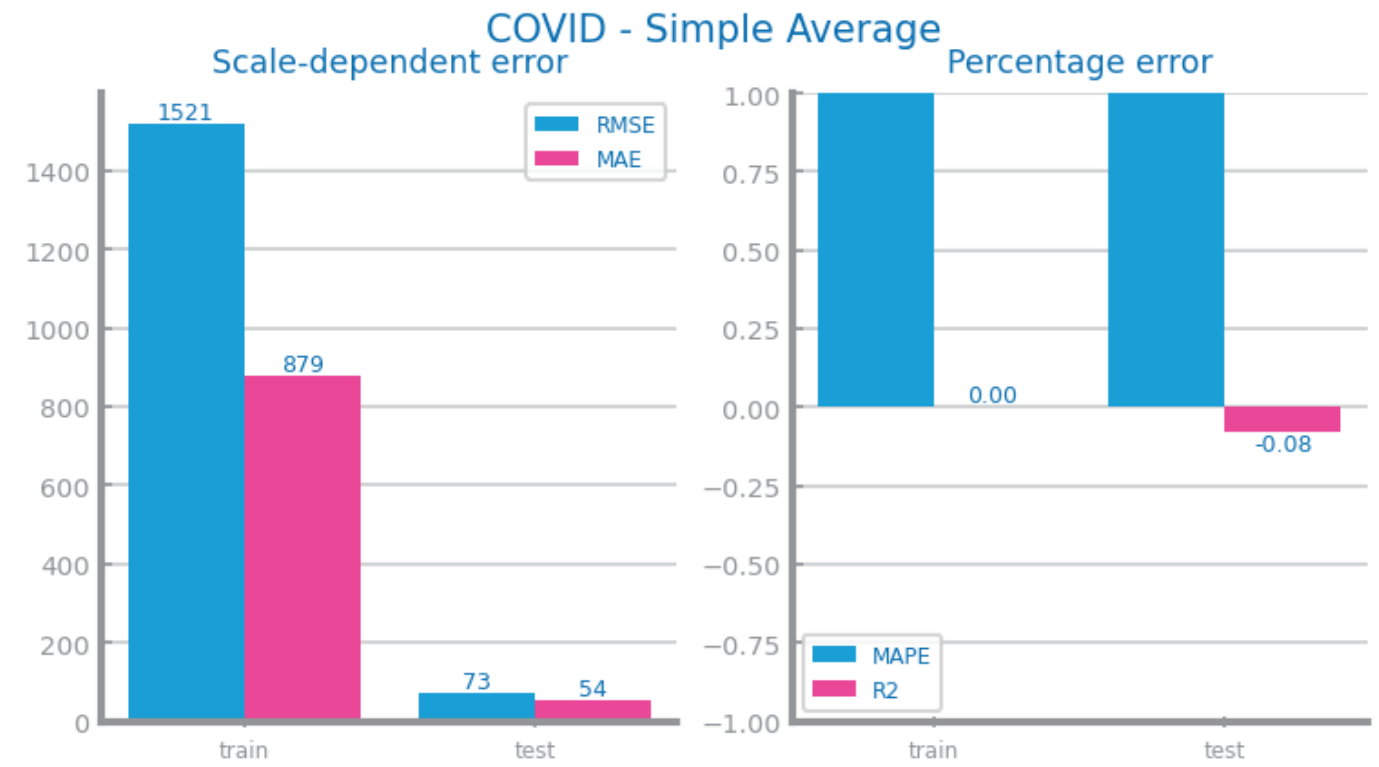


Figure 97 Forecasting results obtained with Simple Average model over time series 1

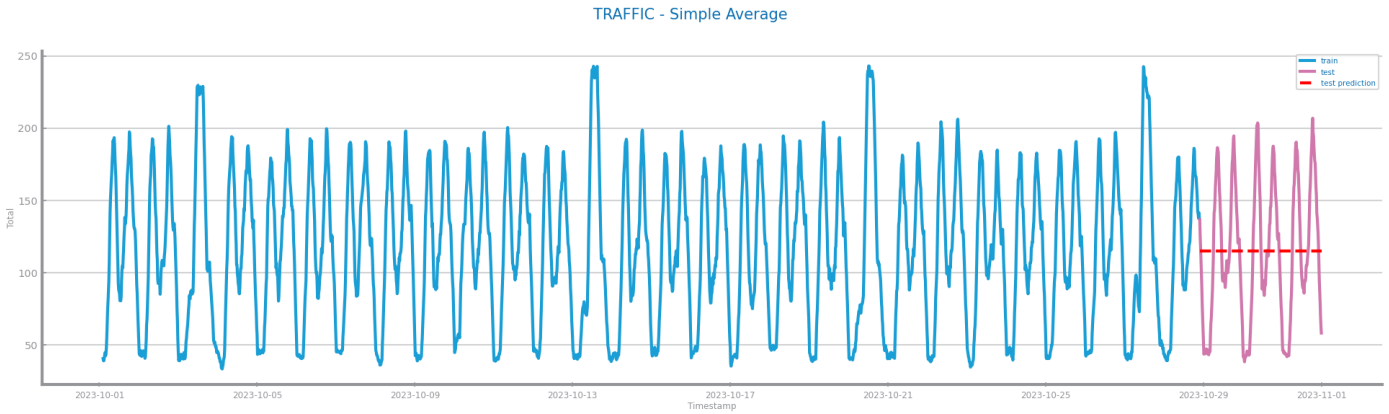


Figure 98 Forecasting plots obtained with Simple Average model over time series 2

