CD Project: 2020-2021, Group 30

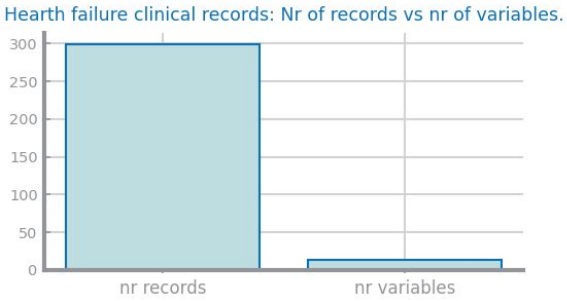
Pedro Alexandre Lopes Correia de Campos, 83951
Pedro de Oliveira Rosa Alves Leitão, 90764
Tomás da Silva Faria Lampreia Gomes, 90782

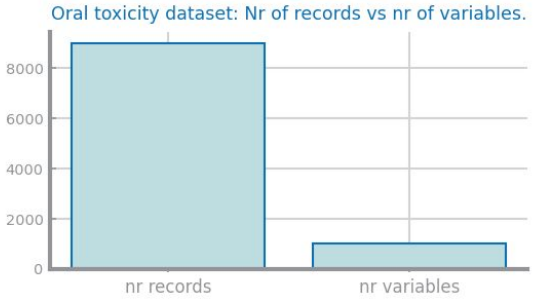
1 Data Profiling

1.1 Data Dimensionality

The Heart Failure Clinical Records dataset has 299 records and a dimensionality of 13 variables. This dataset has no missing values and all his variables are numeric.

The Oral Toxicity dataset has 8991 records and a dimensionality of 1025 variables. Just like the previous one, this dataset has no missing values, but has one symbolic variable (the last variable) and the rest of the variables are numeric.





1.2 Data Distribution

We generated the boxplots for the non-binary variables of the Heart Failure Clinical Records dataset, because the boxplots of the binary variables didn't show usefull information.

From the boxplots generated we can obtain the following values:

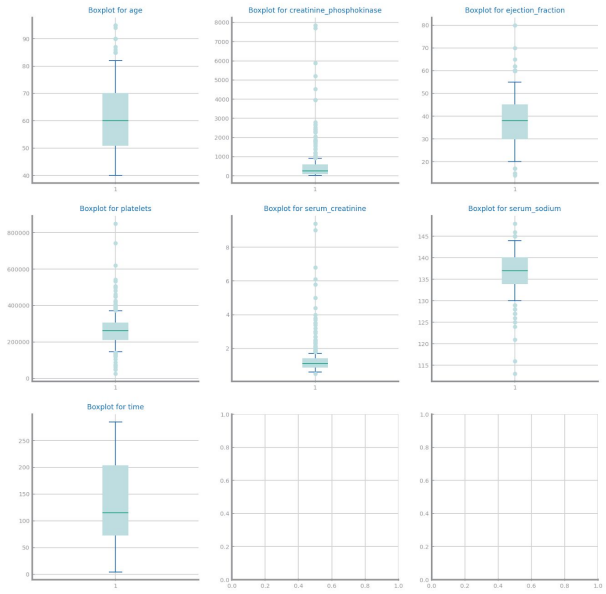
The Outliers for age are all equal or higher than 85 years old. For creatinine_phosphokinase are all equal or higher than 935.9 mcg/L. For ejection_fraction are either equal or lower than 17.00% or equal or higher than 60.02%. For platelets are either lower than 140910 kiloplatelets/mL or equal or higher than 373700 kiloplatelets/mL. For serum_creatinine are either 0.50 mg/dL or equal or higher than 1.80 mg/dL. For serum_sodium are either equal or lower than 129.00 mEq/L or equal or higher than 145.03 mEq/L. The time variable has no Outliers.

We generated barcharts for the binary variables of the Heart Failure Clinical Records dataset and histograms following a Poisson probability distribution for the non-binary variables of the same dataset.

