TÉCNICO LISBOA

DS 2025

Data Science Project

Team nr: 26	**Student 1:**	Alexandre Cachera	**IST nr:** 116285
	Student 2:	Jaime Gosai	**IST nr:** 99239
	Student 3:	Fredrik Preus Dovland	**IST nr:** 116071
	Student 4:	Lukas Bruns	**IST nr:** 116926

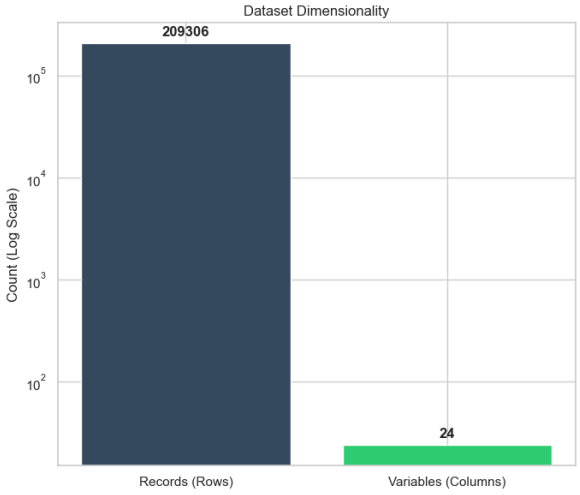
CLASSIFICATION

1 DATA PROFILING

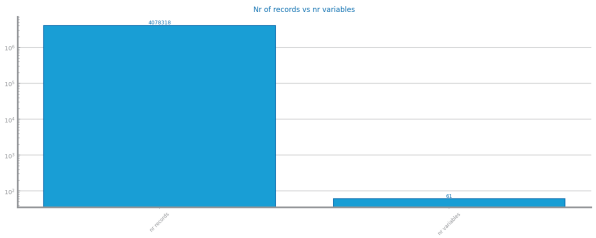
Identifying data leakage in both datasets leads us to remove post-events variables. This ensures the model only uses pre-crash data, preventing artificial performance inflation.

Data Dimensionality

Both datasets show high record-to-variable ratios, supporting model stability. Dataset 1 (Traffic security) contains many categorical variables and implicit missing values (e.g., "UNKNOWN"), requiring encoding and specialized cleaning. Dataset 2 (Flights) is mostly numerical but includes categorical features and 3% missing values in several variables.

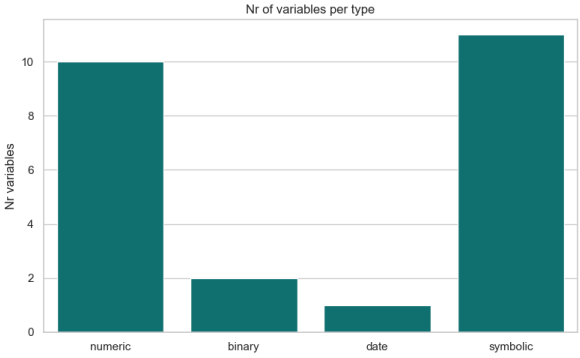


(a) Dataset 1

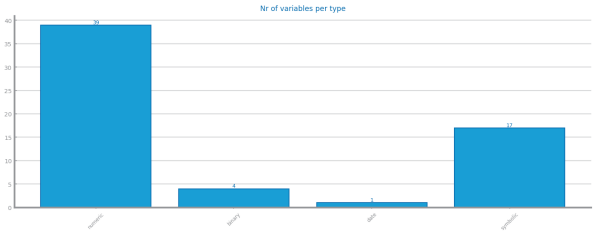


(b) Dataset 2

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

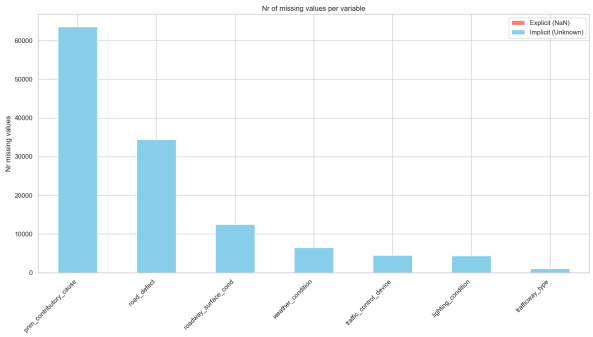


(a) Dataset 1

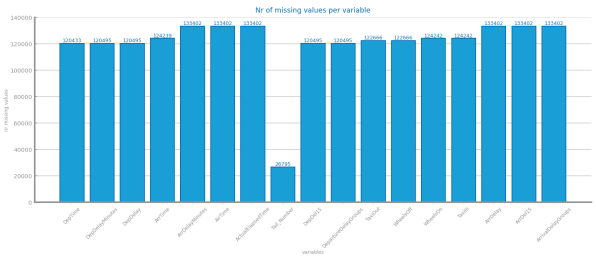


(b) Dataset 2

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



(a) Dataset 1

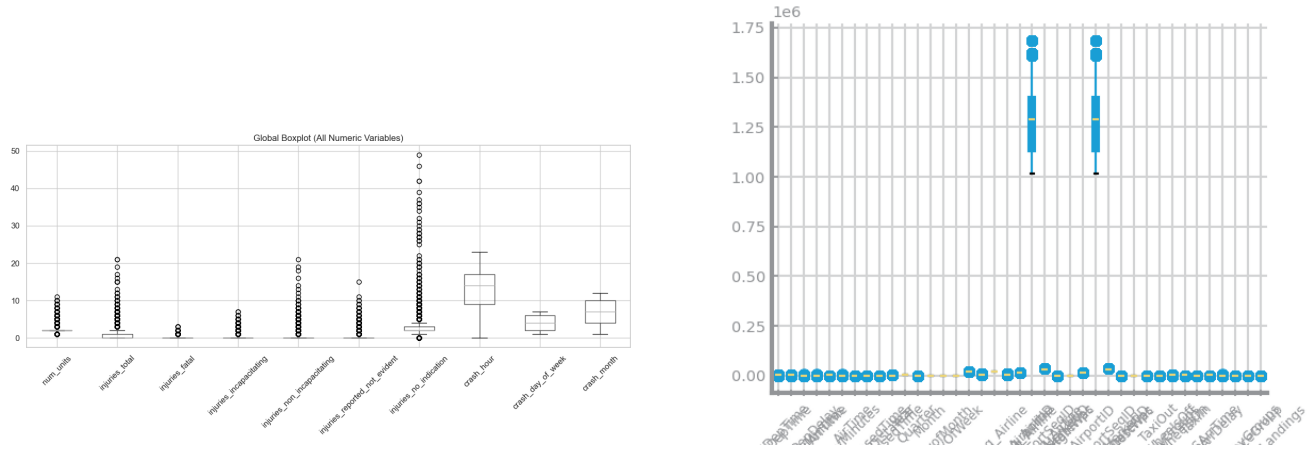


(b) Dataset 2

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

Data Distribution

Shall contain all relevant information and charts respecting to the data distribution perspective, such as each variable distribution, type, domain and range. May be used to describe any useful observation about the data, and that was used in the current project. **Shall not exceed 500 characters.**