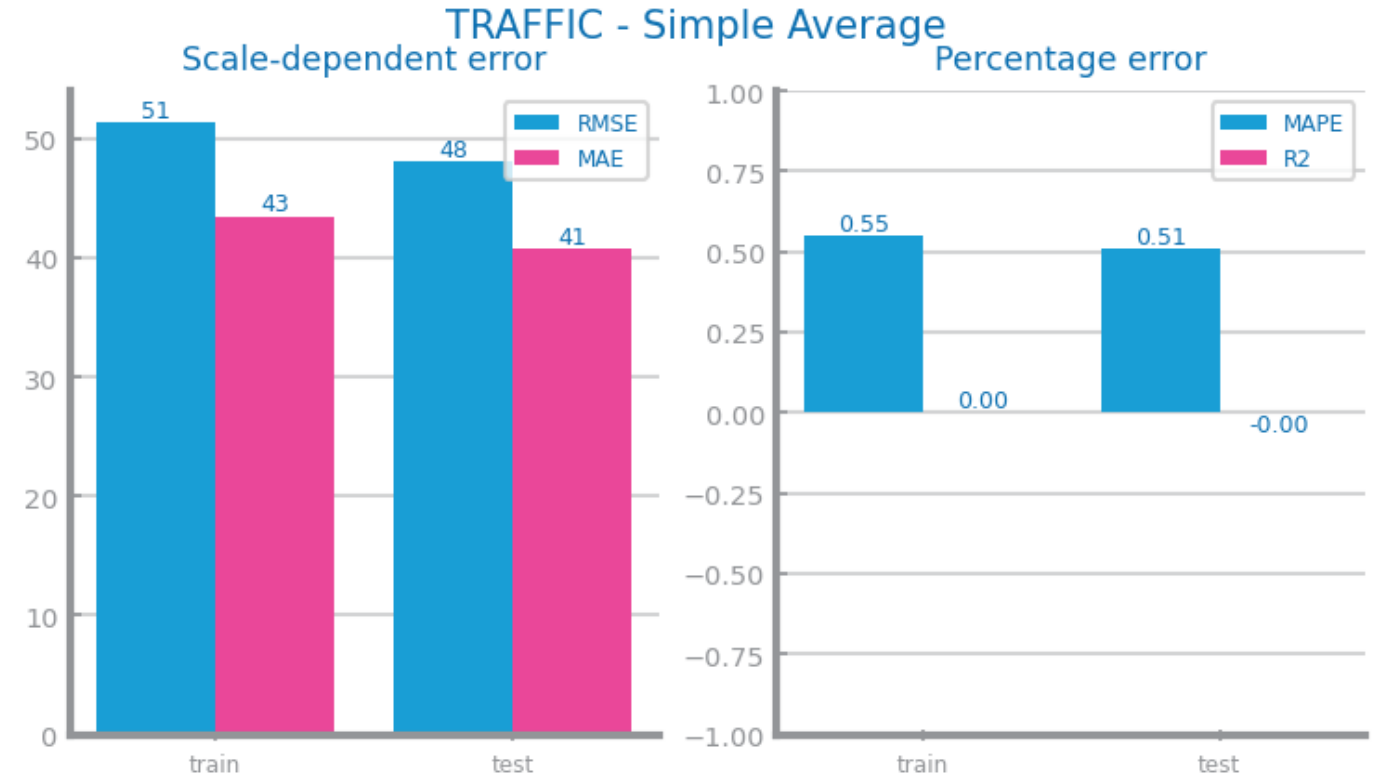


Figure 99 Forecasting results obtained with Simple Average model over time series 2

Persistence Model

For S1, Realist suggests possible overfitting with dramatic error reduction yet poor percentage fit shown by negative R^2. Optimist dramatically reduces errors and shows better fit in test data. For S2, the Realist model slightly improves in scale errors from training to test but worsens in percentage errors, indicating poor fit. Optimist shows good consistency with very low percentage errors, implying a reliable model, we can also see on fig 103 how the predicted values are spot on with the real ones.

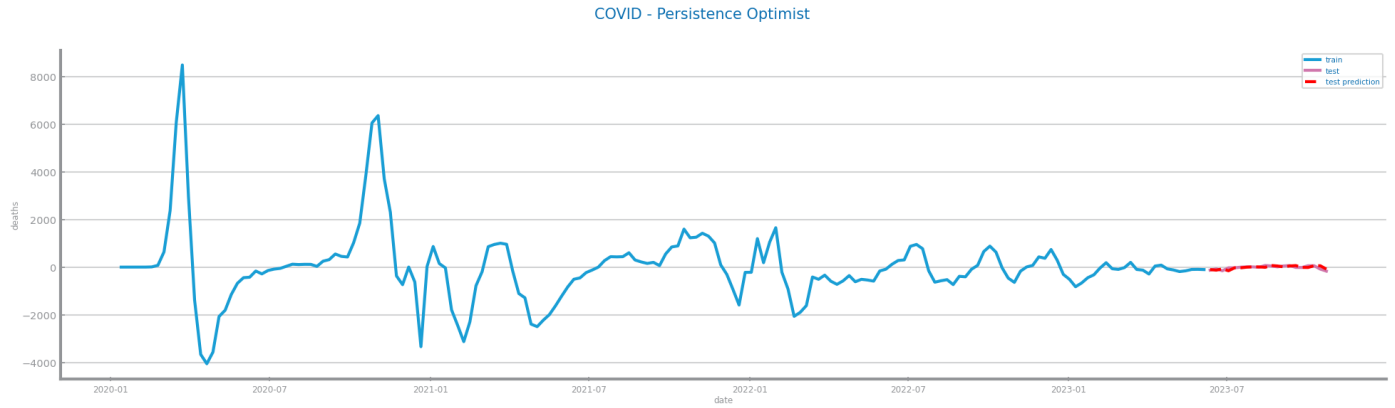


Figure 100 Forecasting plots obtained with Persistence model (long term) over time series 1

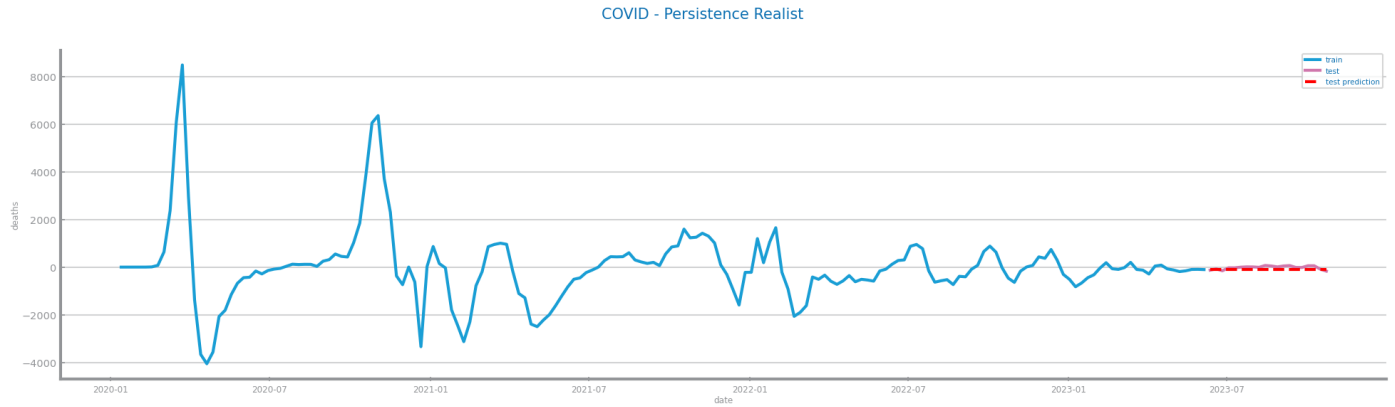


Figure 101 Forecasting plots obtained with Persistence model (next point) over time series 1

Figure 102 Forecasting results obtained with Persistence model in both situations over time series 1