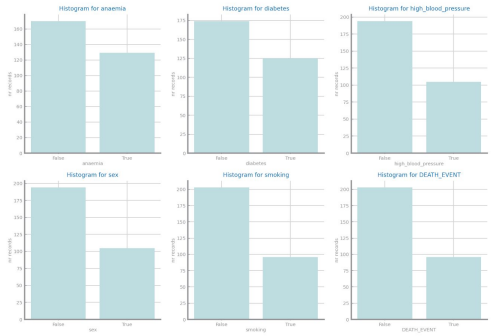
From the barcharts, we can conclude that, from all the records, 170 people don't have anaemia while 129 have. 174 people don't have diabetes while 125 have. 194 people don't have high blood pressure while 105

have. 194 people are a man (True in the barchart) while 105 are a woman (False). 203 people don't smoke while 96 people smoke. And finally, 203 people didn't die during the follow-up period while 96 died.

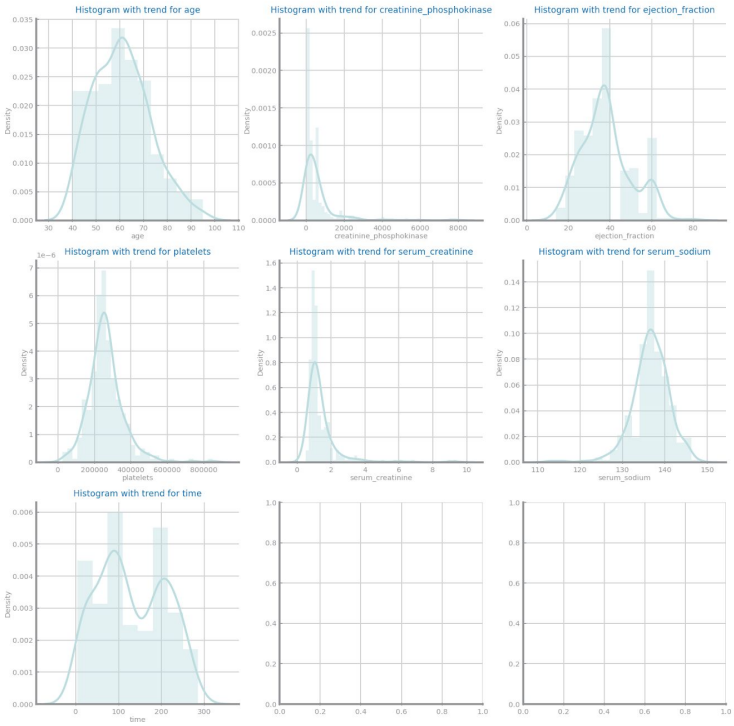
For the Oral Toxicity dataset, we generated two barcharts, one for the distribution of the negative and positive records from the last variable, showing that there are 8250 negative values and 741 positive values. And other for the medium of 1s for both positive and negative values, showing that the medium of 1s of the positive values is 91.40, while the medium of 1s for the negative values is 95.65.



Boxplot of the Heart Failure Clinical Records dataset



Histograms of the distribution of the binary values of the Heart Failure Clinical Records dataset



Histograms of the distribution of the quantitative values of the Heart Failure Clinical Records dataset

