|  |  |  |  |
| --- | --- | --- | --- |
| Logo, company name  Description automatically generated | | **DS 2023** | |
| Data Science Project | | | |
| **Team nr:** 5 | **Student 1 :** Eduardo Lança Lobo | | **IST nr:** 99213 |
| **Student 2 :** Inês Ye Ji | | **IST nr:** 99238 |
| **Student 3 :** Jiqi Wang | | **IST nr:** 99241 |
|  | **Student 4 :** Raquel Filipa Marques Cardoso | | **IST nr:** 99314 |

This document presents a template for the Data Science Project report. It specifies the mandatory format and suggests the structure to follow. All text with grey background shall be replaced with the analysis made over the datasets.

Classification

# 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

A graph showing the number of records

Description automatically generatedA graph showing the number of records

Description automatically generatedShall contain all relevant information and charts respecting to the data dimensionality perspective, such as the number of records and number of dimensions, and their impact on the following analysis. **Shall not exceed 500 characters.**

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

A graph showing the number of variables

Description automatically generated Falta credit

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

A screen shot of a computer screen

Description automatically generated

Falta credit

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

## Data Distribution

A graph with blue and white lines

Description automatically generatedShall 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**.

Falta credit

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

A group of graphs showing the different sizes of the same size

Description automatically generated with medium confidence

Figure Single variable boxplots for dataset 1

Falta credit

Figure Single variable boxplots s for dataset 2

A group of graphs showing different sizes of data

Description automatically generated with medium confidence

Figure Histograms for dataset 1 (with distributions is enough)

Falta credit

Figure Histograms for dataset 2 (with distributions is enough)

Standard ou nao?

Figure Outliers study dataset 1

Standard ou nao?

Figure Outliers study for dataset 2

A graph showing a number of covid-19

Description automatically generated

Figure Class distribution for dataset 1

Falta credit

Figure 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.**

Metemos quais?

Figure Granularity analysis for dataset 1

Metemos quais?

Figure 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.**

A white sheet with lines and dots

Description automatically generated with medium confidence

Figure Sparsity analysis for dataset 1

Falta credit

Figure Sparsity analysis for dataset 2

A graph with a line

Description automatically generated

Figure Correlation analysis for dataset 1

Falta credit

Figure Correlation analysis for dataset 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.**

A graph showing the results of a test

Description automatically generated with medium confidenceA graph of different colored bars

Description automatically generated

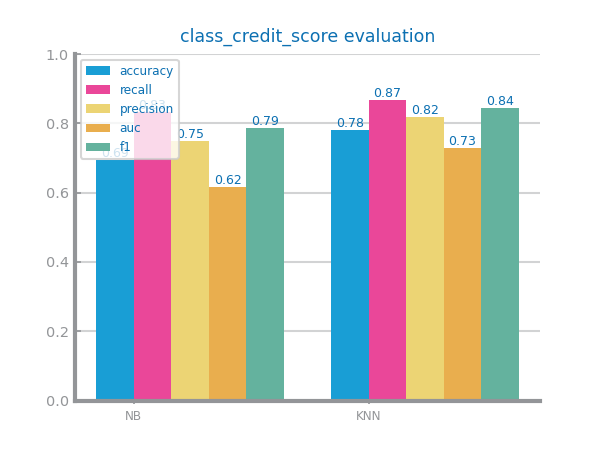
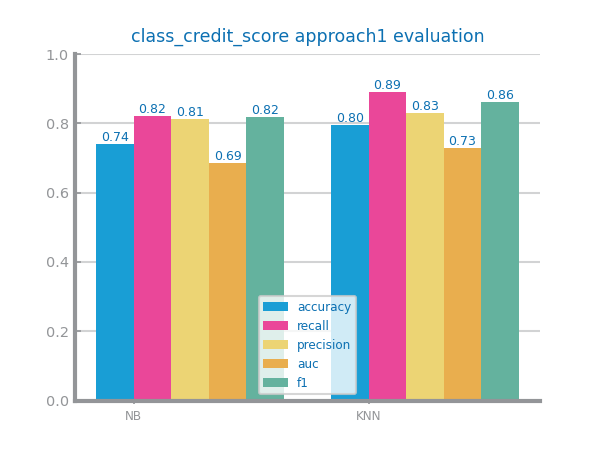
Figure Missing values imputation results with different approaches for dataset 1

Figure 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**.

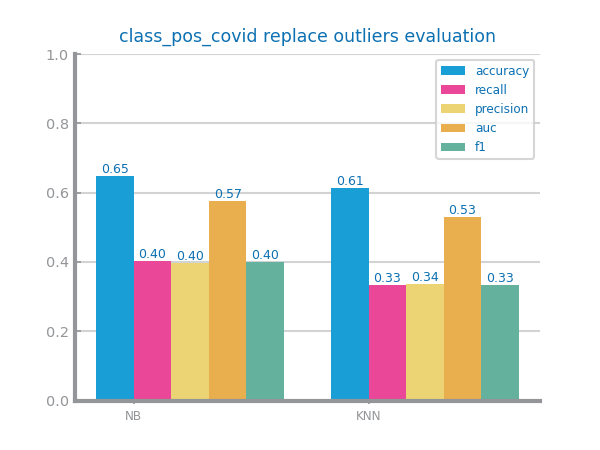
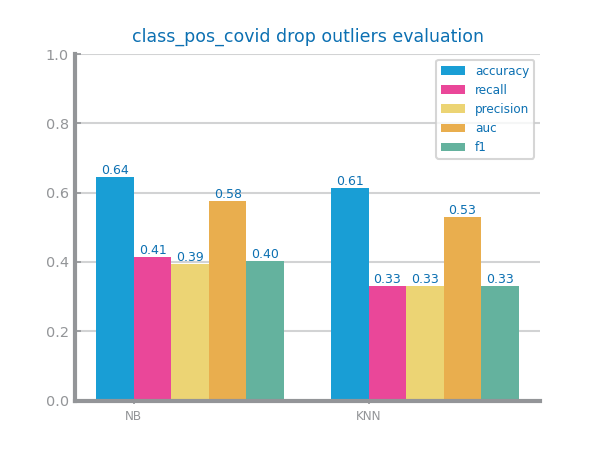
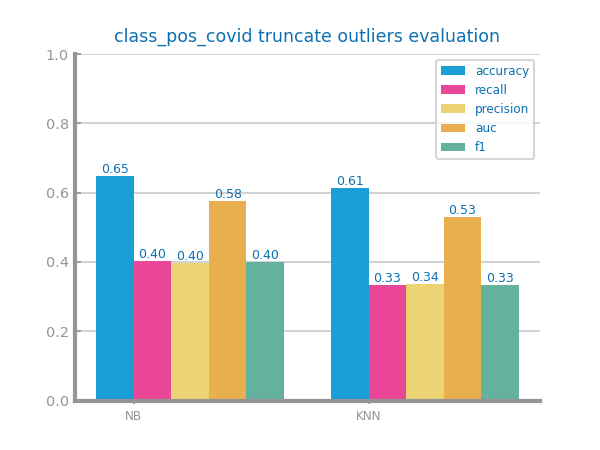


Figure Outliers imputation results with different approaches for dataset 1

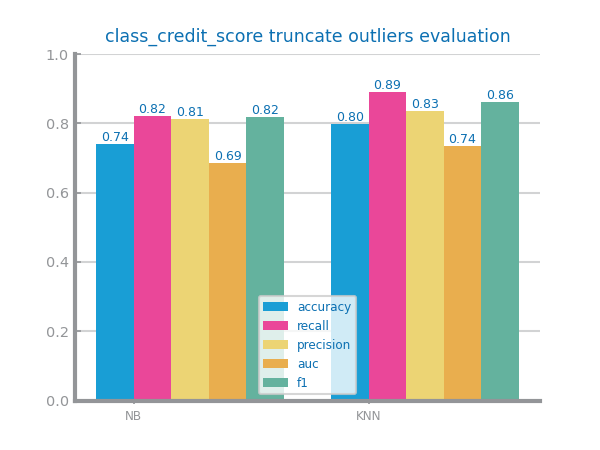
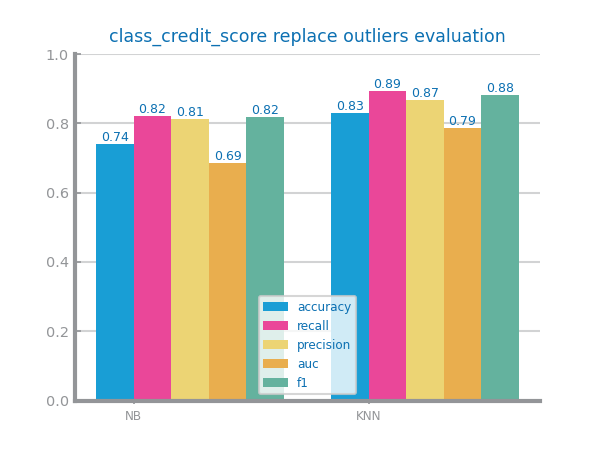
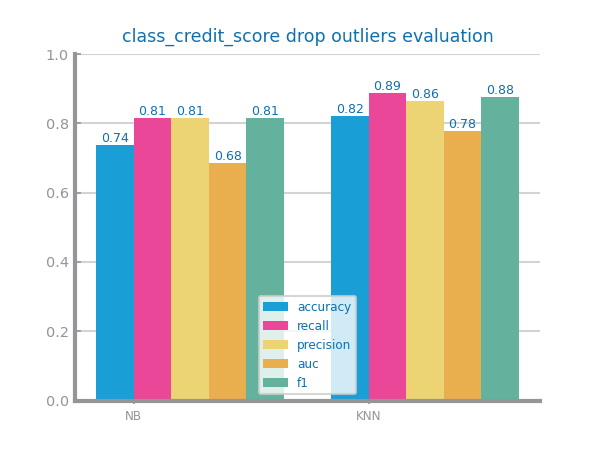


