TÉCNICO LISBOA

DS 2025

Data Science Project

Team nr: 16	**Student 1:**	Antero Morgado	**IST nr:**	1119213
	Student 2:	David Ferreira	**IST nr:**	1107077
	Student 3:	José Fernandes	**IST nr:**	1103727
	Student 4:	Olha Buts	**IST nr:**	1116276

CLASSIFICATION

1 DATA PROFILING

May be used to describe any useful observation about the data, and that was used in the current project. An example is the use of any domain knowledge to process the data or evaluate the results. **Shall not exceed 200 characters.**

Data Dimensionality

Regarding the Missing Values analysis, the completeness metric for Dataset 1 was adjusted to treat 'UNKNOWN' occurrences as missing values.

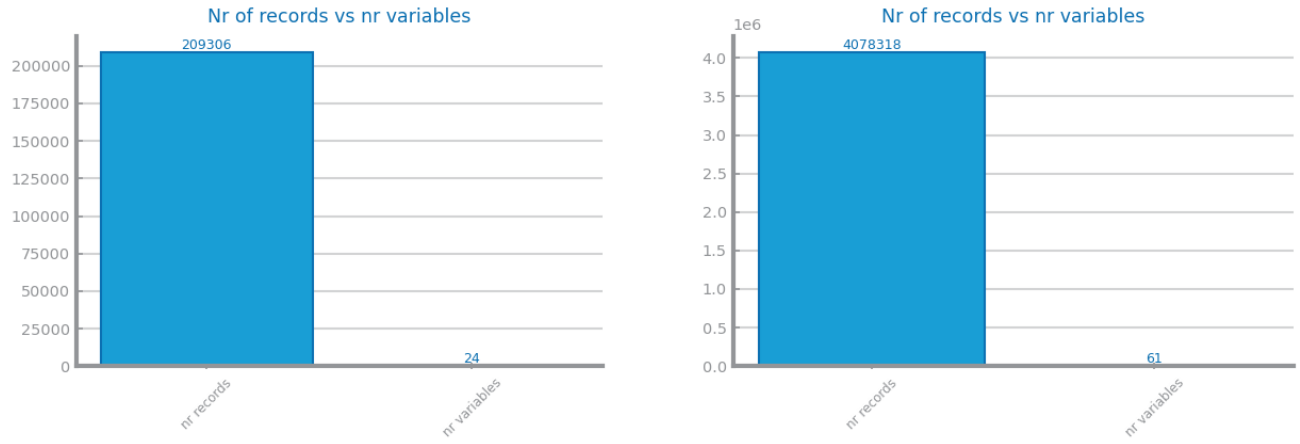


Figure 1: Nr Records x Nr variables for dataset 1 (left) and dataset 2 (right)

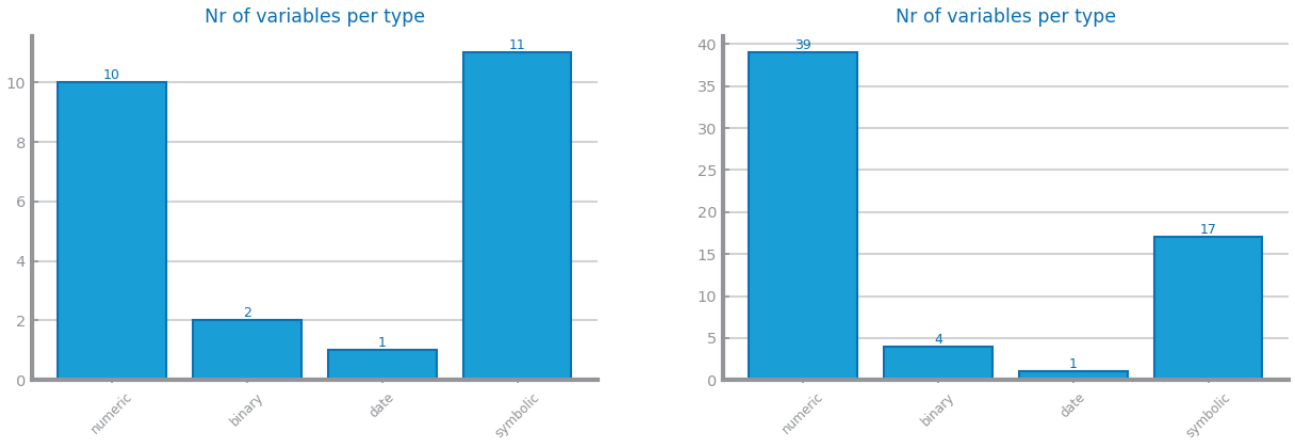


Figure 2: Nr variables per type for dataset 1 (left) and dataset 2 (right)

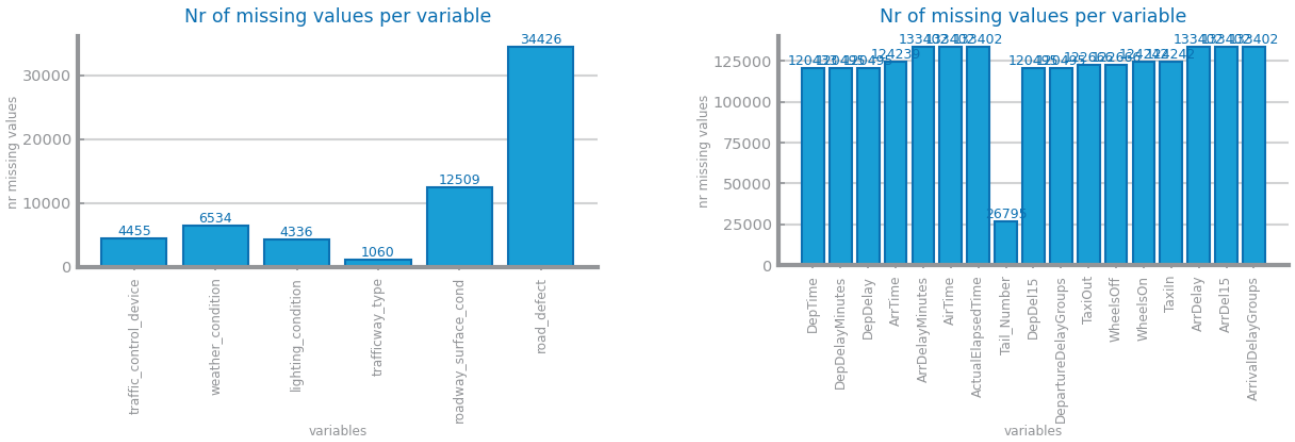


Figure 3: Nr missing values for dataset 1 (left) and dataset 2 (right)

Data Distribution

The distribution analysis reveals that the raw data is significantly dispersed, exhibiting high variance across several features. Certain variables show a clear lack of balance, with distributions heavily skewed toward specific classes or ranges and presence of evident outliers.

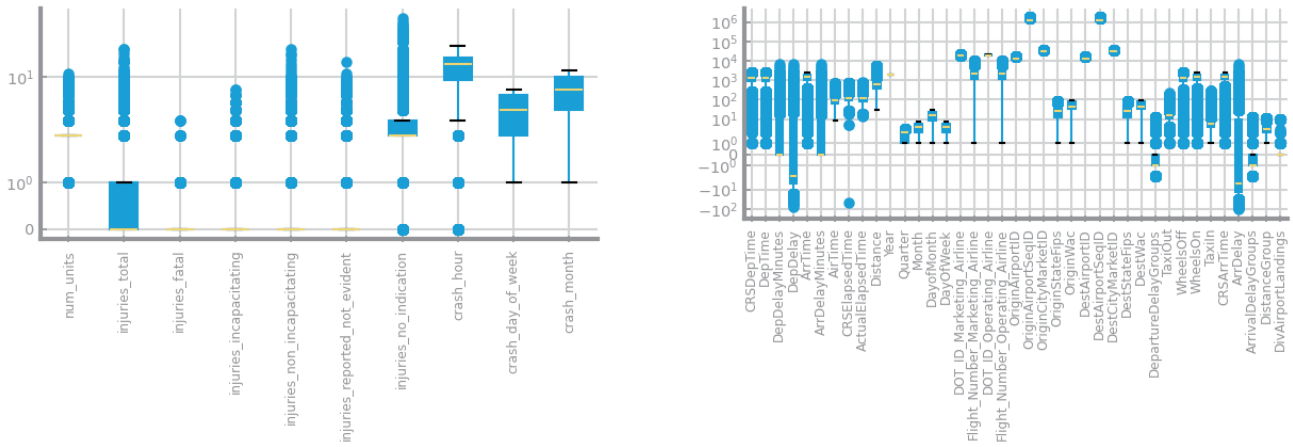


Figure 4: Global boxplots dataset 1 (left) and dataset 2 (right)