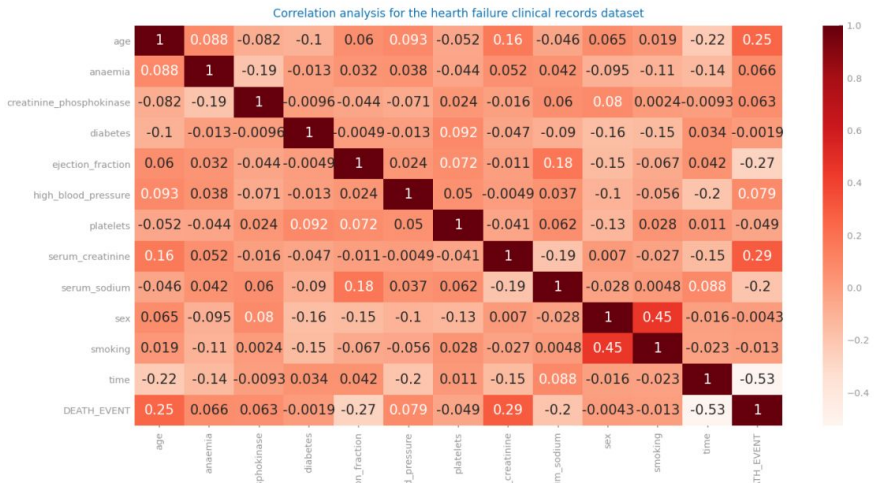
1.3 Data Granularity

Since all our variables in both datasets are numeric (excluding the last variable in the Oral Toxicity dataset) we discretized the data into intervals based on the number of different records per variable. In the histograms shown, we set the bins with values of 10, 100, 1000 or 10000 equivalent to the number of different records per variable.

1.4 Data Sparsity

We generated the scatter plot of the non-binary variables of the Heart Failure Clinical Records dataset (we didn't use the binary variables because their records are less useful to see the sparsity). In the scatter plot we were able to conclude that this dataset is not sparse. Because this dataset has a large dimensionality, we found it better to see the Correlation among variables via the Correlation Analysis Heatmap, where we can see that the highest correlation is 0.45, between sex and smoking, while the lowest is -0.53, between time and DEATH_EVENT, so we concluded that there are no dependencies between variables in this dataset.

We didn't plot any graphics for the Oral Toxicity dataset because the number of variables was too much to generate the graphics.



2 Data Preparation

2.1 Imputations

We didn't need to do any missing values imputation since there are no missing values in any of the two datasets. For the outliers, we removed all the records containing outliers in the Heart Failure Clinical Records dataset, removing 75 records (meaning that out of the 299 records on this dataset, 75 had outliers). It is important to note that because the numeric variables of the Oral Toxicity dataset are all binary, this dataset has no outliers, therefore we didn't need to do an imputation of the outliers in this dataset.

2.2 Scaling

Because the numerical variables from the Oral Toxicity dataset were all binary, we didn't need to scale it.

To scale the Heart Failure Clinical Records dataset, we chose the MinMax normalization, because it was the one that provided a better understanding of the data.

2.3 Data Balancing

Both datasets revealed to be quite unbalanced, especially the Oral Toxicity dataset, with an approximate ratio of 1 positive to 11 negative records (741 to 8250).

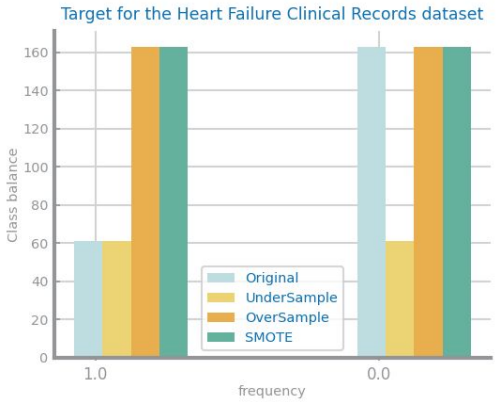
After applying different resampling techniques (UnderSampling, OverSampling and SMOTE), we reached the conclusion that OverSampling was the best resampling technique for the Heart Failure Clinical Records dataset, because this dataset has few records. Meanwhile, for us, UnderSampling was the best for the Oral Toxicity dataset, because this one has too many records, so we concluded that this resampling technique would perform the best for the target variable of the Oral Toxicity dataset.