Figure Outliers imputation results with different approaches for dataset 2

## 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.**

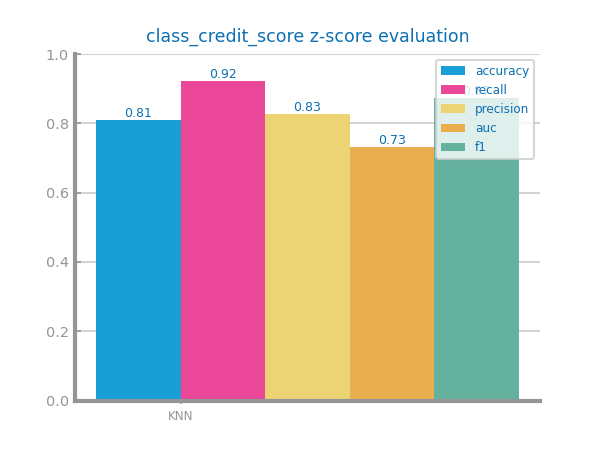
A graph of different colored bars

Description automatically generatedA graph of different colored bars

Description automatically generatedA graph of different colored bars

Description automatically generated

Figure Scaling results with different approaches for dataset 1

A graph of different colored bars

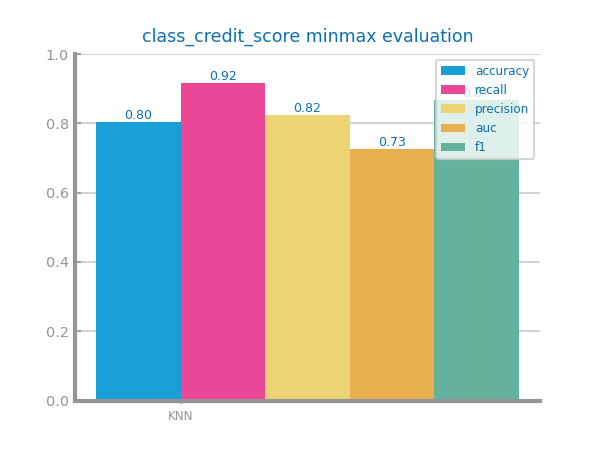
Description automatically generated

Figure 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**.

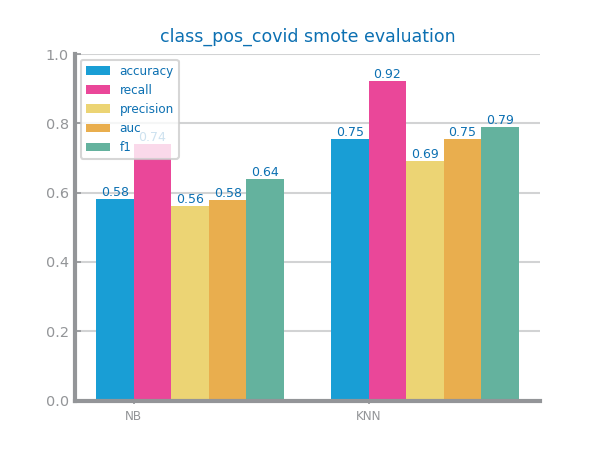
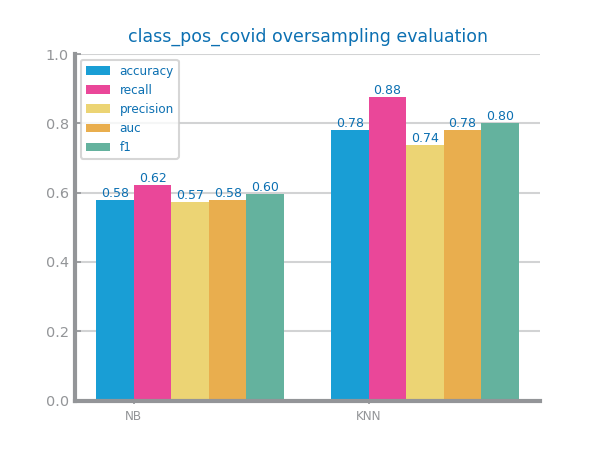
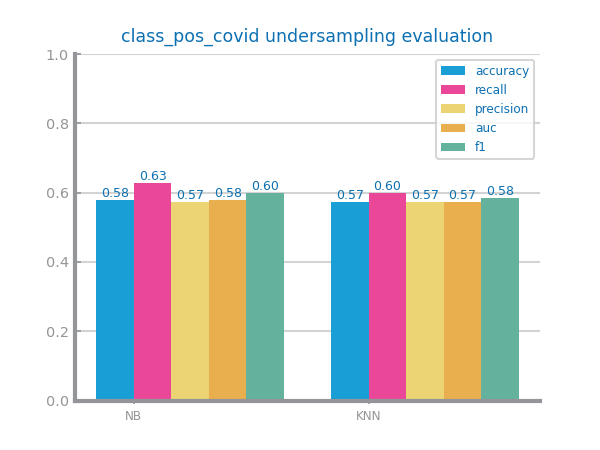


Figure Balancing results with different approaches for dataset 1

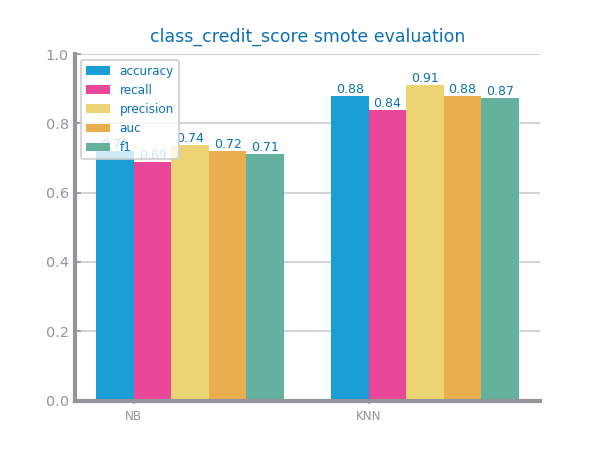
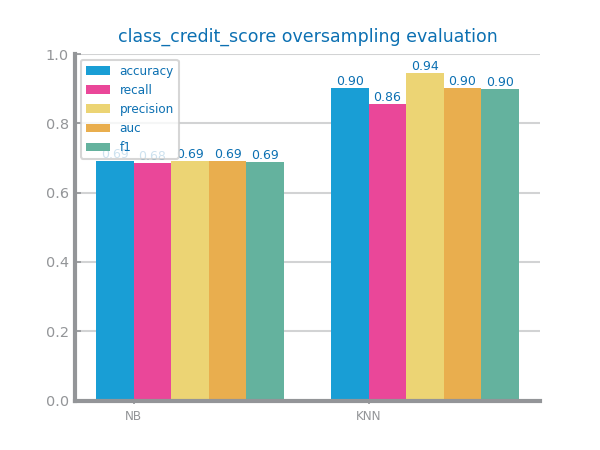
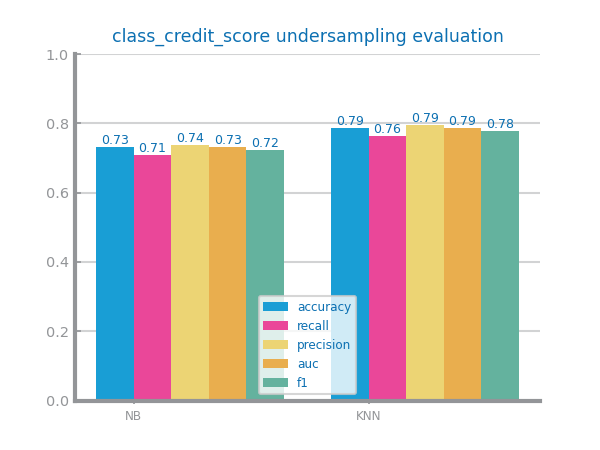


Figure Balancing results with different approaches for dataset 2

## 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.**

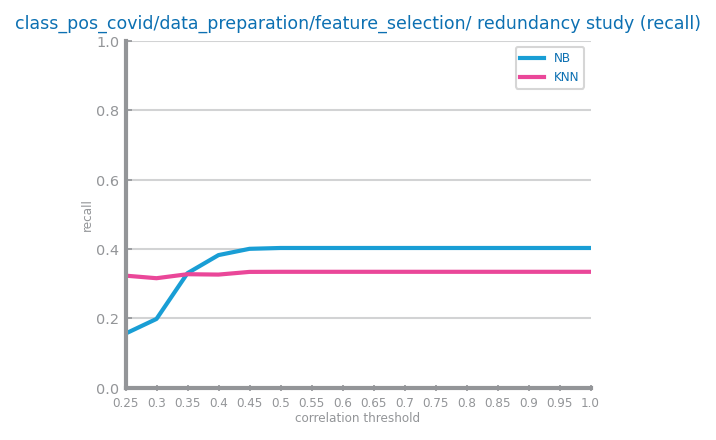


Figure Feature selection of redundant variables results with different parameters for dataset 1

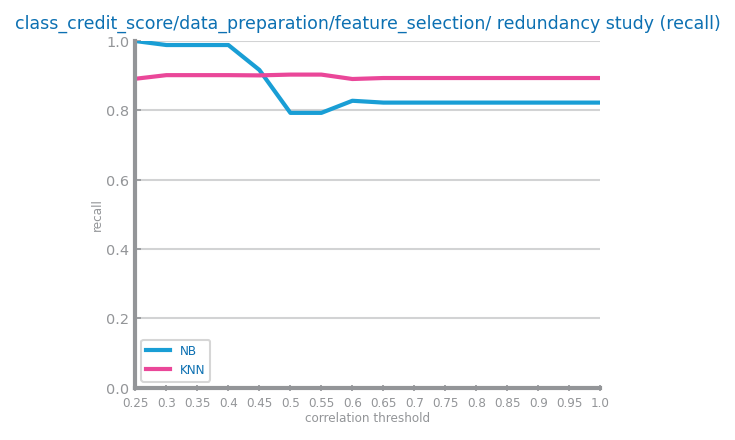


Figure Feature selection of redundant variables results with different parameters for dataset 2

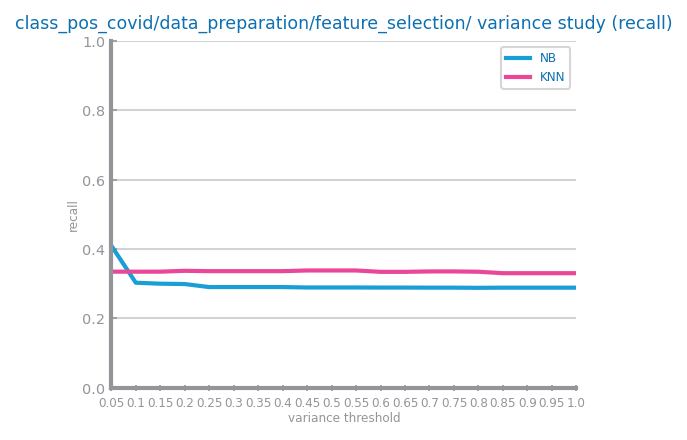


Figure Feature selection of relevant variables results with different parameters for dataset 1 (variance study)