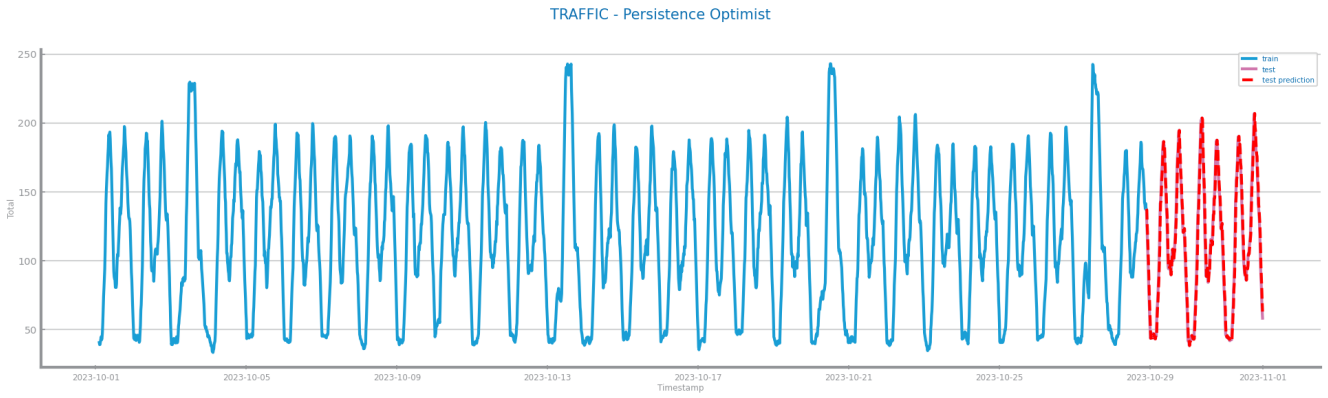


Figure 103 Forecasting plots obtained with Persistence model (long term) over time series 2

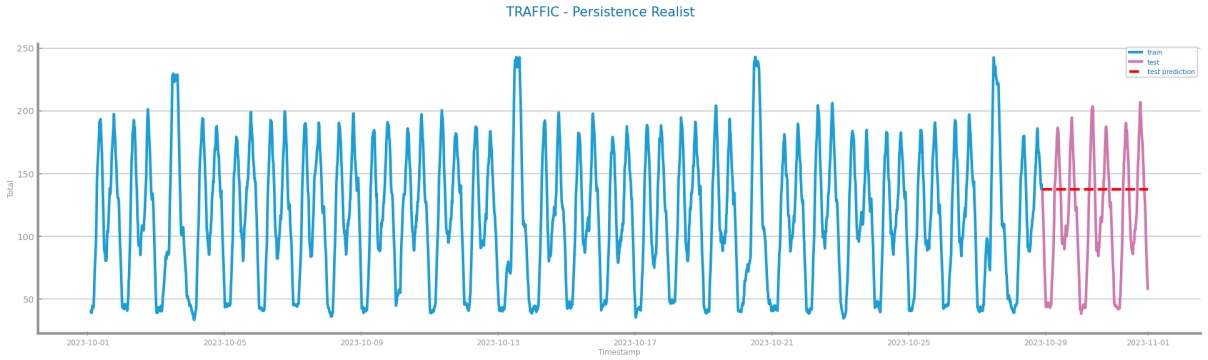


Figure 104 Forecasting plots obtained with Persistence model (next point) over time series 2

Figure 105 Forecasting results obtained with Persistence model in both situation over time series 2

Rolling Mean Model

S1 prediction shows terrible results for MAPE and for R^2. In the parametrizations study the R^2 peaks at a window size of 96 with the value 0 indicating limited predictiveness. For S2, rolling mean forecasting shows reduced scale errors from training to testing, and fairly high percentage errors, which are questionable and might indicate evaluation issues. In the parametrizations study the R^2 stabilizes at 0, also suggesting limited predictiveness.

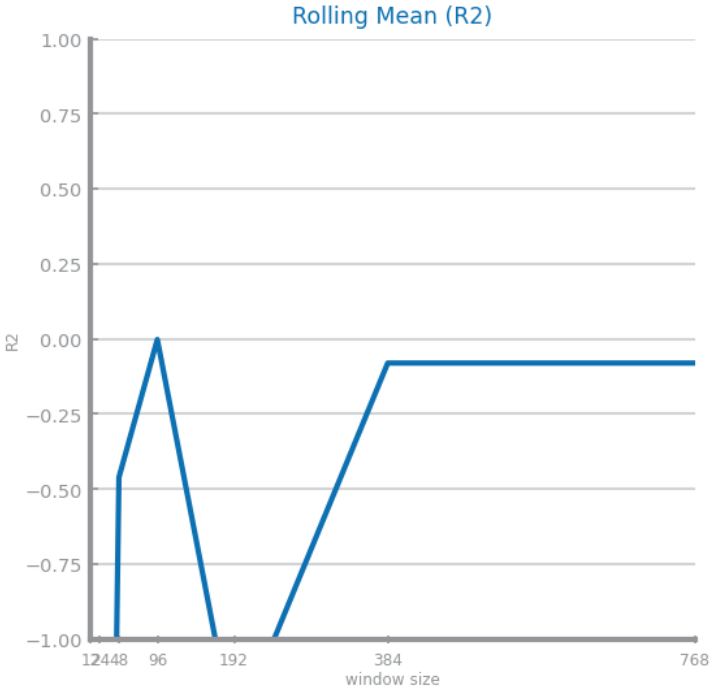


Figure 106 Forecasting study over different parameterisations of the rolling mean algorithm over time series 1

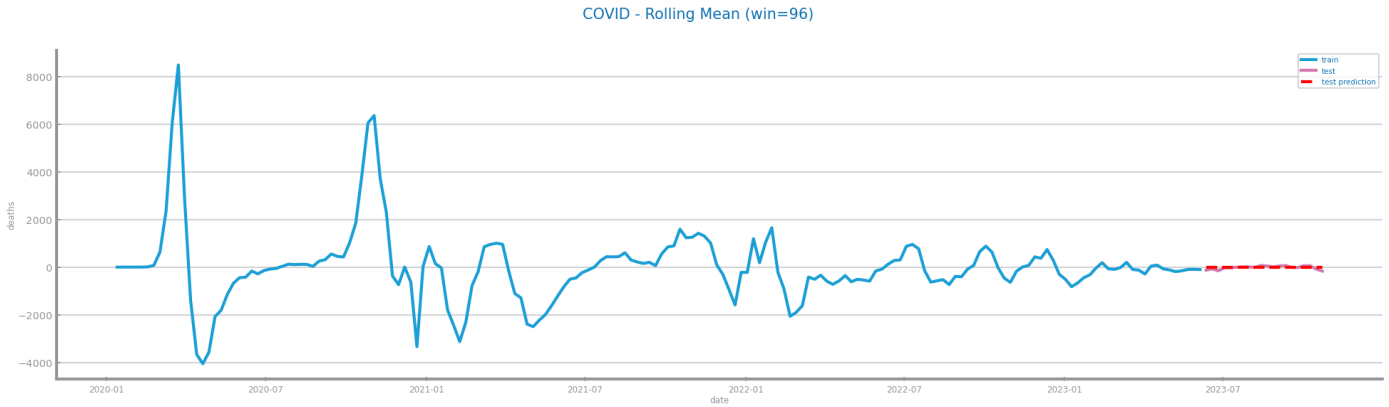


Figure 107 Forecasting plots obtained with the best parameterisation of rolling mean algorithm, over time series 1