(a) Dataset 1

(b) Dataset 2

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

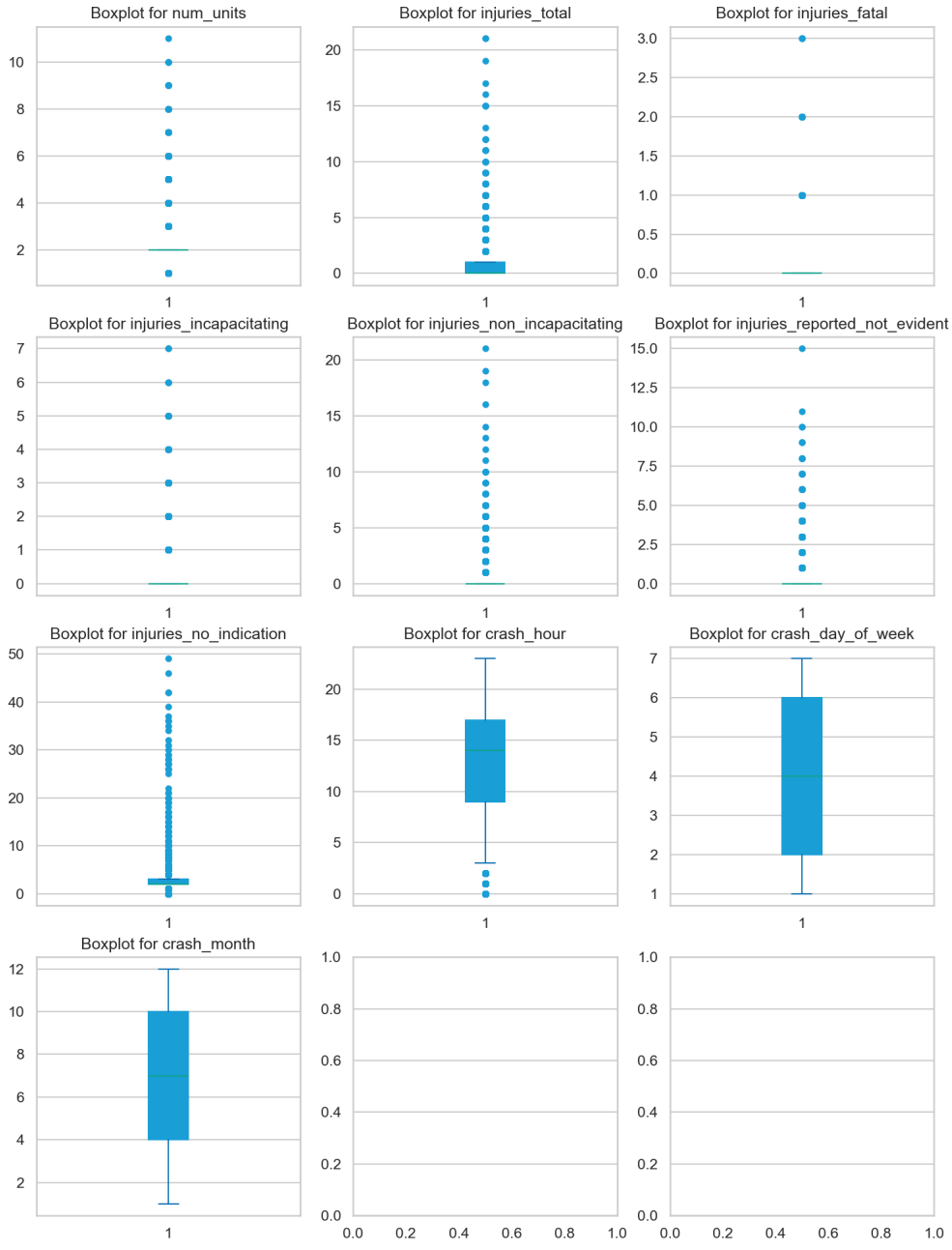


Figure 5: Single variables boxplots for dataset 2

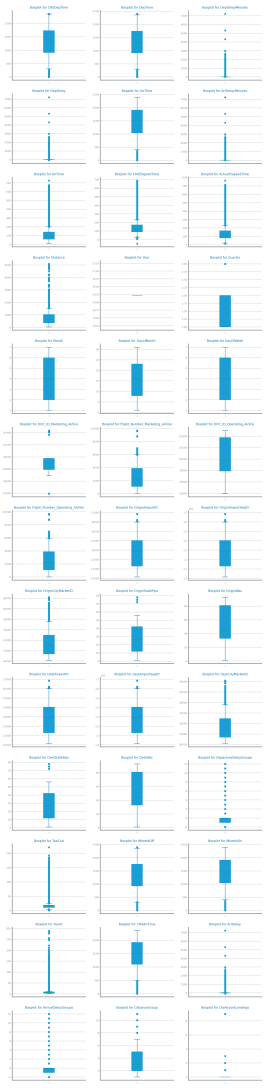


Figure 6: Single variables boxplots for dataset 2

Figure 7: Histograms for dataset 1

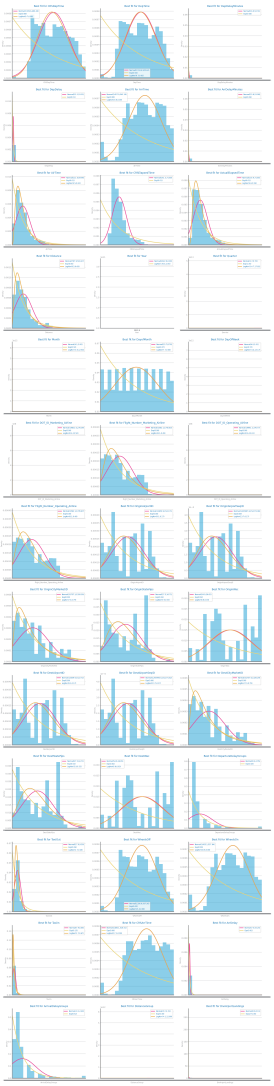


Figure 8: Histograms for dataset 2

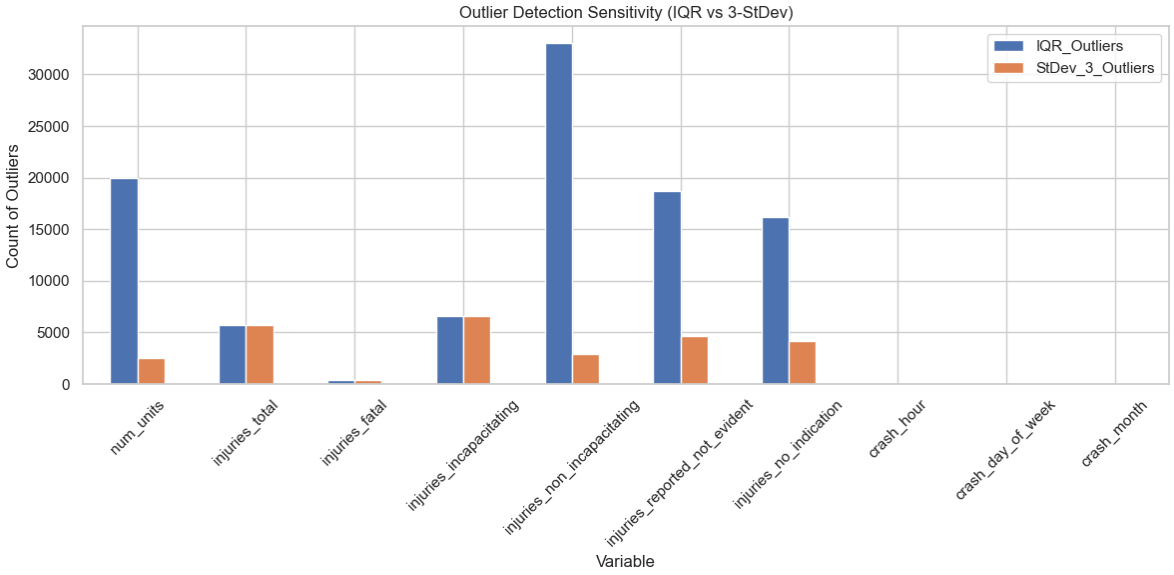


Figure 9: Outliers study dataset 1

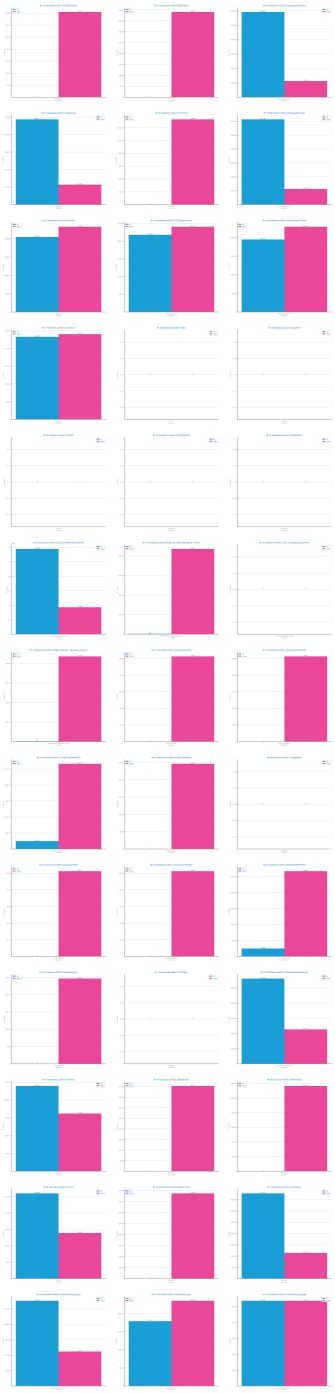


Figure 10: Outliers study dataset 2

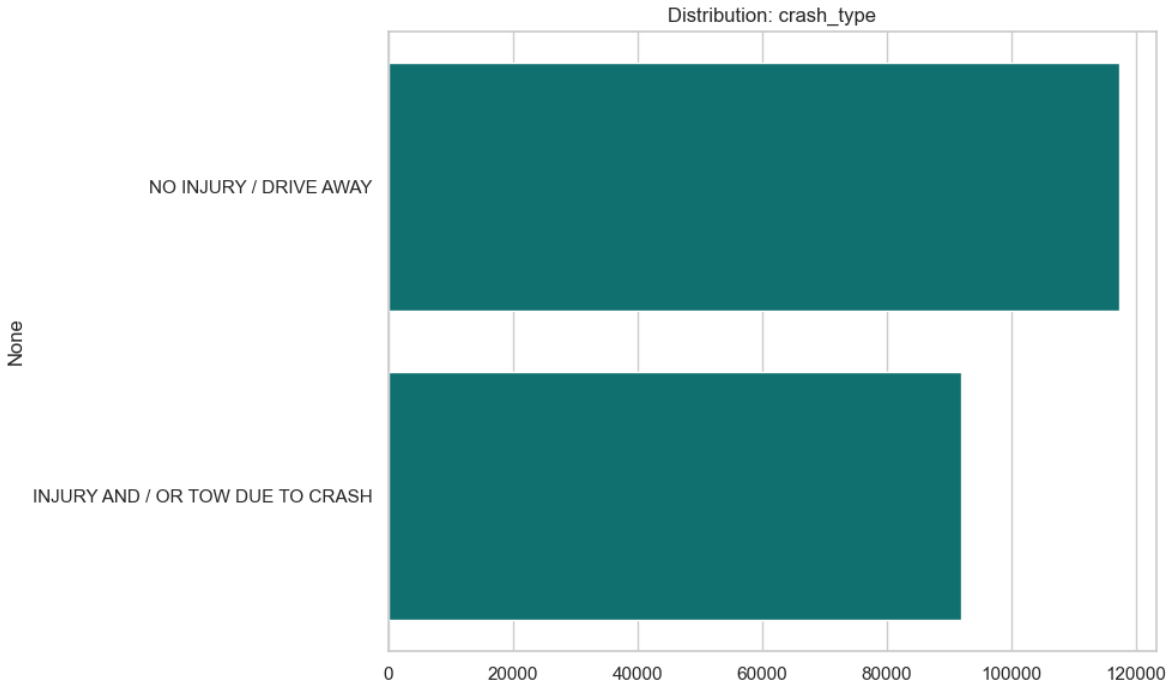


Figure 11: Class distribution for dataset 1

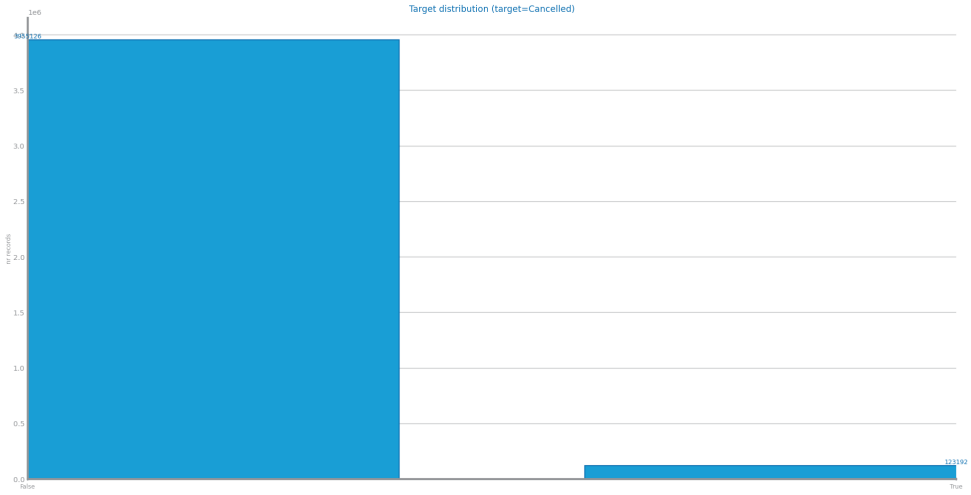


Figure 12: Class distribution for dataset 2

Data Granularity

Shall contain all relevant information and charts respecting to the data granularity perspective, such as the impact of different granularities considered for each variable. May present additional taxonomies if needed. **Shall not exceed 500 characters.**