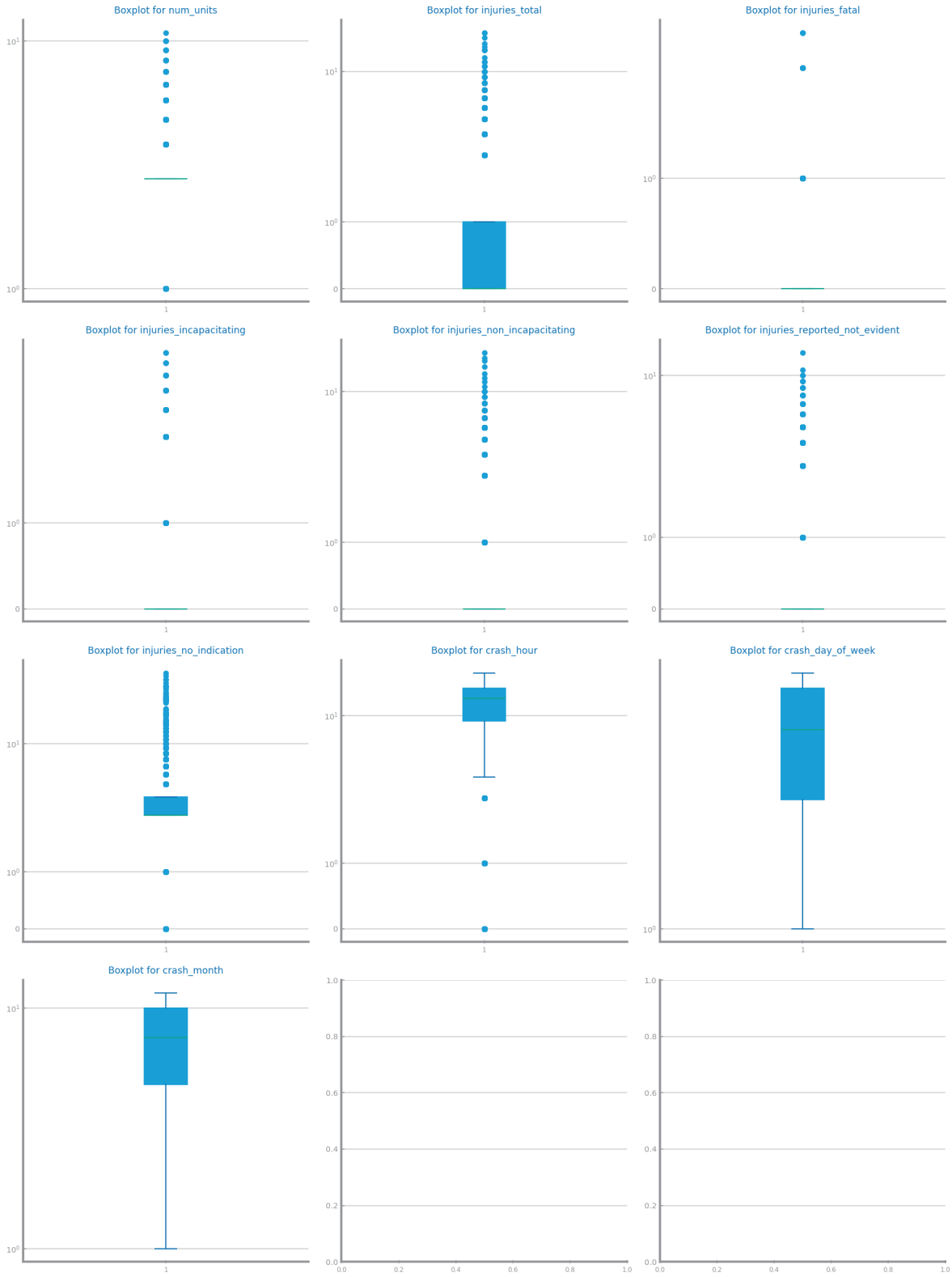


Figure 5: Single variables boxplots for dataset 1

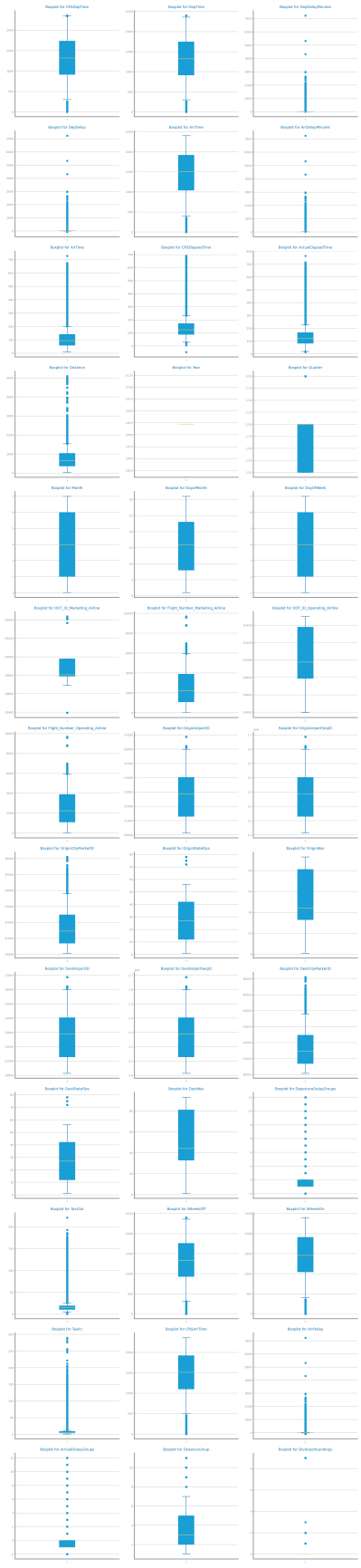


Figure 6: Single variables boxplots for dataset 2

Figure 7: Histograms for dataset 1

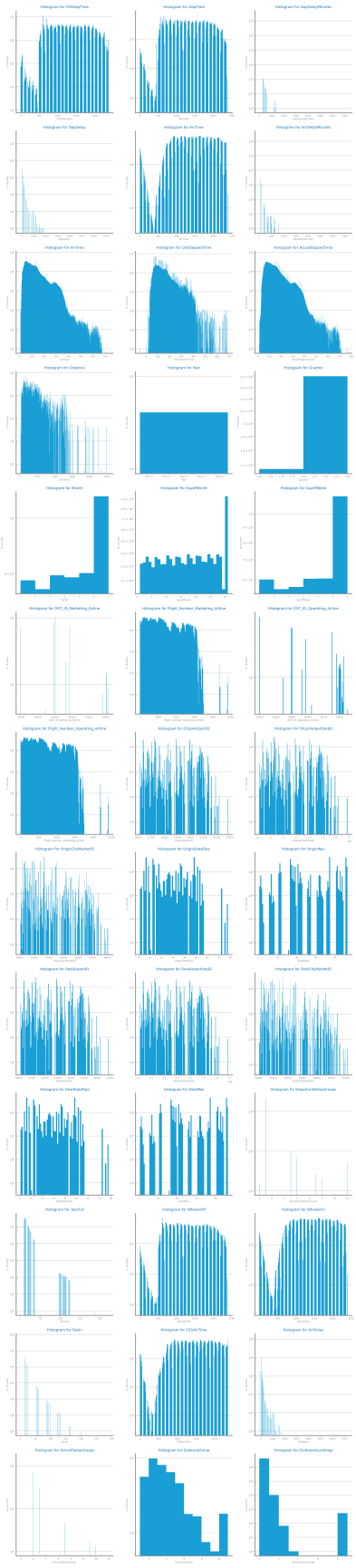


Figure 8: Histograms for dataset 2

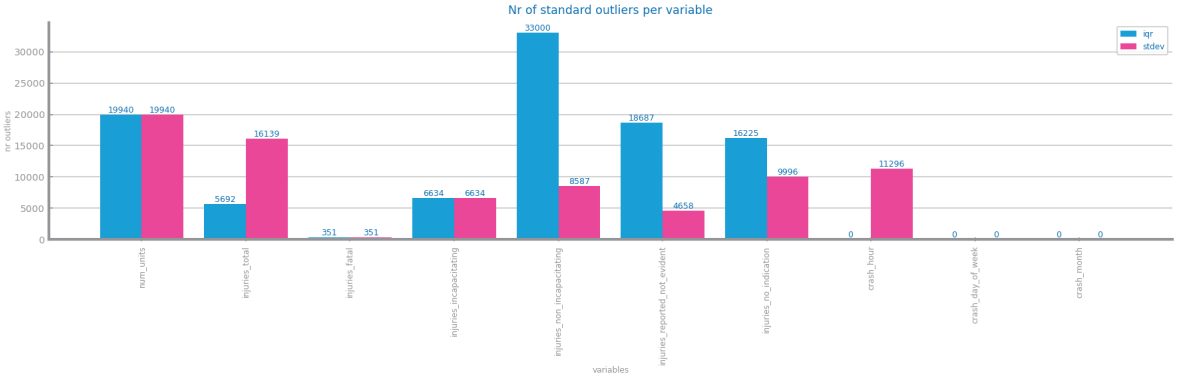


Figure 9: Outliers study dataset 1

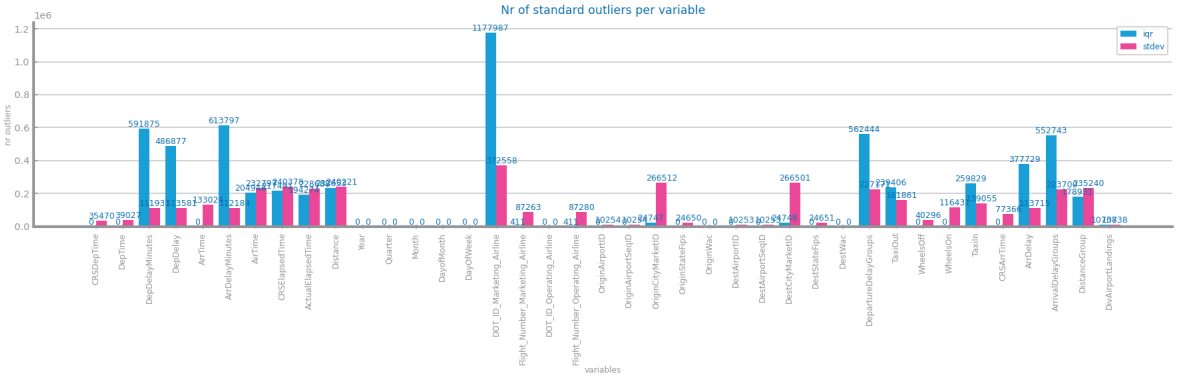


Figure 10: Outliers study dataset 2

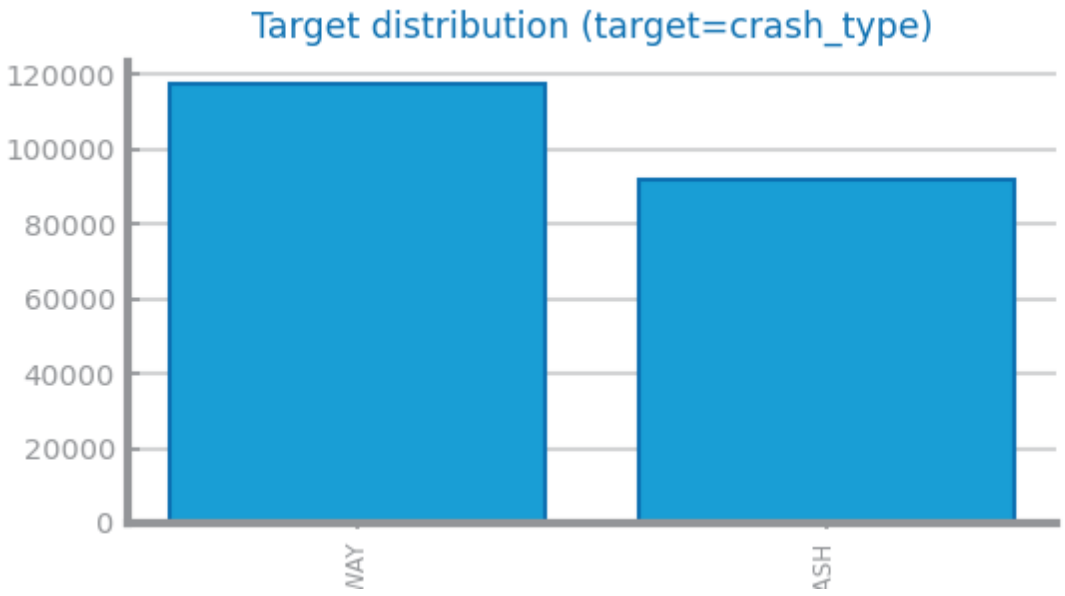


Figure 11: Class distribution for dataset 1

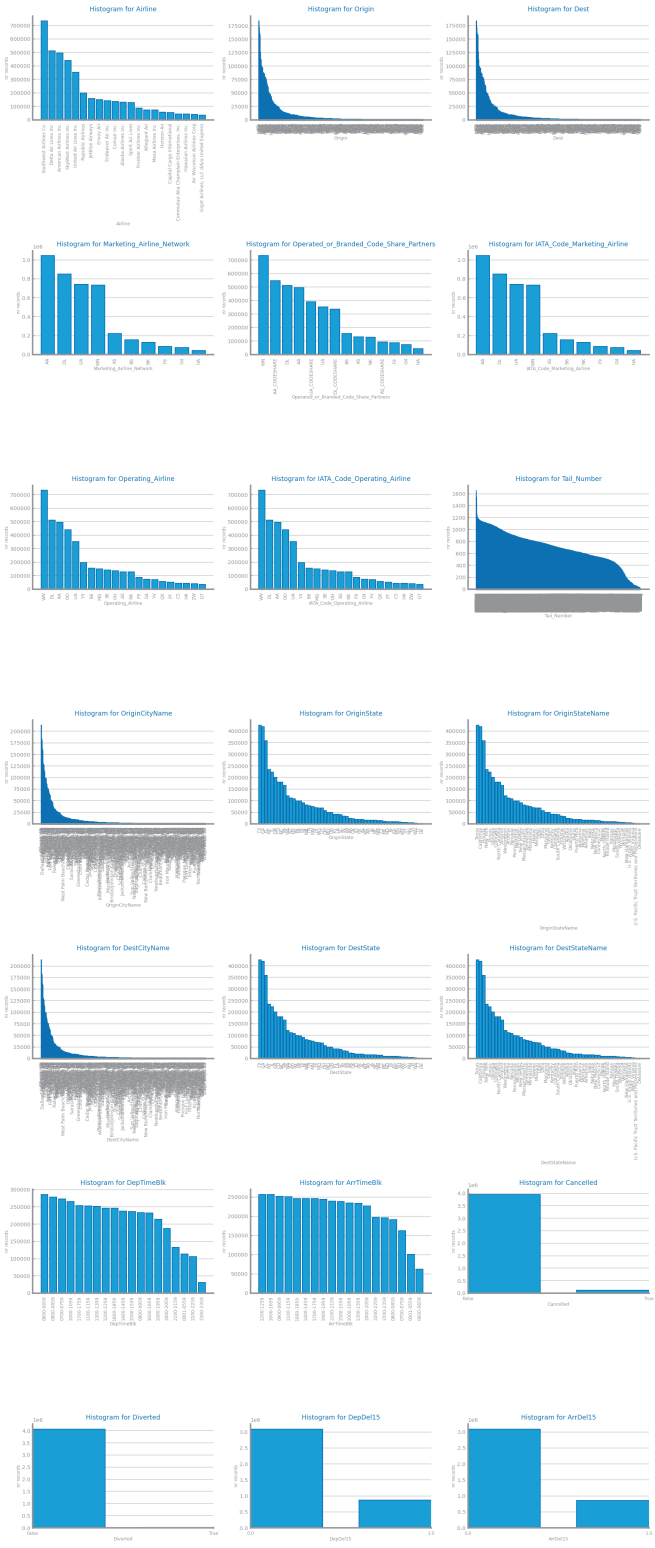
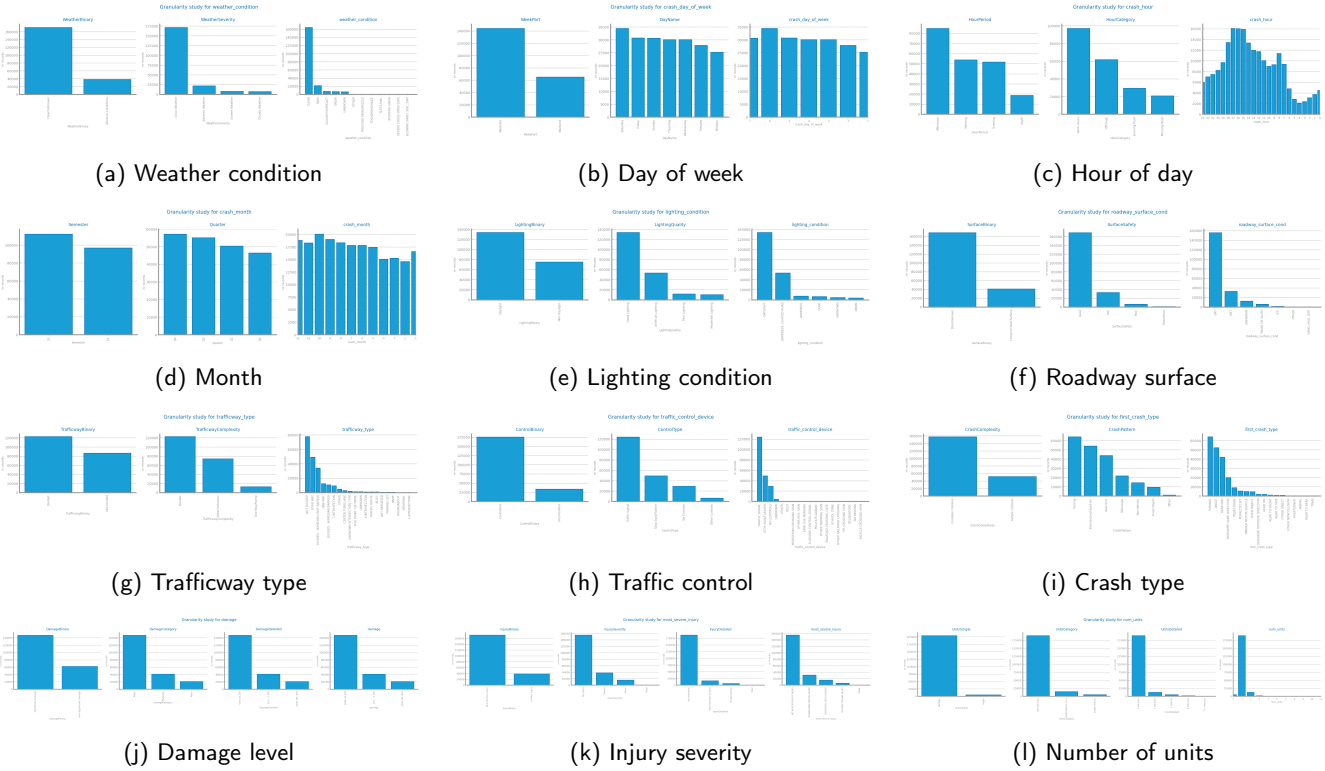


Figure 12: Class distribution for dataset 2

Data Granularity

The granularity analysis highlights significant variability across variables in both datasets. Categorical features such as weather, lighting conditions, airline, and airport present low to medium granularity, while temporal and numerical variables (hour, delays, distance) show higher granularity and dispersion. These differences suggest the need for appropriate grouping or discretization strategies to avoid sparsity and improve downstream analysis and modeling.



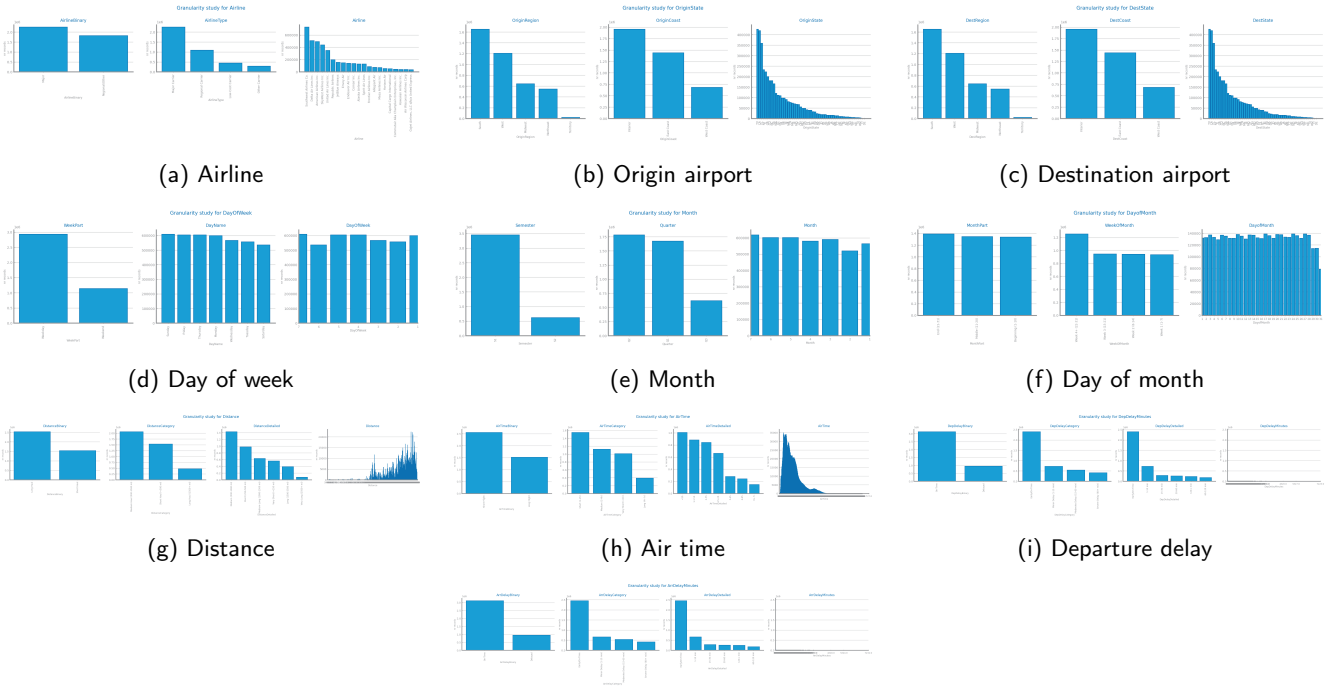
(a) Weather condition (b) Day of week (c) Hour of day

(d) Month (e) Lighting condition (f) Roadway surface

(g) Trafficway type (h) Traffic control (i) Crash type

(j) Damage level (k) Injury severity (l) Number of units

Figure 13: Granularity analysis for dataset 1



(a) Airline (b) Origin airport (c) Destination airport

(d) Day of week (e) Month (f) Day of month

(g) Distance (h) Air time (i) Departure delay

(j) Arrival delay

Figure 14: Granularity analysis for dataset 2

Data Sparsity

The sparsity analysis reveals uneven domain coverage across variables, with several categorical features dominated by a small number of frequent values and many rare ones. Correlation analysis shows limited strong dependencies between most variables, indicating sparse relationships in the feature space.

Figure 15: Sparsity analysis for dataset 1

Figure 16: Sparsity analysis for dataset 2 - View on Google Drive

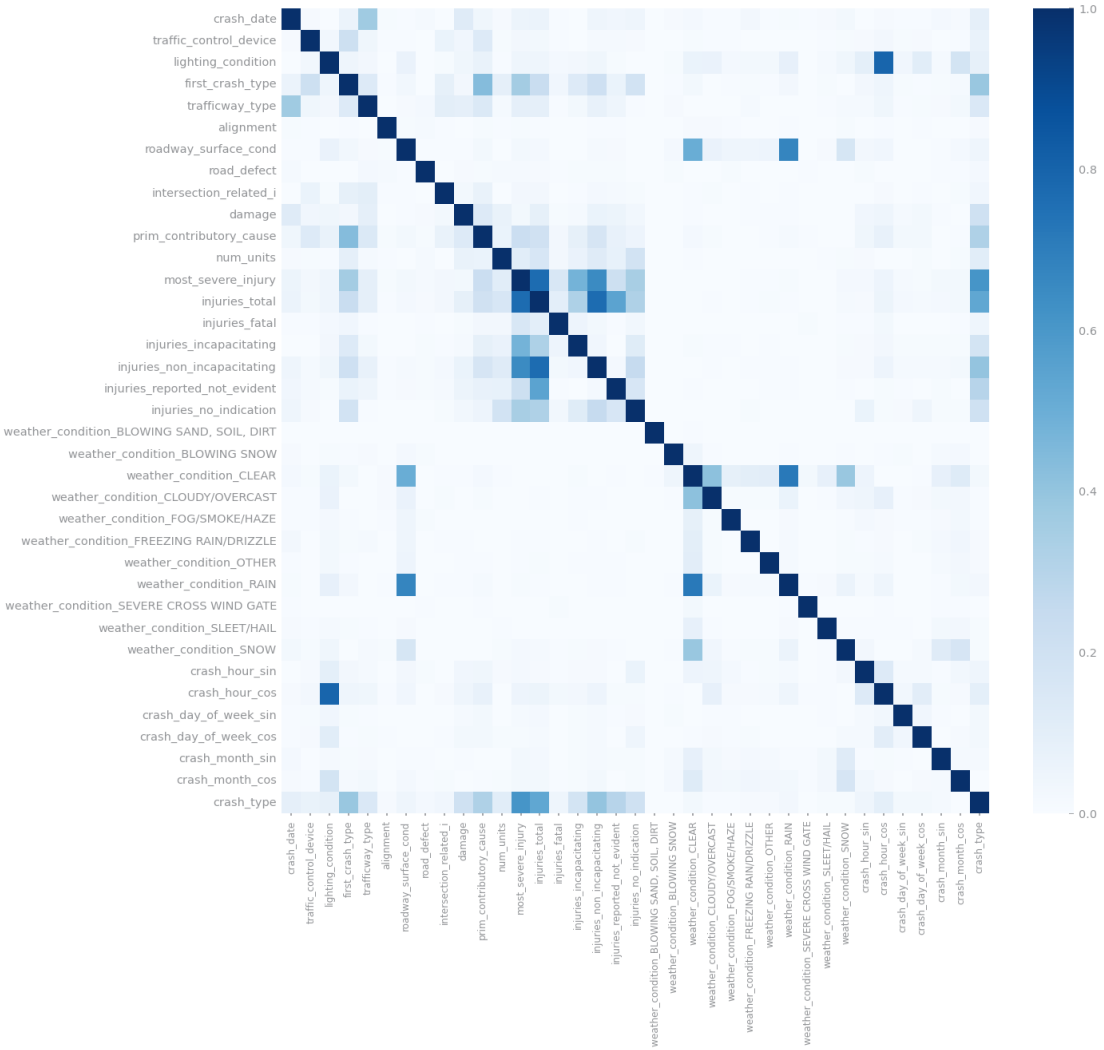


Figure 17: Correlation analysis for dataset 1

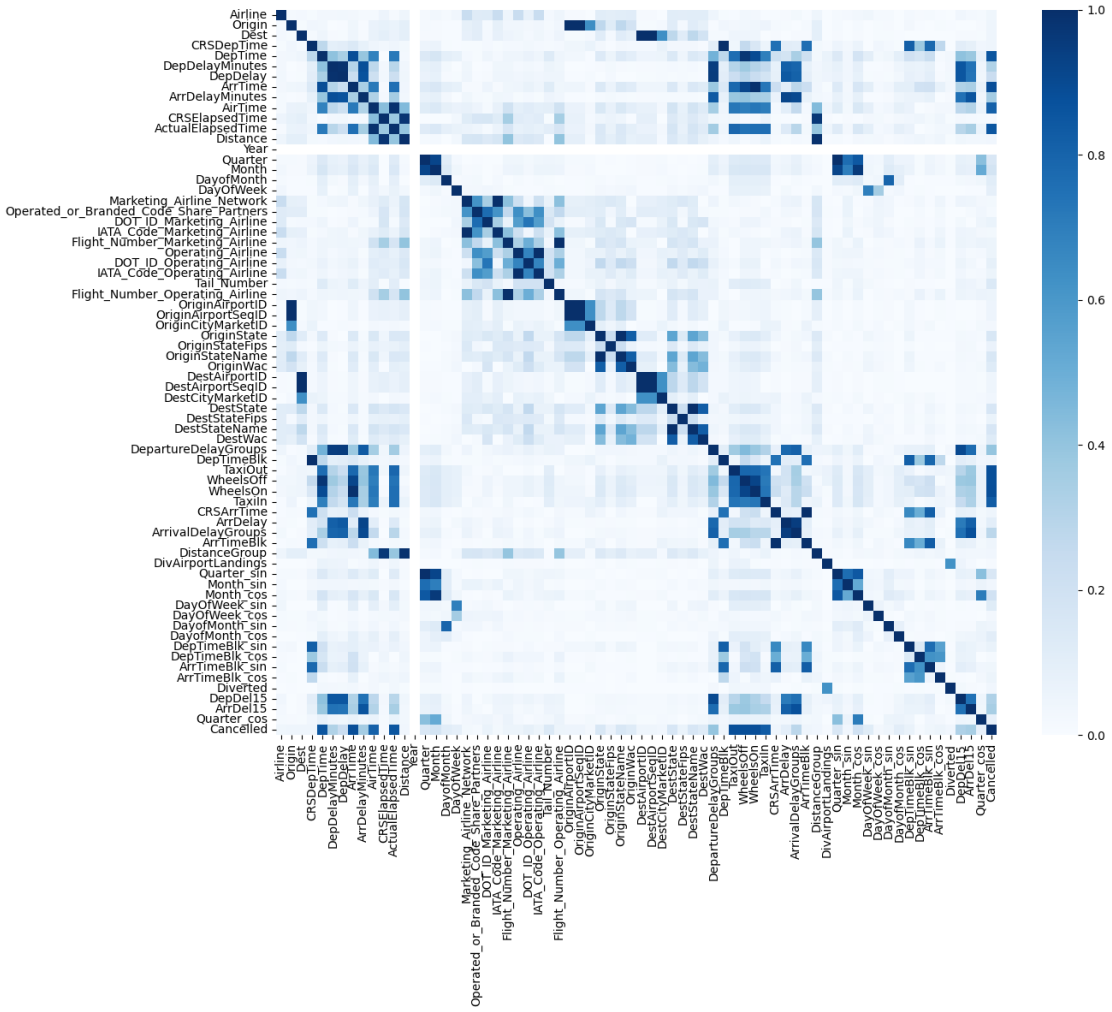


Figure 18: Correlation analysis for dataset 2

2 DATA PREPARATION

Variables Encoding

Shall contain all relevant information respecting to the transformation of variables. The list of variables under each one of the transformations, shall be presented. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 500 characters *for each dataset*.**

Missing Value Imputation

Shall contain all relevant information and charts respecting to missing values imputation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 500 characters.**

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

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

Outliers Treatment

Shall contain all relevant information and charts respecting to outliers imputation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 500 characters.**

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

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

Scaling

Shall contain all relevant information and charts respecting to scaling transformation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 200 characters.**

Figure 23: Scaling results with different approaches for dataset 1

Figure 24: Scaling results with different approaches for dataset 2

Balancing

Shall contain all relevant information and charts respecting to balancing transformation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 500 characters.**

Figure 25: Balancing results with different approaches for dataset 1

Figure 26: Balancing results with different approaches for dataset 2

Feature Selection

Shall contain all relevant information and charts respecting to feature selection based on filtering out redundant (based on correlation) and relevant (based on variation) variables. The different choices and their impact on the modelling results shall be presented and explained. Should also clearly reveal the approach selected to proceed with the processing. All explanations shall be based on data characteristics. **Shall not exceed 500 characters.**

Figure 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)

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

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

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

Additional Feature Generation (if done)

Shall contain all relevant information and charts respecting to feature generation. The different choices and their impact on the modelling results shall be presented and explained. Shall summarise all variables generated and the formula used to derive them (in a table). **Shall not exceed 200 characters.**