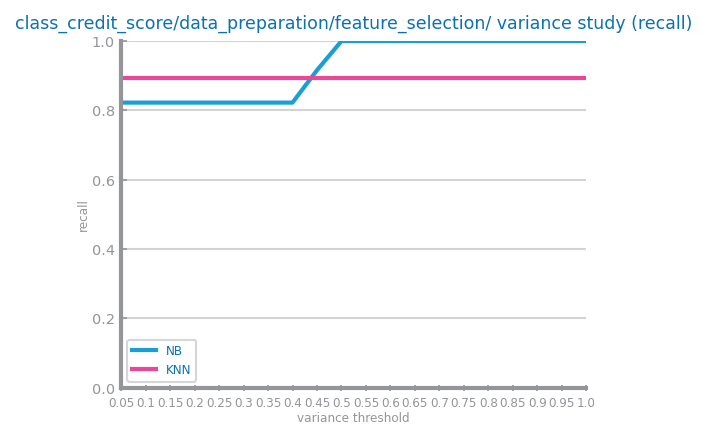


Figure Feature selection of relevant variables results with different parameters for dataset 2 (variance study)

## 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.**

Por ver

Figure Principal components analysis and feature extraction results for dataset 1

Por ver

Figure Principal components analysis and feature extraction results for dataset 2

## Additional Feature Generation (optional)

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.**

Por ver

Figure Feature generation results for dataset 1

Por ver

Figure Feature generation results for dataset 2

# 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.**

A graph of a bar chart

Description automatically generated with medium confidence

Figure Naïve Bayes alternatives comparison for dataset 1

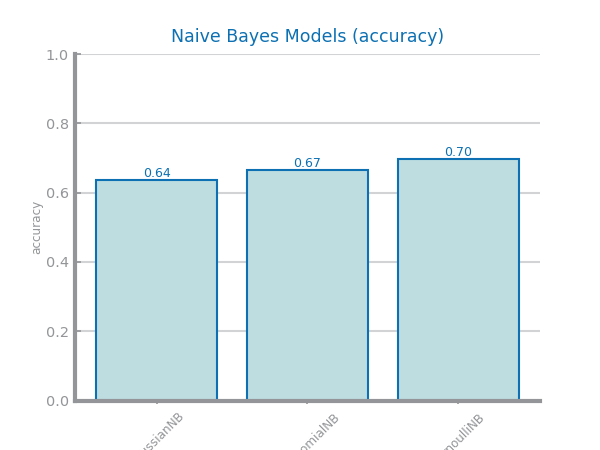


Figure Naïve Bayes alternative comparison for dataset 2

A close-up of a graph

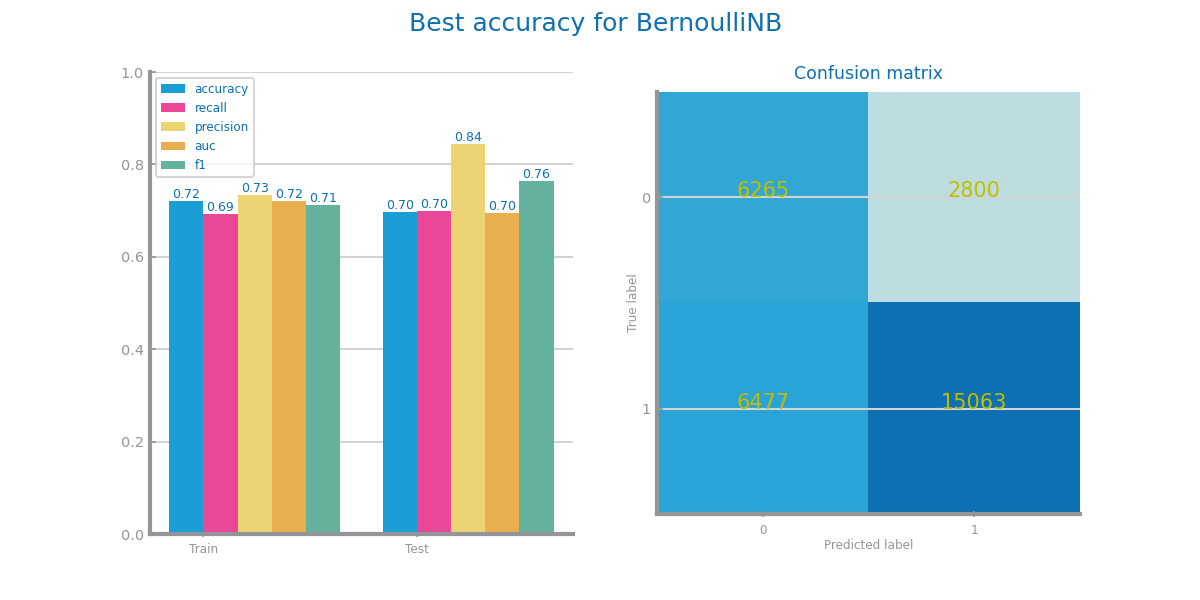
Description automatically generated

Figure 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**

A graph of different colored lines

Description automatically generated

Figure KNN different parameterisations comparison for dataset 1

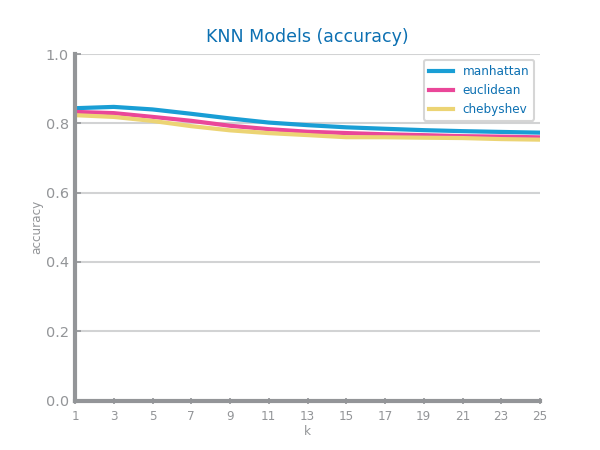


Figure KNN different parameterisations comparison for dataset 2

A graph with a line and a pink line

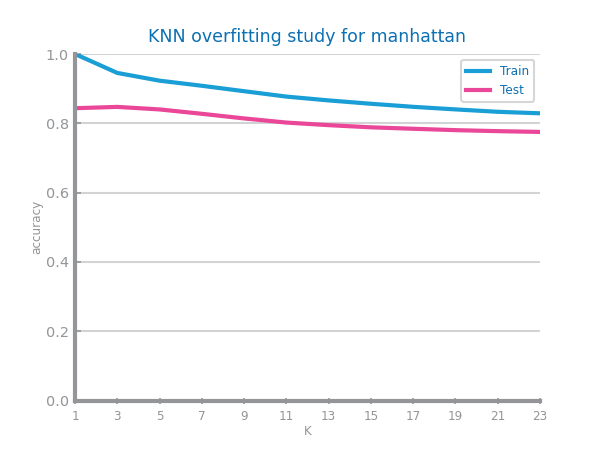
Description automatically generated 

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

A close-up of a graph

Description automatically generated

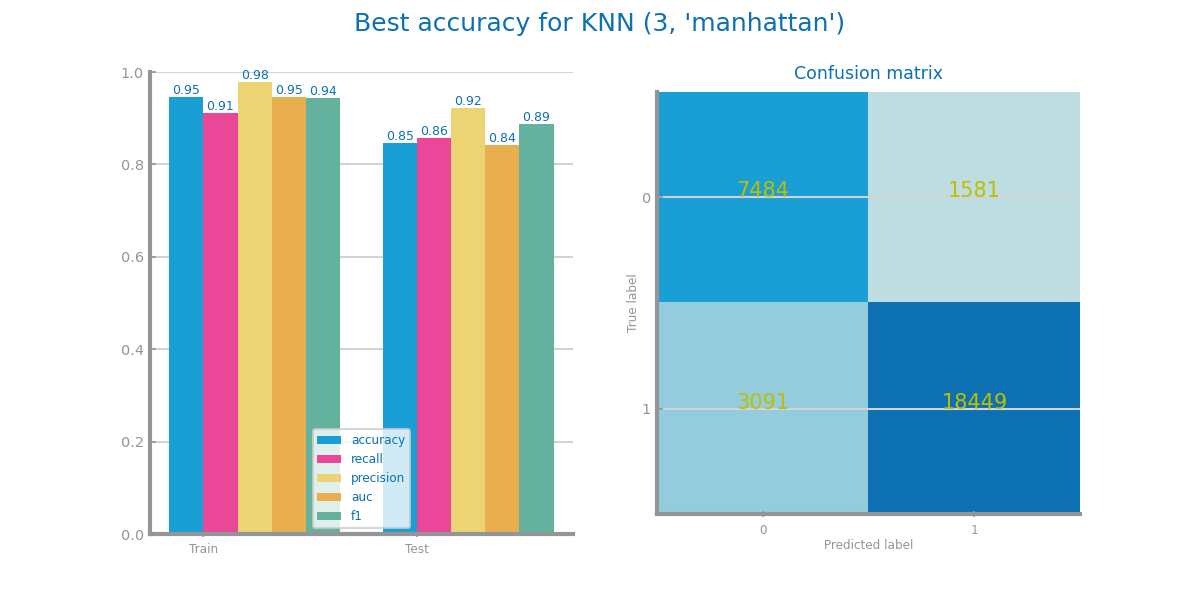


Figure KNN best model results for dataset 1 (left) and dataset 2 (right)