Figure 13: Granularity analysis for dataset 1

Figure 14: Granularity analysis for dataset 2

Data Sparsity

Shall contain all relevant information and charts respecting to the data sparsity perspective, such as domain coverage and correlation among variables. **Shall not exceed 500 characters.**

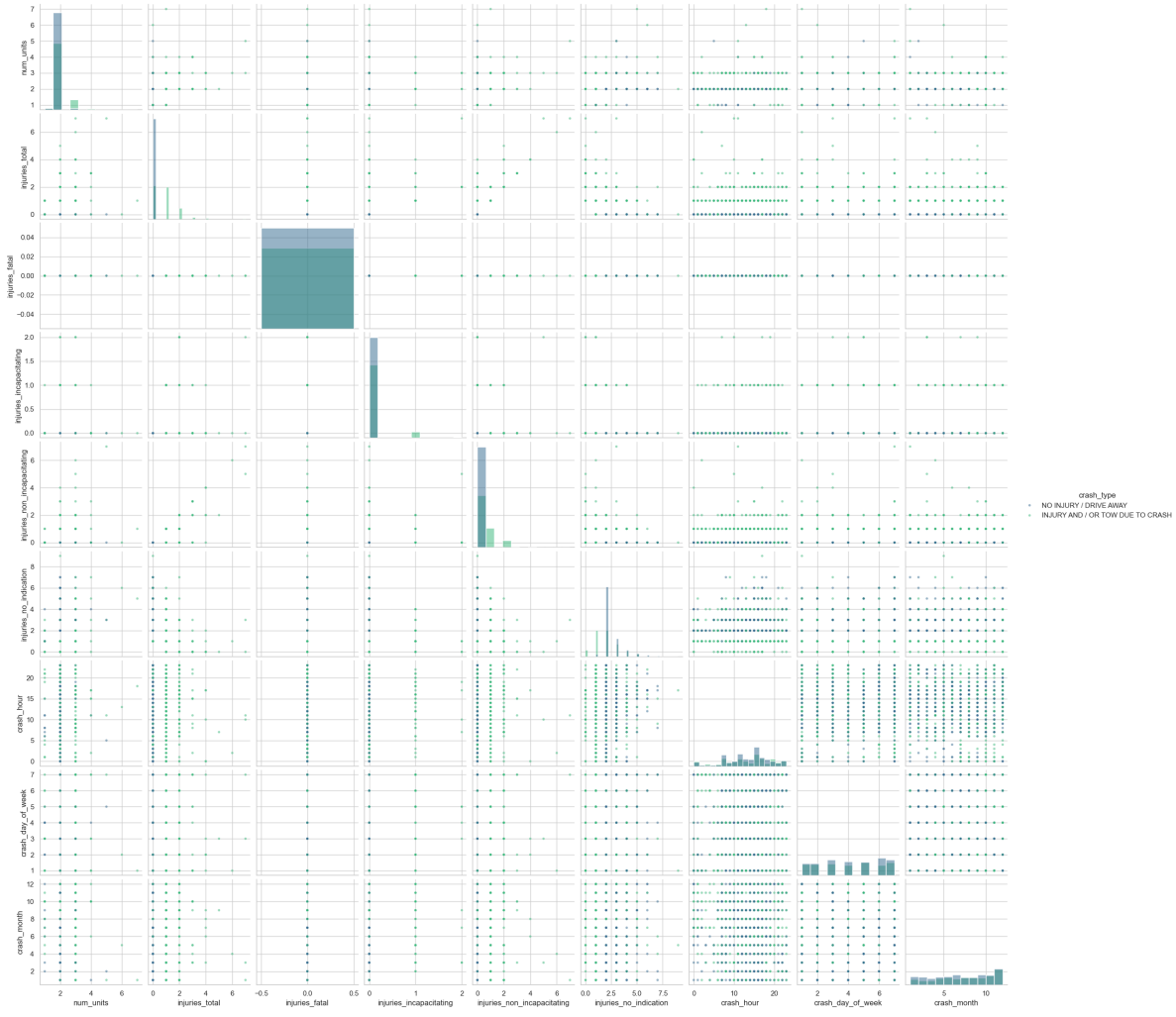


Figure 15: Sparsity analysis for dataset 1

Figure 16: Sparsity analysis for dataset 2

Figure 17: Correlation analysis for dataset 1

Figure 18: Correlation analysis for dataset 2

2 DATA PREPARATION

For the second dataset, we utilized a subsampling technique (200k rows) for the experimental phase. This allowed for rapid iteration and "winner picking" of pipeline steps (scaling, outliers, etc.). The final optimized pipeline is then applied to the full dataset. For evaluation and decision of the best approach, we considered the metric 'balanced accuracy' and the average between results of the two models.

Variables Encoding

Dataset 1 transformations include Binary Mapping for 'intersection related i' and Cyclical Encoding (sine/cosine) for 'crash hour', 'crash month', and 'crash day of week'. Ordinal Encoding was applied to 'damage' and 'most severe injury'

to preserve their inherent hierarchy. Infrequent labels in categorical variables were grouped to mitigate sparsity found during distribution profiling. All remaining categorical variables underwent One-Hot Encoding.

Dataset 2 used Target Encoding for 'Hub Airline', 'Route', 'Airline', 'Origin', and 'Dest' to handle high cardinality effectively. Cyclical Encoding (sine/cosine) was applied to 'Month', 'DayOfWeek', 'DayofMonth', 'CRSDepTime', and 'CRSArrTime' to preserve periodic continuity. One-Hot Encoding was not applied to symbolic variables because the excessive unique values in routes and airports would have created an unmanageable, high-dimensional, and sparse feature space.

Missing Value Imputation

Dataset 1 treats "Unknown" labels as informative features (e.g., signaling hit-and-runs) rather than missing data; thus, no imputation was performed. Dataset 2 required no imputation as pre-flight features remained 100% complete after leakage removal. By bypassing statistical imputation, we avoided synthetic bias, relying on high density and informative categories. This approach ensures data integrity for modeling while respecting observed characteristics.

Additional Feature Generation

New features were logically engineered to provide deeper signals for the target variables. Their ultimate relevance and predictive utility are assessed during the subsequent Feature Selection phase.

Feature	**Formula**
risk x units	prim contributory cause x num units
ped intersection	Pedestrian/Cyclist crashes × Intersection
risk intersection	prim contributory cause * intersection related i
wet dark	poor surface (wet) x poor lighting (dark)
rear end multi	rear end x num units

Table 1: Features created for dataset 1

Feature	Formula
Speed proxy	Distance / (CRSElapsedTime + 1)
Log distance	log(1 + distance)
Time Bin	Converts the raw Sine/Cosine time features back into 4 interpretable buckets: Night, Morning, Afternoon, and Evening
is winter	cosine value of month cose is > 0.5
hub x dest	hub airline * dest

Table 2: Features created for dataset 2

Outliers Treatment

Both datasets were evaluated using Isolation Forest and Z-score. For Dataset 1, Isolation Forest was selected as it yielded superior modeling results by effectively isolating crash-related anomalies. Conversely, Z-score proved more effective for Dataset 2, better handling its specific feature distributions. We proceeded with these optimal choices—Isolation Forest for D1 and Z-score for D2—to reduce noise and enhance overall model robustness.

	Isolation Forest		**Z-Score**	
	KNN	NB	KNN	NB
Accuracy	0.6195	0.6932	0.6197	0.6860
Recall	0.5093	0.5683	0.5110	0.4859
F1	0.5404	0.6194	0.5414	0.5762
Balanced Accuracy	0.6076	0.6797	0.6080	0.6643

Table 3: Outliers imputation results with different approaches for dataset 1

	Replace		**Truncate**	
	KNN	NB	KNN	NB
Accuracy				
Recall				
F1				
Balanced Accuracy				

Table 4: Outliers imputation results with different approaches for dataset 2

Scaling

Compared Z-score and Min-Max via KNN. Z-score won for D1; Min-Max for D2. These were selected to optimize feature scales, resulting in superior modeling performance for both datasets.

	Standardization		**Normalization**	
	KNN	NB	KNN	NB
Accuracy	0.6589	0.6926	0.6739	0.6926
Recall	0.5805	0.5759	0.5936	0.0.5759
F1	0.5992	0.6220	0.6152	0.6220
Balanced Accuracy	0.6504	0.6799	0.6652	0.6799

Table 5: Scaling results with different approaches for dataset 1

	Standardization		**Normalization**	
	KNN	NB	KNN	NB
Accuracy				
Recall				
F1				
Balanced Accuracy				

Table 6: Scaling 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.**

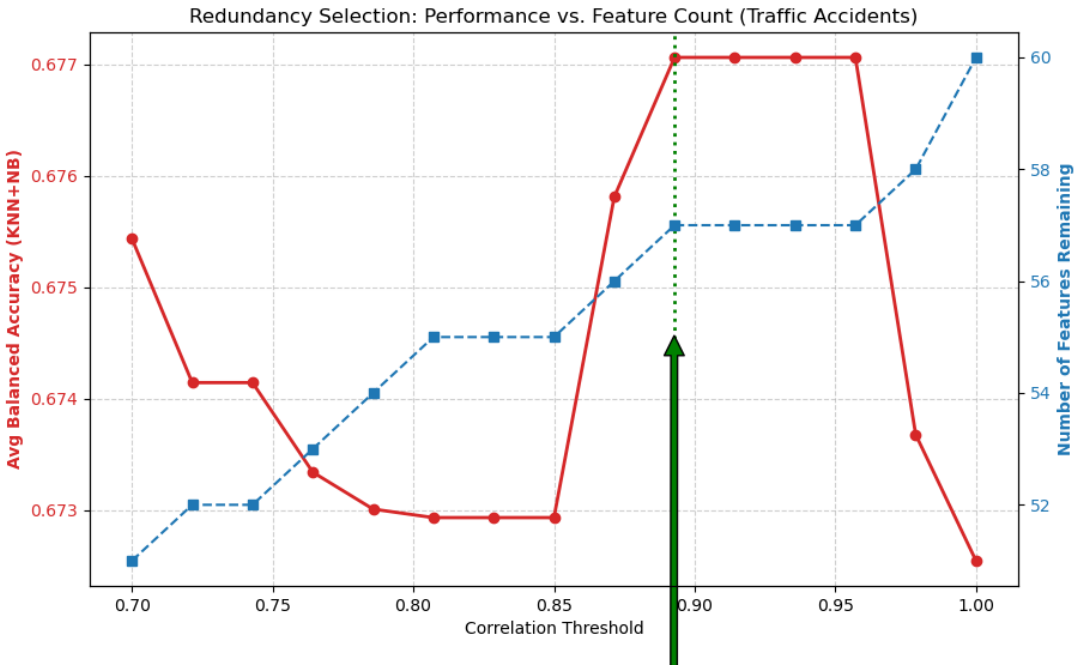


Figure 19: Feature selection of redundant variables results with different parameters for dataset 1

Figure 20: Feature selection of redundant variables results with different parameters for dataset 2

	Chi 2		**Random Forest**	
	KNN	NB	KNN	NB
Accuracy	0.6725	0.6894	0.6746	0.7021
Recall	0.5976	0.7313	0.6012	0.6479
F1	0.6158	0.6741	0.6741	0.6564
Balanced Accuracy	0.66643	0.6939	0.6666	0.6963

Table 7: Feature selection for relevant variables results for dataset 1

	Chi 2		Random Forest	
	KNN	NB	KNN	NB
Accuracy	0.6725	0.6894	0.6746	0.7021
Recall	0.5976	0.7313	0.6012	0.6479
F1	0.6158	0.6741	0.6741	0.6564
Balanced Accuracy	0.66643	0.6939	0.6666	0.6963

Table 8: Feature selection for relevant variables results for dataset 2

Balancing

Undersampling and SMOTE were evaluated as balancing alternatives. Dataset 1 showed no improvement and remained in its original state, likely because its initial distribution was already sufficiently balanced. For Dataset 2, Undersampling was selected as it yielded the highest performance gains over SMOTE and the baseline. This choice resolves the majority class bias identified in Dataset 2 without introducing synthetic noise in Dataset 1.