Figure 33: Feature generation results for dataset 1

Figure 34: Feature generation results for dataset 2

3 MODELS' EVALUATION

Shall be used to point out any important decision taken during the training, including training strategy and evaluation measures used. **Shall not exceed 500 characters.**

Naïve Bayes

Shall be used to present the results achieved with each one of Naïve Bayes implementations, comparing and proposing explanations for them. If any of the implementations is not used, a justification for it shall be presented. Shall be used to present the evaluation of the best model achieved. **Shall not exceed 300 characters.**

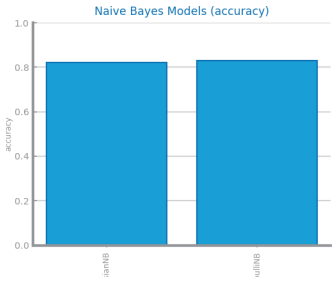


Figure 35: Naïve Bayes alternatives comparison for dataset 1

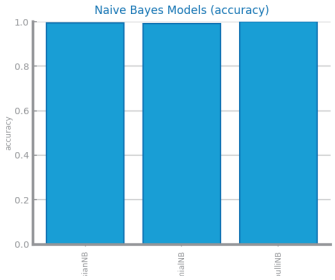


Figure 36: Naïve Bayes alternative comparison for dataset 2

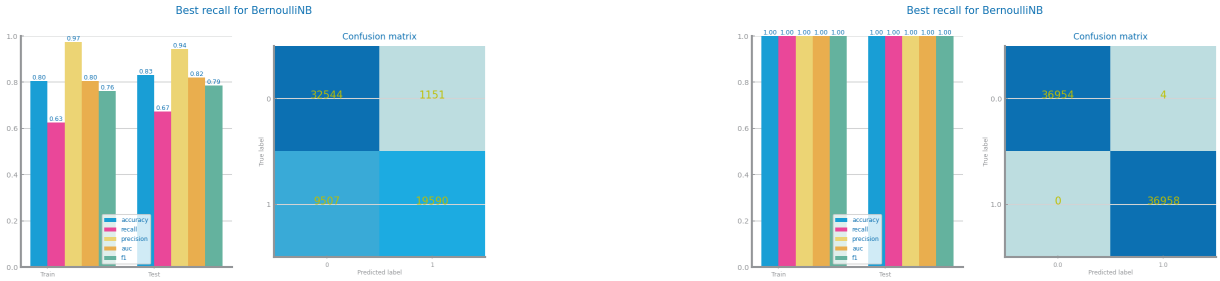


Figure 37: Naïve Bayes best model results for dataset 1 (left) and dataset 2 (right)

KNN

Shall be used to present the results achieved through different similarity measures and KNN parameterisations. The results shall be compared and explanations for them shall be presented. The justification for the chosen similarity measures shall be presented. Shall be used to address the overfitting phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. **Shall not exceed 500 characters.**

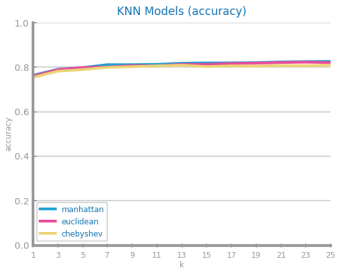


Figure 38: KNN different parameterisations comparison for dataset 1

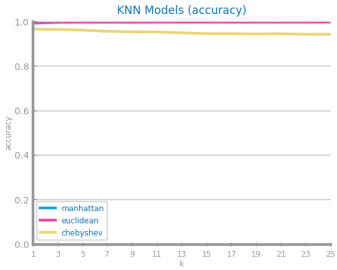


Figure 39: KNN different parameterisations comparison for dataset 2

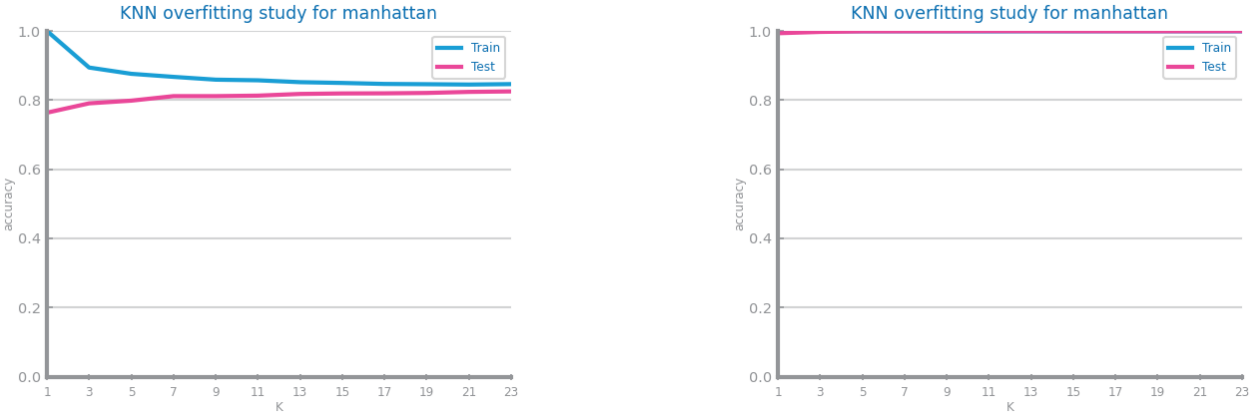


Figure 40: KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

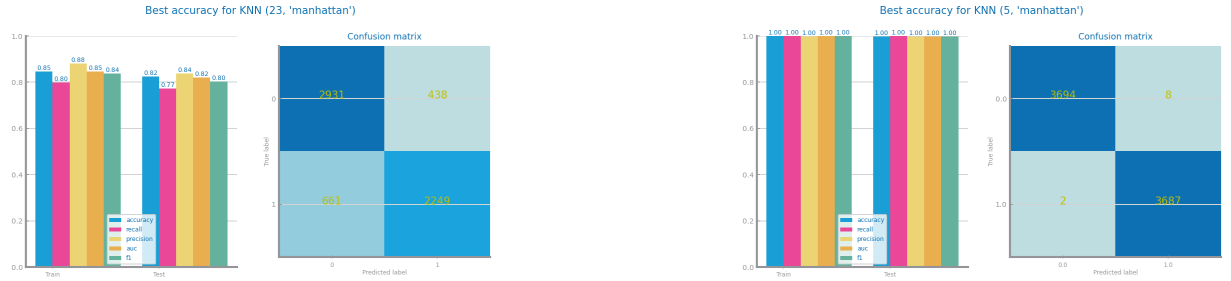


Figure 41: KNN best model results for dataset 1 (left) and dataset 2 (right)

Logistic Regression

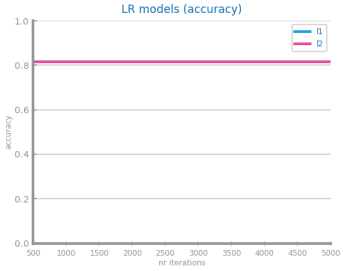


Figure 42: Logistic Regression different parameterisations comparison for dataset 1

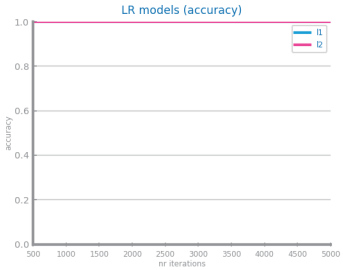


Figure 43: Logistic Regression different parameterisations comparison for dataset 2

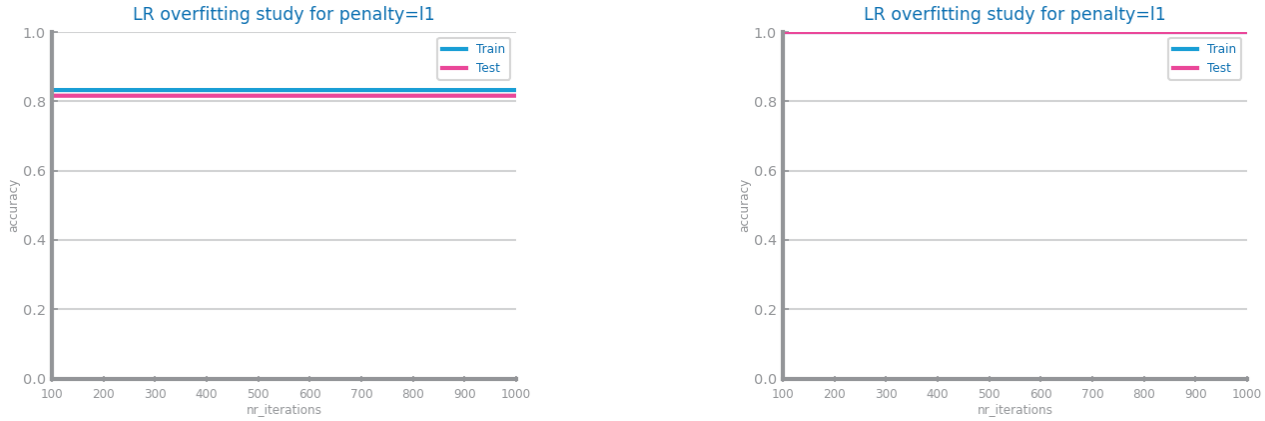


Figure 44: Logistic Regression overfitting analysis for dataset 1 (left) and dataset 2 (right)

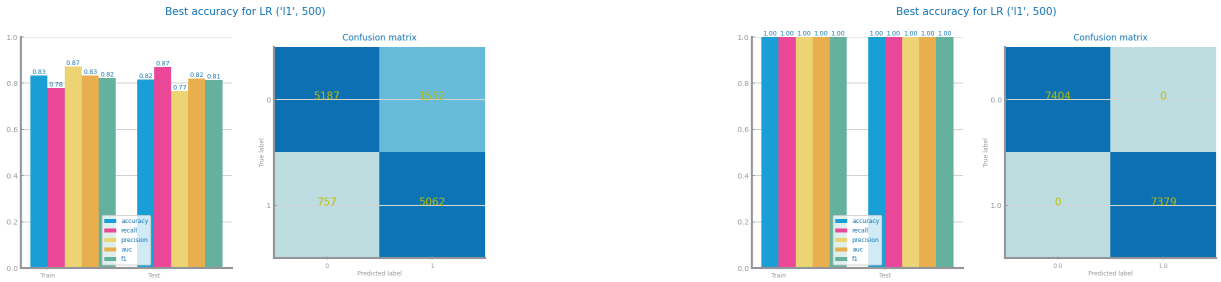


Figure 45: Logistic Regression best model results for dataset 1 (left) and dataset 2 (right)

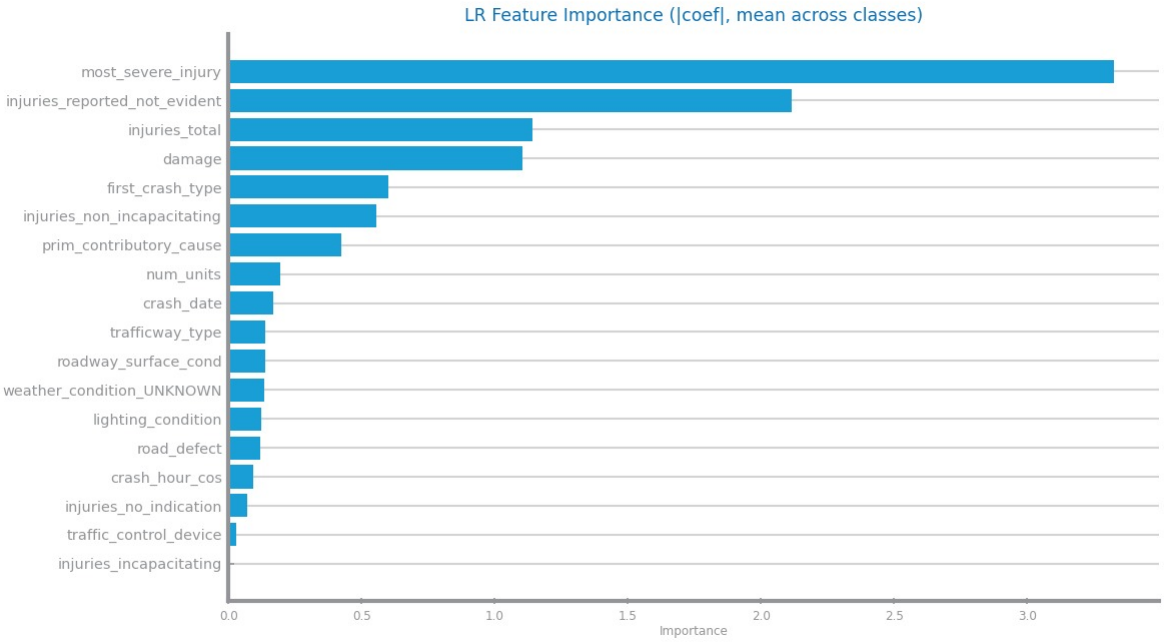


Figure 46: Logistic Regression feature importance for dataset 1

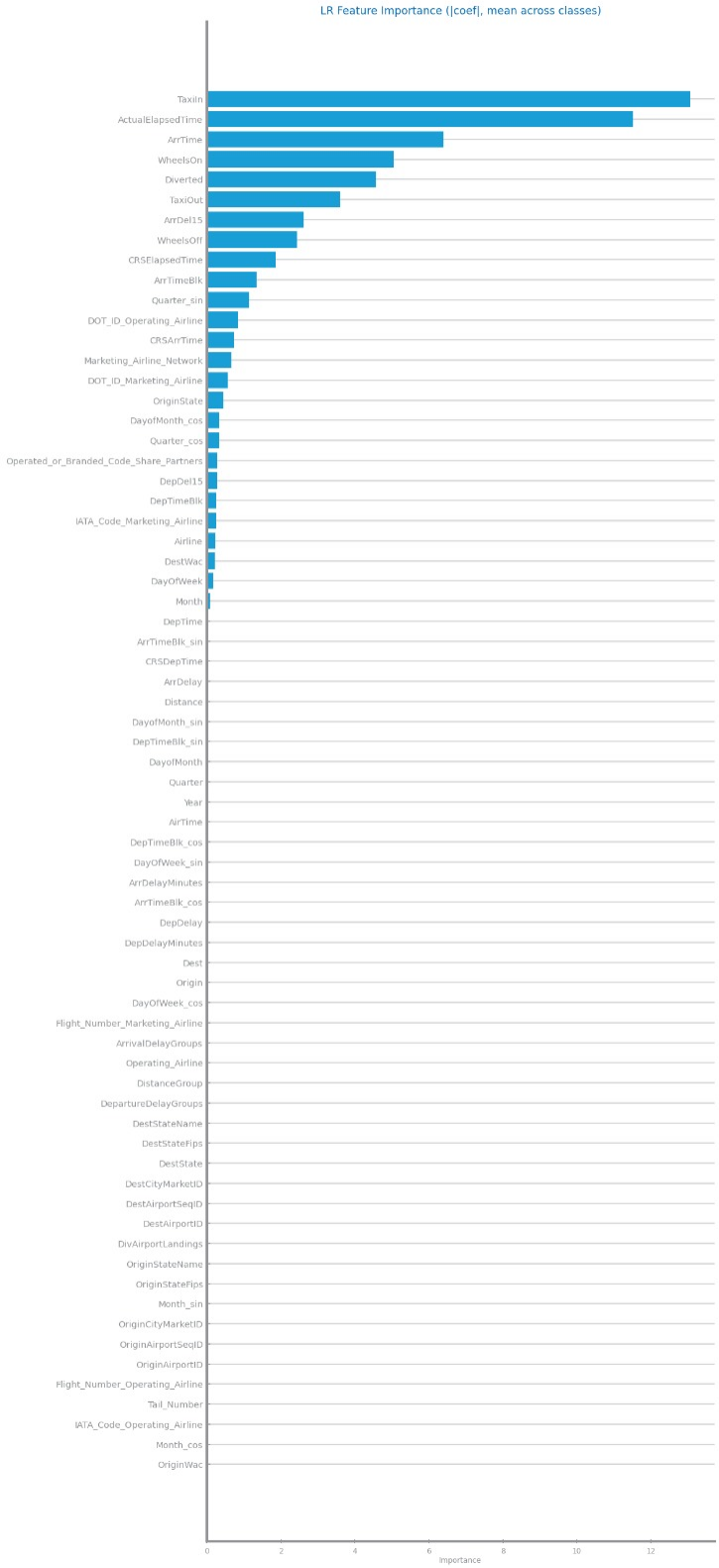


Figure 47: Logistic Regression feature importance for dataset 2

Decision Trees

Shall be used to present the results achieved through different parameterisations for the train of decision trees. The results shall be compared and explanations for them shall be presented. Shall be used to address the overfitting phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. Shall be used to present the best tree achieved and its succinct description. **Shall not exceed 500 characters.**

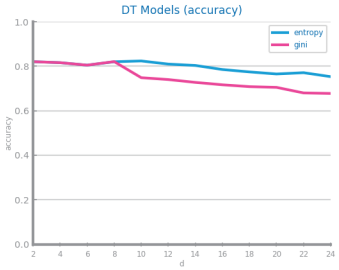


Figure 48: Decision Trees different parameterisations comparison for dataset 1

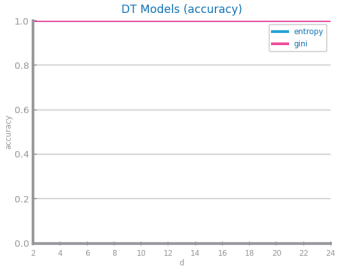


Figure 49: Decision Trees different parameterisations comparison for dataset 2

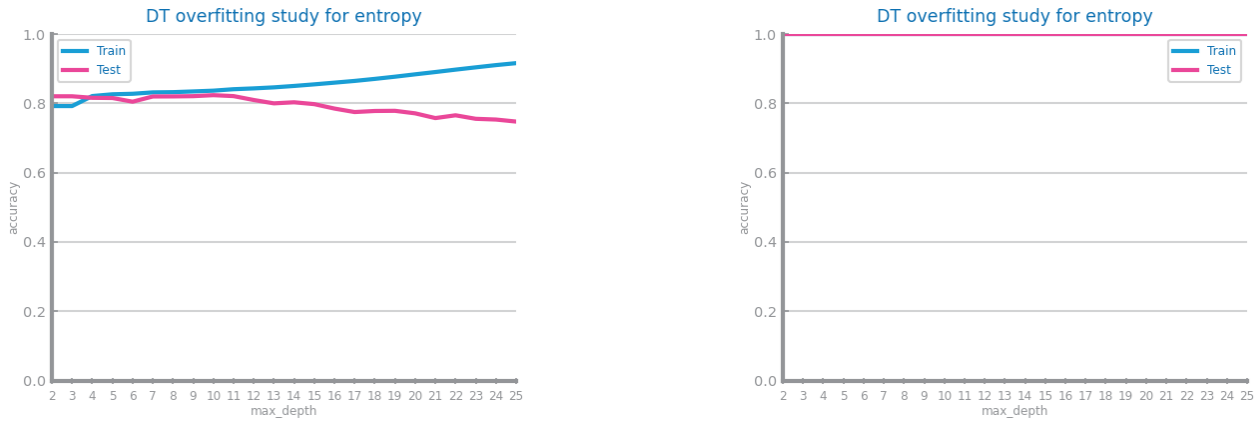


Figure 50: Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

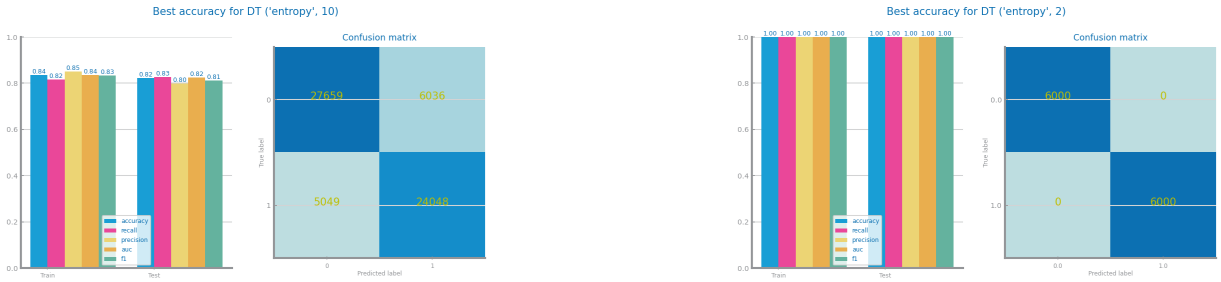


Figure 51: Decision trees best model results for dataset 1 (left) and dataset 2 (right)