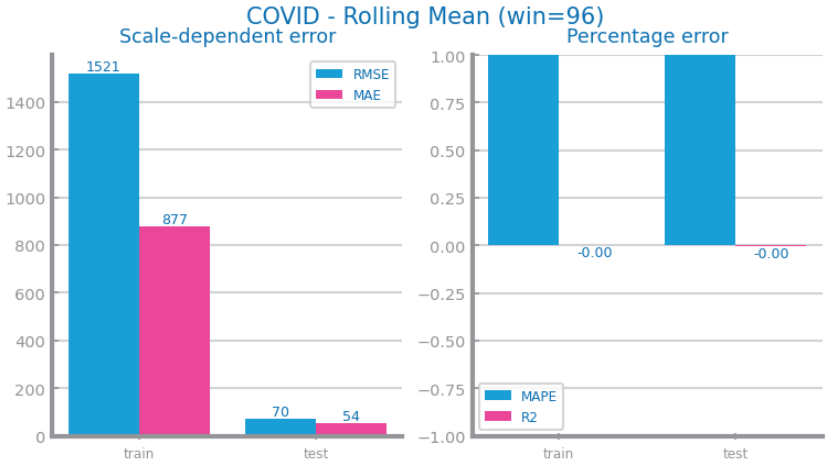


Figure 108 Forecasting results obtained with the best parameterisation of rolling mean algorithm, over time series 1

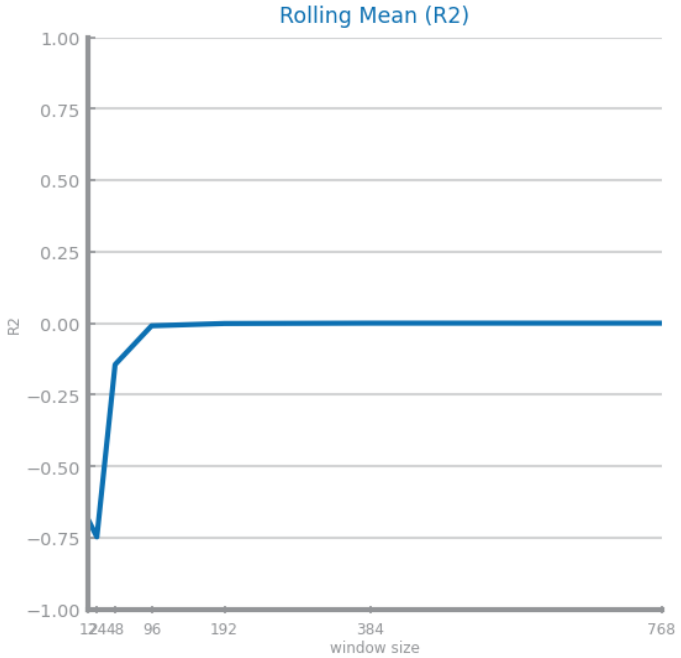


Figure 109 Forecasting study over different parameterisations of the rolling mean algorithm over time series 2

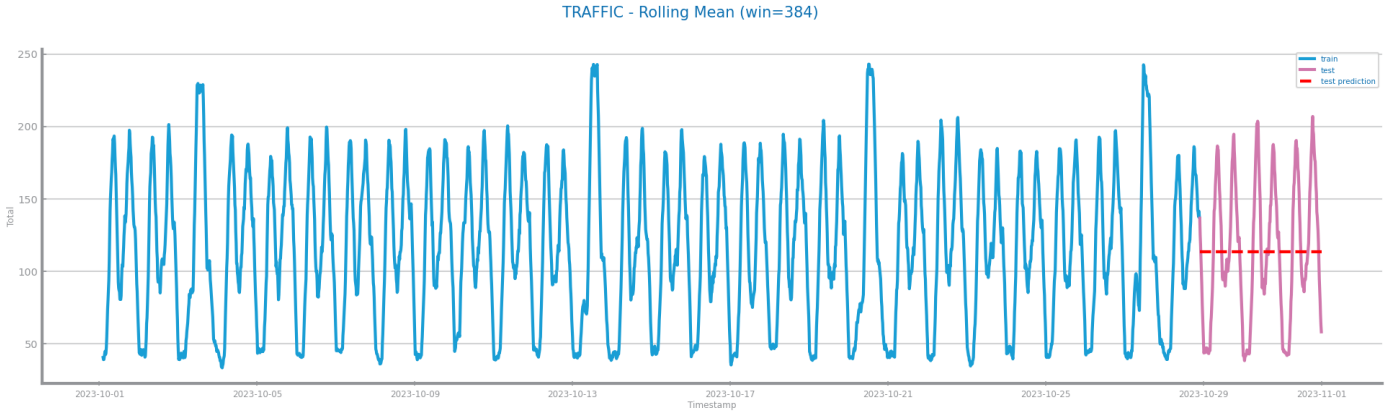


Figure 110 Forecasting plots obtained with the best parameterisation of rolling mean algorithm, over time series 2

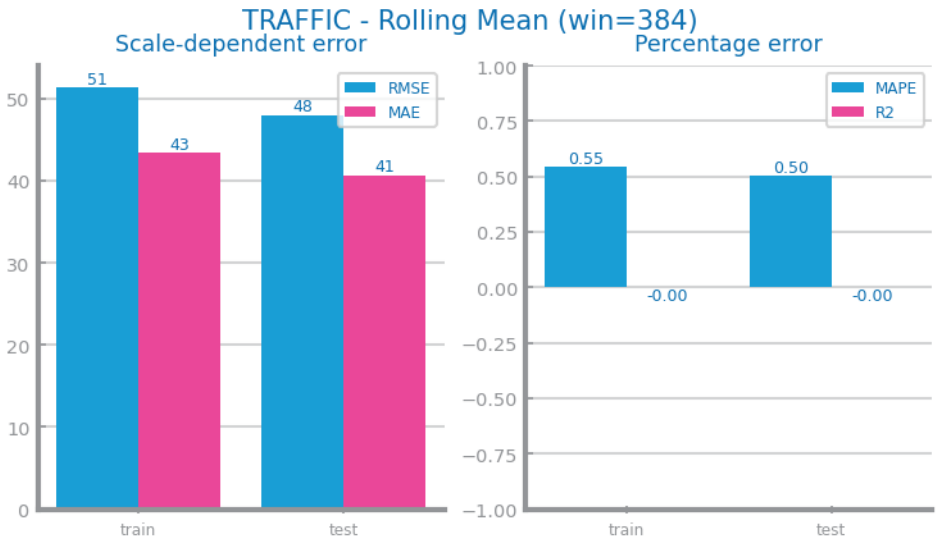


Figure 111 Forecasting results obtained with the best parameterisation of rolling mean algorithm, over time series 2

ARIMA Model

For S1, ARIMA reduced scale errors dramatically, indicating strong test performance; percentage errors showed a slight improvement, with a notable rise in R^2, suggesting enhanced fit. However, R^2's variability across configurations, highlights the model's parameter sensitivity. S2's ARIMA model increased scale errors, suggesting overfitting, while percentage errors improved, indicating a moderate fit; R^2 values fluctuated significantly.