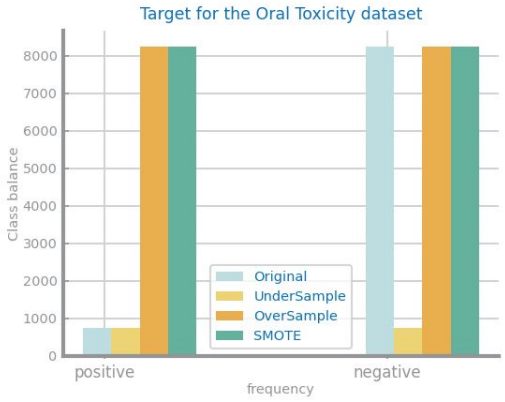
2.4 Feature Engineering

2.4.1 Feature Selection

By checking the heatmap of the Correlation Analysis of the Heart Failure Clinical Records dataset, we saw that the highest correlation was 0.45, so we decided that we wouldn't select any variable to remove.

For the Oral Toxicity dataset, we decided to remove every variable that had a correlation higher than 0.8, ending up removing 64 variables (the dataset ended with 960 numeric variables).





2.4.2 Feature Generation

For the Heart Failure Clinical Records dataset, we created a new variable, called platelets_ejection_fraction (that represents the quantity of platelets in the blood leaving the heart at each contraction), where every record is calculated by the product between the platelets and the ejection_fraction variables in that record.

We didn't apply Feature Generation in the Oral Toxicity dataset because it would increase the number of variables and we already have too many variables in that dataset.

3 Unsupervised Learning

3.1 Association Rules

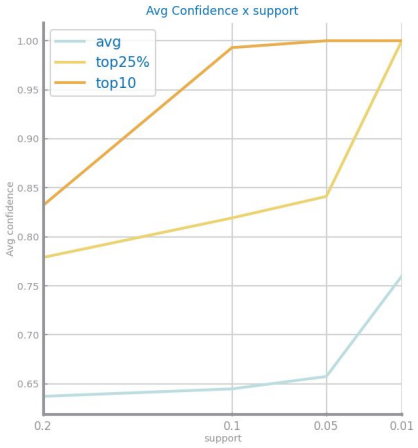
We did the Pattern Mining only for the Heart Failure Clinical Records dataset, since the Oral Toxicity was taking too much time to compute.

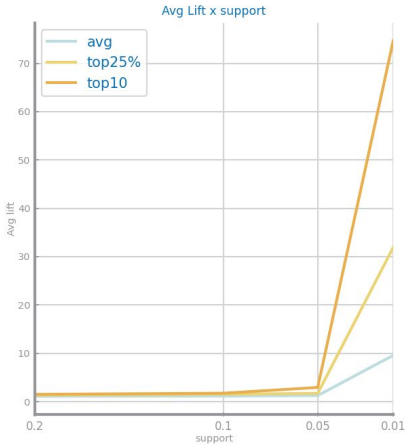
By checking the graphics generated from the Heart Failure Clinical Records dataset, from the graphic of the Number of Patterns as a function of Support, we concluded that the dataset has almost zero combinations of values with a support of 0.2, increasing that number a little bit from 0.2 to 0.1 of support, and increasing more from 0.1 to 0.05 of support, having the biggest increase starting at 0.05, and so we can conclude that there are 6373 combinations of values that are present in at least 1% of the records (the support of 0.01 that we defined as minimum support).

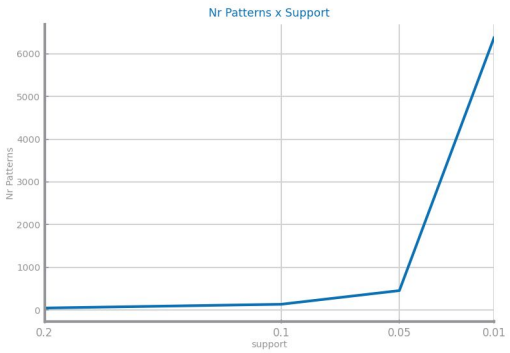
The top 10 combinations with the highest confidence for the minimum support defined (0.01) have all 100% of confidence, being 9 of them around the age conditioning different variables, so we can't conclude anything yet.