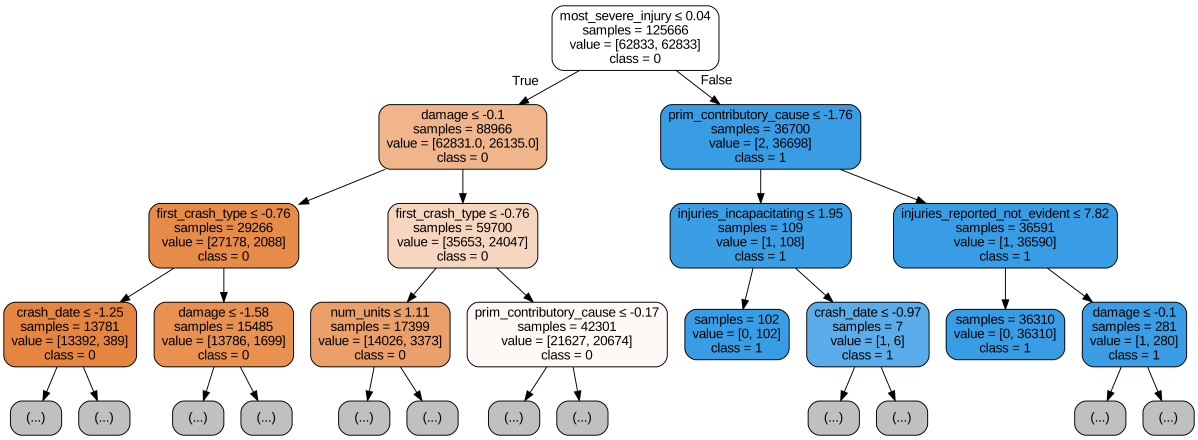


Figure 52: Best tree for dataset 1

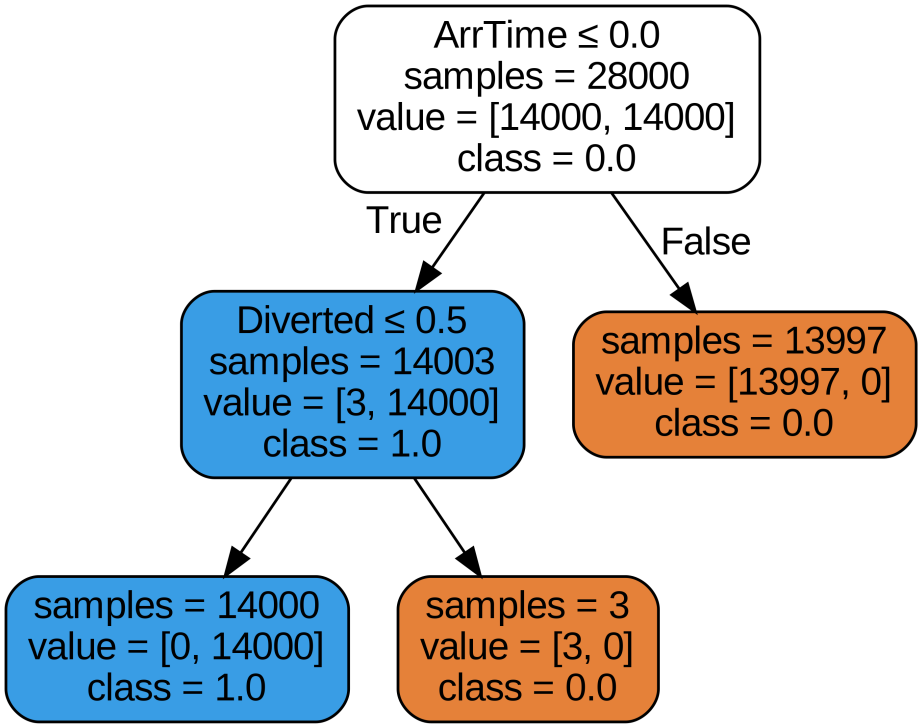


Figure 53: Best tree for dataset 2

Random Forests

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

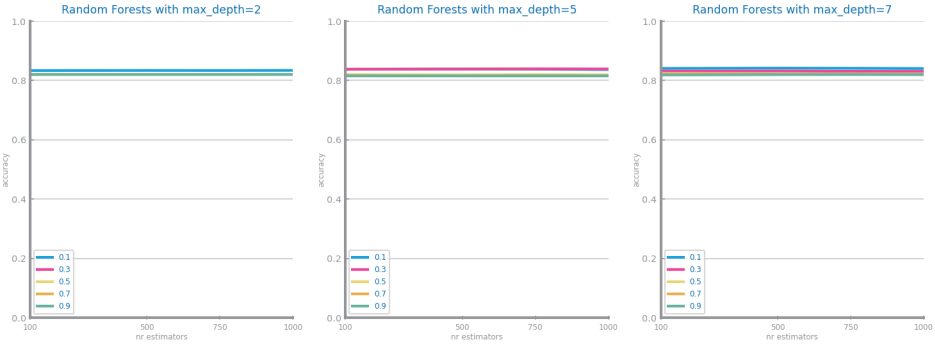


Figure 54: Random Forests different parameterisations comparison for dataset 1

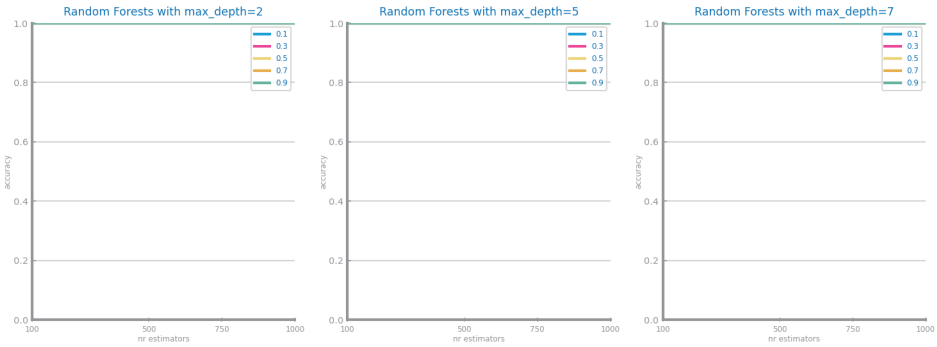


Figure 55: Random Forests different parameterisations comparison for dataset 2

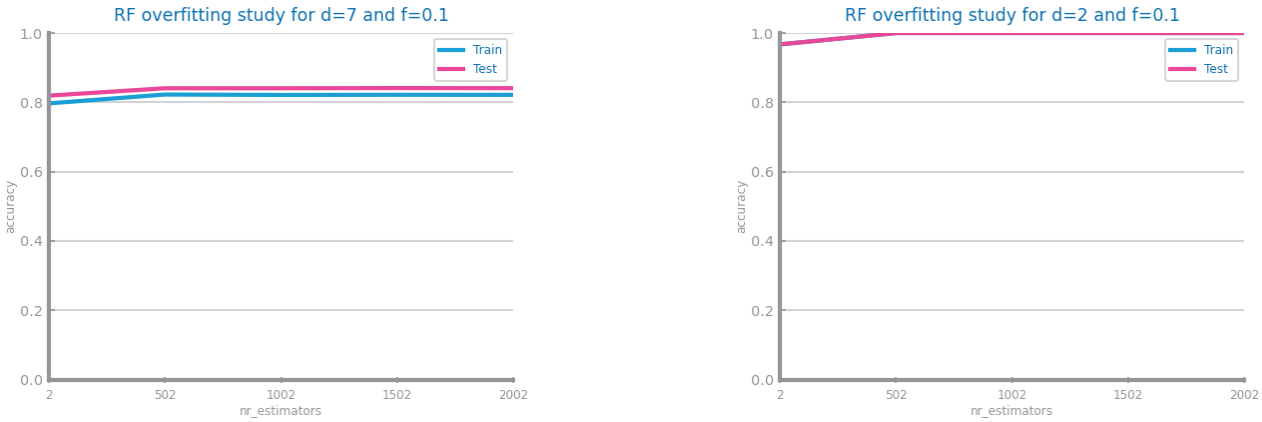


Figure 56: Random Forests overfitting analysis for dataset 1 (left) and dataset 2 (right)

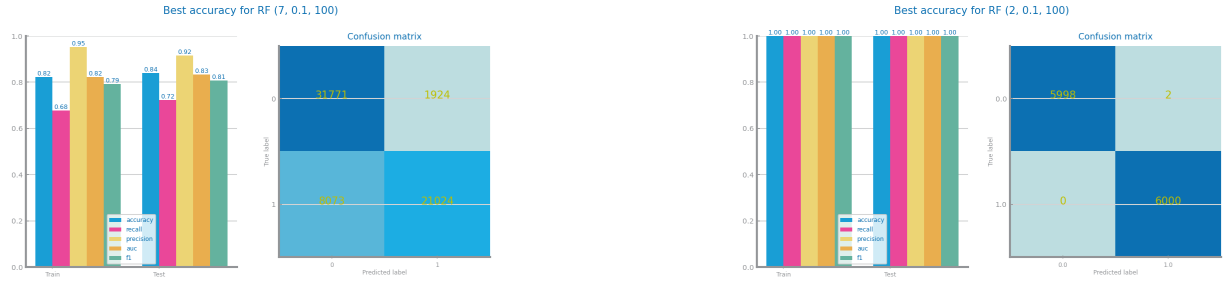


Figure 57: Random Forests best model results for dataset 1 (left) and dataset 2 (right)

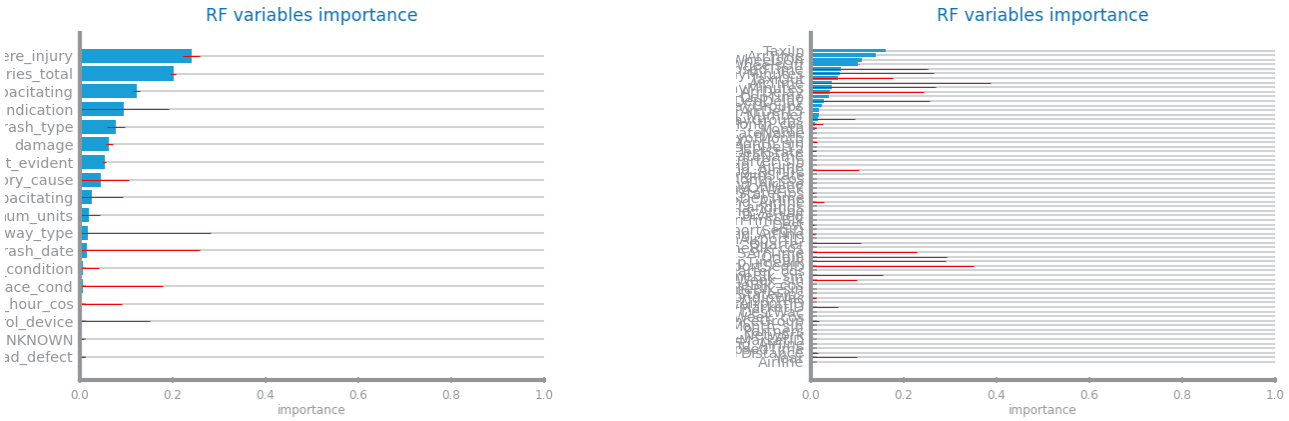


Figure 58: Random Forests variables importance for dataset 1 (left) and dataset 2 (right)

Gradient Boosting

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

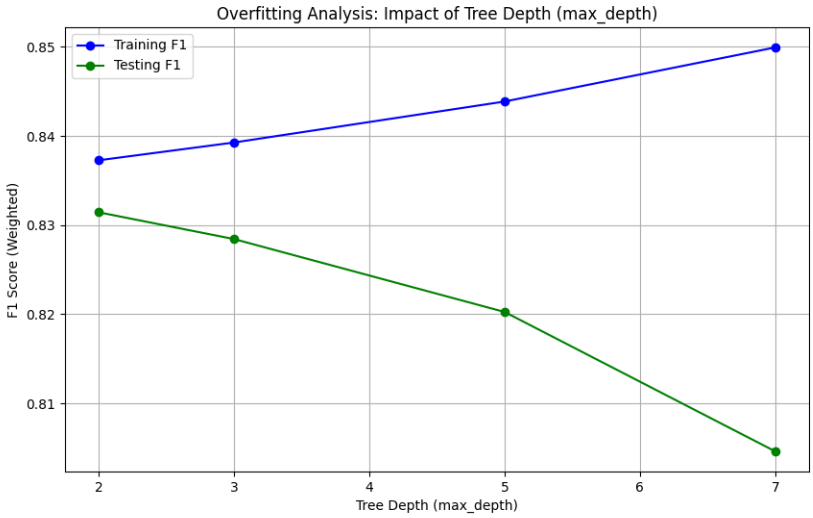


Figure 59: Gradient boosting different parameterisations comparison for dataset 1

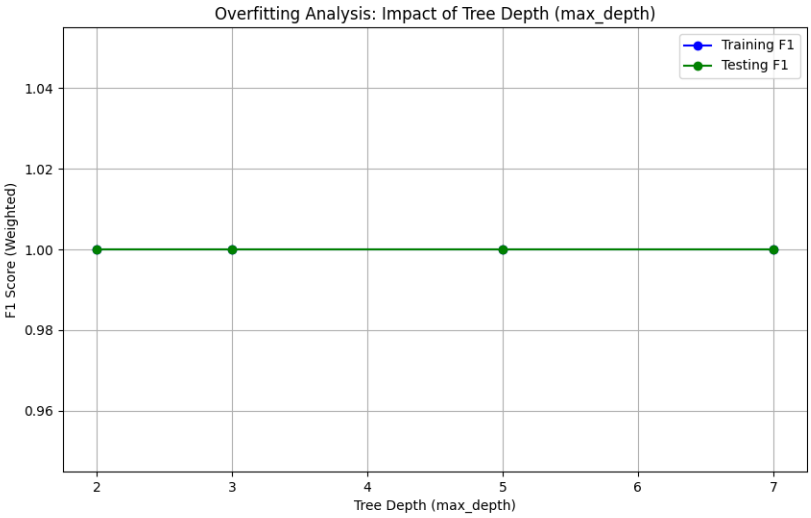


Figure 60: Gradient boosting different parameterisations comparison for dataset 2

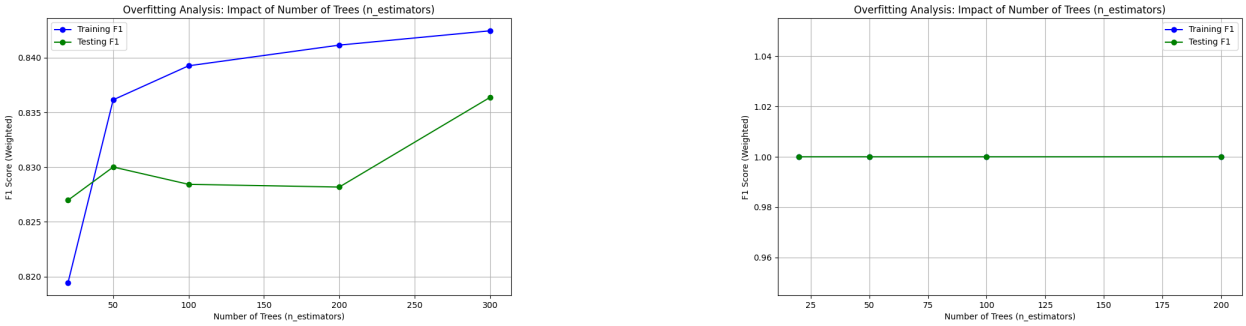


Figure 61: Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

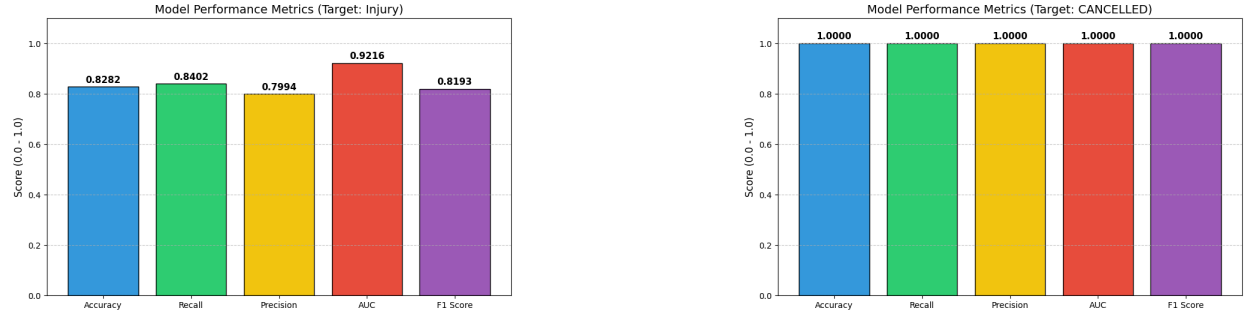


Figure 62: Gradient boosting best model results for dataset 1 (left) and dataset 2 (right)

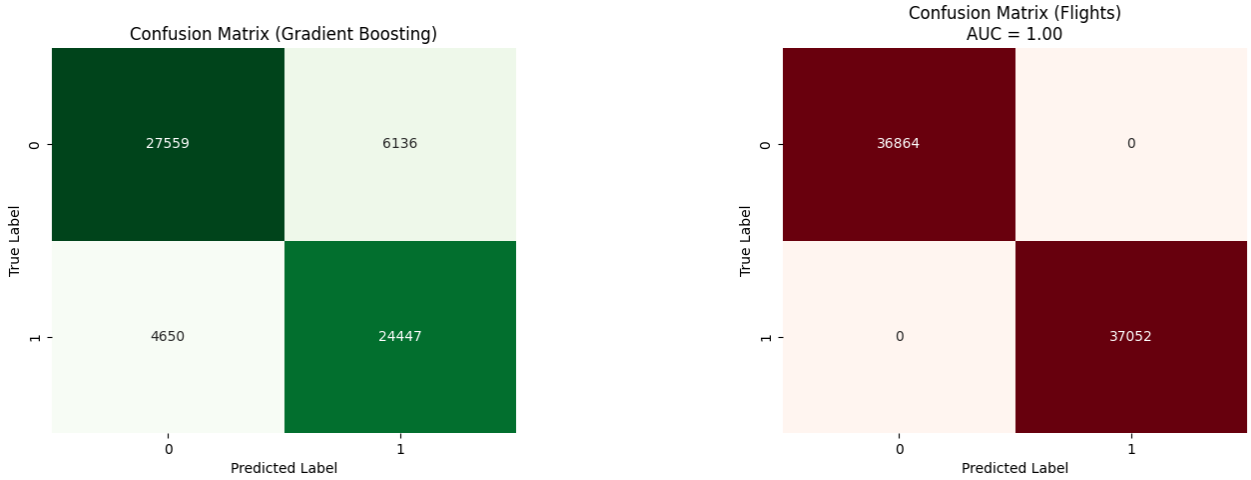


Figure 63: Gradient boosting variables importance for dataset 1 (left) and dataset 2 (right)