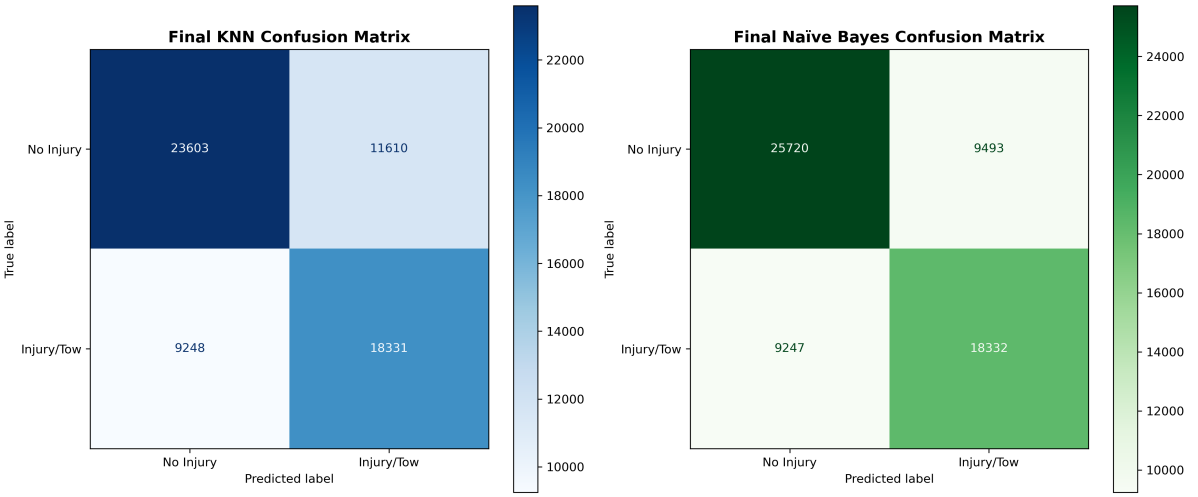
	SMOTE		Undersampling	
	KNN	NB	KNN	NB
Accuracy	0.6746	0.7021	0.6634	0.7016
Recall	0.6012	0.6479	0.6601	0.6662
F1	0.6187	0.6564	0.6327	0.6623
Balanced Accuracy	0.6666	0.6963	0.6631	0.6978

Table 9: Balancing results with different approaches for dataset 1

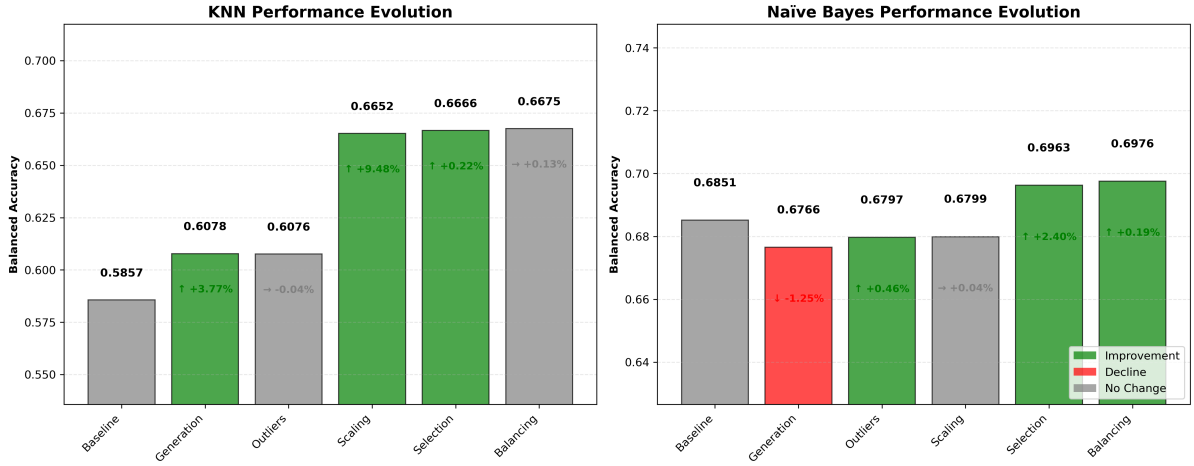
	SMOTE		Undersampling	
	KNN	NB	KNN	NB
Accuracy				
Recall				
F1				
Balanced Accuracy				

Table 10: Balancing results with different approaches for dataset 2

Summary of the pipeline

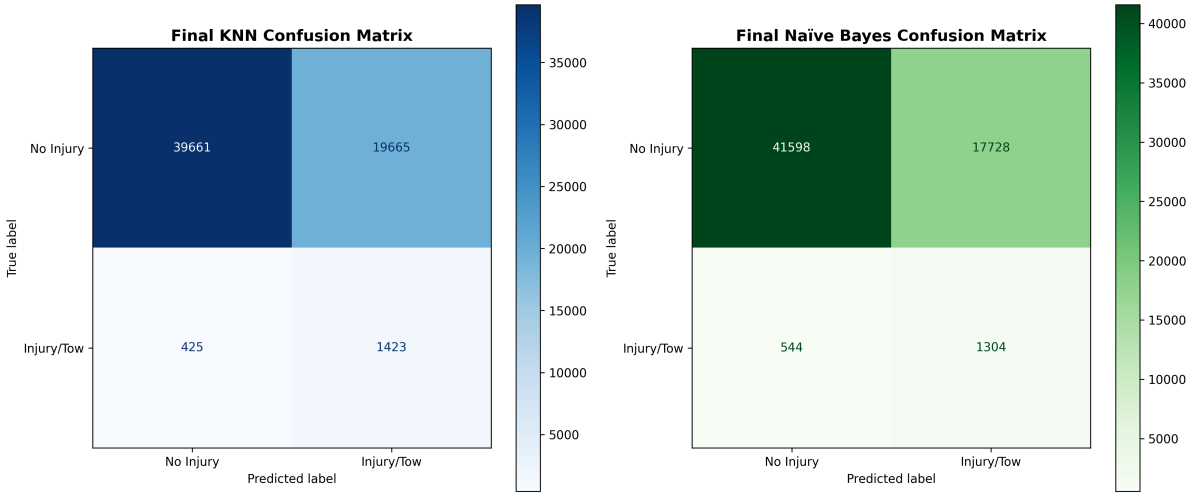


(a) Confusion Matrix

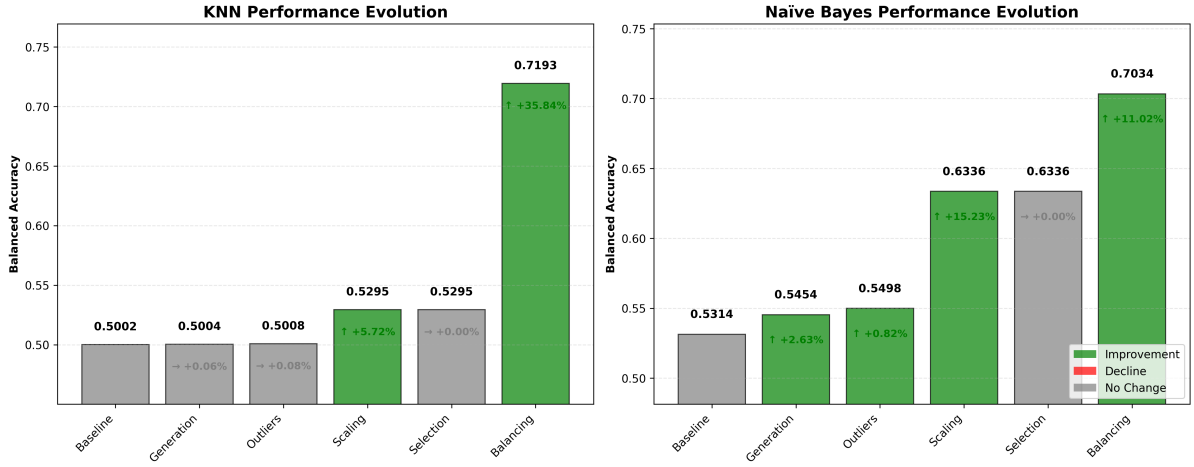


(b) Performance evolution

Figure 21: Summary of the pipeline for dataset 1



(a) Confusion Matrix



(b) Performance evolution

Figure 22: Summary of the pipeline 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 23: Naïve Bayes alternatives comparison for dataset 1

Figure 24: Naïve Bayes alternative comparison for dataset 2

Figure 25: 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 26: KNN different parameterisations comparison for dataset 1

Figure 27: KNN different parameterisations comparison for dataset 2

Figure 28: KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 29: KNN best model results for dataset 1 (left) and dataset 2 (right)

Decision Trees

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

Figure 30: Decision Trees different parameterisations comparison for dataset 1

Figure 31: Decision Trees different parameterisations comparison for dataset 2

Figure 32: Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 33: Decision trees best model results for dataset 1 (left) and dataset 2 (right)

Figure 34: Best tree for dataset 1

Figure 35: 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 36: Random Forests different parameterisations comparison for dataset 1

Figure 37: Random Forests different parameterisations comparison for dataset 2

Figure 38: Random Forests overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 39: Random Forests best model results for dataset 1 (left) and dataset 2 (right)

Figure 40: 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 41: Gradient boosting different parameterisations comparison for dataset 1

Figure 42: Gradient boosting different parameterisations comparison for dataset 2

Figure 43: Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 44: Gradient boosting best model results for dataset 1 (left) and dataset 2 (right)

Figure 45: 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 46: MLP different parameterisations comparison for dataset 1

Figure 47: MLP different parameterisations comparison for dataset 2

Figure 48: MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 49: Loss curve analysis for dataset 1 (left) and dataset 2 (right)

Figure 50: 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.**

5 DEPLOYMENT

deployment part

TIME SERIES ANALYSIS

6 DATA PROFILING

Data Dimensionality and Granularity

Atomic granularity is set at 15-minute intervals, capturing high-frequency traffic counts. To identify recurring patterns, we explored hourly and daily granularities, which revealed distinct rush-hour peaks and commuting cycles. Weekly aggregation was also analyzed to distinguish between weekday and weekend behaviors.