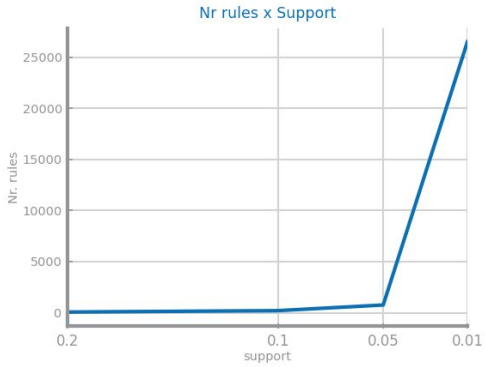
When looking at the average lift and support, the average, top 10 and top 25% were low and stable up until the support reached values lower or equal to 0.05. where we noticed a big increase in the slope, as the average lift rapidly increased.

The graphic of the Number of Rules as a function of Support also has almost zero values when the support is higher than 0.2, also increasing the value from 0.2 to 0.1, increasing more from 0.1 to 0.05 and increasing the most from 0.05 to 0.01 support, so we can conclude that there are 26490 rules that are present in at least 1% of the records.









3.2 **Clustering**

We only analysed the graphics of the Clustering for the Heart Failure Clinical Records dataset, because the Oral Toxicity dataset's numeric variables are all binary, so their graphics didn't show useful information. The variables we found to better give clustering information were serum_creatinine (var 7) and serum_sodium (var 8).

There are 4 kinds of Clustering approaches: Hierarchical, Partition-Based, Model-Based and Density-Based.

With the Hierarchical approach using MSE, it was hard to apply the elbow method, as there wasn't a clear turning point in the curvature of the graph, but we're inclined to believe it was at around 5 clusters (k=5), while with SC the best k value was 25, as it had the highest silhouette score. Regarding metrics, euclidean had the highest silhouette score when looking at complete links, whereas cityblock prevailed in average links.

When it comes to the Partition-Based approach (K-means), we faced the same problem applying the elbow method, but 11 appeared to be a good value with MSE, and 21 with SC.

For Model-Based (EM), k=19 seemed to us to be the best value for both MSE and SC.

Lastly, following the Density-Based approach (DBSCAN), we weren't able to correctly plot the DBSCAN MSE and DBSCAN SC graphs, but the chebyshev metric presented the highest SC.

4 Training Strategies

For the Heart Failure Clinical Records dataset, since the number of records is low, we used cross-validation with n/2 splits. On the other hand, for the Oral Toxicity dataset, we decided to use hold-out, with a 70/100 and 30/100 split, since this dataset has thousands of records. We tried to use leave-one-out for the Heart Failure Clinical Records dataset, but didn't manage to do so without errors.