Multi-Layer Perceptrons

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

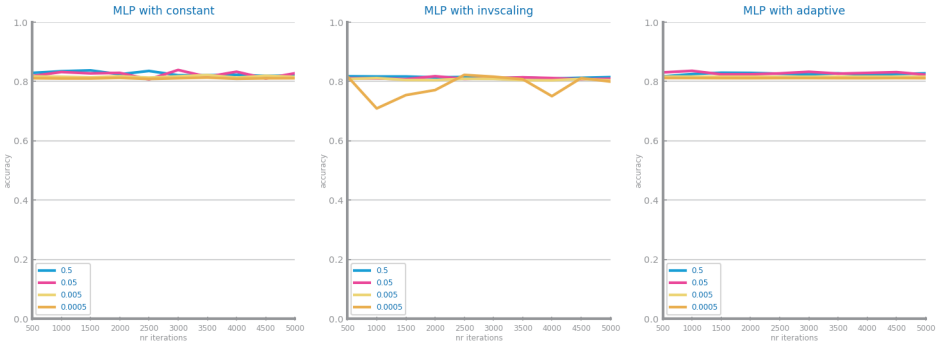


Figure 64: MLP different parameterisations comparison for dataset 1

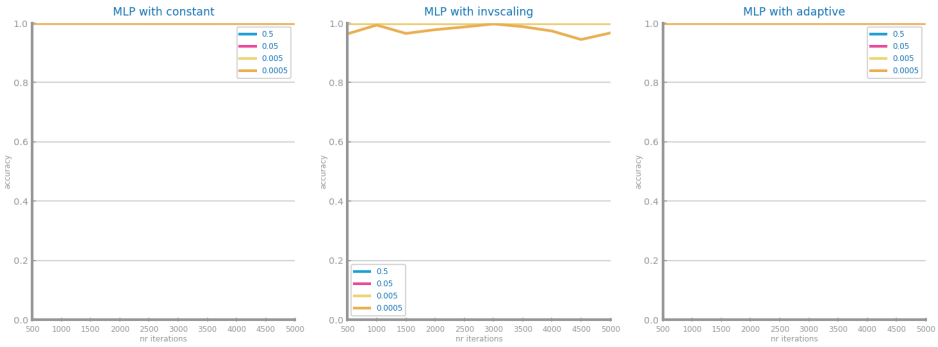


Figure 65: MLP different parameterisations comparison for dataset 2

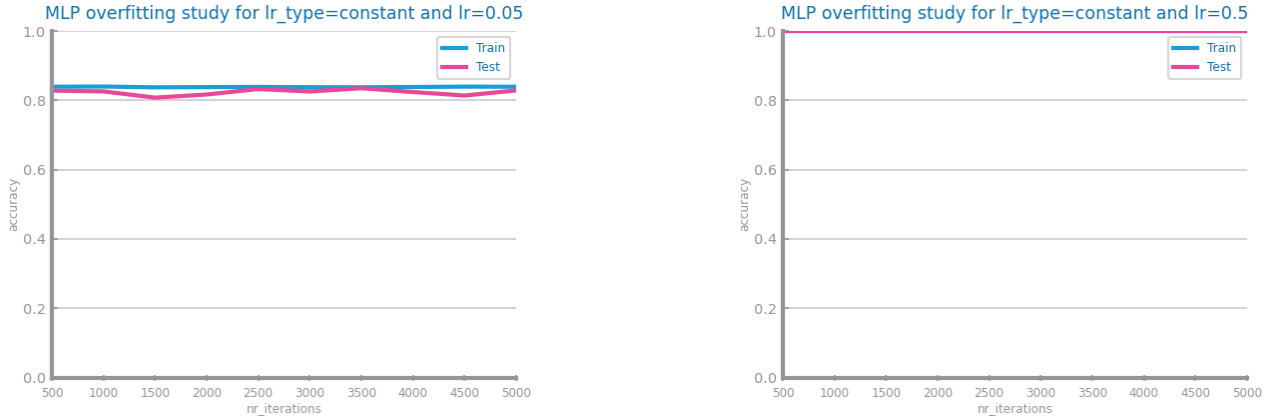


Figure 66: MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

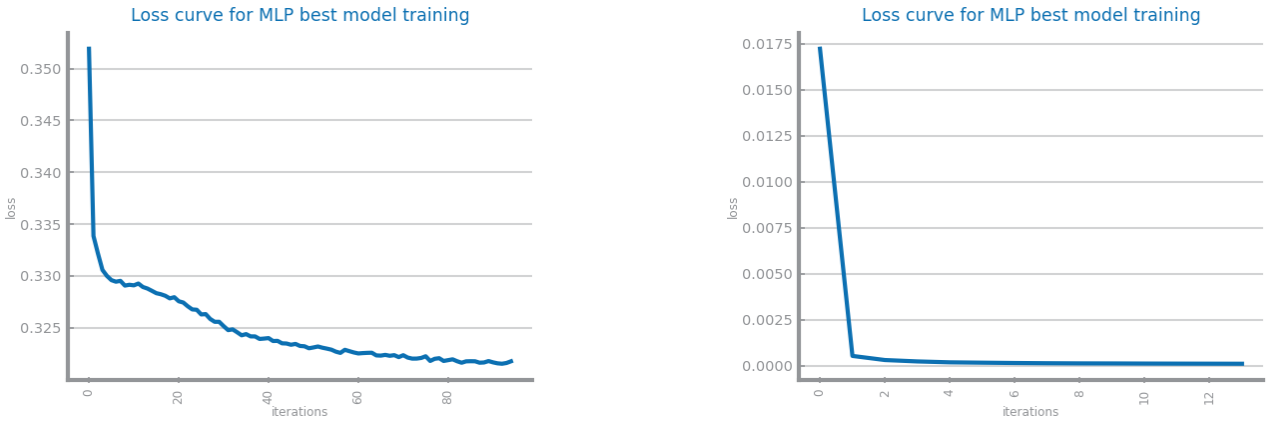


Figure 67: Loss curve analysis for dataset 1 (left) and dataset 2 (right)

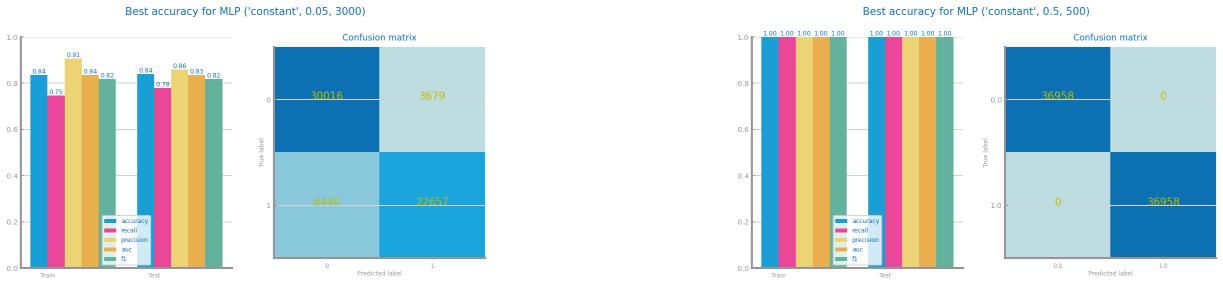


Figure 68: MLP best model results for dataset 1 (left) and dataset 2 (right)

4 CRITICAL ANALYSIS

Shall be used to present a summary of the results achieved with the different modelling techniques, and the impact of the different preparation tasks on their performance. A cross-analysis of the different models may also be presented, identifying the most relevant variables common to all of them (when possible) and the relation among the patterns identified within the different classifiers. A critical assessment of the best models shall be presented, clearly stating if the models seem to be good enough for the problem at hand. **Additional charts may be presented here. Shall not exceed 2000 characters.**

TIME SERIES ANALYSIS

5 DATA PROFILING

Data Dimensionality and Granularity

The time series was analyzed at three temporal granularities: 15-minute intervals (most atomic), daily aggregation, and weekly aggregation. The most granular level captures short-term variability and peaks, while daily and weekly granularities smooth fluctuations and highlight medium- and long-term trends.

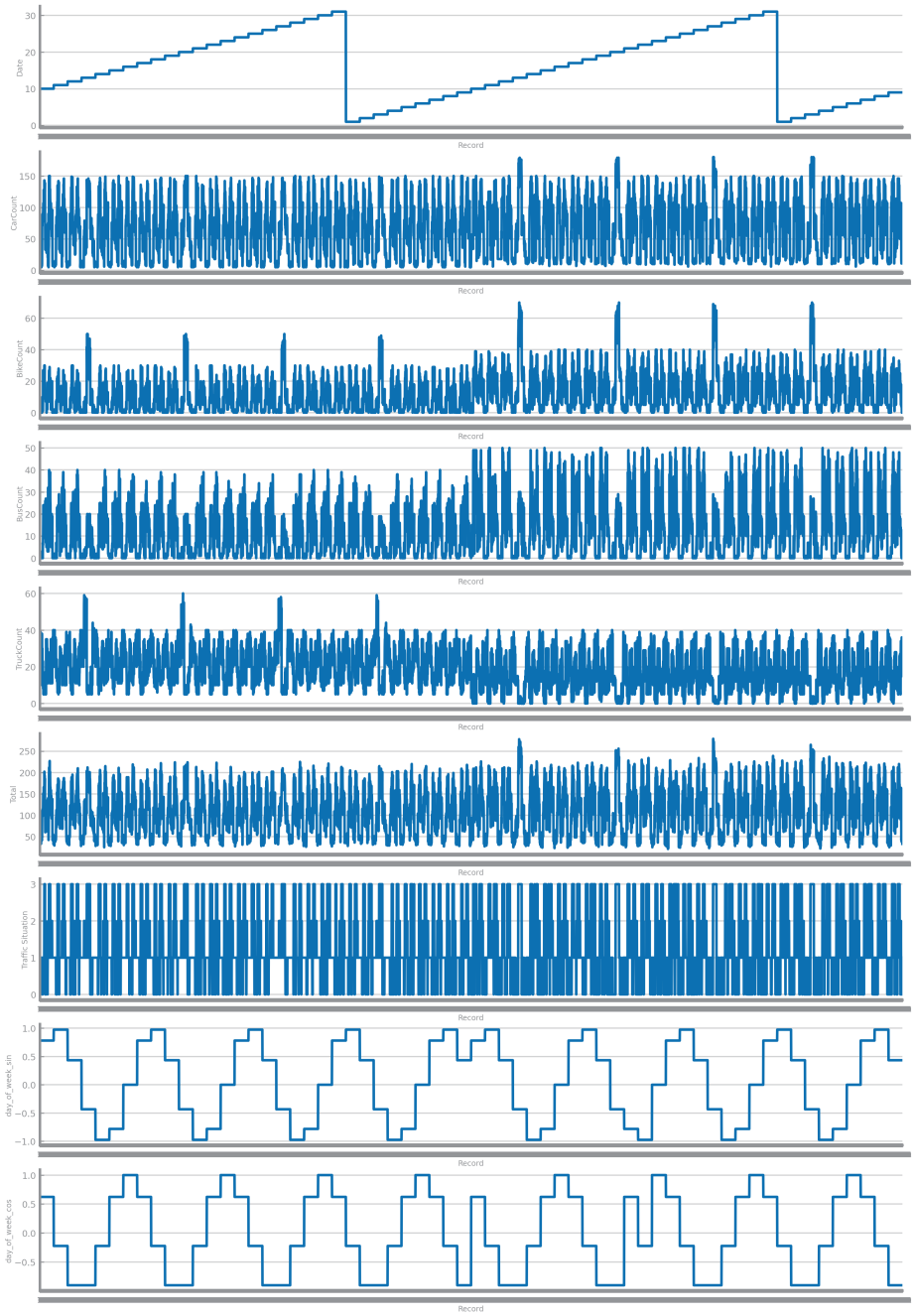


Figure 69: Time series 1 at the most granular detail

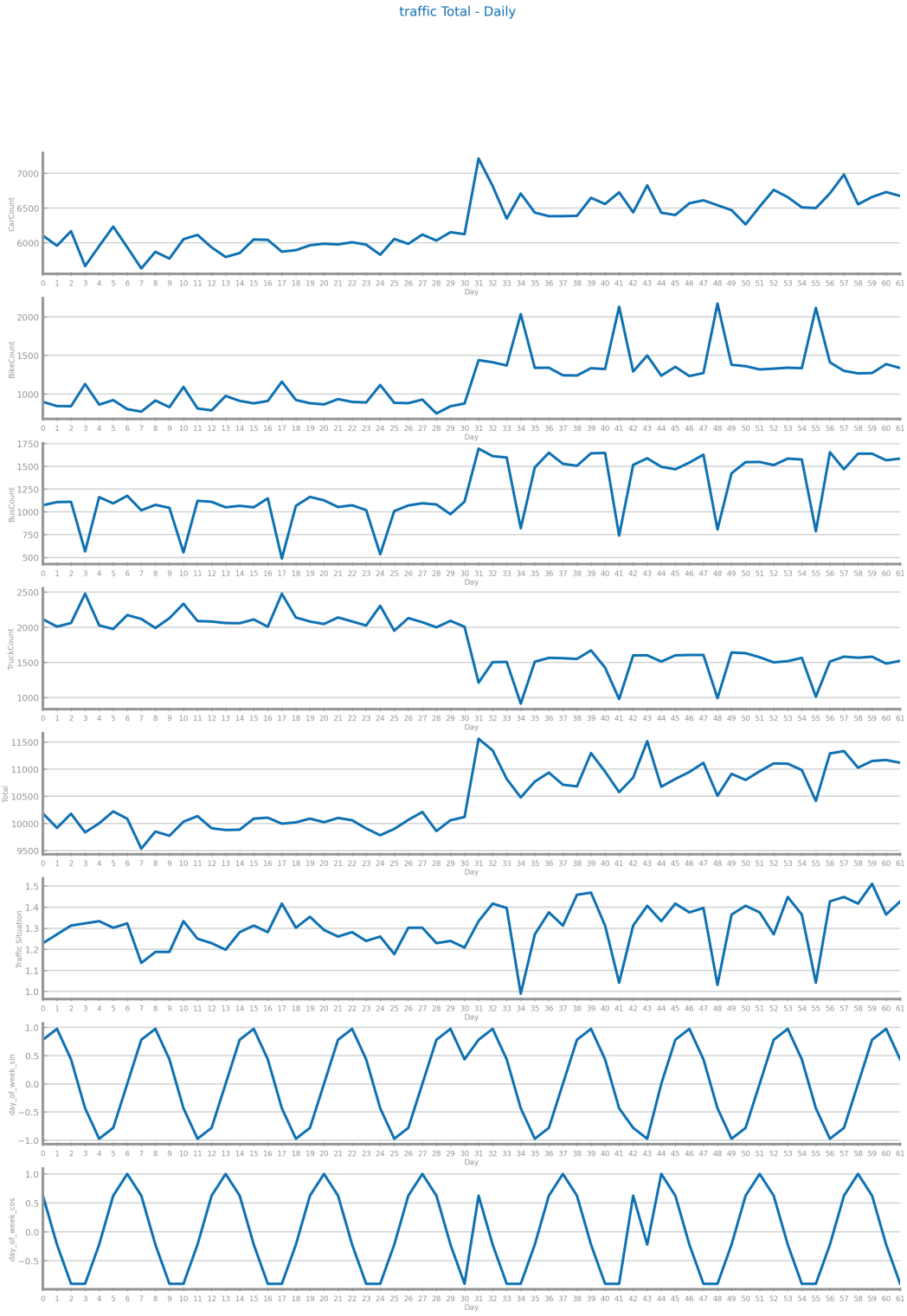


Figure 70: Time series 1 at the second chosen granularity

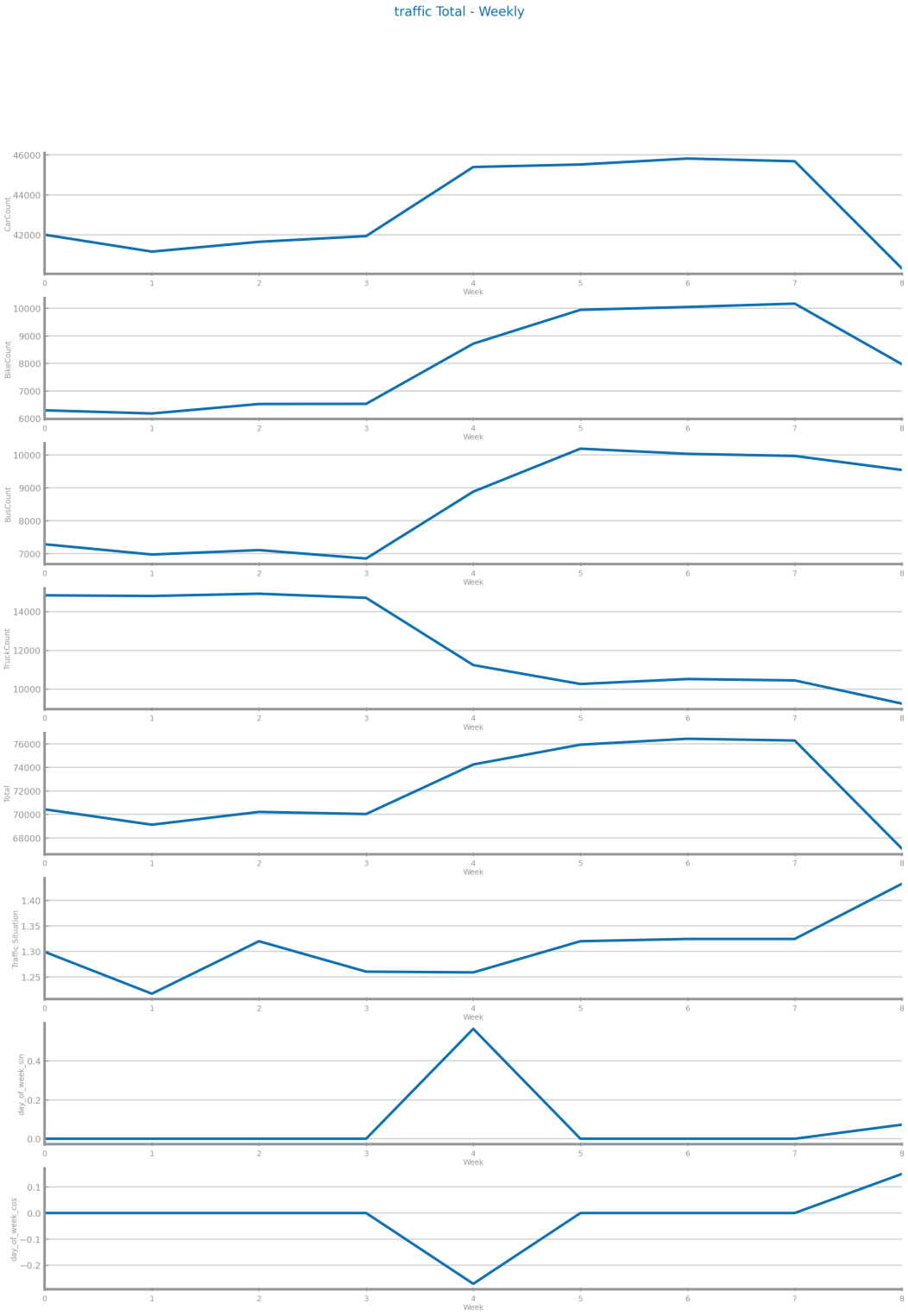


Figure 71: Time series 1 at the third chosen granularity

Data Distribution

The distribution analysis across granularities shows that aggregation reduces variability and outliers. The 15-minute series presents higher dispersion and skewness, while hourly and daily series exhibit smoother distributions.