## 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**

A graph with lines and numbers

Description automatically generated

Figure Decision Trees different parameterisations comparison for dataset 1

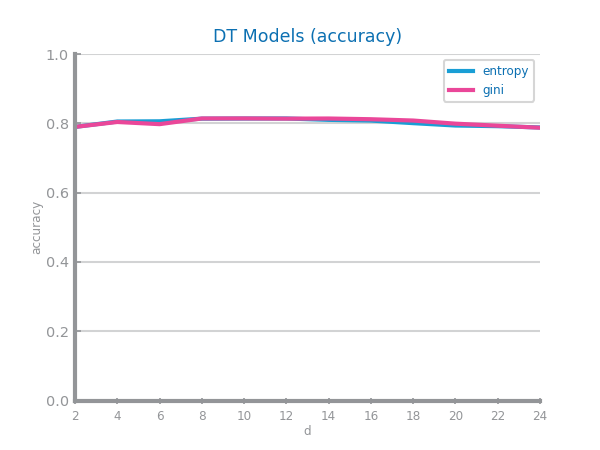


Figure Decision Trees different parameterisations comparison for dataset 2

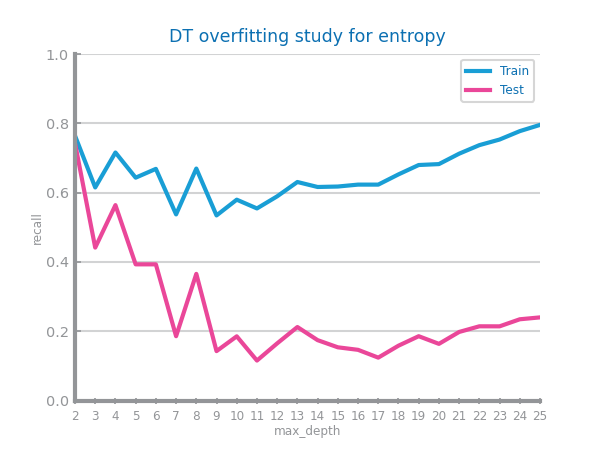
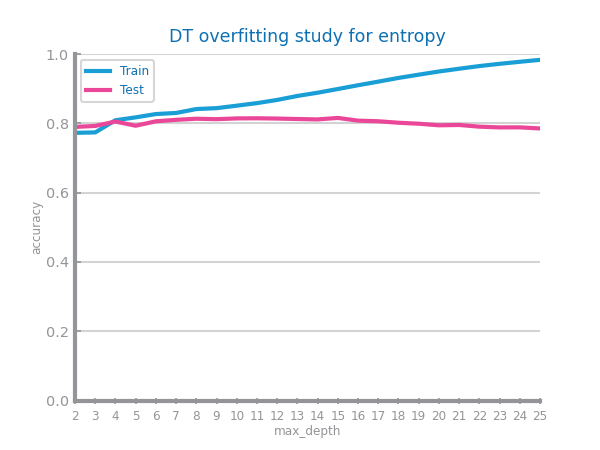
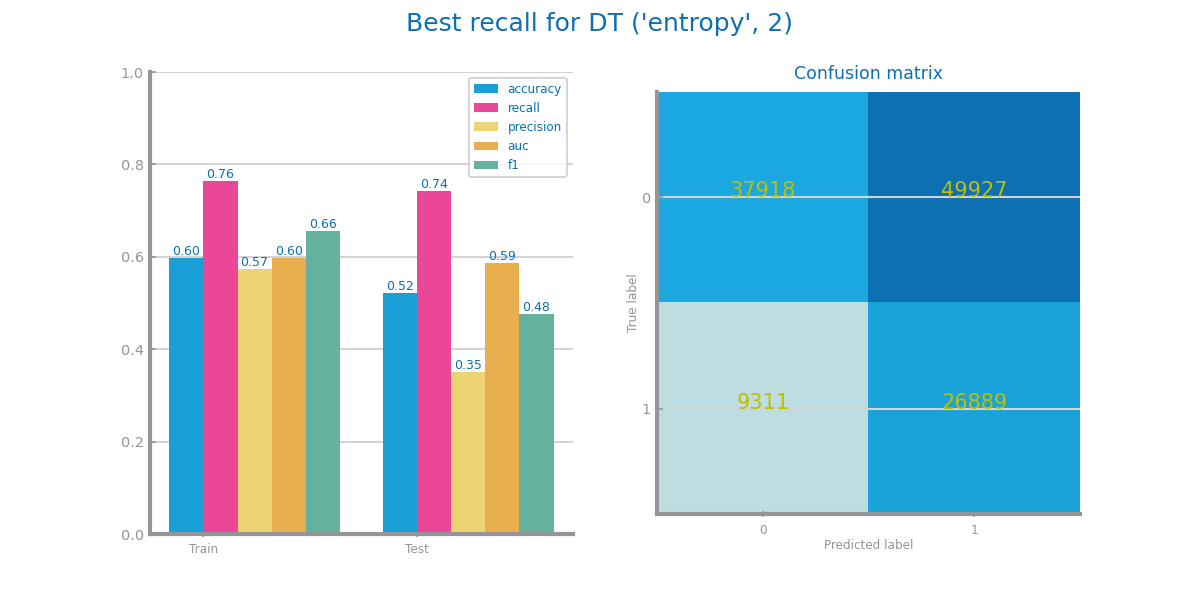
 

Figure Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)



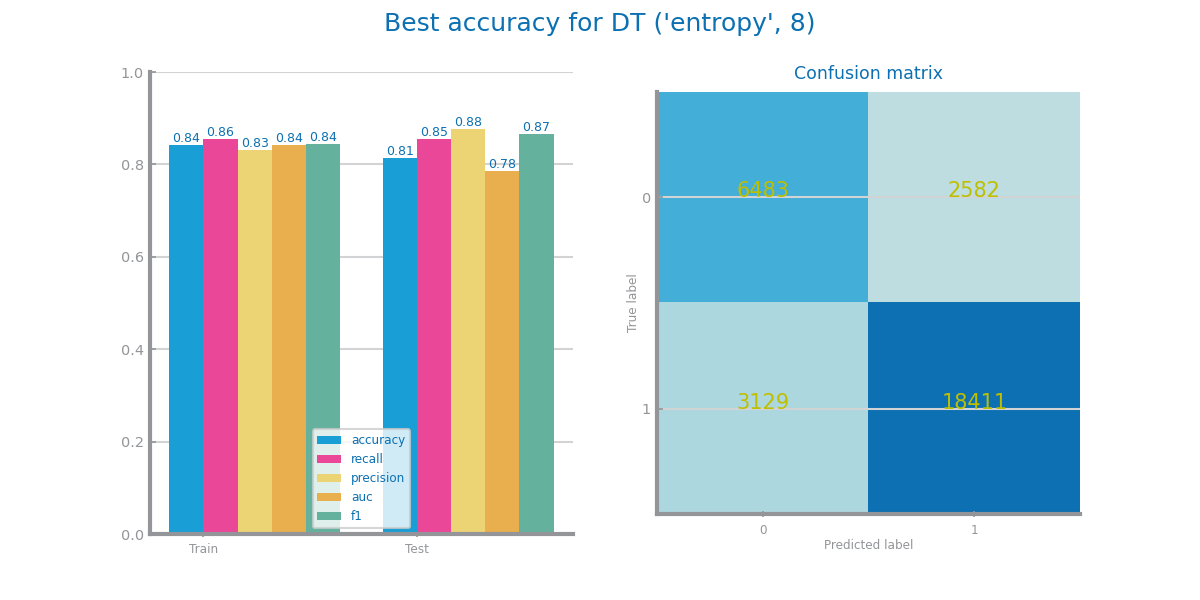


Figure Decision trees best model results for dataset 1 (left) and dataset 2 (right)