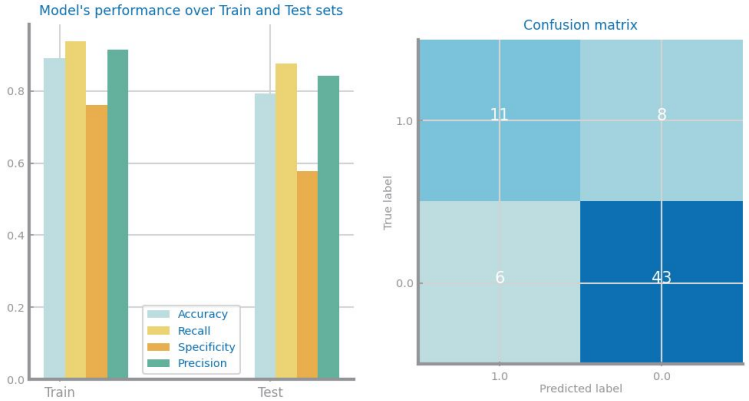
5 Classification

5.1 Naive Bayes

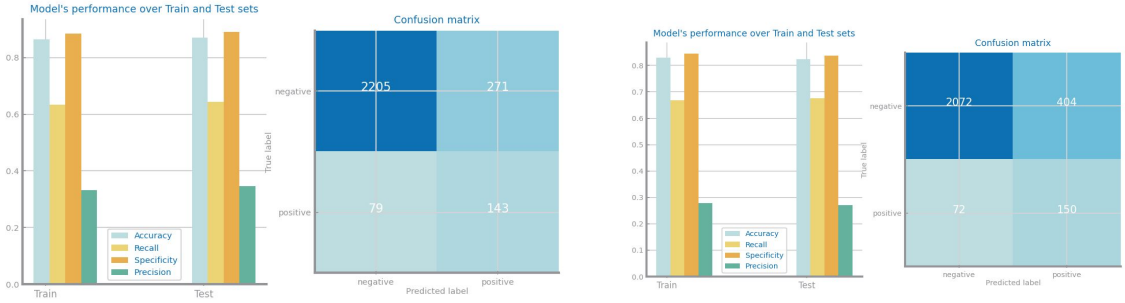
There are three models of Naive Bayes classifiers: Gaussian, Multinomial and Bernoulli.

For the Heart Failure Clinical Records dataset, since our dataset has a majority of variables with normal distribution, we used the Gaussian model.

For the Oral Toxicity dataset, since we have a binary distribution for all features, the most suitable model would be the Bernouly. However the Multinomial appears to show better results for our data (separated by train and test).



Heart Failure Clinical Records: Naive bayes using Gaussian



Oral Toxicity: Naive bayes using Multinomial

Oral Toxicity: Naive bayes using Bernoulli

5.2 KNN

To apply the KNN classifier, we have three different ways to measure distance: Manhattan, Euclidean and Chebyshev.

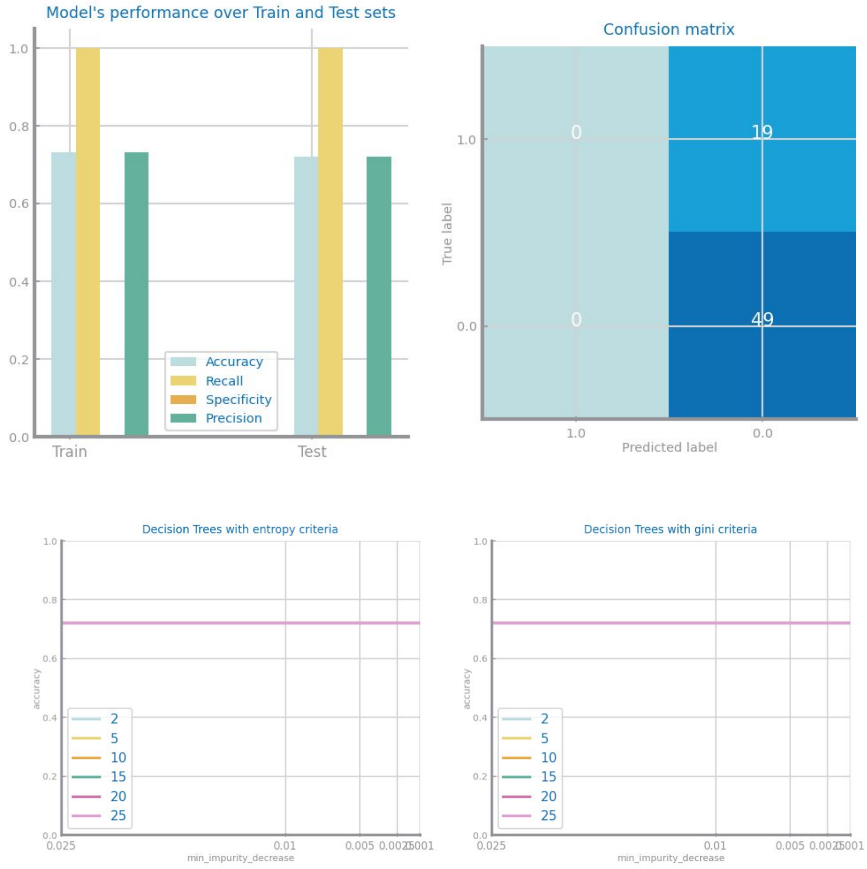
For the Heart Failure Clinical Records dataset, by analyzing the graph that compares these three different types, we can conclude that the best results are obtained when Manhattan is used.