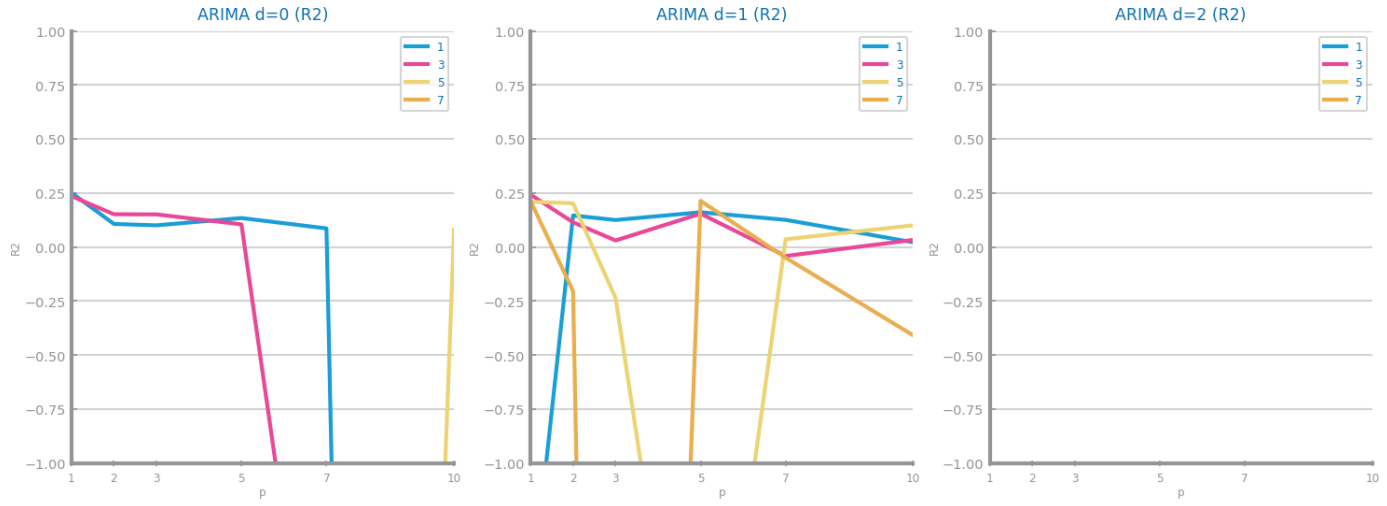


Figure 112 Forecasting study over different parameterisations of the ARIMA algorithm over time series 1

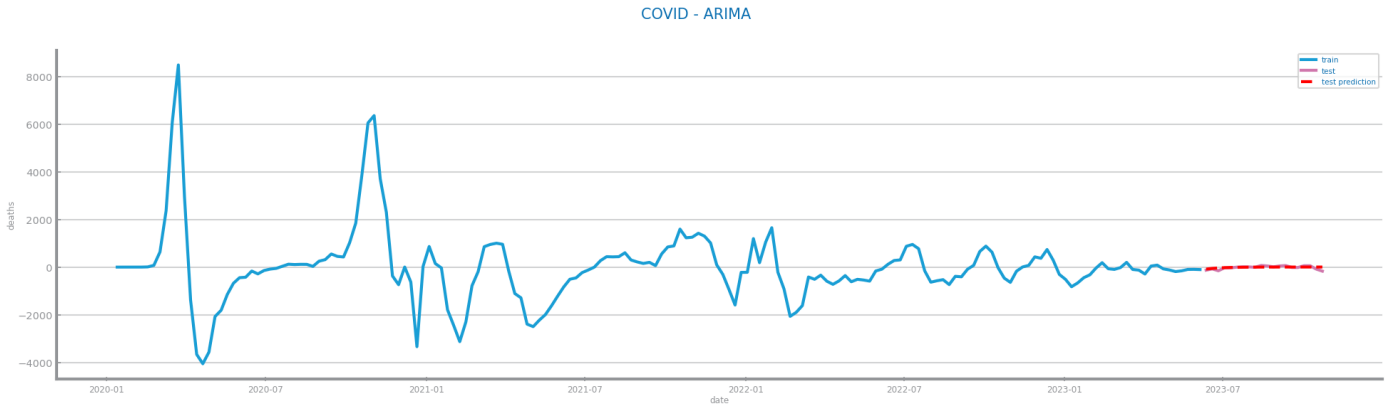


Figure 113 Forecasting plots obtained with the best parameterisation of ARIMA algorithm, over time series 1

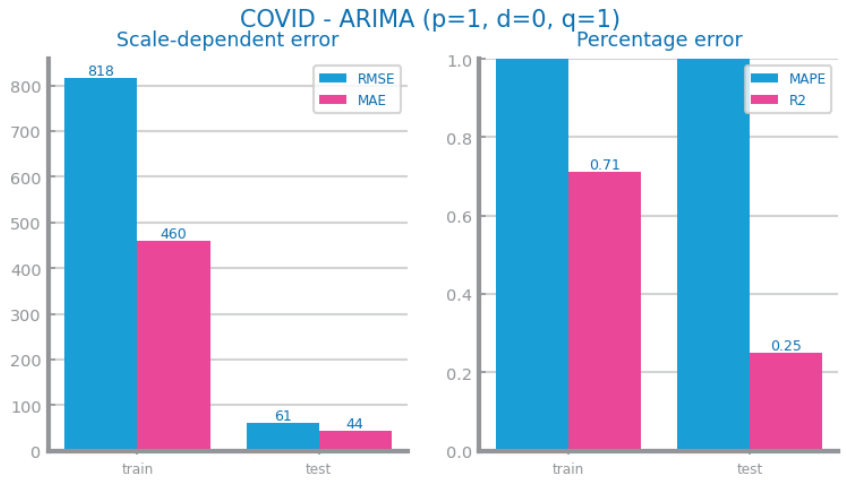


Figure 114 Forecasting results obtained with the best parameterisation of ARIMA algorithm, over time series 1

