TÉCNICO LISBOA

DS 2025

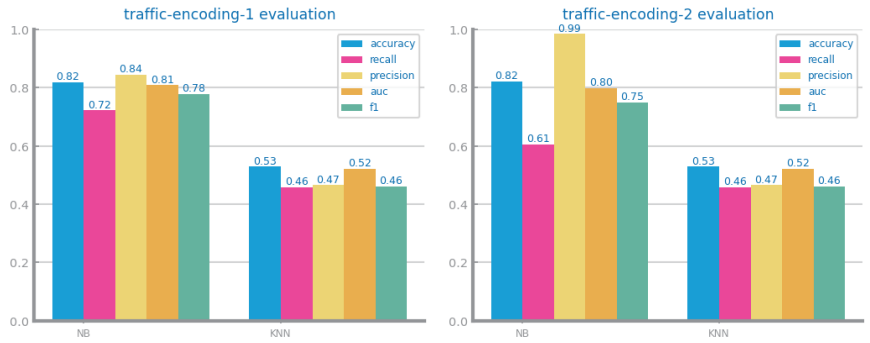
Data Science Lab 3

Team nr: 2	**Student 1:** Åmund Grimstad	**IST nr:** 1116675
	Student 2: Arthur de Arruda Chau	**IST nr:** 1116090
	Student 3: Benjamin Raymond Kuhn	**IST nr:** 1115778
	Student 4: João Rafael Freitas Lourenço	**IST nr:** 425699

CLASSIFICATION - DATA PREPARATION

1 Traffic Accident Data

Variables Encoding



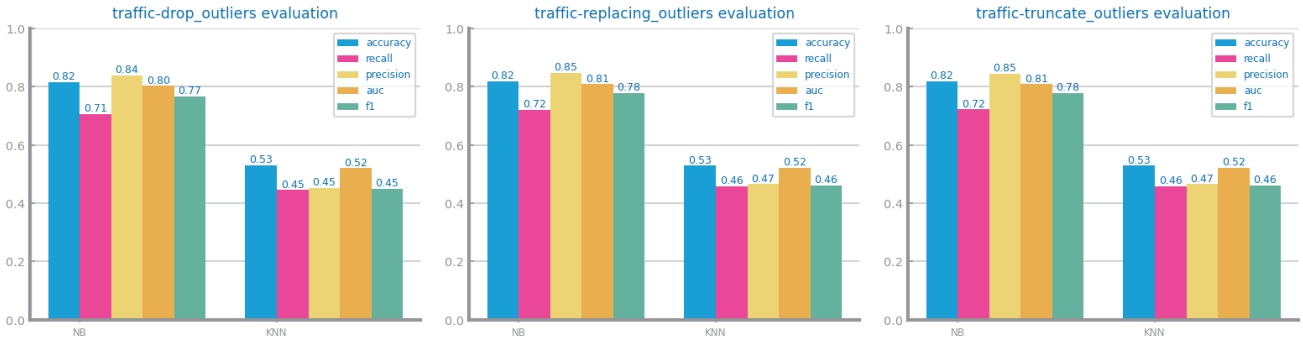
(a) (b)

Figure 1: Encoding results with different approaches for Traffic Accident Data

Missing Value Imputation

There are no missing values in the Traffic Accident Dataset.

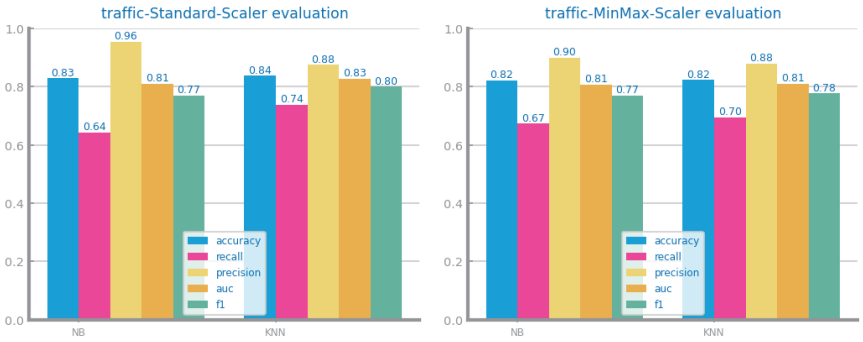
Outliers Treatment



(a) (b) (c)