5.3 Decision Trees

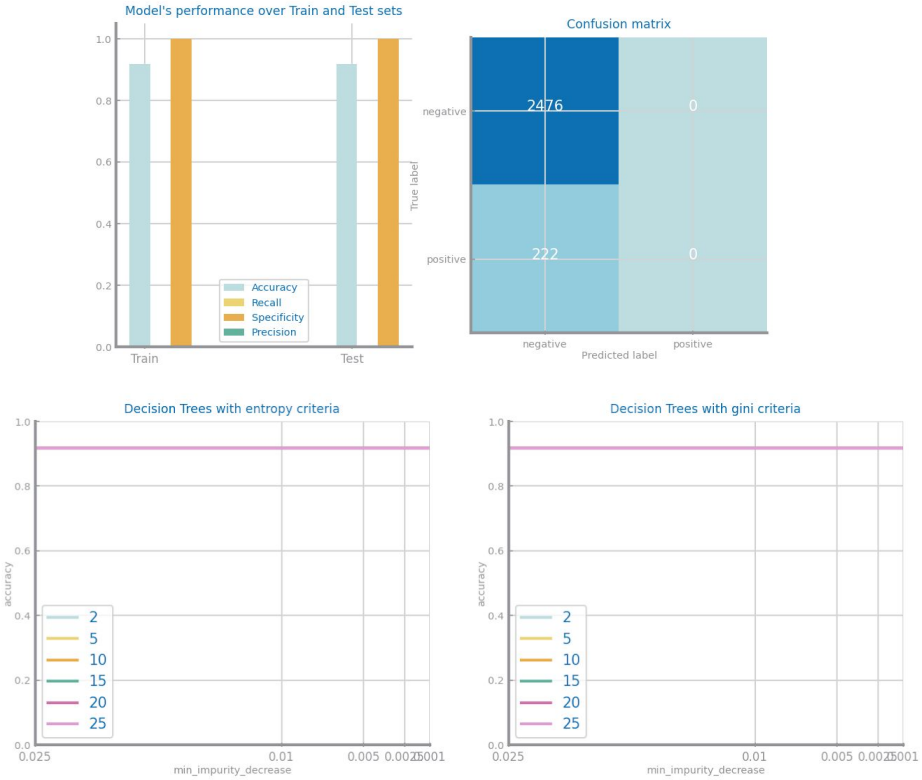
For Heart Failure Clinical Records dataset: the best results are achieved with entropy criteria and min_impurity_decrease=0.03 ==> accuracy=0.7.

For the Oral Toxicity dataset, the best results are achieved with entropy criteria and min_impurity_decrease=0.03 ==> accuracy=0.92.

In both, the depth doesn't seem to influence the result. Still, it give us a depth of 2 in both datasets.



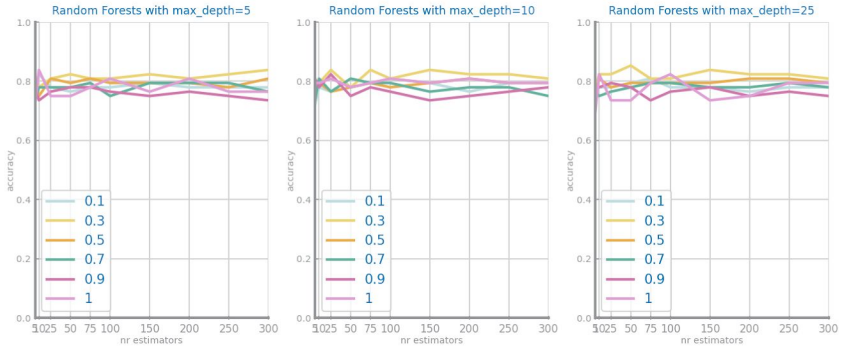
Graphs related to the decision tree do Heart Failure Clinical Records dataset.