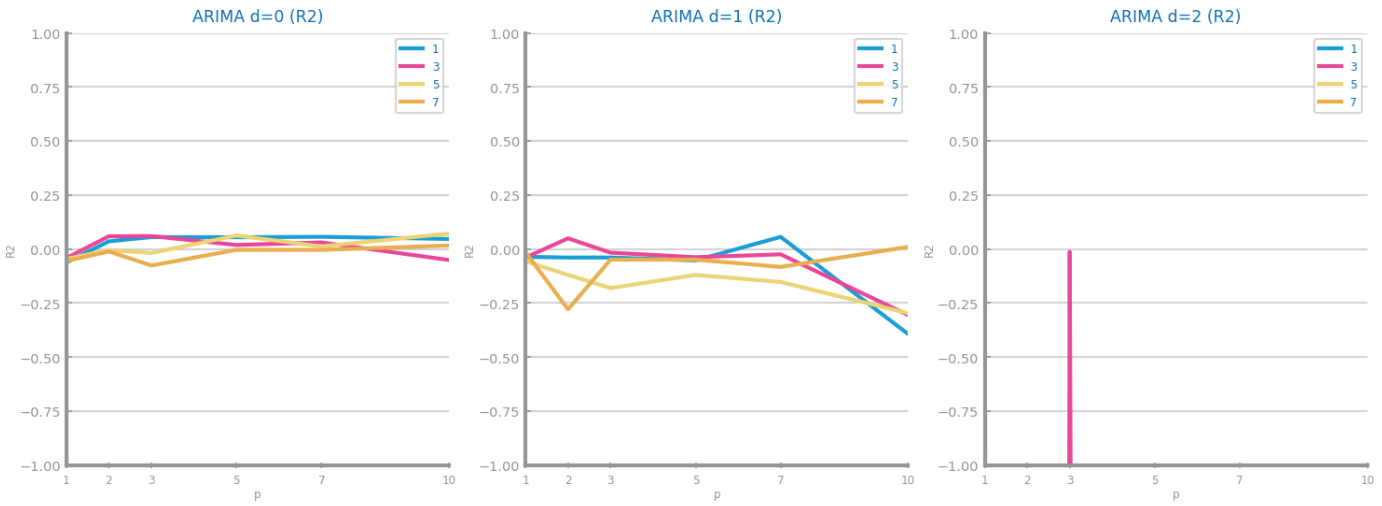


Figure 115 Forecasting study over different parameterisations of the ARIMA algorithm over time series 2

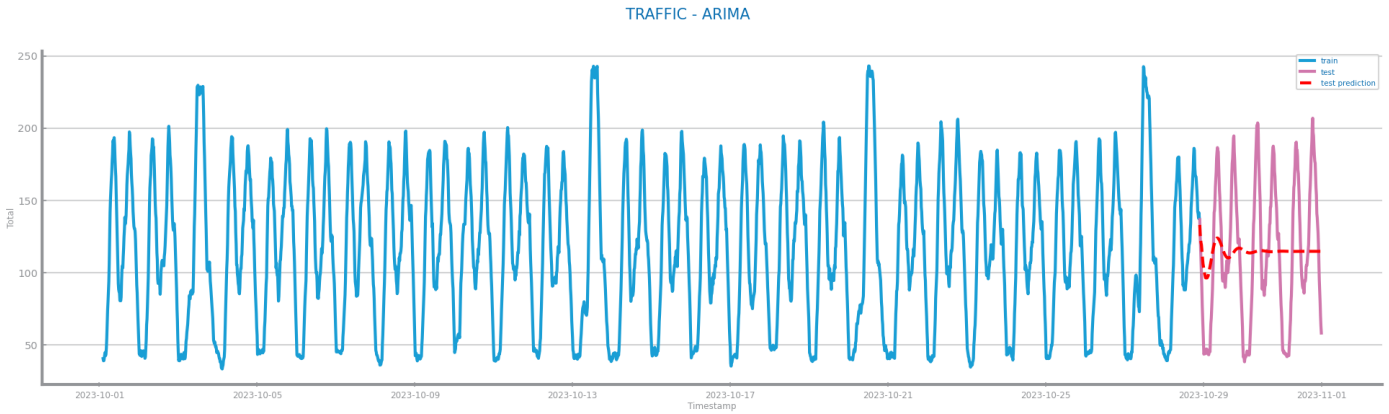


Figure 116 Forecasting plots obtained with the best parameterisation of ARIMA algorithm, over time series 2

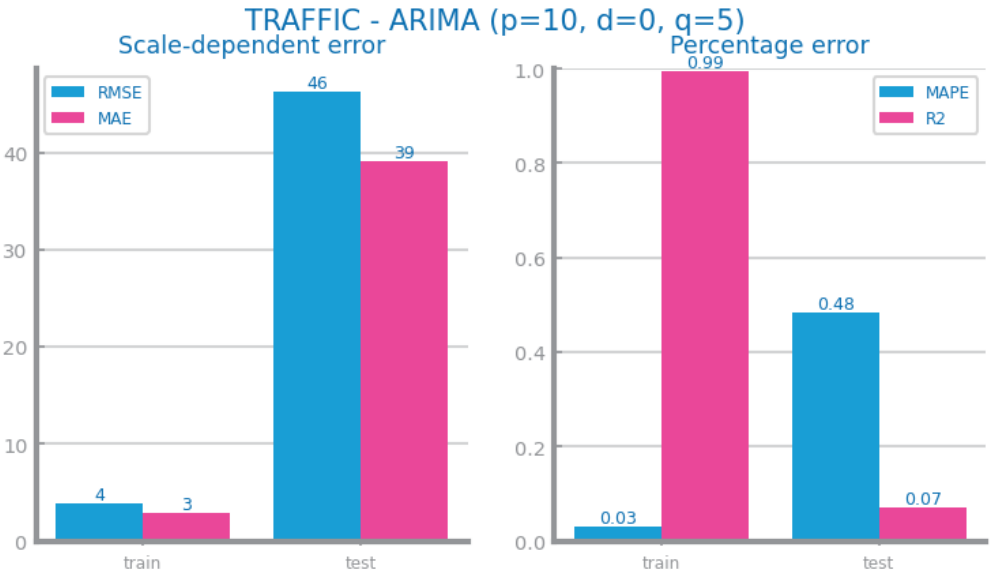


Figure 117 Forecasting results obtained with the best parameterisation of ARIMA algorithm, over time series 2

LSTMs Model

S1's LSTM demonstrates a significant drop in scale errors and minimal percentage errors, implying strong generalization to test data; yet, R^2 values decrease with increased training, suggesting overfitting may occur with longer sequences. S2's LSTM model shows reduced scale errors (RMSE, MAE) from training to test and insignificant percentage errors (MAPE, R^2), indicating excellent model performance and fit, with consistent R^2 across different stages.