Figure 2: Outliers imputation results with different approaches for Traffic Accident Data

Scaling



(a) (b)

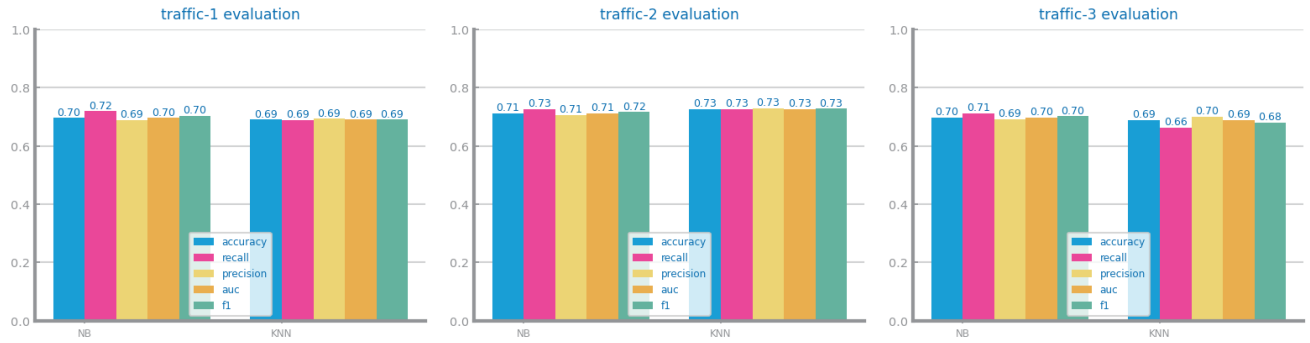
Figure 3: Scaling results with different approaches for Traffic Accident Data

Balancing

(a) (b)

Figure 4: Balancing results with different approaches for Traffic Accident Data

Feature Selection



(a) (b) (c)

Figure 5: Feature selection results for Traffic Accident Data

Figure 6: Feature selection of redundant variables results with different parameters for Traffic Accident Data

Figure 7: Feature selection of relevant variables results with different parameters for Traffic Accident Data (variance study)

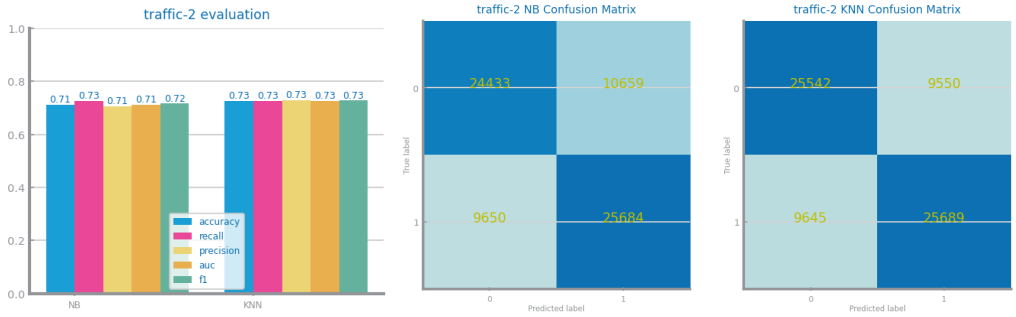
Feature Extraction (optional)

Figure 8: Principal components analysis and feature extraction results for Traffic Accident Data

Additional Feature Generation (if done)

Figure 9: Feature generation results for Traffic Accident Data

Best approach

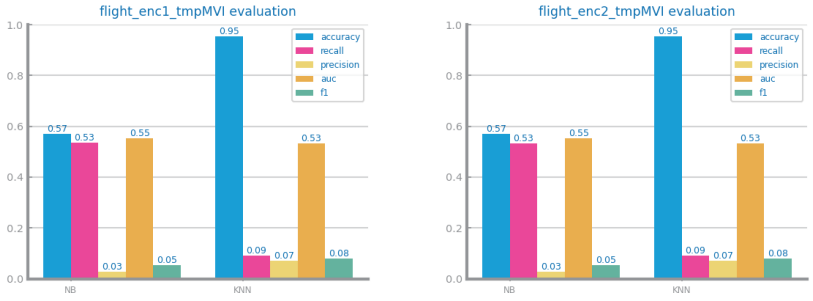


(a) (b)