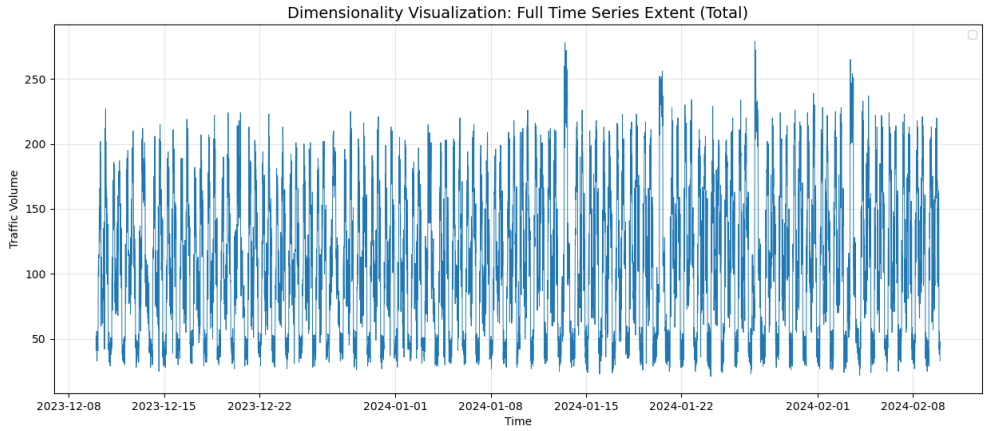


Figure 51: Time series 1 at the most granular detail

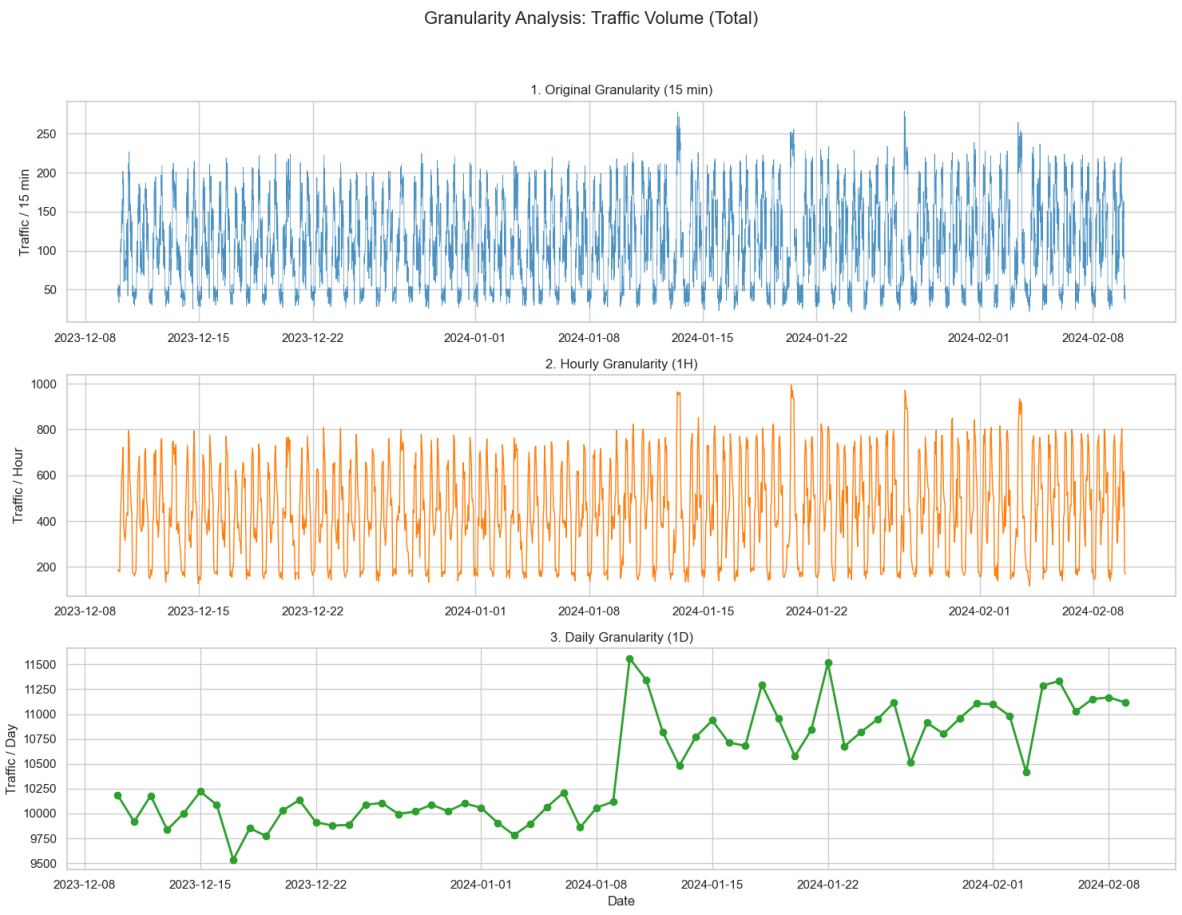


Figure 52: Time series 1 granularities

Data Distribution

Analysis across granularities shows multimodal distributions in Original (O) and Hourly (H) scales, representing distinct traffic regimes (peak/off-peak). Daily (D) aggregation shifts volume to higher clusters. No significant outliers are present, as shown by the continuous KDE and tight distributions. Strong positive autocorrelation and linear lag plots for the atomic series confirm high temporal dependency; past volume strongly influences future values, justifying a forecasting approach.