A diagram of a diagram

Description automatically generated with medium confidence

Figure Best tree for dataset 1

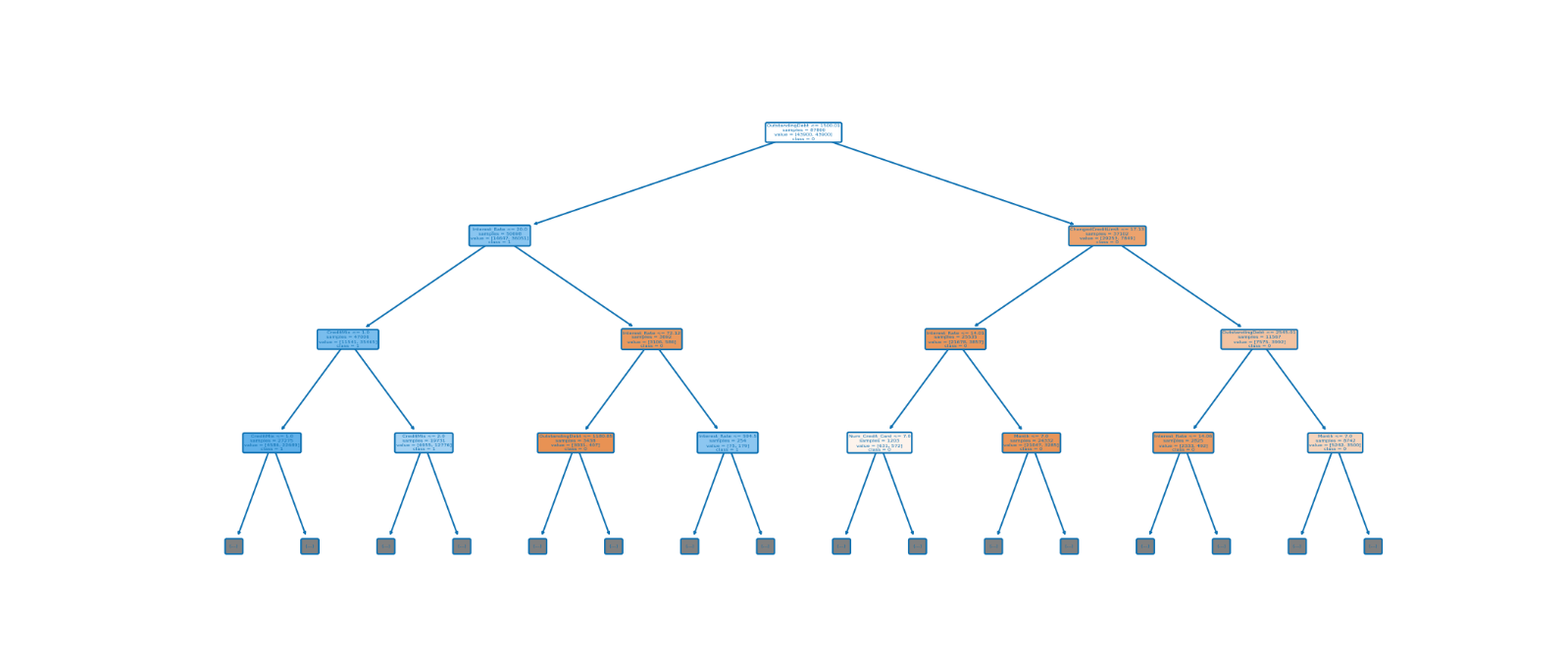


Figure Best trees for dataset 2

## 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**

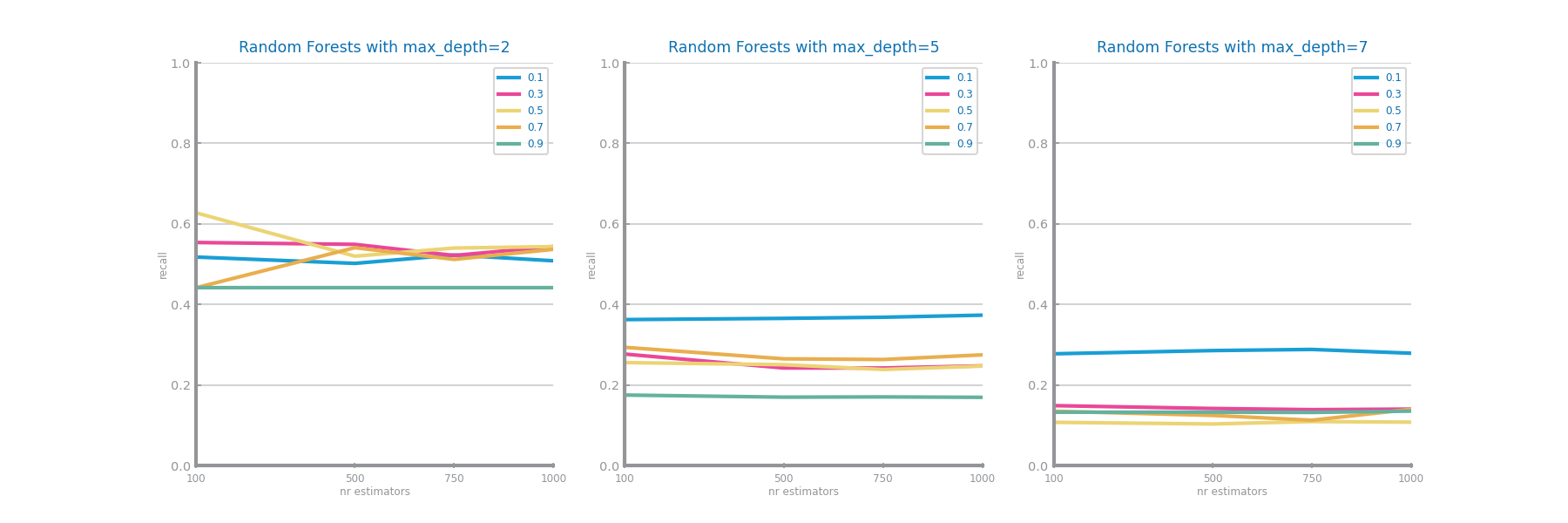


Figure Random Forests different parameterisations comparison for dataset 1

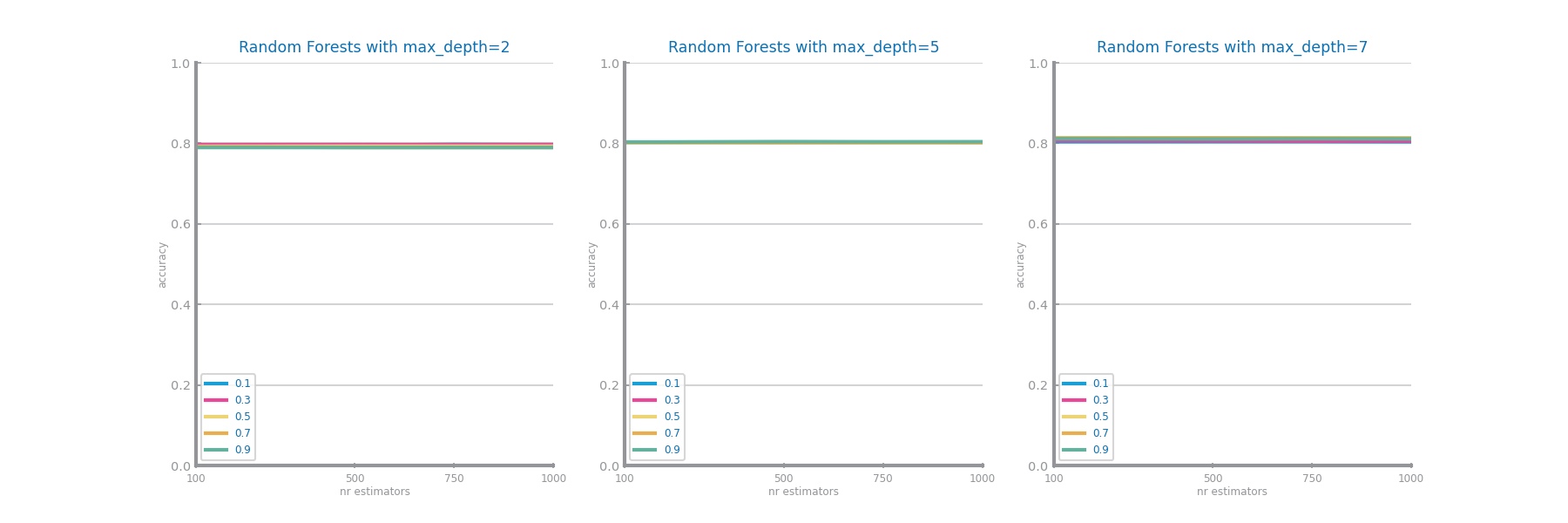
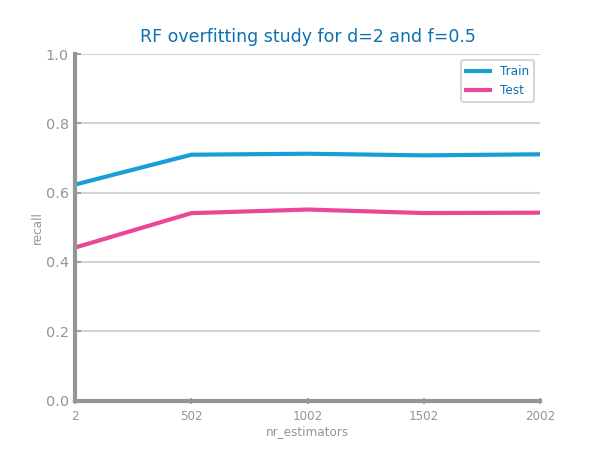
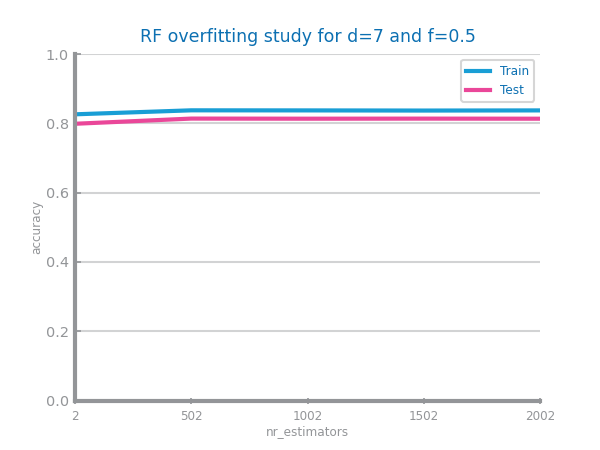
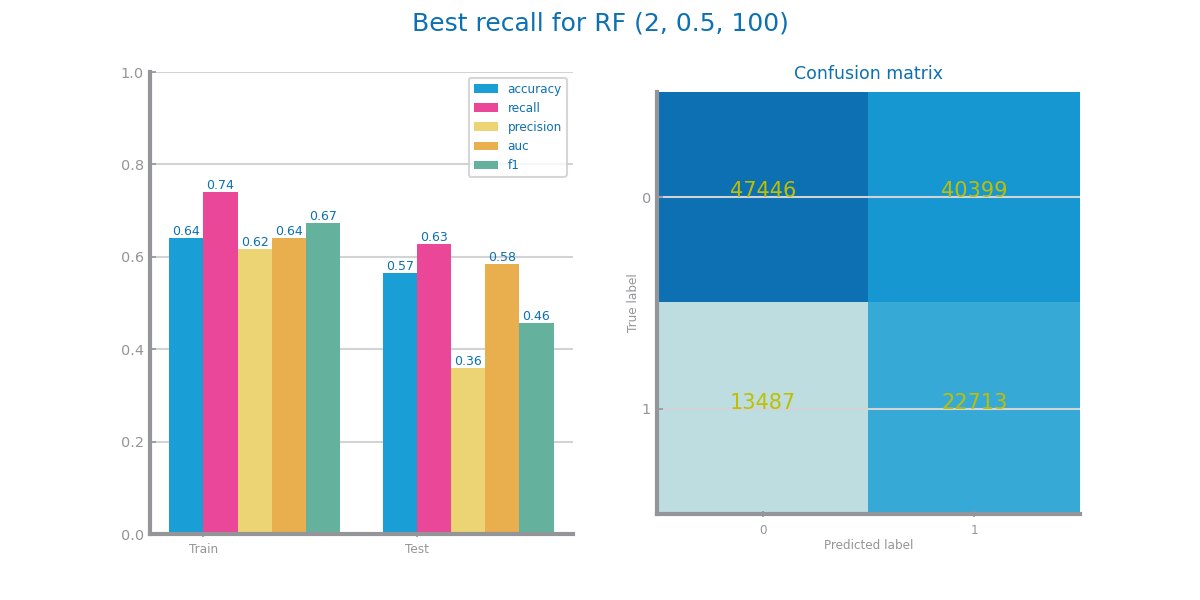