Figure 10: Performance and confusion matrix for best approach for Traffic Accident Data

2 Flight Cancellation Data

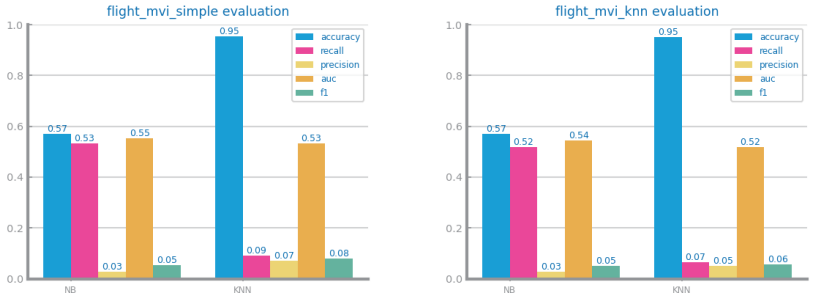
Variables Encoding



(a) (b)

Figure 11: Encoding results with different approaches for Flight Cancellation Data

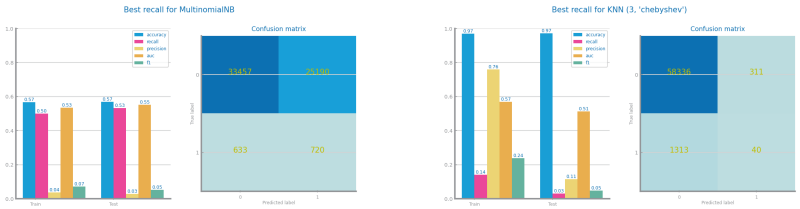
Missing Value Imputation



(a) (b)

Figure 12: Missing values imputation results with different approaches for Flight Cancellation Data

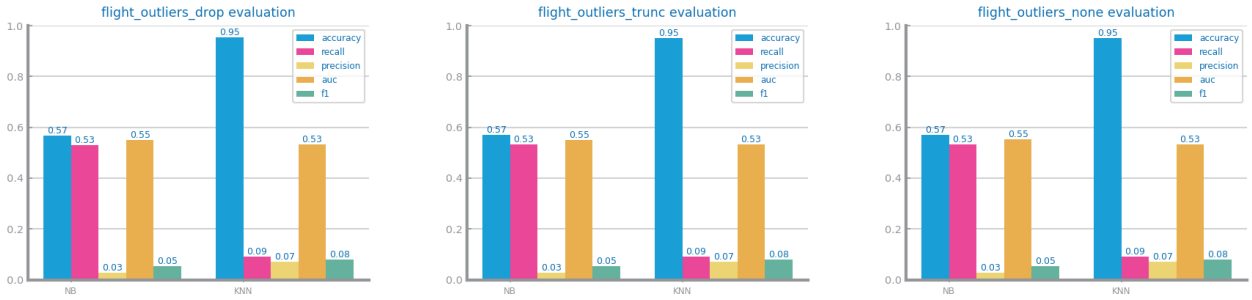
Best approach for Missing Value Imputation



(a) (b)

Figure 13: Best Bayes and KNN Models applied to best Missing Value Imputation approach for Flight Cancellation Data