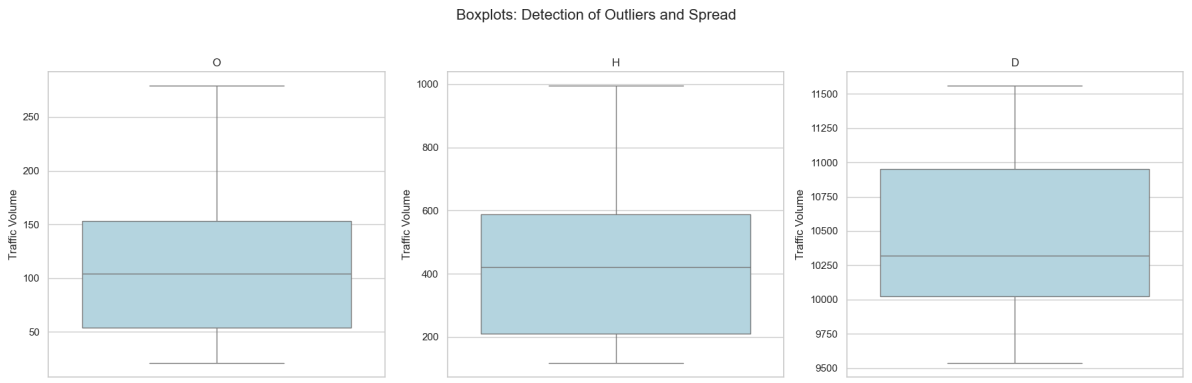


Figure 53: Boxplot(s) for time series 1

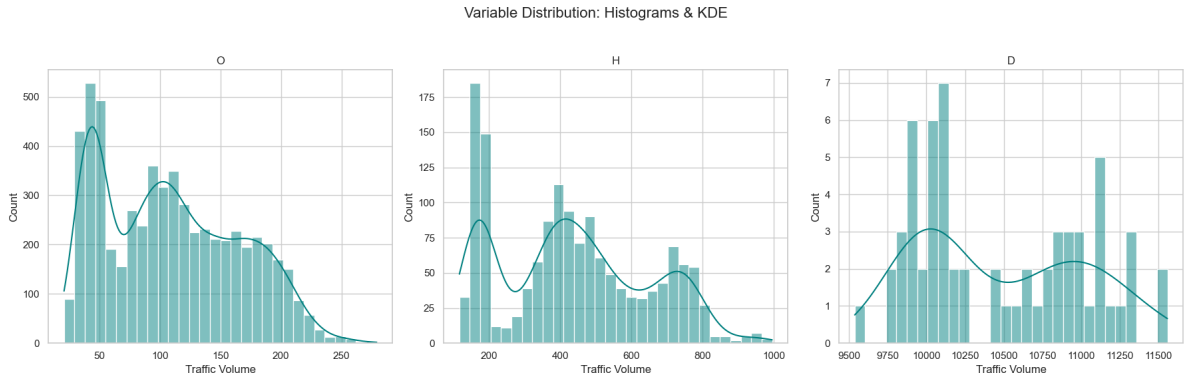


Figure 54: Histogram(s) for time series 1

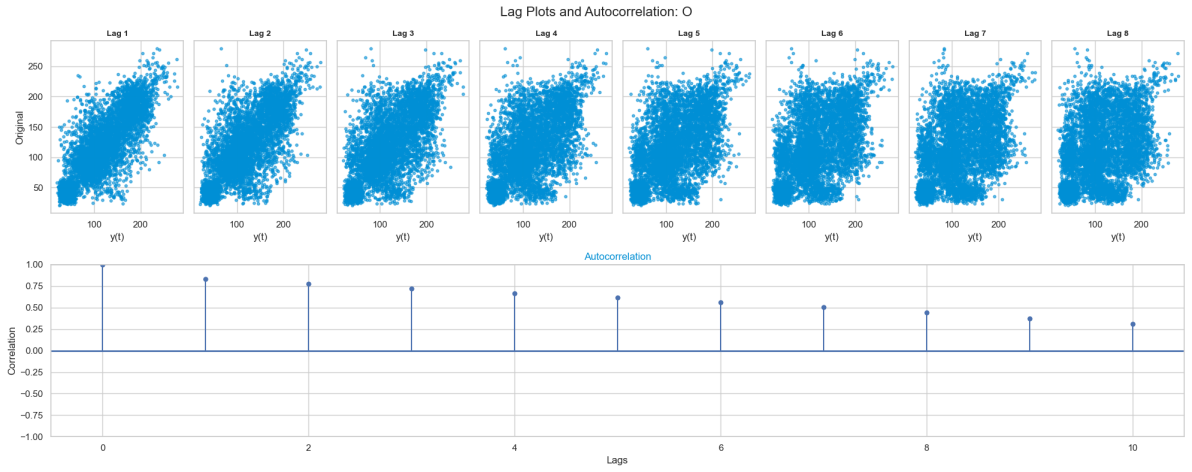
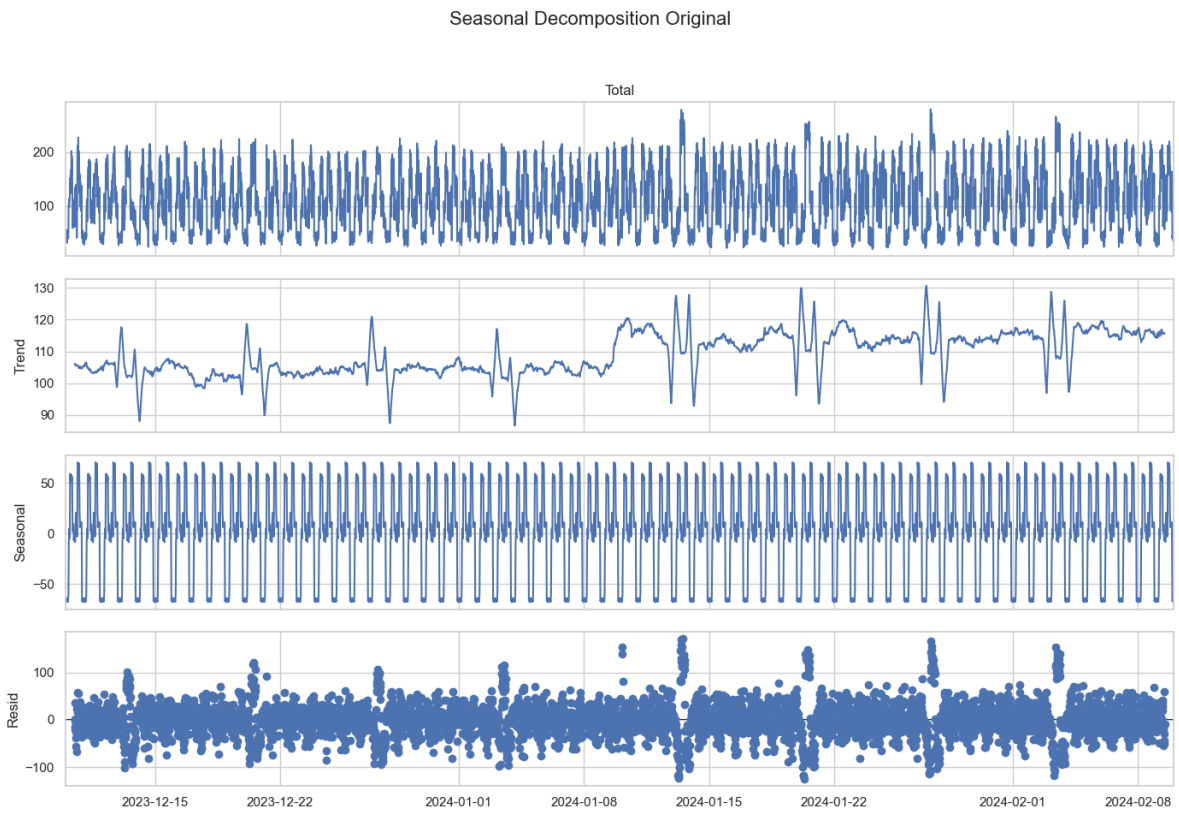


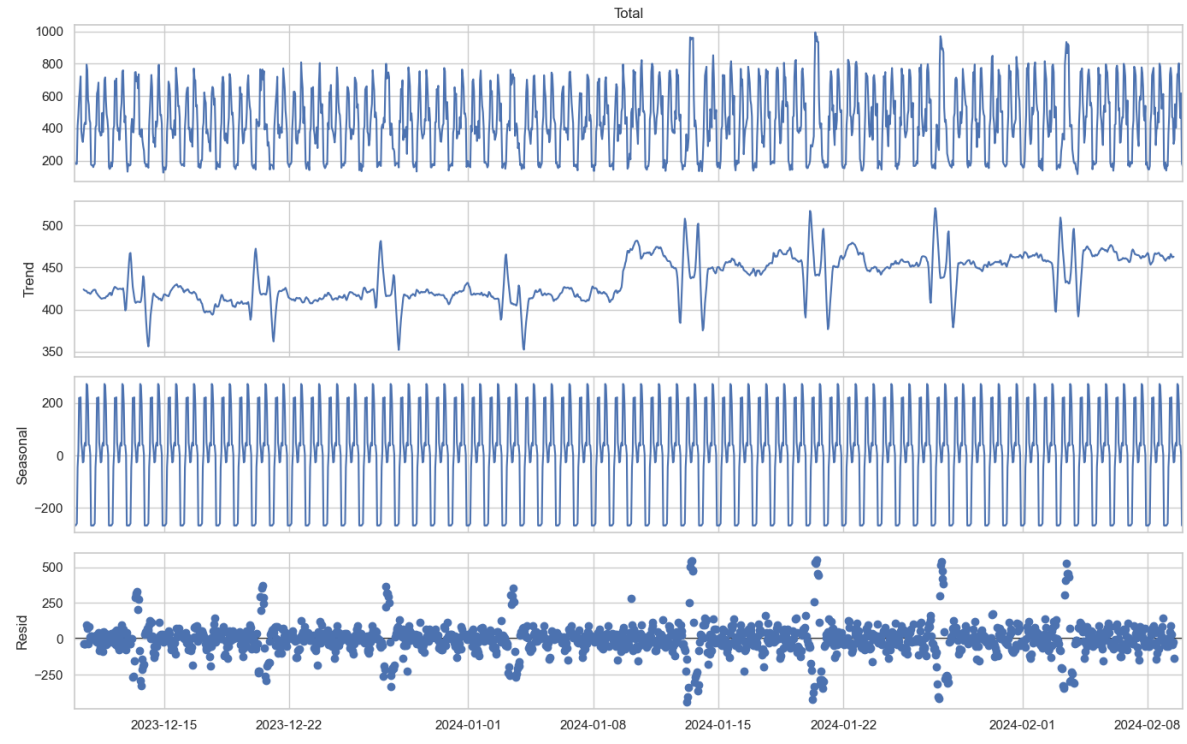
Figure 55: Autocorrelation lag-plots and correlogram for original time series 1

Data Stationarity

ADF tests confirm stationarity for Original and Hourly series (p-value 0.000), while the Daily series is non-stationary (p-value 0.811) due to a late-period trend shift. Decompositions reveal consistent seasonality across scales, though residual noise increases at finer granularities.

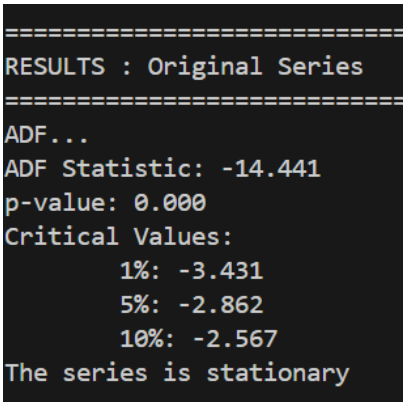


(a) Original

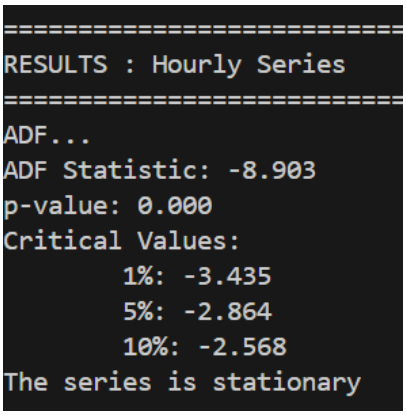


(b) Hourly

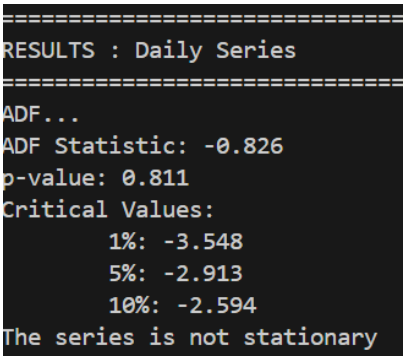
Seasonal Decomposition Daily



(a) Original



(b) Hourly



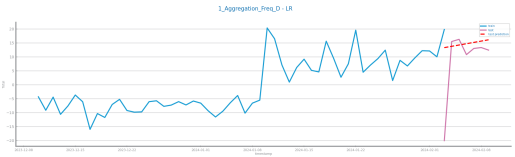
(c) Daily

Figure 57: Stationarity study for time series 1

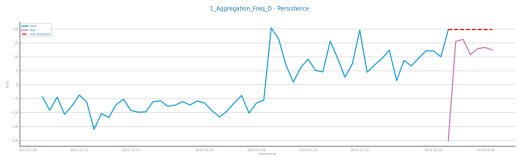
7 DATA TRANSFORMATION

Aggregation

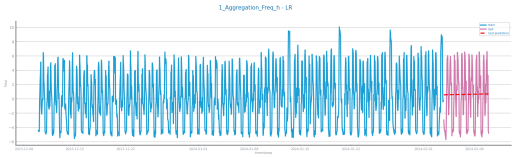
Shall describe the results of applying three different aggregations over both datasets, and identifying the granularity chosen to proceed. **Shall not exceed 300 characters.**