Figure Random Forests different parameterisations comparison for dataset 2

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

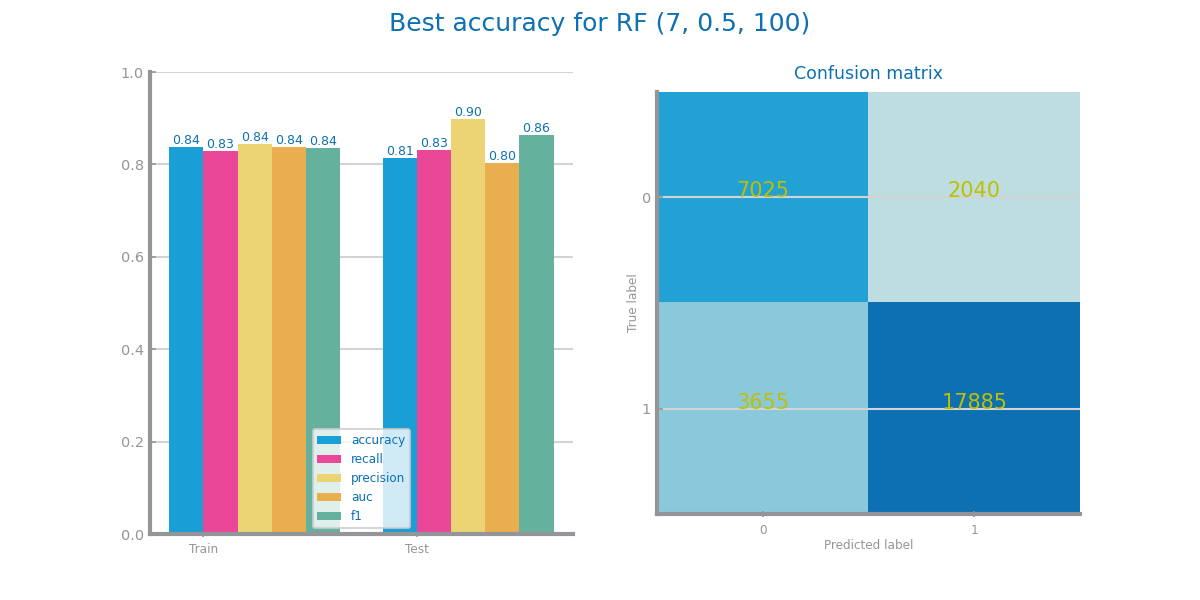
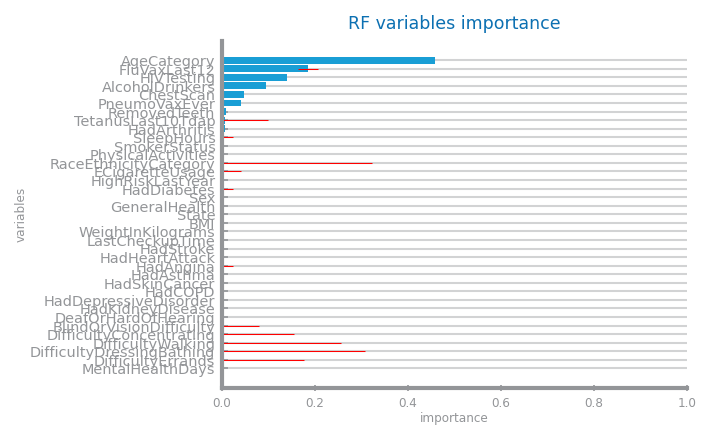
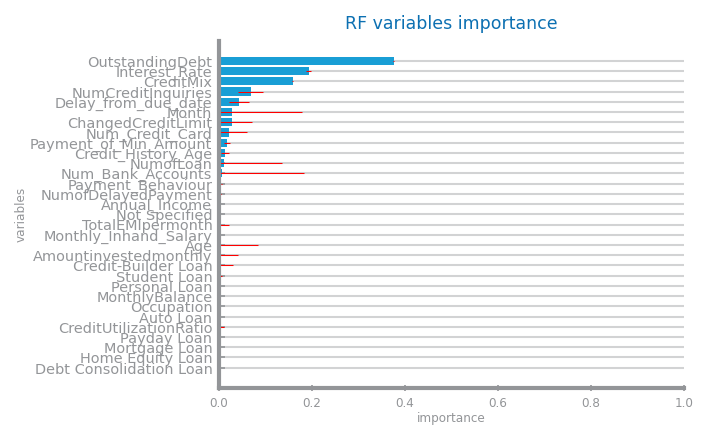


Figure Random Forests best model results for dataset 1 (left) and dataset 2 (right)