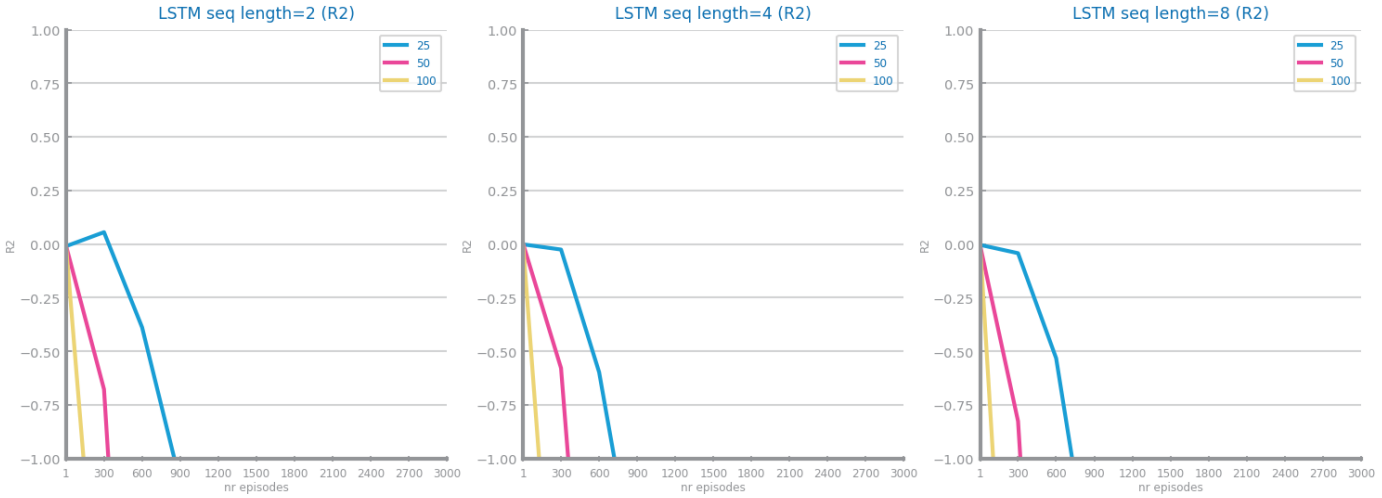


Figure 118 Forecasting study over different parameterisations of LSTMs over time series 1

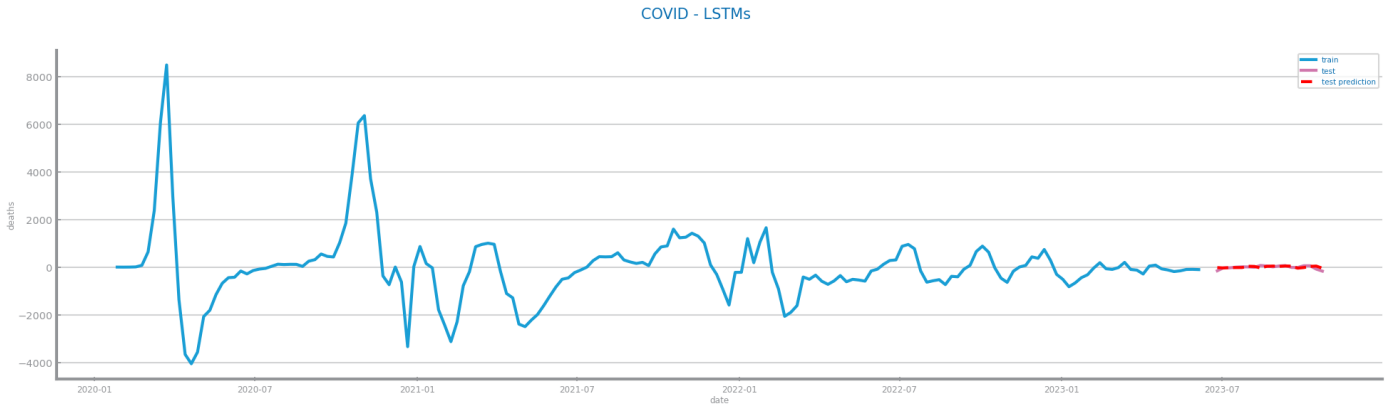


Figure 119 Forecasting plots obtained with the best parameterisation of LSTMs, over time series 1

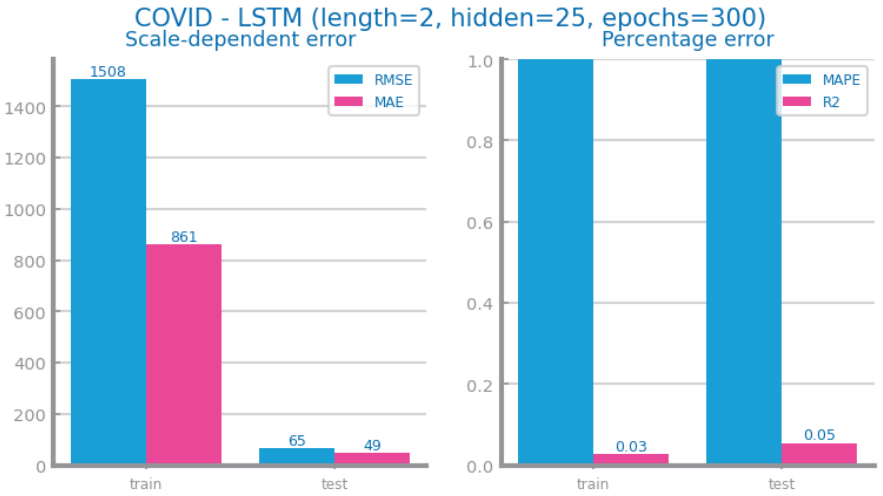


Figure 120 Forecasting results obtained with the best parameterisation of LSTMs, over time series 1

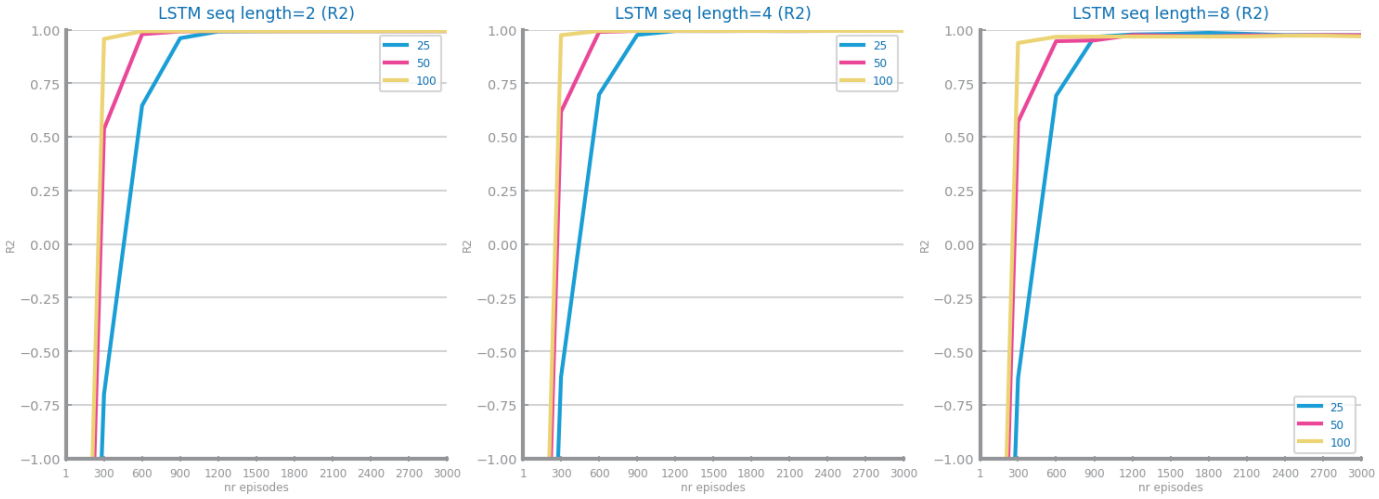


Figure 121 Forecasting study over different parameterisations of the LSTMs over time series 2