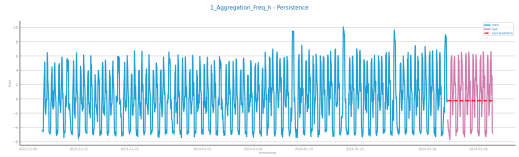
(a) Daily - LR



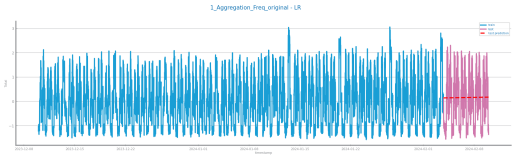
(b) Daily - Persistence



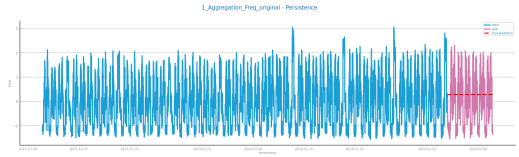
(c) Hourly - LR



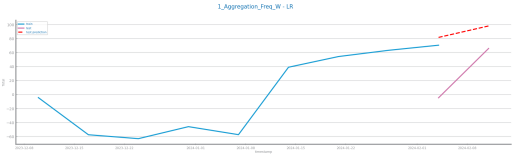
(d) Hourly - Persistence



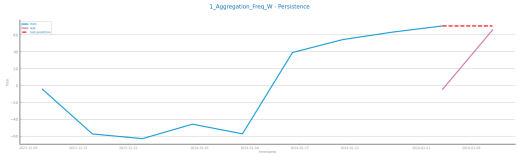
(e) Original - LR



(f) Original - Persistence

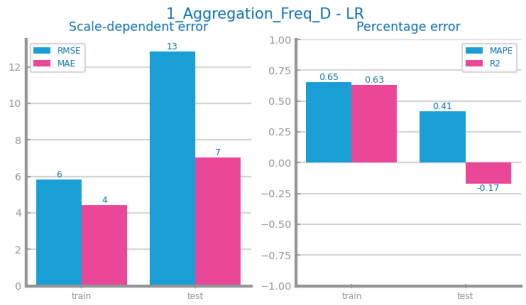


(g) Weekly - LR

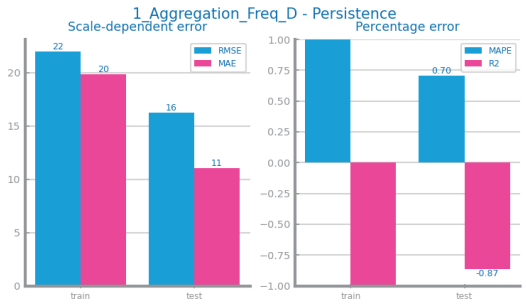


(h) Weekly - Persistence

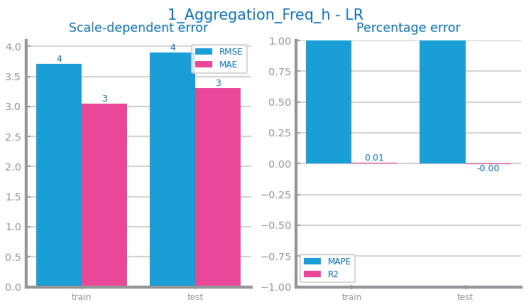
Figure 58: Forecasting plots after different aggregations on time series 1



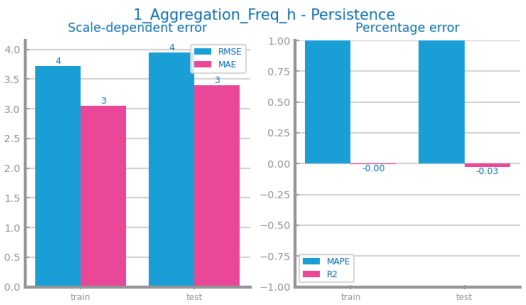
(a) Daily - LR



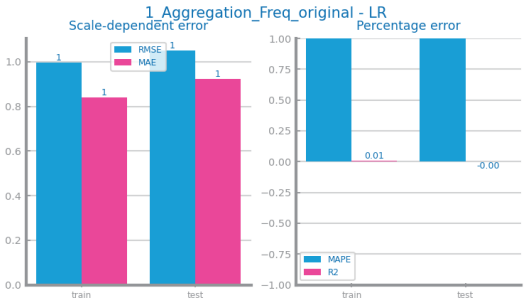
(b) Daily - Persistence



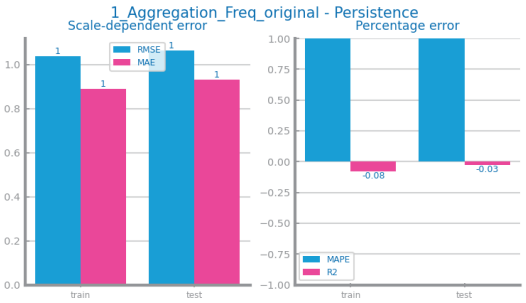
(c) Hourly - LR



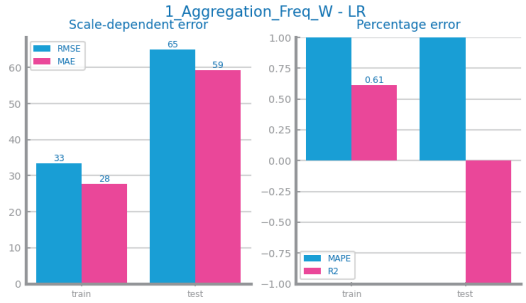
(d) Hourly - Persistence



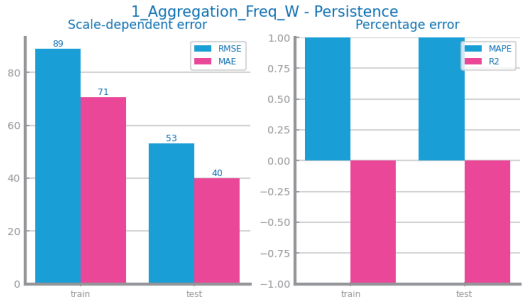
(e) Original - LR



(f) Original - Persistence



(g) Weekly - LR



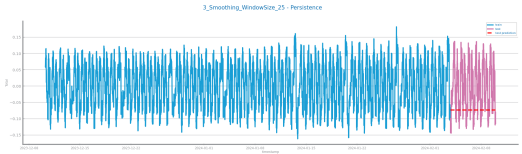
(h) Weekly - Persistence

Figure 59: Forecasting results after different aggregations on time series 1

Smoothing

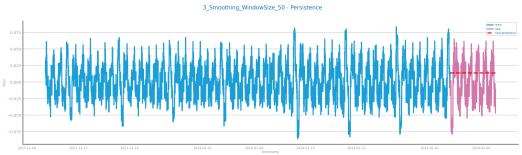
Shall describe the results of applying smoothing transformations over both datasets, and identifying the best result to proceed. **Shall not exceed 300 characters.**

(a) WindowSize 25 - LR



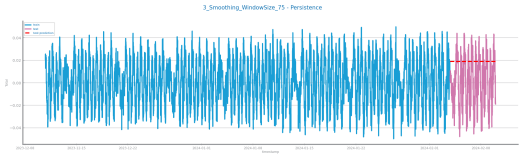
(b) WindowSize 25 - Persistence

(c) WindowSize 50 - LR

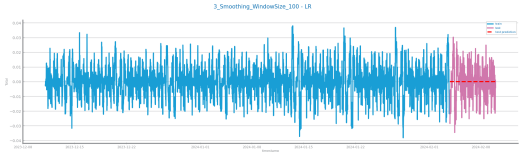


(d) WindowSize 50 - Persistence

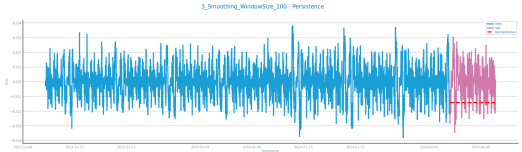
(e) WindowSize 75 - LR



(f) WindowSize 75 - Persistence

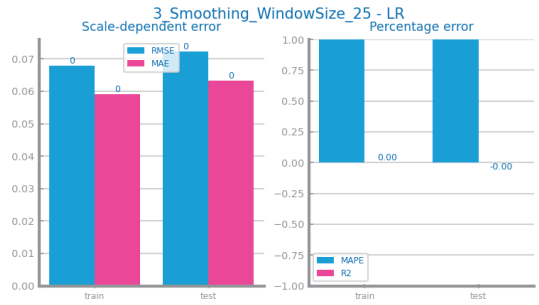


(g) WindowSize 100- LR

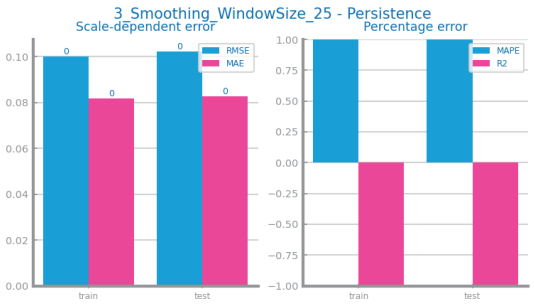


(h) WindowSize 100 - Persistence

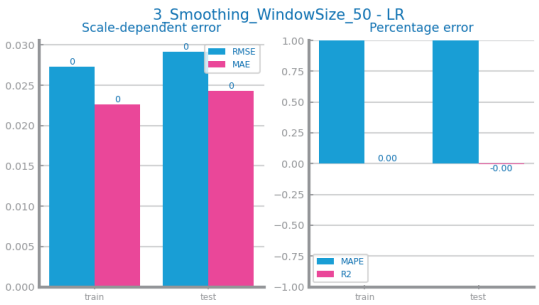
Figure 60: Forecasting plots after different smoothing parameterisations on time series 1