Figure 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**

A graph of different colored lines

Description automatically generated

Figure Gradient boosting different parameterisations comparison for dataset 1

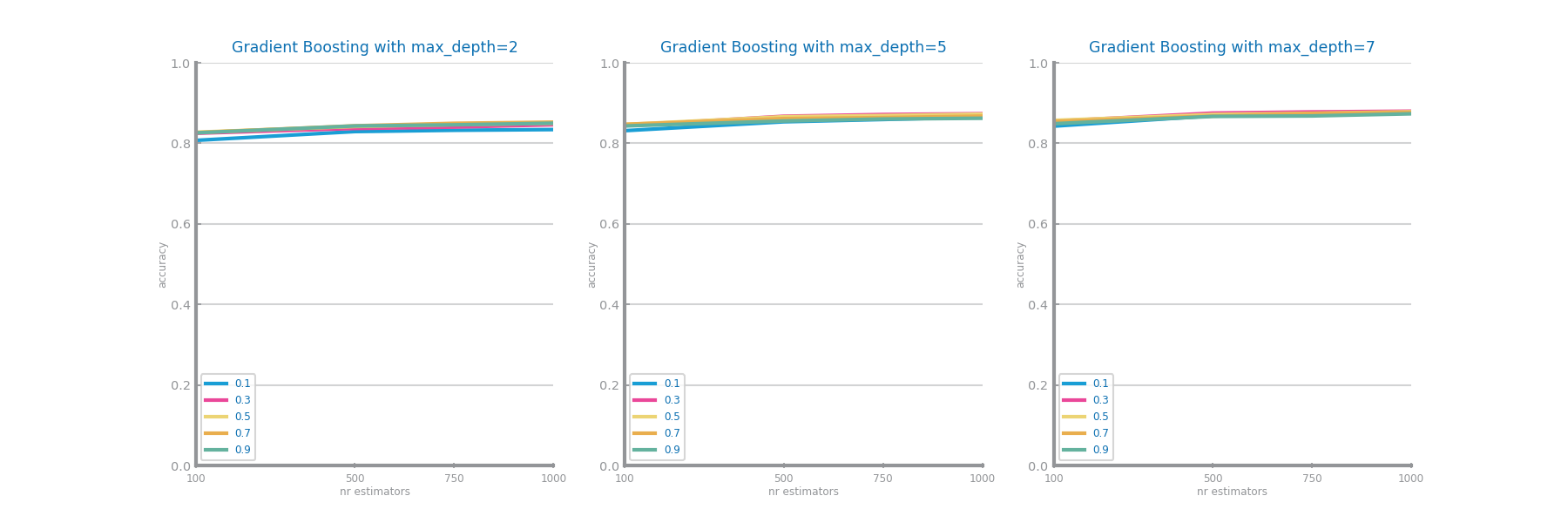


Figure Gradient boosting different parameterisations comparison for dataset 2

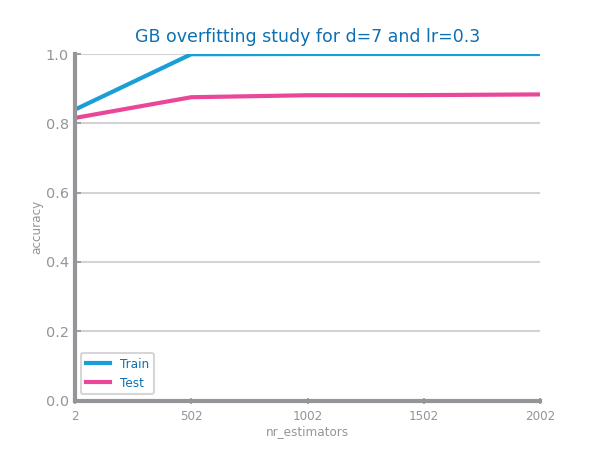
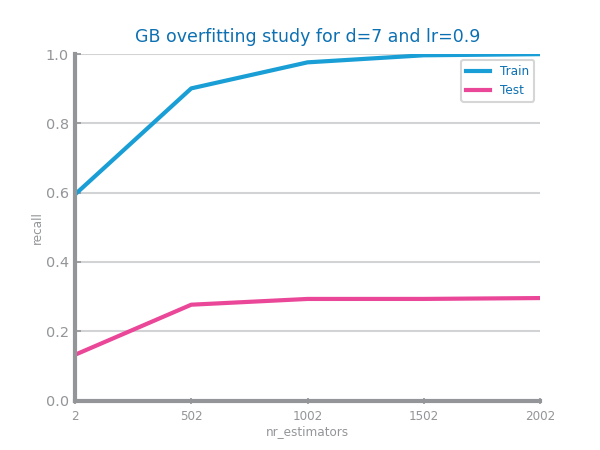
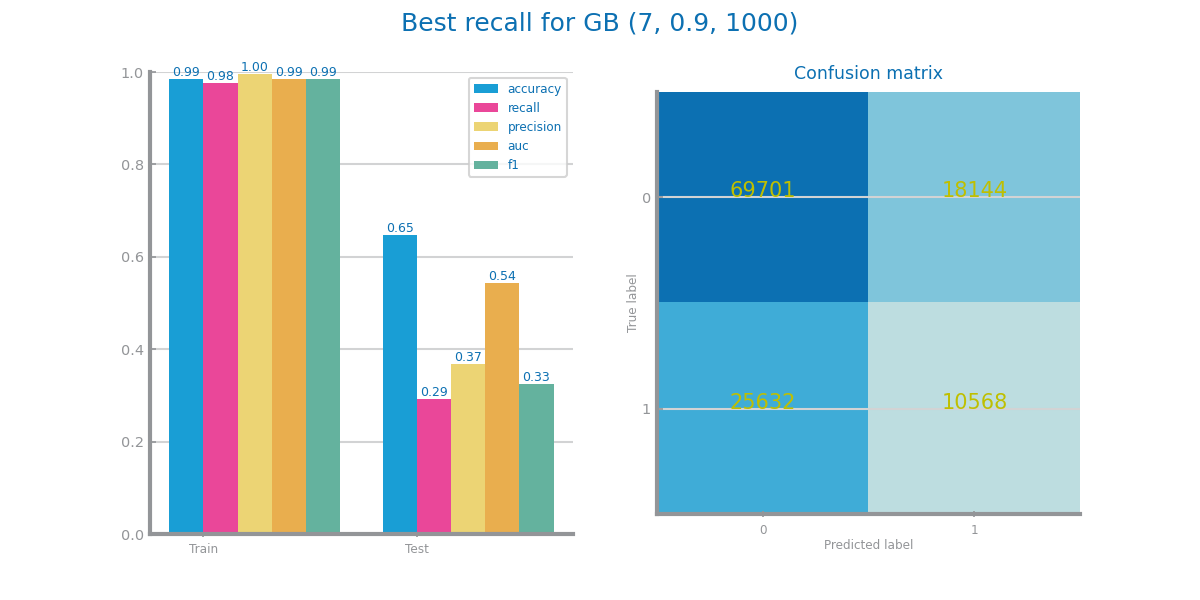
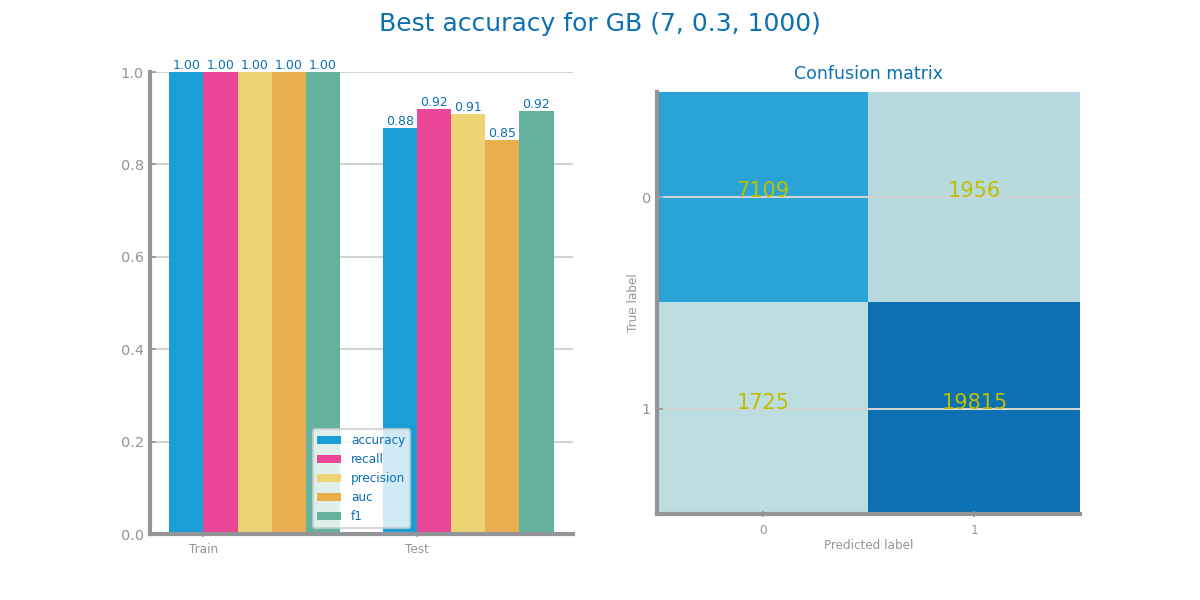
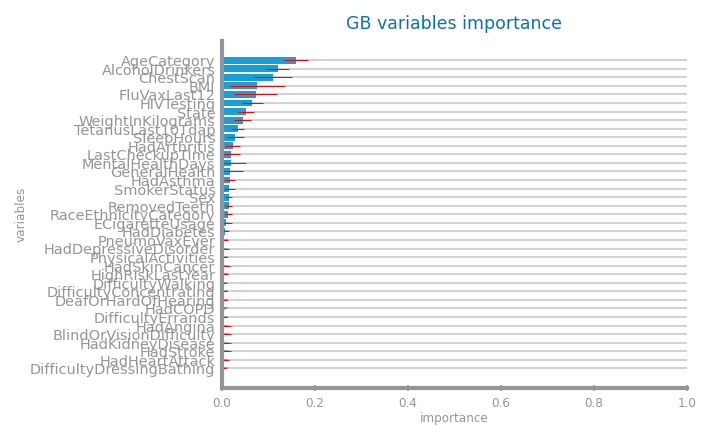


Figure Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)





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

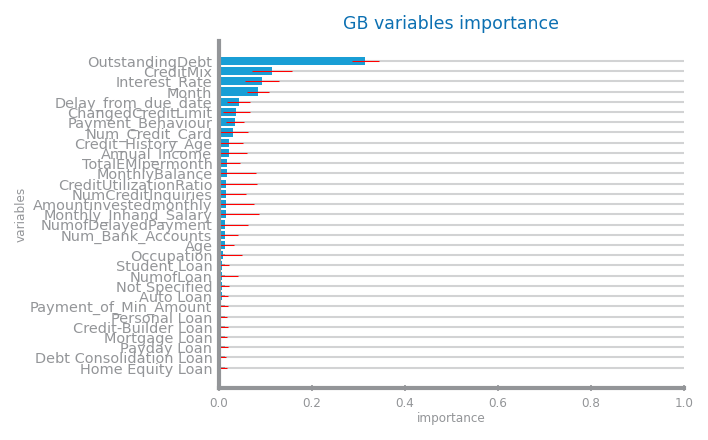


Figure Gradient boosting variables importance for dataset 1 (left) and dataset 2 (right)