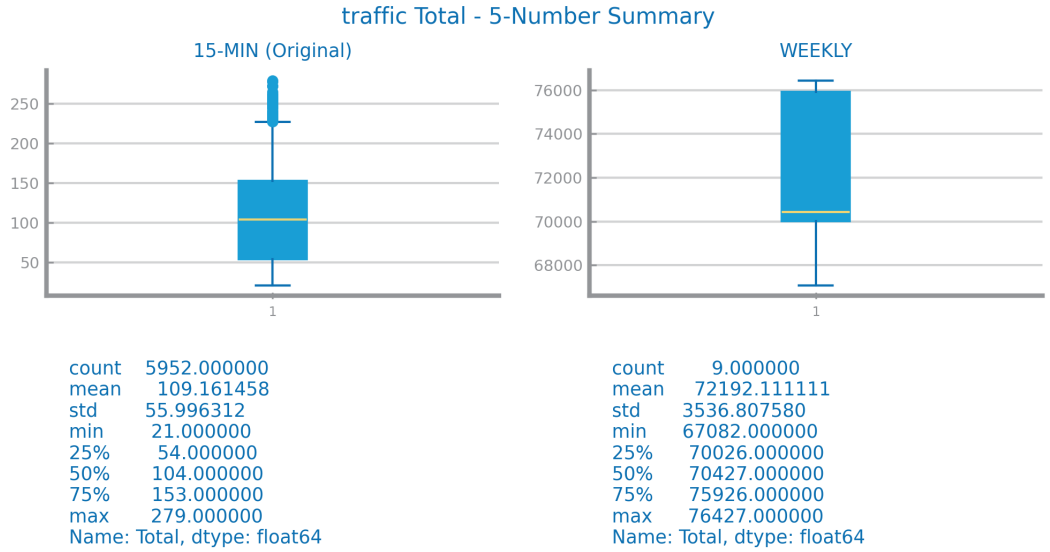


Figure 72: Boxplot(s) for time series 1

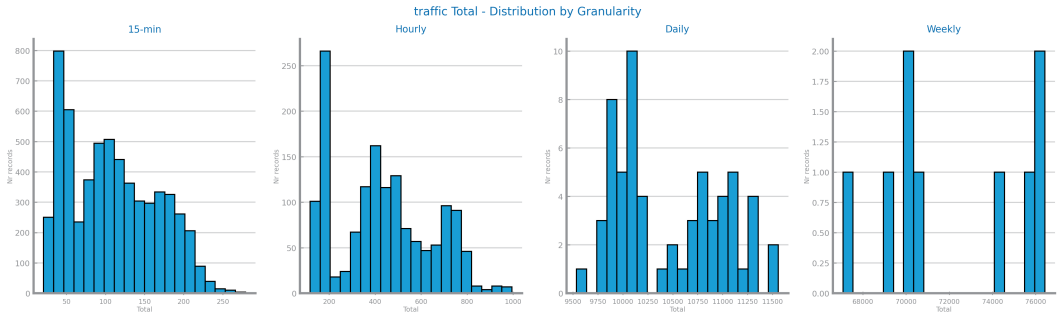


Figure 73: Histogram(s) for time series 1

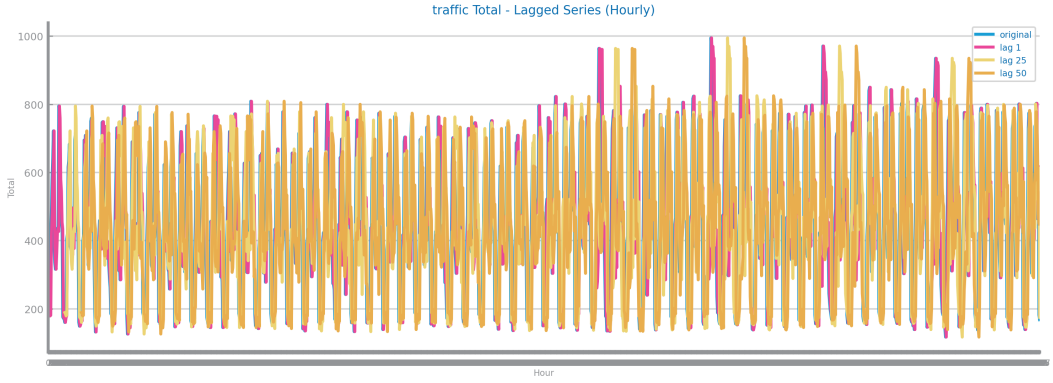


Figure 74: Autocorrelation lag-plots for original time series 1

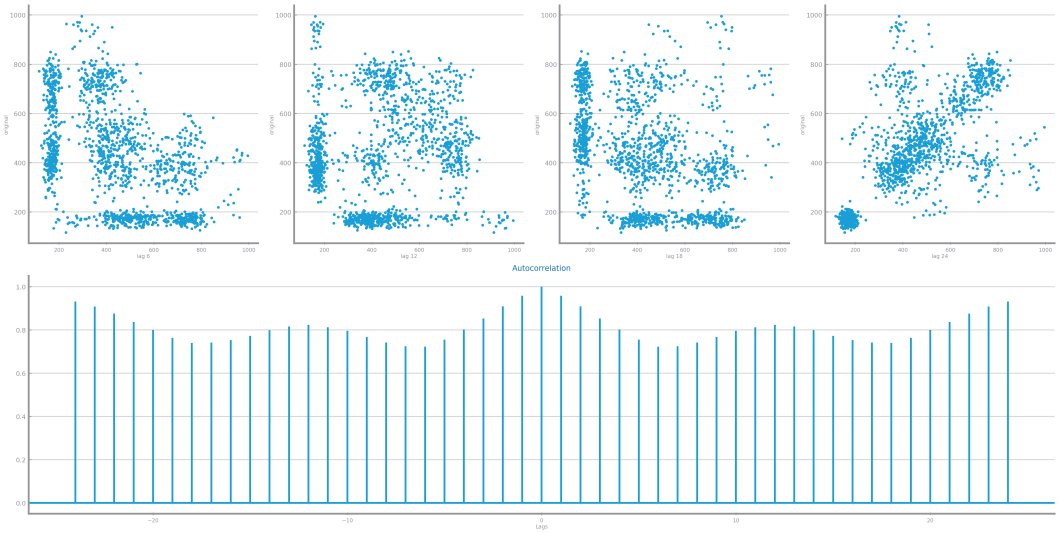


Figure 75: Autocorrelation correlogram for original time series 1

Data Stationarity

Stationarity analysis shows that the series is stationary at finer granularities but loses stationarity when aggregated daily. Augmented Dickey-Fuller Test confirm that trend components become more pronounced at coarser granularities, indicating the need for detrending or differencing when modeling aggregated series.

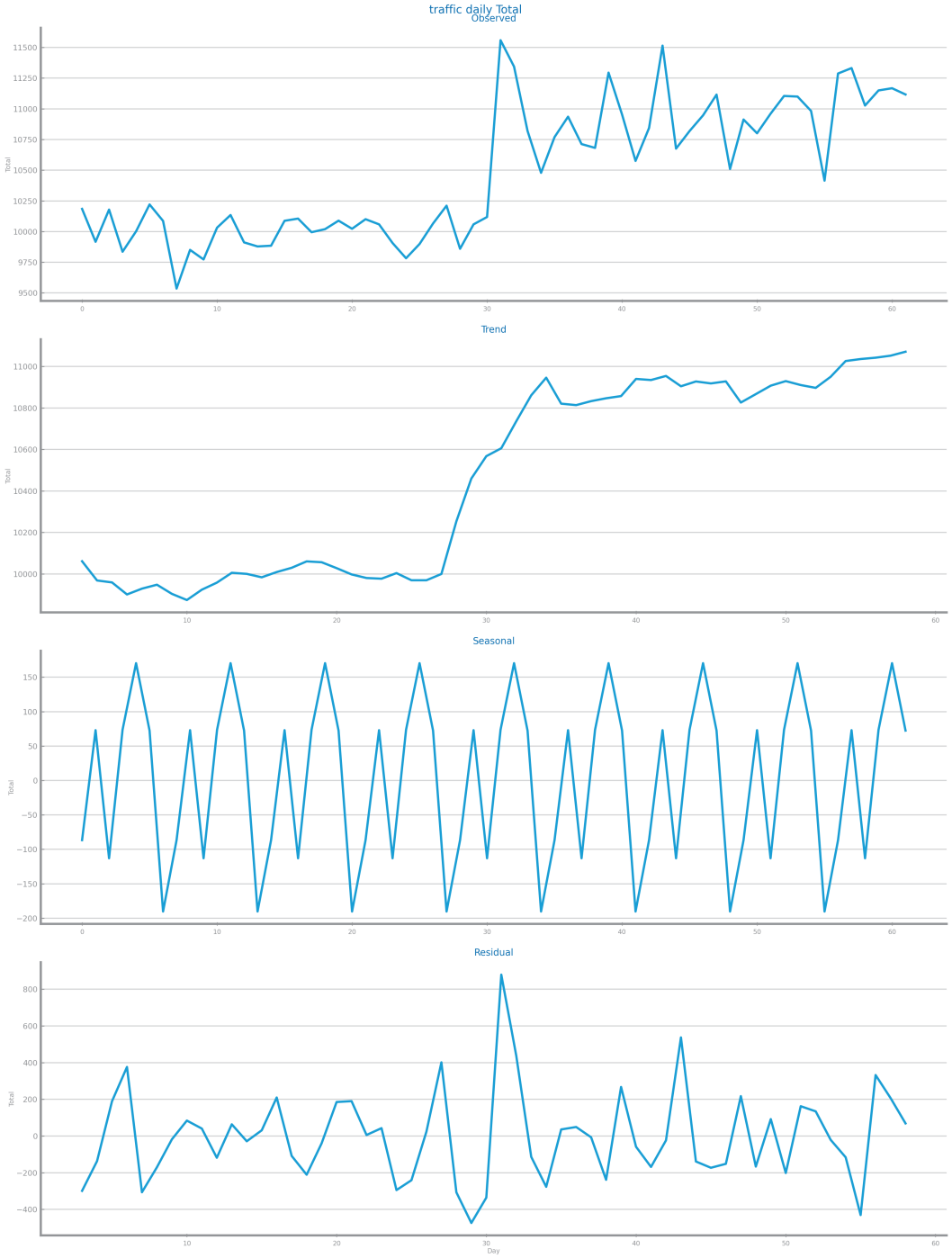


Figure 76: Components study for time series 1

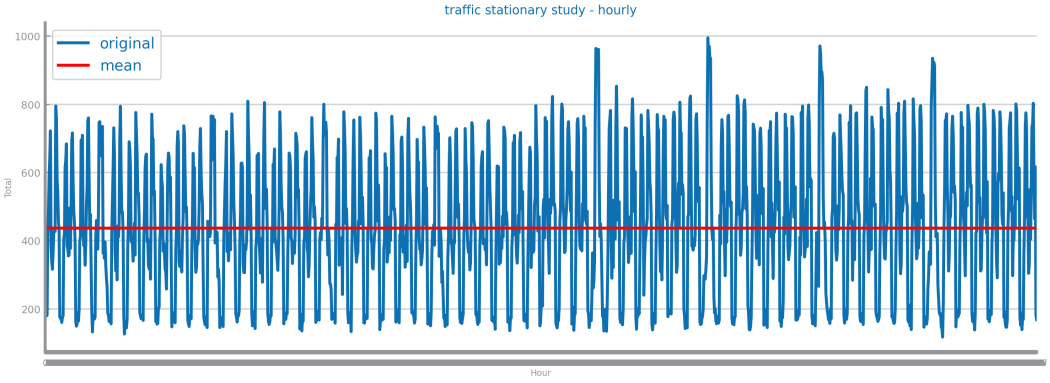


Figure 77: Stationarity study for time series 1

Augmented Dickey-Fuller Test Results:

Original (15-min):

- ADF Statistic: -14.441
- p-value: 0.000
- Critical Values: 1%: -3.431, 5%: -2.862, 10%: -2.567
- **The series IS stationary**

Hourly:

- ADF Statistic: -8.903
- p-value: 0.000
- Critical Values: 1%: -3.435, 5%: -2.864, 10%: -2.568
- **The series IS stationary**

Daily:

- ADF Statistic: -0.826
- p-value: 0.811
- Critical Values: 1%: -3.548, 5%: -2.913, 10%: -2.594
- **The series IS NOT stationary**

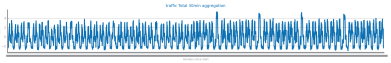
6 DATA TRANSFORMATION

Aggregation

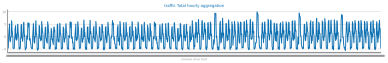
We tested 30-minute, hourly, daily, and weekly aggregations to see how each level reduces noise and how much detail is lost.The plots (see Figures 78, 79, and 80) reveal that no aggregation keeps high-frequency noise, while daily and weekly

levels smooth the data too much.On the test set, linear regression produced constant predictions with R^2 close to 0, signaling a poor fit to the data.Meanwhile, the optimistic persistence model performed better, and the 30-minute level gave the best R^2 (0.67), so we chose it.

(a) No Aggregation (original after scaling)

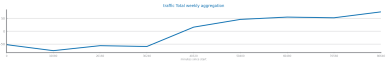


(b) 30min Aggregation



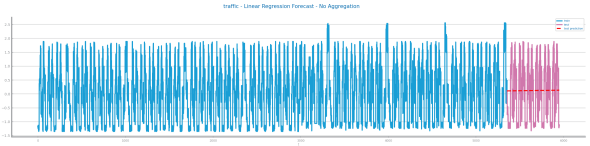
(c) Hourly Aggregation

(d) Daily Aggregation

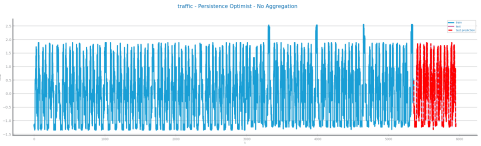


(e) Weekly Aggregation

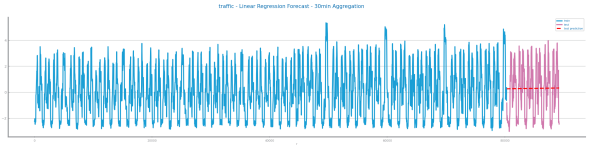
Figure 78: Time series plots after different levels of aggregation



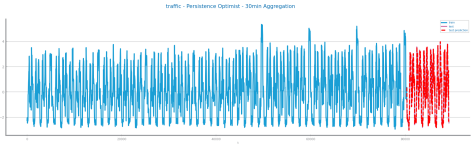
(a) Linear Regression – No Aggregation



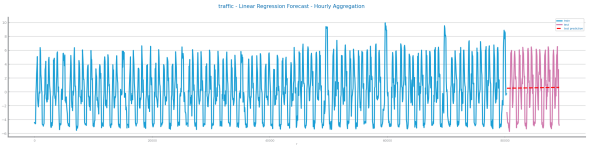
(b) Persistence Optim – No Aggregation



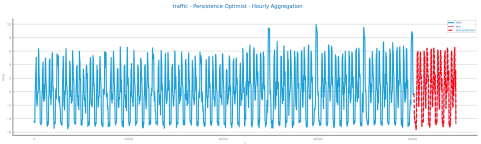
(c) Linear Regression – 30min Aggregation



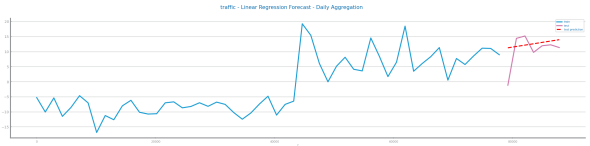
(d) Persistence Optim – 30min Aggregation



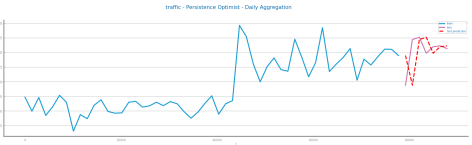
(e) Linear Regression – Hourly Aggregation



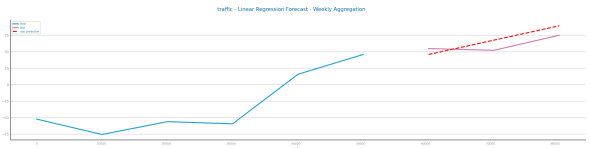
(f) Persistence Optim – Hourly Aggregation



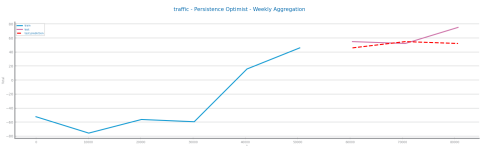
(g) Linear Regression – Daily Aggregation



(h) Persistence Optim – Daily Aggregation



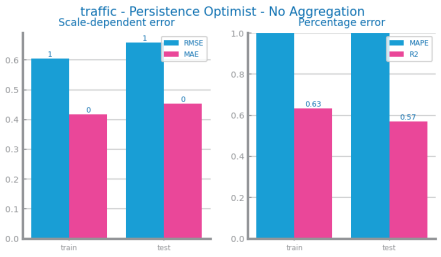
(i) Linear Regression – Weekly Aggregation



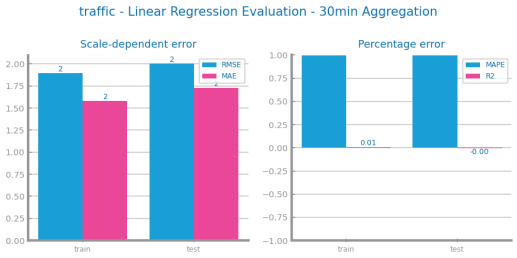
(j) Persistence Optim – Weekly Aggregation

Figure 79: Forecasting plots for Linear Regression and Persistence Optim after different aggregations

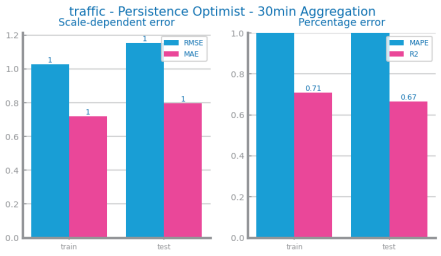
(a) Linear Regression Evaluation – No Aggregation



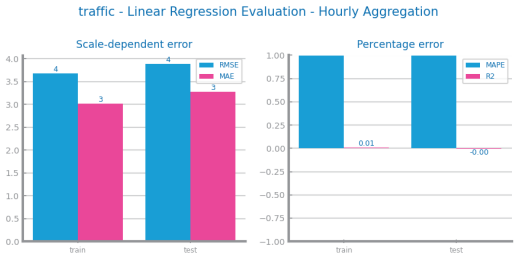
(b) Persistence Optim Evaluation – No Aggregation



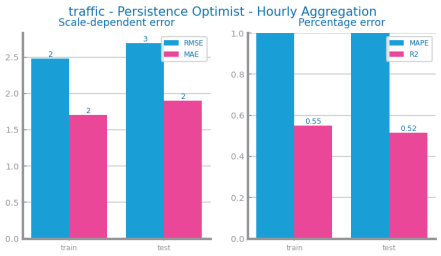
(c) Linear Regression Evaluation – 30min Aggregation



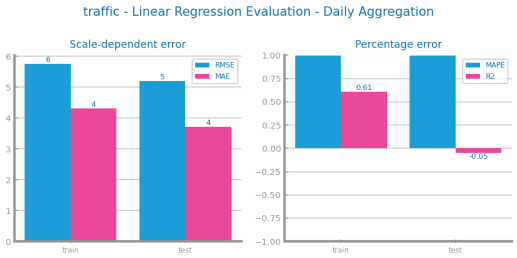
(d) Persistence Optim Evaluation – 30min Aggregation



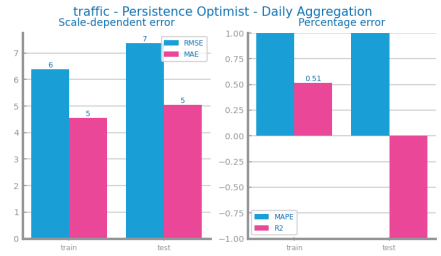
(e) Linear Regression Evaluation – Hourly Aggregation



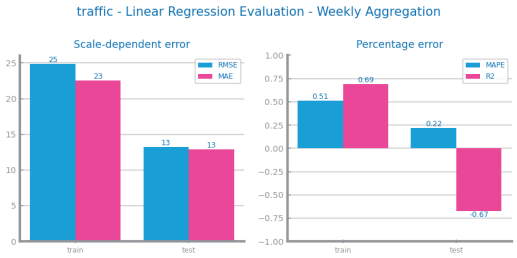
(f) Persistence Optim Evaluation – Hourly Aggregation



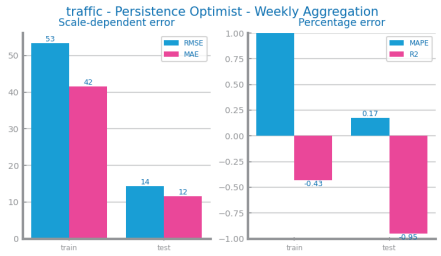
(g) Linear Regression Evaluation – Daily Aggregation



(h) Persistence Optim Evaluation – Daily Aggregation



(i) Linear Regression Evaluation – Weekly Aggregation

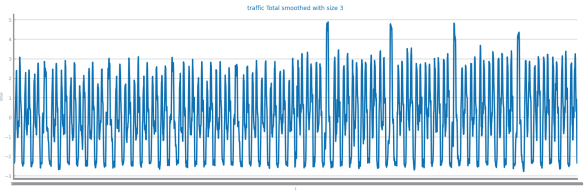


(j) Persistence Optim Evaluation – Weekly Aggregation

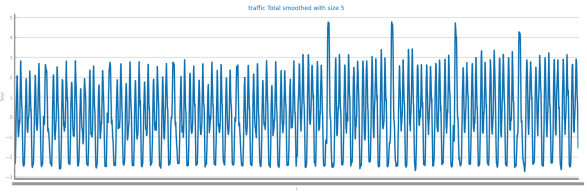
Figure 80: Evaluation results for Linear Regression and Persistence Optim after different aggregations

Smoothing

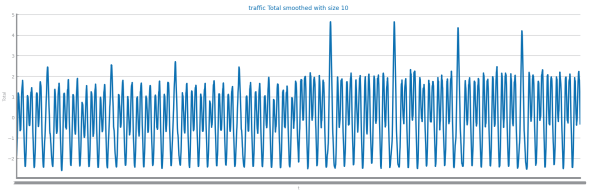
We tested moving-average smoothing with window sizes of 3, 5, 10, and 15 to see how noise reduction affects forecasting without distorting the data.Smaller windows kept more noise, while larger ones over-smoothed.The window of 15 worked best, giving RMSE 0.3, MAE 0.25, and R^2 0.96 in the optimistic persistence model.Linear regression stayed flat with low R^2, so we chose size 15.