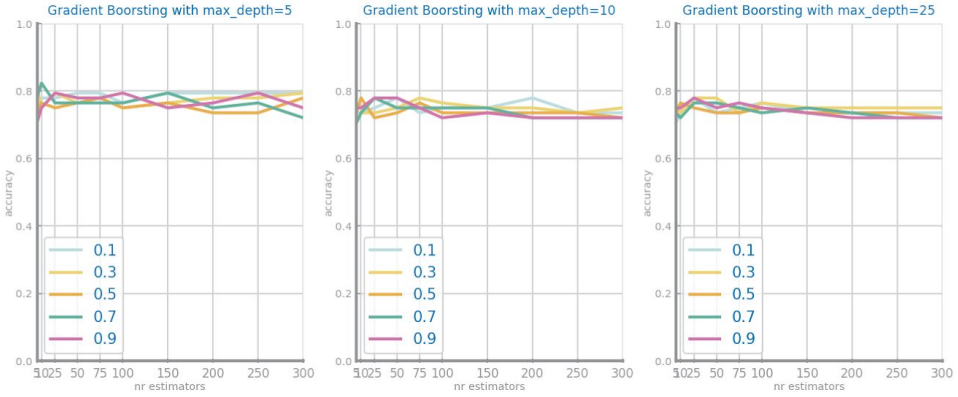
Graphs related to the decision tree do Oral Toxicity dataset.

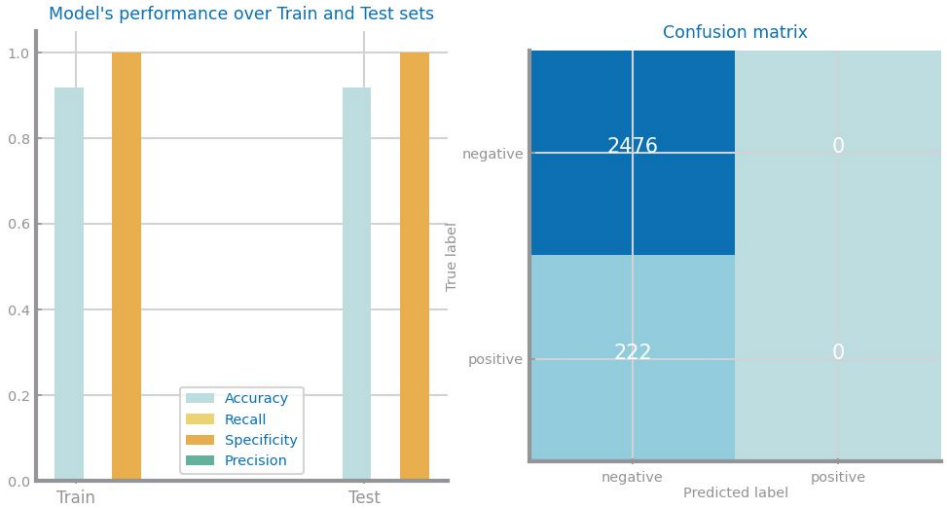
5.4 **Random Forests**

We only generated the graphics of the Random Forests for the Heart Failure Clinical Records dataset, because the Oral Toxicity dataset took too long to generate. By analysing the Model's performance graphic and the Confusion matrix, we concluded that the model had maximum values in training phase, but lower in test phase.



5.5 **Gradient Boosting**





Graphs related to the Gradient Boosting do Heart Failure Clinical Records dataset.