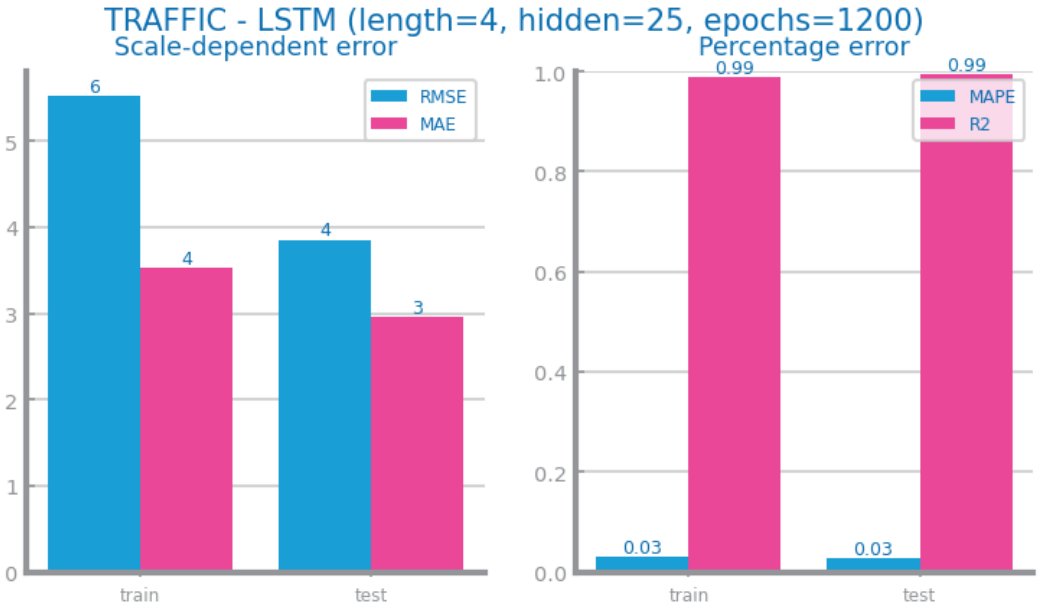


Figure 122 Forecasting plots obtained with the best parameterisation of LSTMs, over time series 2

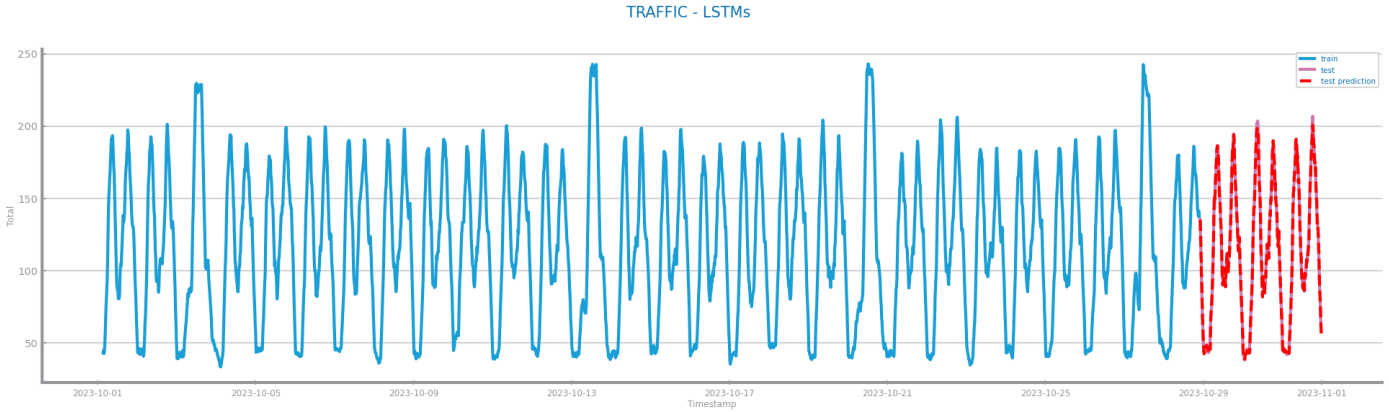


Figure 123 Forecasting results obtained with the best parameterisation of LSTMs, over time series 2

8 Critical Analysis

The study in time series forecasting revealed key insights through data profiling, transformation, and model evaluation. In data profiling, granularity markedly influenced data characteristics. Both S1 and S2 showed no new trends on the granularity study. Significant autocorrelation was observed in S1, especially in weekly aggregations, capturing COVID death trends with a right skew. S2 demonstrated consistent hourly autocorrelation, emphasizing the need for finer granularity for accurate traffic pattern modeling.

Stationarity analysis found S1 to be non-stationary, necessitating differencing, whereas S2 was stationary. This distinction required different preprocessing approaches for each dataset.

In data transformation, the study identified optimal aggregation levels: weekly for S1 and hourly for S2. Smoothing, despite slight improvements in R^2, was not adopted to preserve data integrity in S1 and adopted with window size = 10 in S2. Differentiation showed high RMSE and MAE, but the first differentiation was chosen for lower error rates despite limited R^2 values.

Model evaluation yielded interesting results. For S1 all the models fit the data pretty well, mainly because of the basic nature of the test set, or even the whole data, which converges to zero. For S2 the results were more varied, just a few models to be worth mentioning. ARIMA model started to predict the curves in the data, however was not able to fully grasp the pattern, perhaps SARIMA for seasonal data would be a better fit, since the data was seasonal, then Persistence Model Optimist was able to fit the test data really well with fantastic RMSE and R^2 error metrics, finally the LSTM also predicted test set beautifully with even lower RMSE and higher R^2 taking the best model spot and showing its efficacy and power.

Unfortunately in S1 the accurate predictions caused a “bug” in the RMSE metric making it always high, so it is a challenge to point out which models had the best results in this scenario for S1. Nonetheless we can look at R^2 and say that both ARIMA and Persistence Optimist yielded the best results.

Overall, the analysis highlighted a need for careful model selection and robust validation techniques. The balance between capturing detailed data patterns and generalizing to new data emerged as a crucial challenge. Future efforts could explore ensemble methods, more robust cross-validation, and feature engineering to enhance model accuracy.