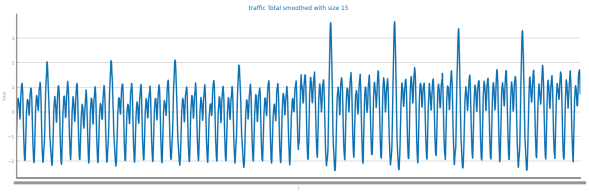
(a) Smoothing Size 3



(b) Smoothing Size 5

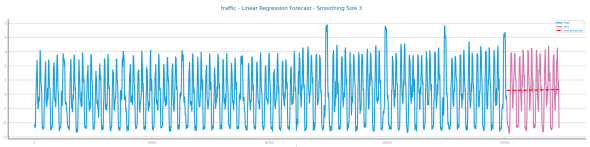


(c) Smoothing Size 10

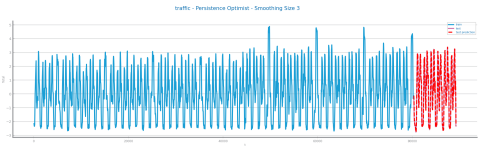


(d) Smoothing Size 15

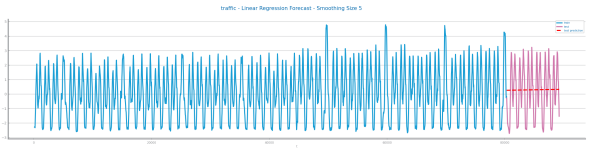
Figure 81: Time series plots after applying moving average smoothing with different window sizes



(a) Linear Regression – Size 3

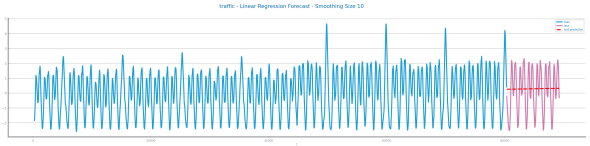


(b) Persistence Optim – Size 3



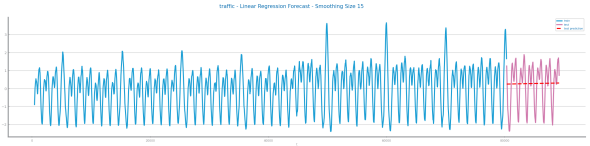
(c) Linear Regression – Size 5

(d) Persistence Optim – Size 5

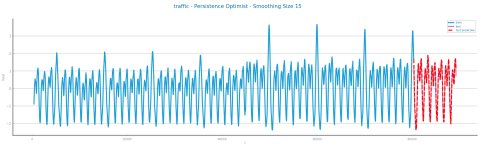


(e) Linear Regression – Size 10

(f) Persistence Optim – Size 10

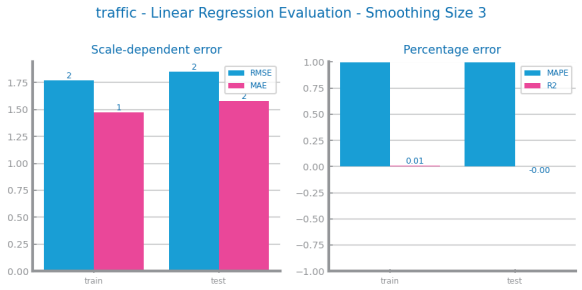


(g) Linear Regression – Size 15

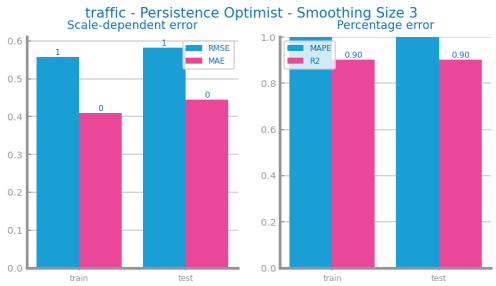


(h) Persistence Optim – Size 15

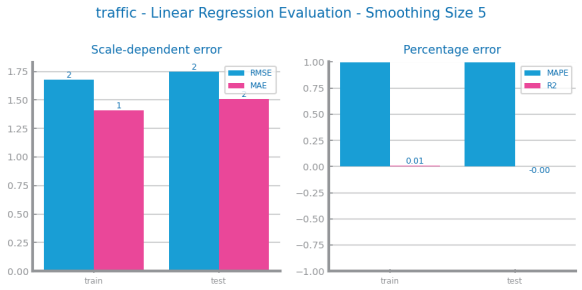
Figure 82: Forecasting plots for Linear Regression and Persistence Optim after different smoothing window sizes



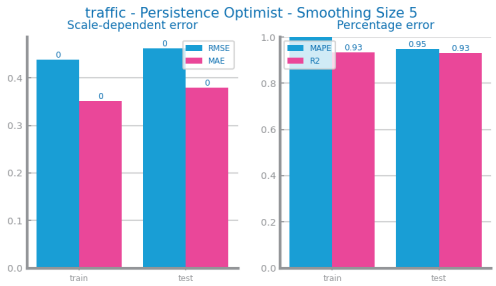
(a) Linear Regression Evaluation – Size 3



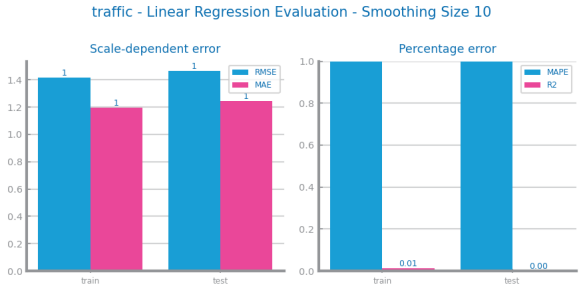
(b) Persistence Optim Evaluation – Size 3



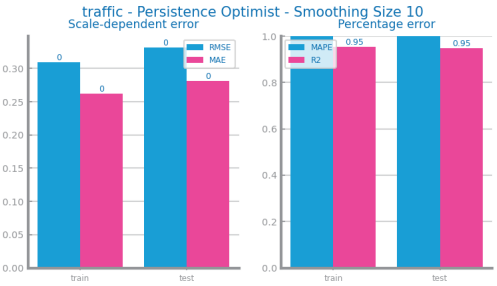
(c) Linear Regression Evaluation – Size 5



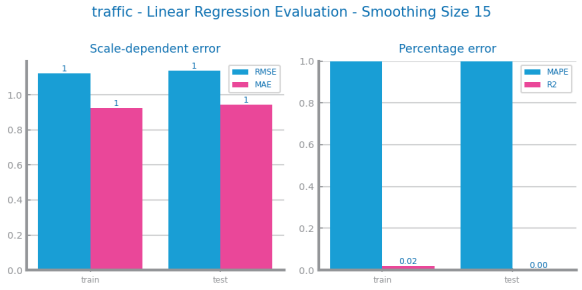
(d) Persistence Optim Evaluation – Size 5



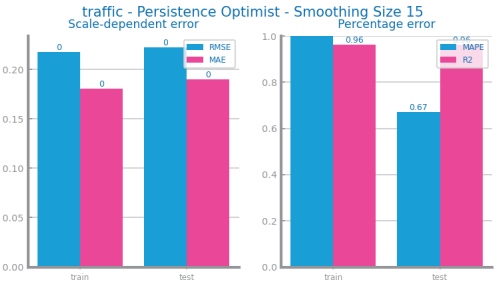
(e) Linear Regression Evaluation – Size 10



(f) Persistence Optim Evaluation – Size 10



(g) Linear Regression Evaluation – Size 15



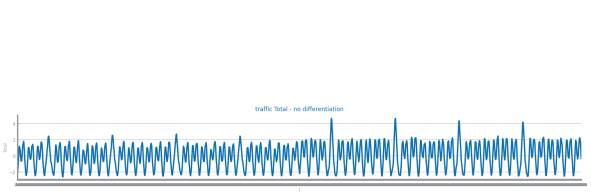
(h) Persistence Optim Evaluation – Size 15

Figure 83: Evaluation results for Linear Regression and Persistence Optim after different smoothing window sizes

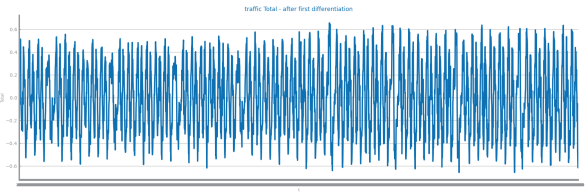
Differentiation

To handle possible non-stationarity, we tested no differencing, first-order, and second-order differencing.No differencing kept the trend, first-order removed linear trends and centered the data, and second-order added noise without helping.The

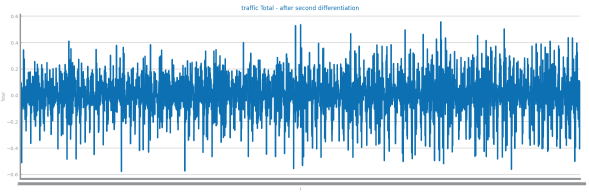
optimistic persistence model performed best with no differencing, reaching R^2 0.95.Linear regression improved only slightly, so we chose no differencing.



(a) No Differentiation

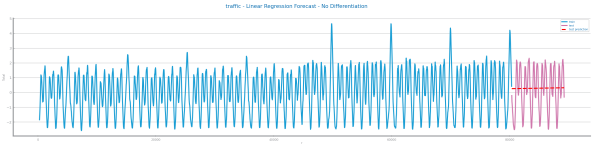


(b) First Differentiation

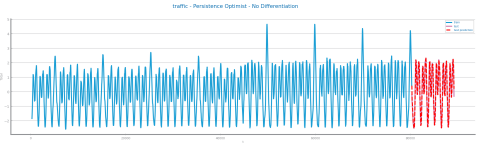


(c) Second Differentiation

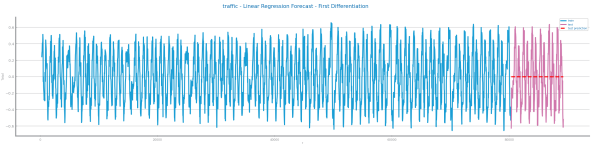
Figure 84: Time series plots without and after applying first, and second differentiation



(a) Linear Regression – No Differentiation

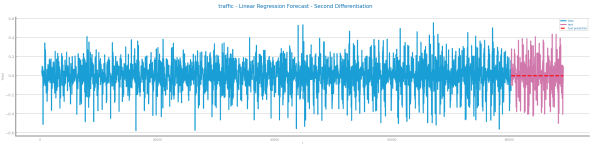


(b) Persistence Optim – No Differentiation

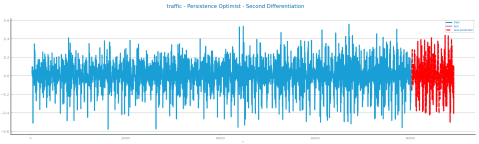


(c) Linear Regression – First Differentiation

(d) Persistence Optim – First Differentiation

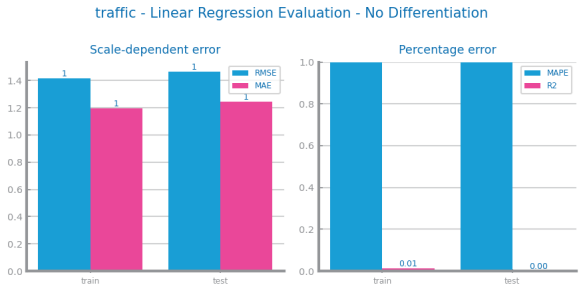


(e) Linear Regression – Second Differentiation

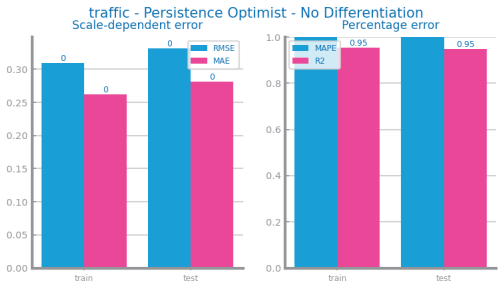


(f) Persistence Optim – Second Differentiation

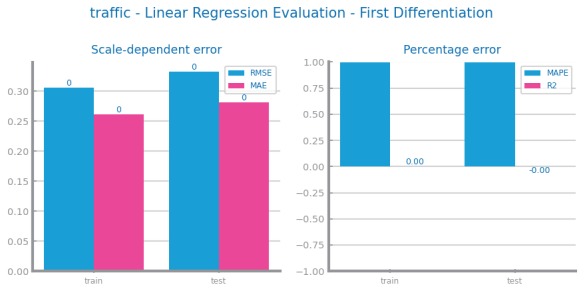
Figure 85: Forecasting plots for Linear Regression and Persistence Optim after different levels of differentiation



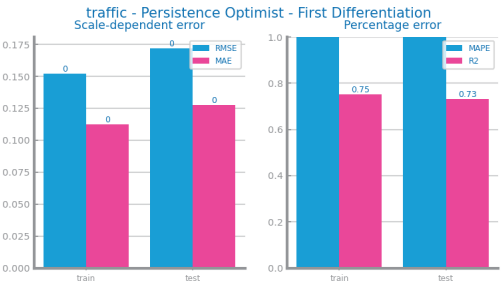
(a) Linear Regression Evaluation – No Differentiation



(b) Persistence Optim Evaluation – No Differentiation

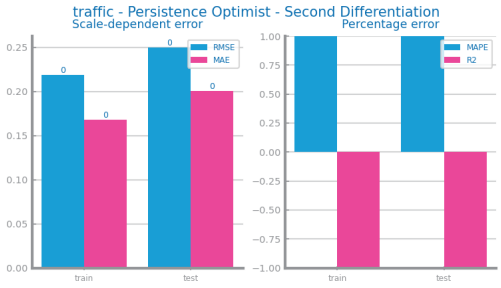


(c) Linear Regression Evaluation – First Differentiation



(d) Persistence Optim Evaluation – First Differentiation

(e) Linear Regression Evaluation – Second Differentiation

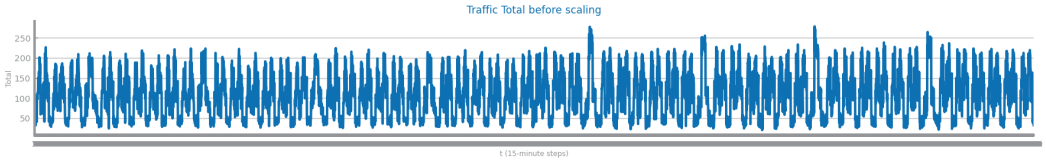


(f) Persistence Optim Evaluation – Second Differentiation

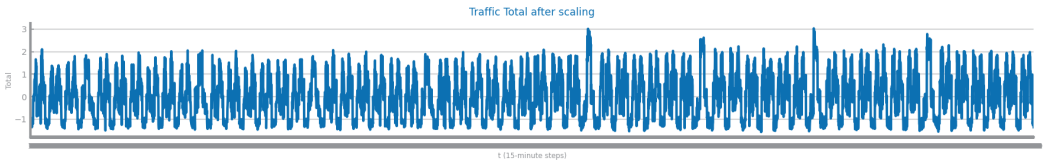
Figure 86: Evaluation results for Linear Regression and Persistence Optim after different levels of differentiation

Scaling

We used StandardScaler to normalize the data, converting it to a mean of 0 and a standard deviation of 1.Before scaling, the values ranged from about 25 to 250 with high variance; after scaling, they fell roughly between –1 and 3.This improved model stability,and sped up convergence as shown in the before-and-after plots.We did not test other scaling methods because StandardScaler solved the scale issues without adding complexity.The final preparation pipeline for the traffic dataset is: 30-minute aggregation, smoothing with window size 15, no differencing, and scaling with StandardScaler.



(a) Before scaling



(b) After StandardScaler()

Figure 87: Effect of StandardScaler on the original 15-minute time series

7 MODELS' EVALUATION

Exponential Smoothing Model

We applied simple exponential smoothing and tuned alpha from 0.1 to 0.9. The hyperparameter study (Figure 88) shows R^2 starts near 0.0 at alpha = 0.1 and drops sharply to −1.0 by alpha = 0.3, indicating the model overreacts to recent changes.

The forecast (Figure 89) follows the overall trend but smooths out sharp variations. The evaluation (Figure 90) shows poor test performance: RMSE ≈ 0.24, MAE ≈ 0.20, MAPE ≈ 100%, R^2 ≈ −0.01, revealing weak accuracy and poor generalization.

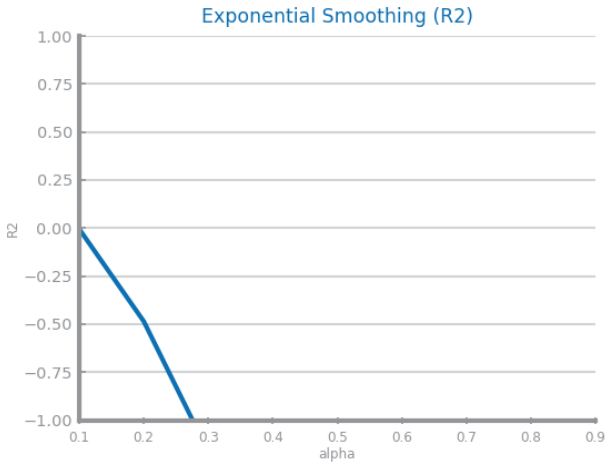


Figure 88: Hyperparameter study: R^2 as a function of alpha for Exponential Smoothing

Figure 89: Forecasting plots obtained with the best Exponential Smoothing model (predictions in red vs actual test data in pink)

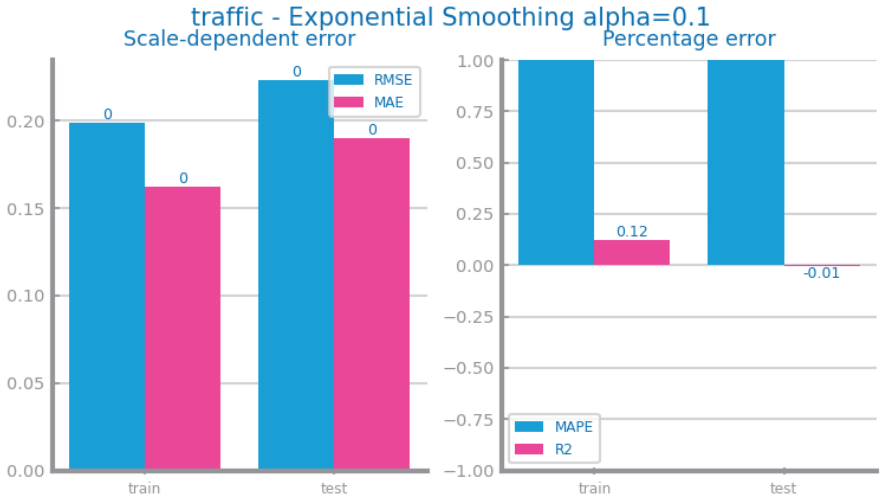


Figure 90: Performance metrics of the best Exponential Smoothing model (RMSE, MAE, MAPE, R^2)

Multi-layer Perceptrons Model

For the MLP model, we explored various hidden layer architectures, including single layers with 50 or 100 neurons, and multi-layers such as (50, 50) and (100, 50), trained with the Adam optimizer and MSE loss.The best setup was (50, 50), reaching test $R^2 \approx 0.86$ after $\sim$800 epochs with stable training and no overfitting.The forecasting plot (Figure 92) closely matched the test data, capturing nonlinear patterns better than simpler models.Performance metrics (Figure 93) show RMSE ≈ 0.25, MAE ≈ 0.20, MAPE $\approx 18\%$, and $R^2 \approx 0.86$, confirming solid medium-term performance despite higher computational cost.