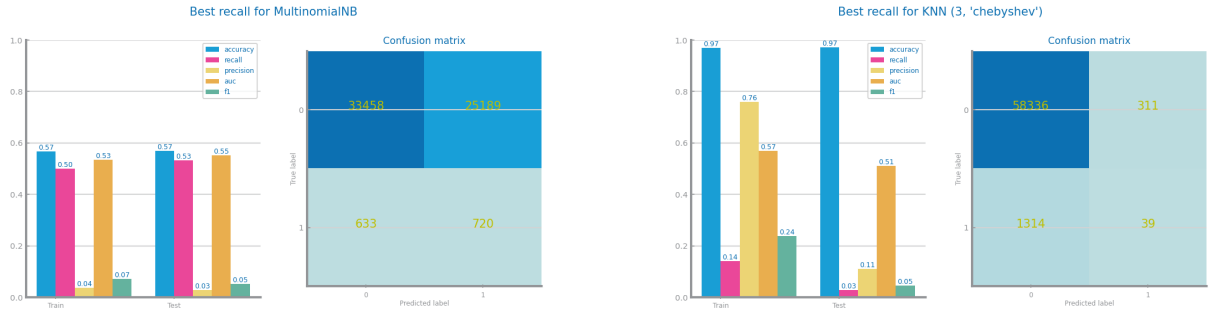
Outliers Treatment



(a) (b) (c)

Figure 14: Outliers imputation results with different approaches for Flight Cancellation Data

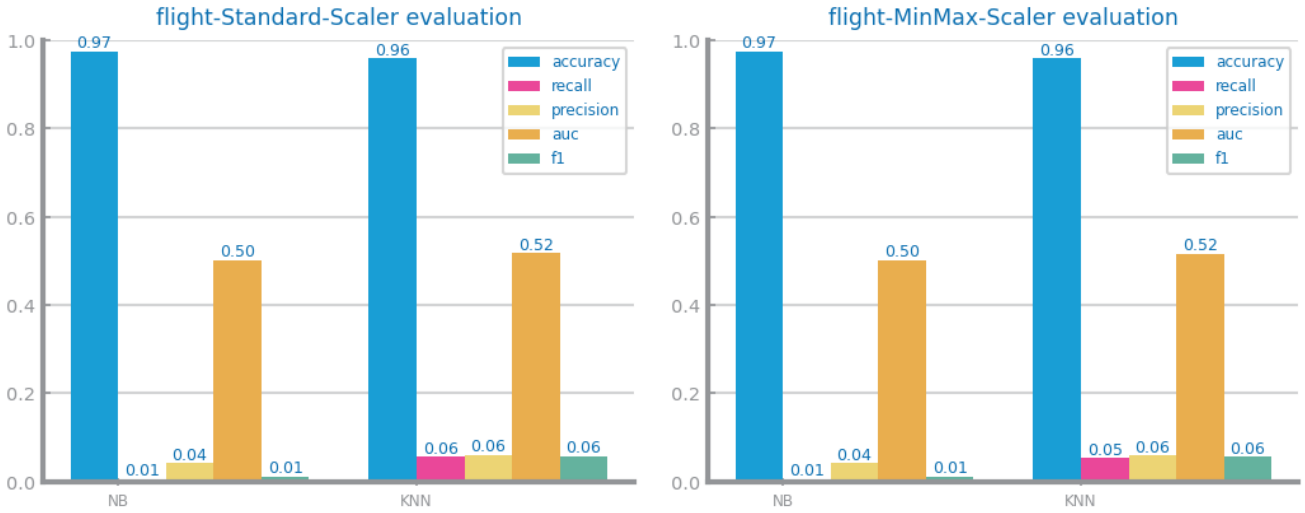
Best approach for Outliers Treatment. No treatment yielded highest recall.



(a) (b)

Figure 15: Best Bayes and KNN Models applied to best Outlier treatment for Flight Cancellation Data

Scaling



(a) (b)

Figure 16: Scaling results with different approaches for Flight Cancellation Data

Balancing

Figure 17: Balancing results with different approaches for Flight Cancellation Data

Feature Selection

Figure 18: Feature selection of redundant variables results with different parameters for Flight Cancellation Data

Figure 19: Feature selection of relevant variables results with different parameters for Flight Cancellation Data (variance study)

Feature Extraction (optional)

Figure 20: Principal components analysis and feature extraction results for Flight Cancellation Data

Additional Feature Generation (if done)

Figure 21: Feature generation results for Flight Cancellation Data

Best approach

Figure 22: Performance and confusion matrix for best approach for Flight Cancellation Data