## 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**

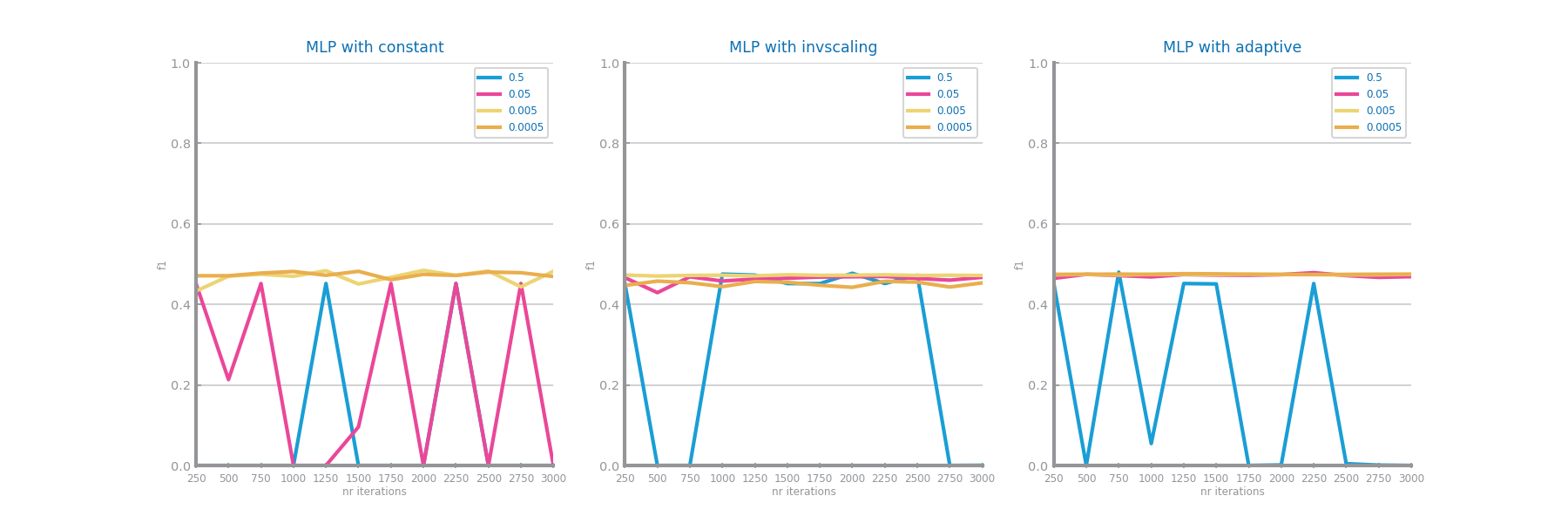


Figure MLP different parameterisations comparison for dataset 1

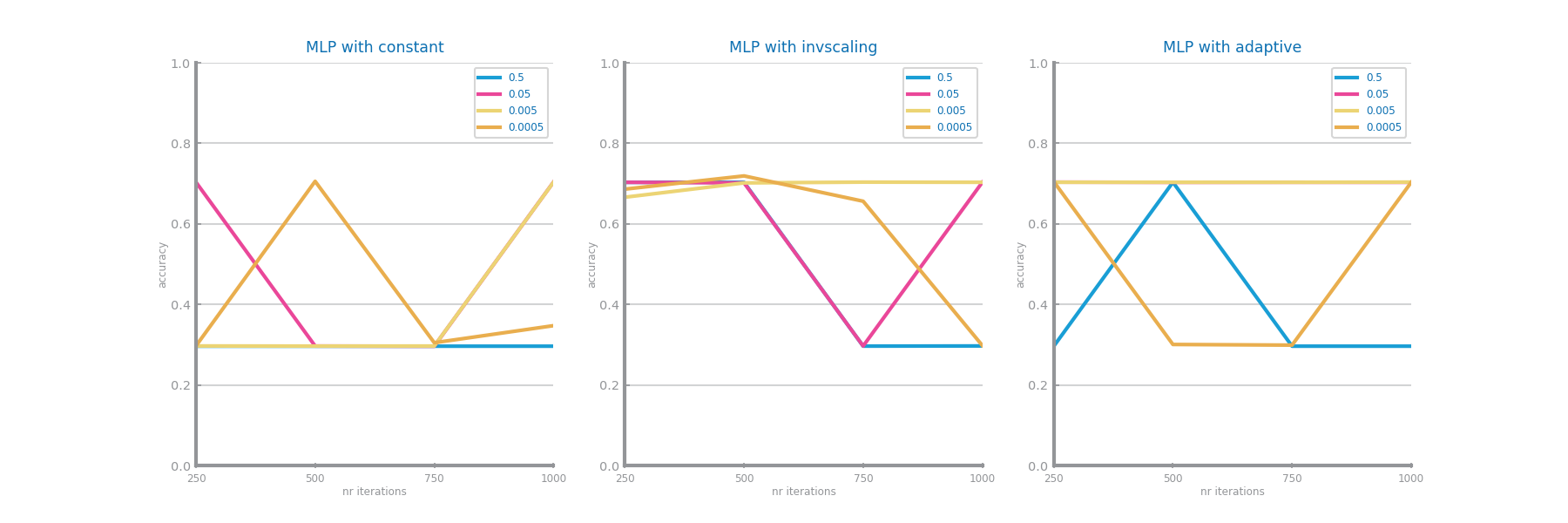


Figure MLP different parameterisations comparison for dataset 2

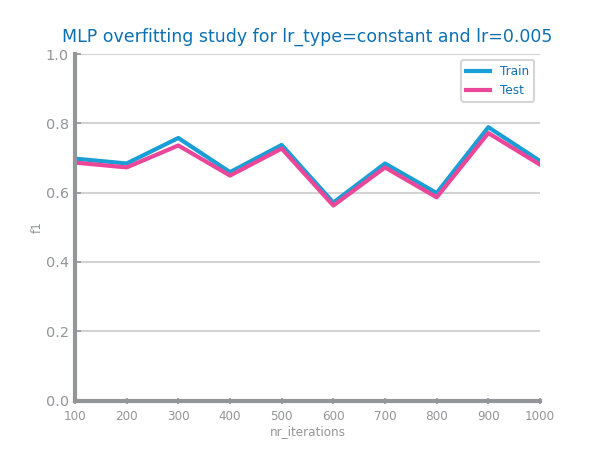
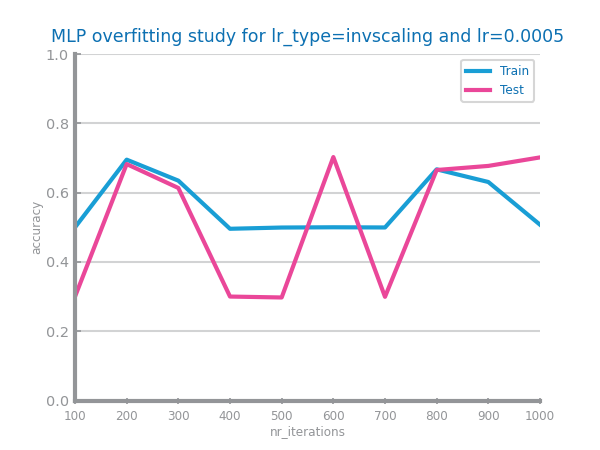
 

Figure MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

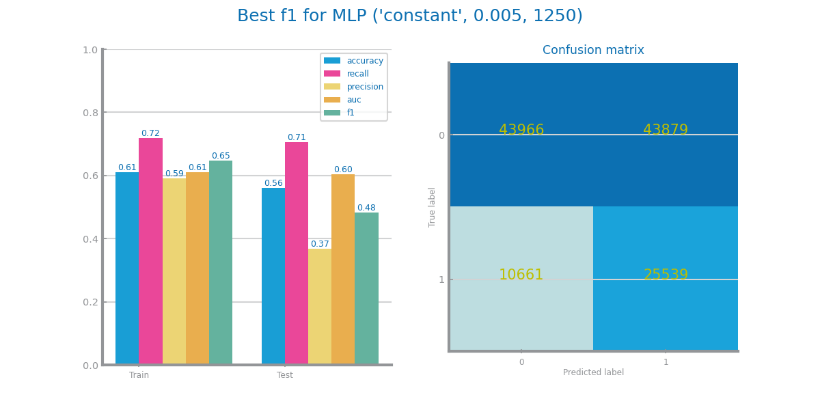
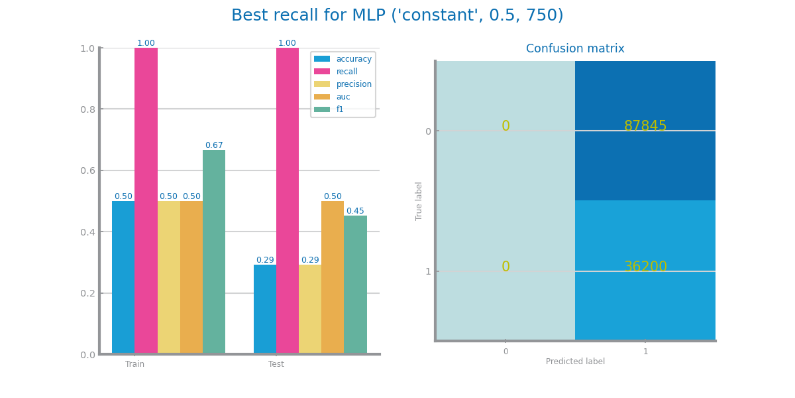
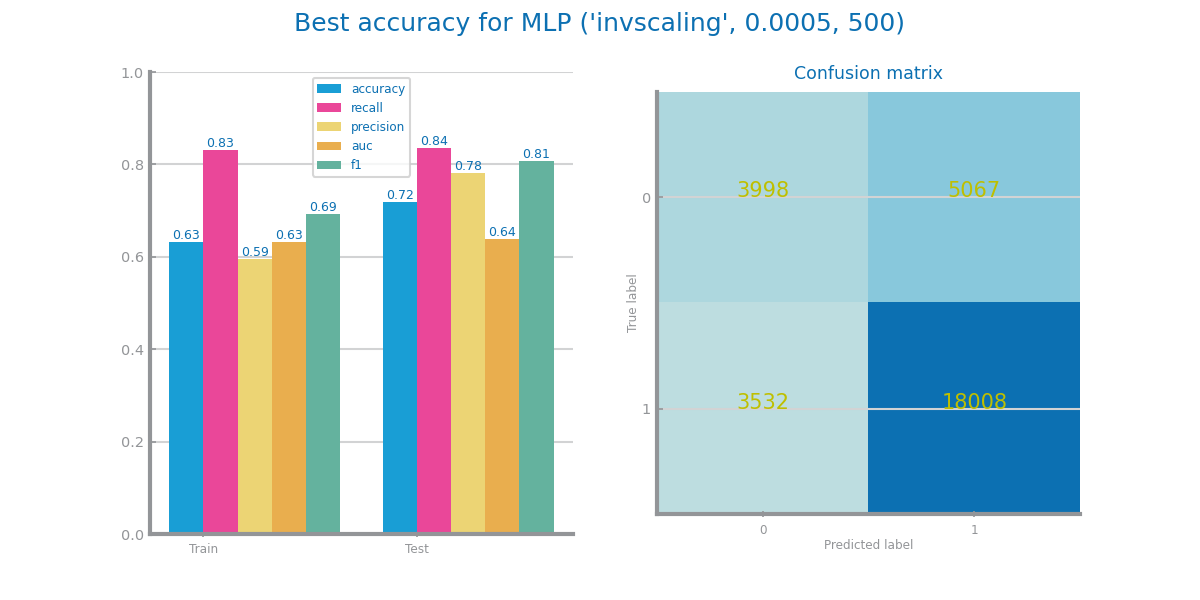


Figure MLP best model results for dataset 1 (left) and dataset 2 (right)

# Critical Analysis

Shall be used to present a summary of the results achieved with the different modeling 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 Forecasting

# Data Profiling

## Data Dimensionality and Granularity

May be used to identify the most atomic granularity and two other different granularities to consider. **Shall not exceed 500 characters.**

Figure Original time series 1 (the most atomic detail)

Figure Time series 1 at the second chosen granularity

Figure Time series 1 at the third chosen granularity

Figure Original time series 2 (the most atomic detail)

Figure Time series 2 at the second chosen granularity

Figure Time series 2 at the third chosen granularity

## Data Distribution

Shall be used to perform the data analysis at those three different granularities, concerning the series distribution. **Shall not exceed 500 characters.**

Figure Boxplots for time series 1 at different granularities

Figure Boxplots for time series 2 at different granularities

Figure Histograms for time series 1 at different granularities

Figure Histograms for time series 2 at different granularities

Figure Autocorrelation lag-plots for original time series 1

Figure Autocorrelation lag-plots for original time series 2

Figure Autocorrelation correlogram for original time series 1

Figure Autocorrelation correlogram for original time series 2

## Data Stationarity

Shall be used to perform the stationarity analysis at those three different granularities. **Shall not exceed 500 characters.**

Figure Components study for time series 1

Figure Stationarity study for time series 1

Figure Components study for time series 2

Figure Stationarity study for time series 2

# Data Transformation

## Aggregation

Shall describe the results of applying three different aggregations over both datasets and identifying the granularity chosen to proceed, based on the performance of Linear Regression models trained over the different time series. **Shall not exceed 300 characters.**

Figure Forecasting plots after different aggregations on time series 1

Figure Forecasting results after different aggregations on time series 1

Figure Forecasting plots after different aggregations on time series 2

Figure Forecasting results after different aggregations on time series 2

## Smoothing

Shall describe the results of applying different smoothing transformations over both datasets and identifying the best result to proceed, based on the performance of Linear Regression models trained over the different time series. **Shall not exceed 300 characters.**

Figure Forecasting plots after different smoothing parameterisations on time series 1

Figure Forecasting results after different smoothing parameterisations on time series 1

Figure Forecasting plots after different smoothing parameterisations on time series 2

Figure Forecasting results after different smoothing parameterisations on time series 2

## Differentiation

Shall describe the results of applying two consecutive differentiations of both datasets and identifying the best result to proceed, based on the performance of Linear Regression models trained over the different time series. **Shall not exceed 300 characters.**

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

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

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

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

## Other transformations (optional)

Shall describe the results of applying other transformations over both datasets and identifying the best result to proceed, based on the performance of Linear Regression models trained over the different time series. **Shall not exceed 500 characters.**

Figure Forecasting plots after applying other transformations over time series 1

Figure Forecasting results after applying other transformations over time series 1

Figure Forecasting plots after applying other transformations over time series 2

Figure Forecasting results after applying other transformations over time series 2

# Models’ Evaluation

Shall be used to summarise the transformations done over the original time series and state which measure will be maximized. **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 Forecasting plots obtained with Simple Average model over time series 1

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

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

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

## Persistence Model

Shall be used to present the results achieved through the persistence model, using both evaluations (long term and one-step behind). **Shall not exceed 500 characters.**

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

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

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

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

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

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

## Rolling Mean Model

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

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

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

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

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

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

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

## ARIMA Model

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

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

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

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

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

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

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

## LSTMs Model

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

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

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

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

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

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

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

# 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.**