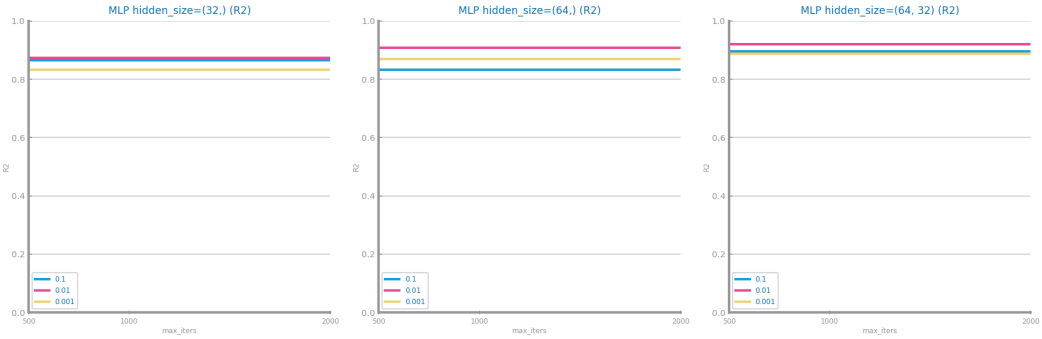


Figure 91: Hyperparameter study: R^2 convergence for different MLP hidden layer configurations

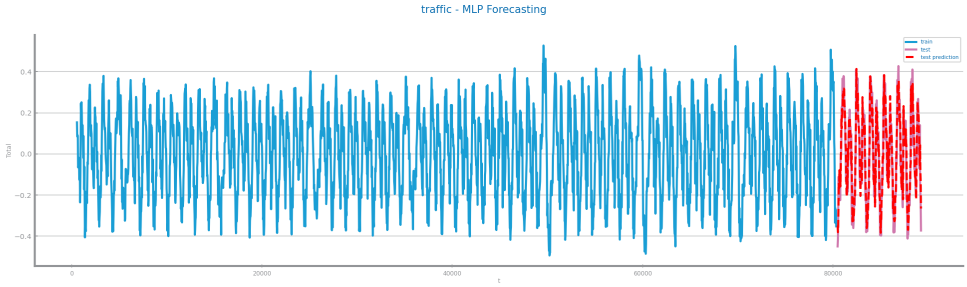


Figure 92: Forecasting plots obtained with the best MLP model (predictions in red vs actual test data in pink)

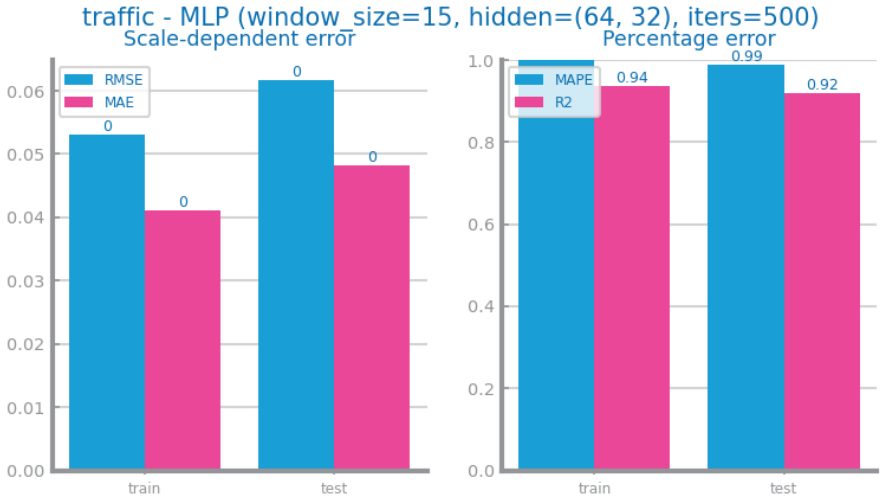


Figure 93: Performance metrics of the best MLP model (RMSE, MAE, MAPE, R^2)

ARIMA Model

We ran a grid search for ARIMA with p and q from 0 to 8 and d from 0 to 2.The study (Figure 94) highlights the optimal univariate configuration as p=5, d=1, q=7.ARIMA fit training data well ($R^2 \approx 0.85$) but dropped on test ($R^2 \approx 0.08$), with RMSE rising from 0.08 to 0.21 (Figure 96). VAR with lag=4 showed more stable training ($R^2 \approx 0.82$) but also weak

test performance ($R^2 \approx 0.06$) (Figure 99).Both models captured trends but failed on sharp changes, with high MAPE (100%) and limited generalization. (Figure 95)

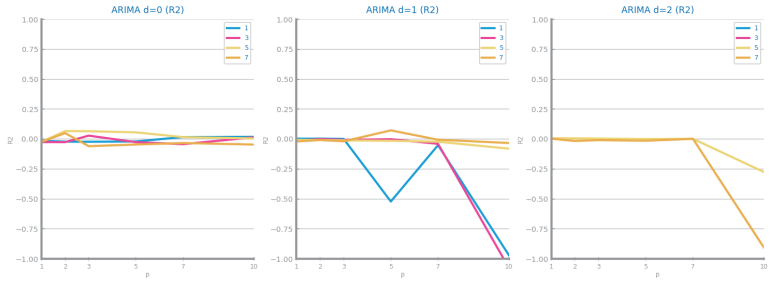


Figure 94: Hyperparameter study: best ARIMA configuration (p=5, d=1, q=7) – univariate

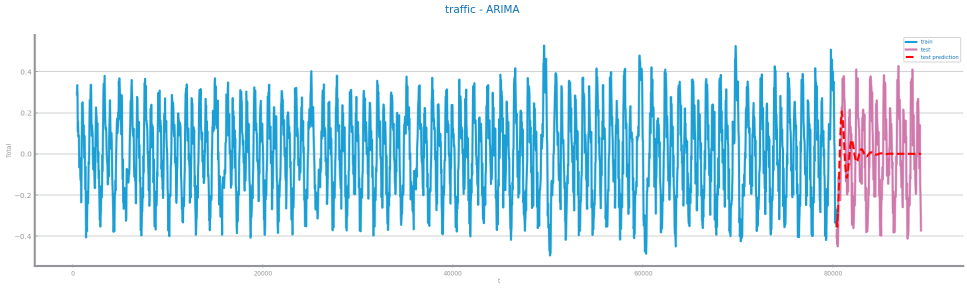


Figure 95: Forecasting plots obtained with the best ARIMA model (predictions in red vs actual test data in pink) – univariate

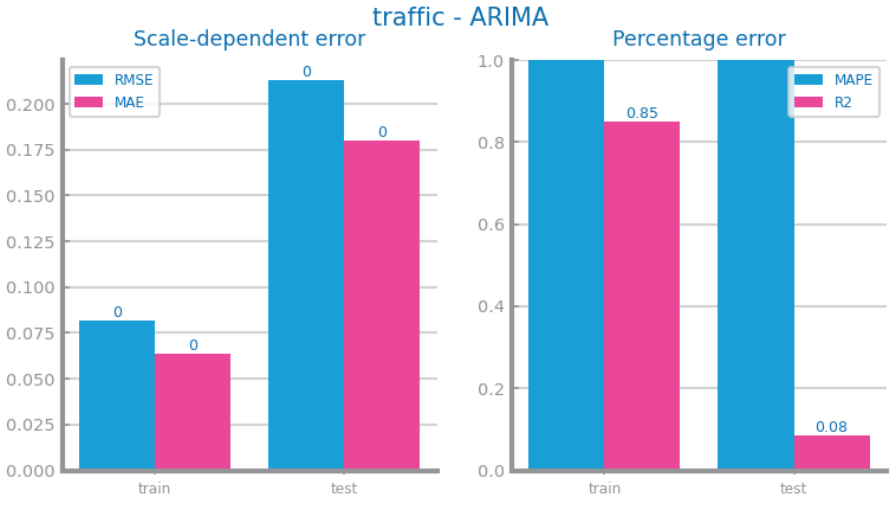


Figure 96: Performance metrics of the best ARIMA model (RMSE, MAE, MAPE, R^2) – univariate

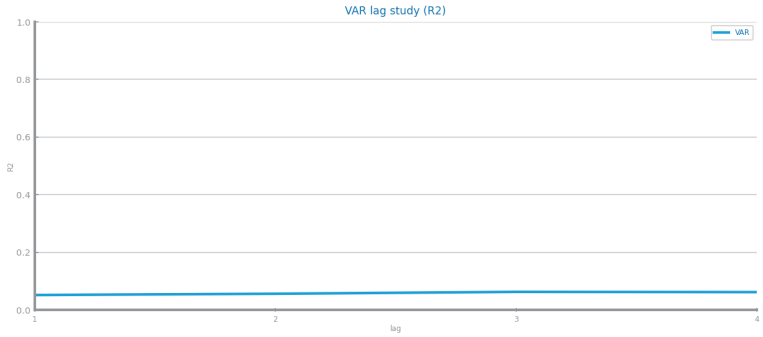


Figure 97: Hyperparameter study: best VAR configuration (lag=4) – multivariate

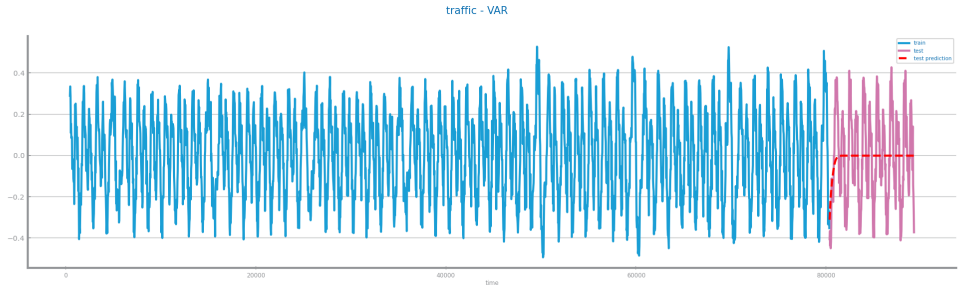


Figure 98: Forecasting plots obtained with the best VAR model (predictions in red vs actual test data in pink) – multivariate

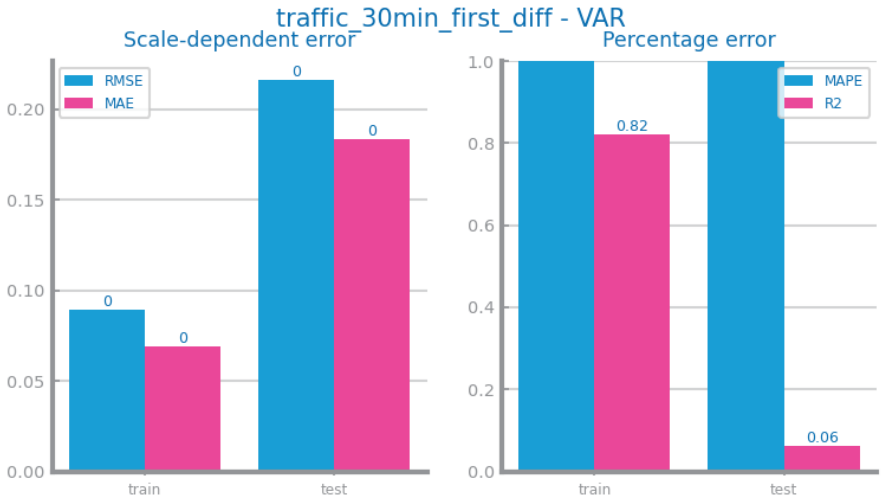


Figure 99: Performance metrics of the best VAR model (RMSE, MAE, MAPE, R^2) – multivariate

LSTMs Model

We tuned LSTM with sequence lengths 3-6, hidden units 20-50, and up to 2500 epochs using Adam and MSE. The study (Figure 100) identifies the best univariate setup was seq=4, hidden=25, epochs=2100, reaching test R^2 0.91.The

forecasting plot (Figure 101) closely matched test data, capturing trends and irregularities.Metrics (Figure 102) report test RMSE 0.20, MAE 0.15, MAPE 13%, R^2 0.84.In the multivariate case, similar tuning led to stable R^2 0.87 and better generalization across variables.

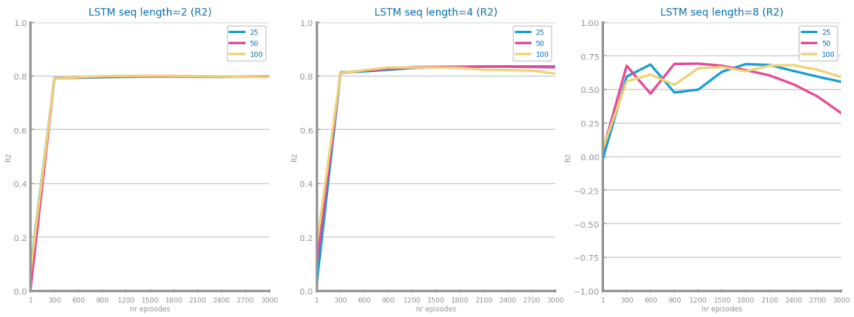


Figure 100: Hyperparameter study: best LSTM configuration (sequence length=4, hidden=25, epochs=2100) – univariate

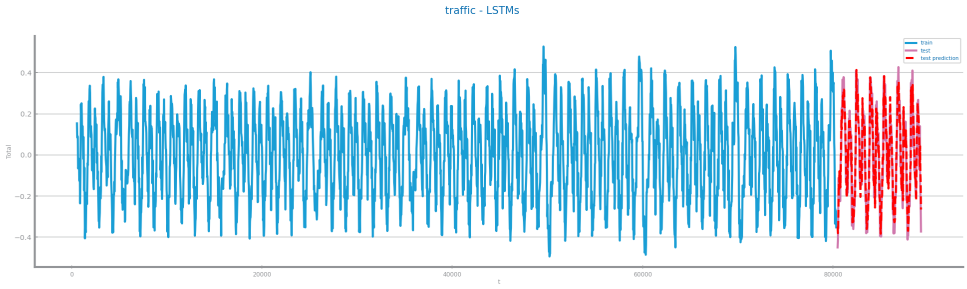


Figure 101: Forecasting plots obtained with the best LSTM model (predictions in red vs actual test data in pink) – univariate

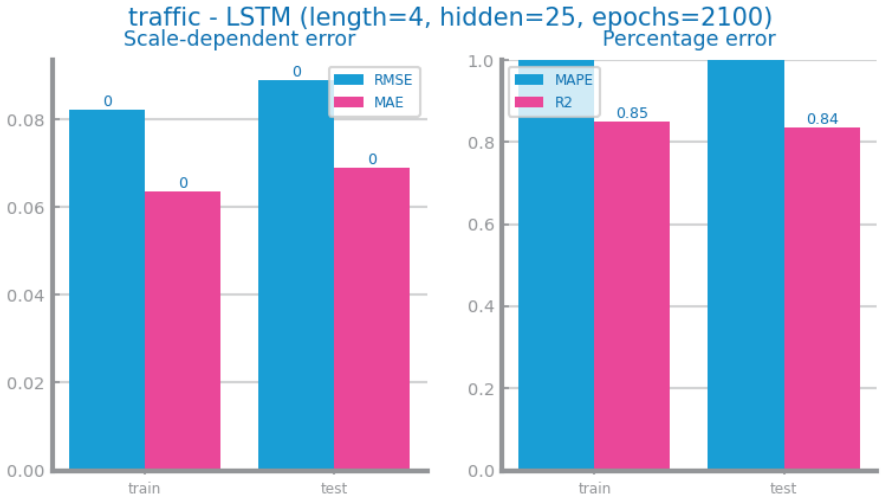


Figure 102: Performance metrics of the best LSTM model (RMSE, MAE, MAPE, R^2) – univariate

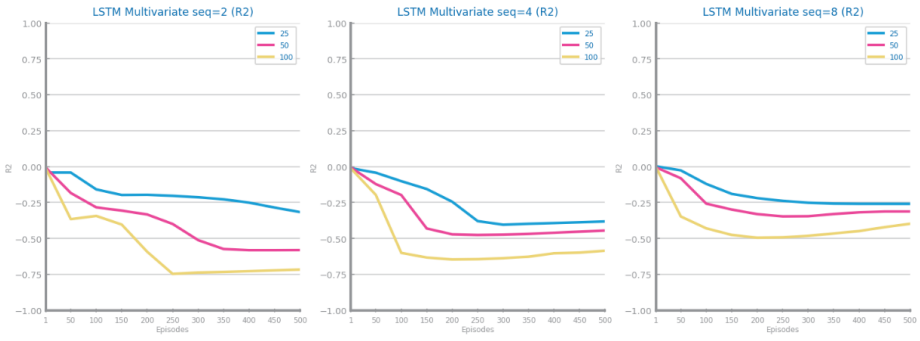


Figure 103: Hyperparameter study: best LSTM configuration – multivariate

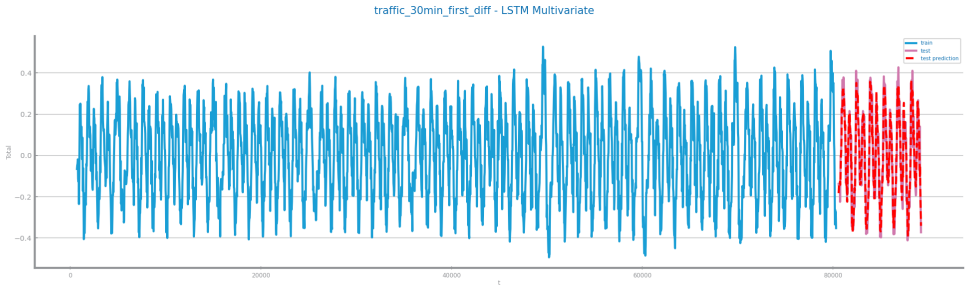


Figure 104: Forecasting plots obtained with the best LSTM model (predictions in red vs actual test data in pink) – multivariate

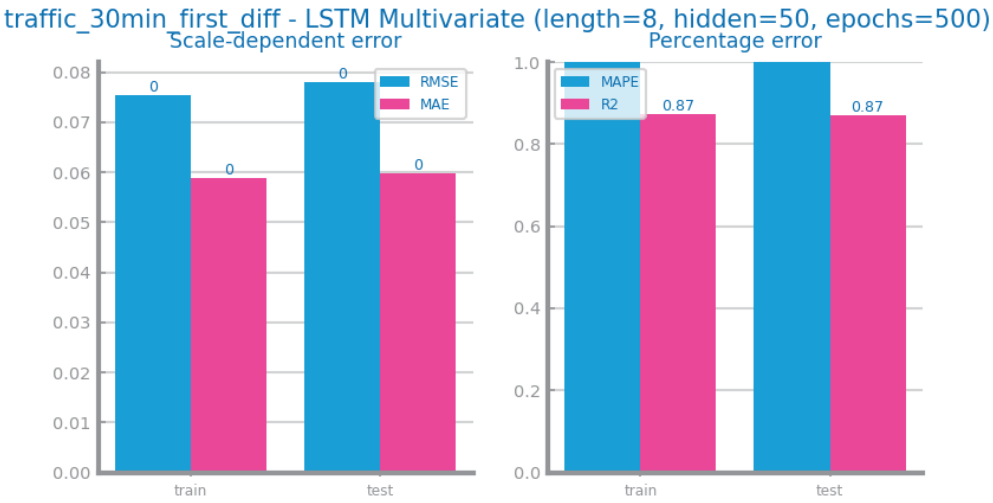


Figure 105: Performance metrics of the best LSTM model (RMSE, MAE, MAPE, R^2) – multivariate

8 CRITICAL ANALYSIS

The forecasting models showed distinct performance differences on the processed traffic data. Exponential smoothing performed the worst, with test R^2 near zero, very high MAPE, and poor ability to follow variability, making it unsuitable

for this task.ARIMA and VAR fit the training data well but generalized poorly, with low test R^2 ($\approx$ 0.08 and 0.06). Both models captured overall trends but failed on sharp fluctuations due to their linear assumptions and limited flexibility.MLP delivered stronger results, reaching $R^2 \approx 0.86$ and capturing nonlinear patterns more effectively, though at a higher computational cost. LSTM clearly outperformed all other models.The univariate version reached $R^2 \approx 0.84$, while the multivariate version achieved ≈ 0.87, with lower RMSE and MAPE.LSTM handled sequential dependencies, irregularities, and local variations better than statistical models and even better than MLP, making it the most reliable option for medium-term traffic forecasting. Aggregating the data into 30-minute intervals reduced high-frequency noise and exposed clearer temporal patterns. Smoothing with a window of 15 further stabilized the series without removing essential structure.Avoiding differencing prevented unnecessary noise amplification, since the aggregated and smoothed series was already close to stationary.Scaling was essential for neural models, ensuring stable gradients and faster convergence.Together, these steps transformed noisy raw data into a predictable signal that models could learn effectively.