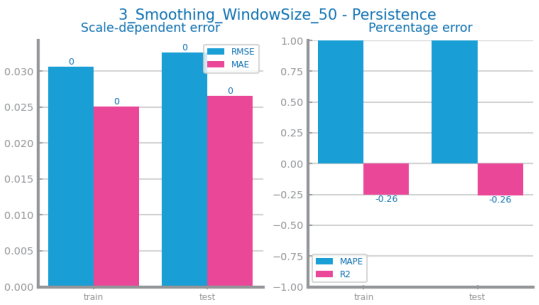
(a) WindowSize 25 - LR



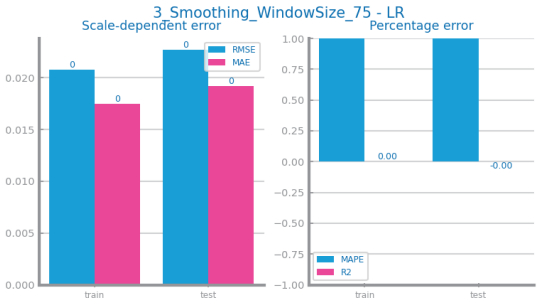
(b) WindowSize 25 - Persistence



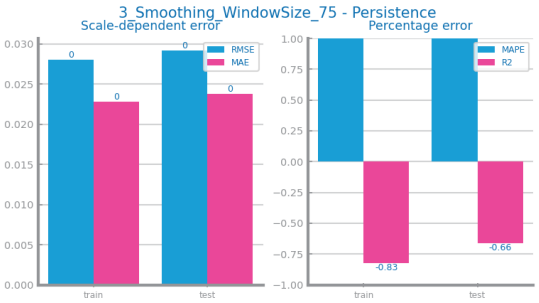
(c) WindowSize 50 - LR



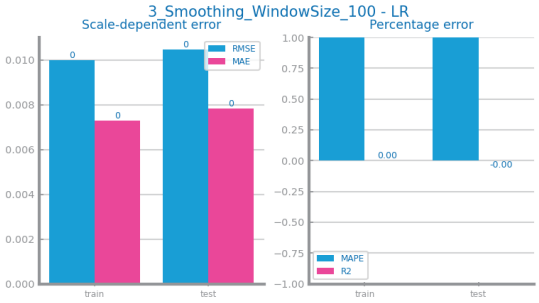
(d) WindowSize 50 - Persistence



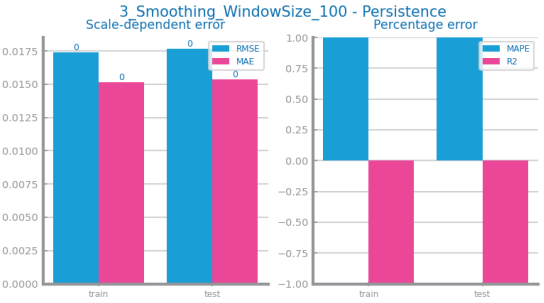
(e) WindowSize 75 - LR



(f) WindowSize 75 - Persistence



(g) WindowSize 100 - LR

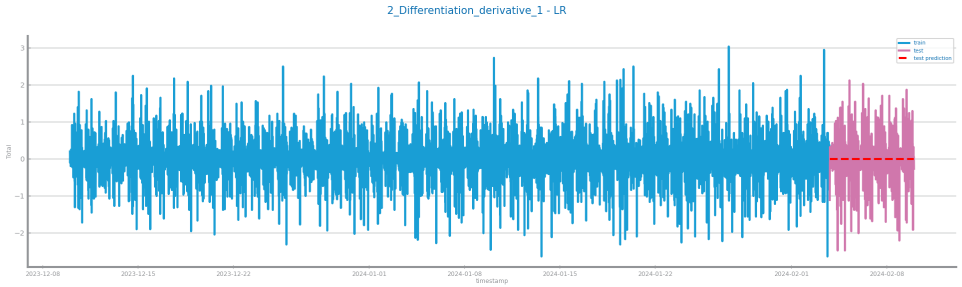


(h) WindowSize 100 - Persistence

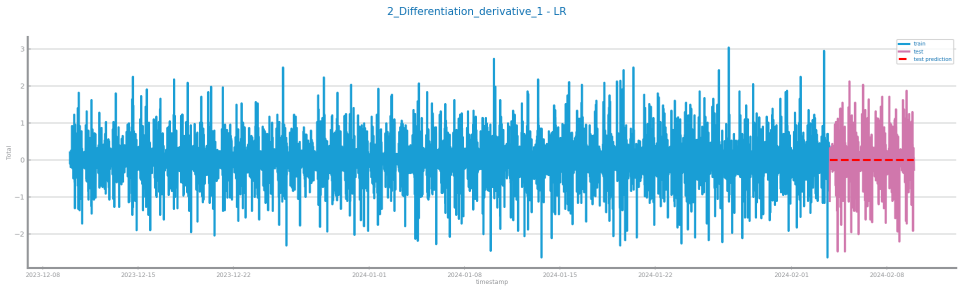
Figure 61: Forecasting results after different smoothing parameterisations on time series 1

Differentiation

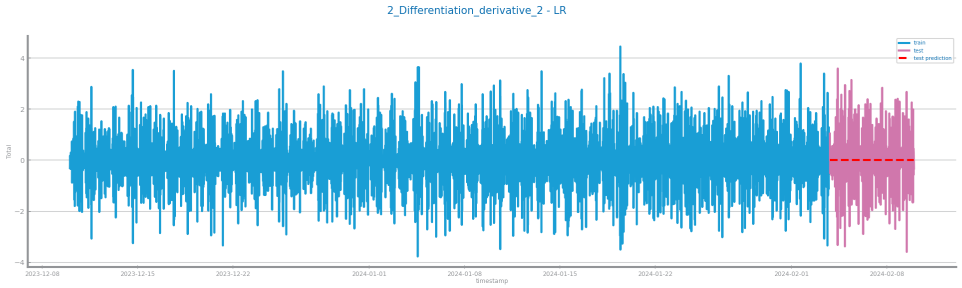
Shall describe the results of applying two consecutive differentiation of both datasets, and identifying the best result to proceed. **Shall not exceed 300 characters.**



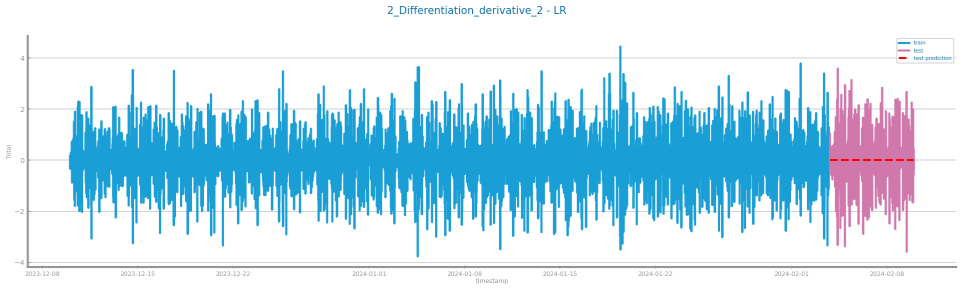
(a) Rist derivative - LR



(b) First derivative - Persistence

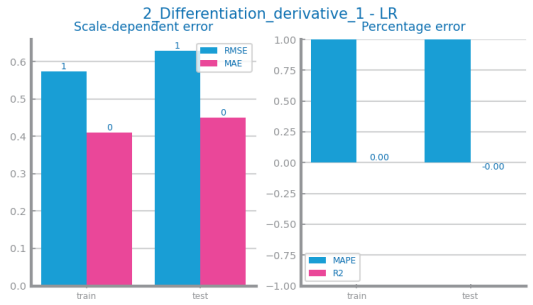


(c) Second derivative - LR

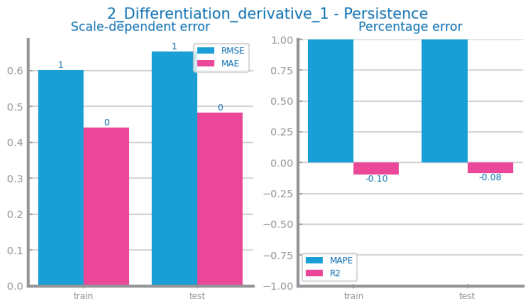


(d) Second derivative - Persistence

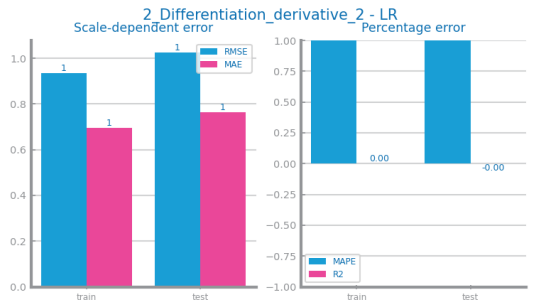
Figure 62: Forecasting plots after first and second differentiation of time series 1



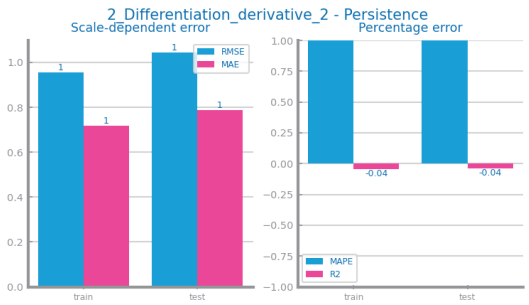
(a) First derivative - LR



(b) First derivative - Persistence



(c) Second derivative - LR



(d) Second derivative - Persistence

Figure 63: Forecasting results after first and second differentiation of time series 1

8 MODELS' EVALUATION

Shall be used to summarise the transformations done over the original time series. **Shall not exceed 500 characters.**

Simple Average Model

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

Figure 64: Forecasting plots obtained with Simple Average model over time series 1

Figure 65: Forecasting results obtained with Simple Average model over time series 1

Persistence Model

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

Figure 66: Forecasting plots obtained with Persistence model (long term) over time series 1

Figure 67: Forecasting plots obtained with Persistence model (one-set-behind) over time series 1

Figure 68: Forecasting results obtained with Persistence model in both situations over time series 1

Rolling Mean Model

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

Figure 69: Forecasting study over different parameterisations of the Rolling Mean algorithm over time series 1

Figure 70: Forecasting plots obtained with the best parameterisation of Rolling Mean algorithm, over time series 1

Figure 71: Forecasting results obtained with the best parameterisation of Rolling Mean algorithm, over time series 1

Exponential Smoothing Model

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

Figure 72: Forecasting study over different parameterisations of the Exponential Smoothing algorithm over time series 1

Figure 73: Forecasting plots obtained with the best parameterisation of Exponential Smoothing algorithm, over time series 1

Figure 74: Forecasting results obtained with the best parameterisation of Exponential Smoothing algorithm, over time series 1

Linear Regression Model

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

Figure 75: Forecasting plots obtained with Linear Regression model over time series 1

Figure 76: Forecasting results obtained with Linear Regression model over time series 1

ARIMA Model

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

Figure 77: Forecasting study over different parameterisations of the ARIMA algorithm over time series 1, only with the target variable

Figure 78: Forecasting plots obtained with the best parameterisation of ARIMA algorithm, over time series 1, only with the target variable

Figure 79: Forecasting results obtained with the best parameterisation of ARIMA algorithm, over time series 1, only with the target variable

Figure 80: Forecasting study over different parameterisations of the ARIMA algorithm with multiple variables over time series 1

Figure 81: Forecasting plots obtained with the best parameterisation of ARIMA algorithm with multiple variables over time series 1

Figure 82: Forecasting results obtained with the best parameterisation of ARIMA algorithm with multiple variables over time series 1

LSTMs Model

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

Figure 83: Forecasting study over different parameterisations of LSTMs over time series 1, only with the target variable

Figure 84: Forecasting plots obtained with the best parameterisation of LSTMs, over time series 1, only with the target variable

Figure 85: Forecasting results obtained with the best parameterisation of LSTMs, over time series 1, only with the target variable

Figure 86: Forecasting study over different parameterisations of LSTMs with multiple variables over time series 1

Figure 87: Forecasting plots obtained with the best parameterisation of LSTMs with multiple variables over time series 1

Figure 88: Forecasting results obtained with the best parameterisation of LSTMs with multiple variables over time series 1

9 CRITICAL ANALYSIS

Shall be used to present a summary of the results achieved with the different forecasting techniques, and the impact of the different preparation tasks on their performance. A critical assessment of the best models shall be presented, clearly stating if the models seem to be good enough for the problem at hand. Additional charts may be presented here. **Shall not exceed 2000 characters.**