# Data Science

*by Cláudia Antunes*

# Lab Models Evaluation

## A. Exam 2021-01-19

Consider the problem of diagnosing arrhythmia in patients, through the use of a dataset with 452 medical records, described by 250 variables. One of these variables, call it Z, contains the type of arrhythmia detected in each positive patient, and 0 if the problem was not diagnosed. From it, the variable class was derived assuming the value regular whenever Z=0 (245) and abnormal (207) otherwise. Consider the original dataset and the presented tree and the chart on Figure 2, reporting the accuracy and recall collected for different decision trees, trained with some algorithm with different pre-pruning requirements based on the maximum depth of the trees learned.

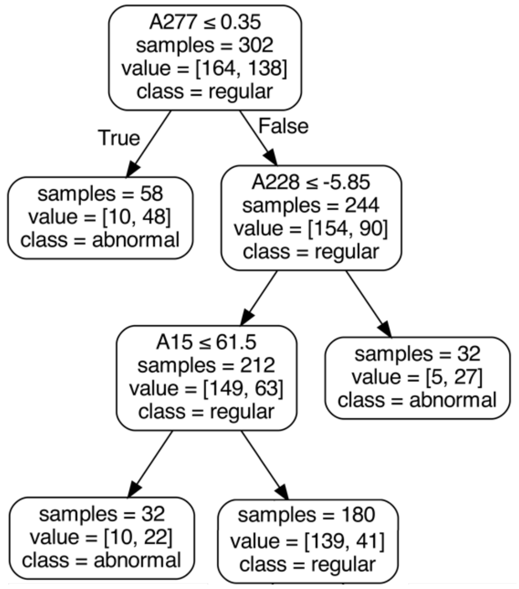

*Figure 1 Decision Tree trained over 302 records*

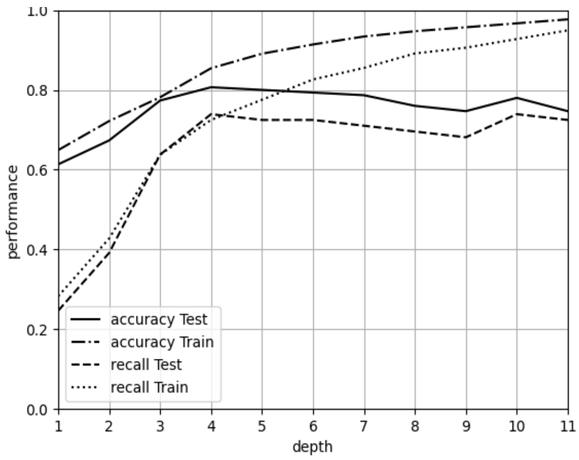

*Figure 2 Performance of different decision trees specializations*

1. The number of True Positives is higher than the number of True Negatives for the presented tree.
2. The number of False Positives reported in the same tree is 25.
3. The recall for the same tree is less than 70%
4. We are able to identify the existence of overfitting for models with less than 4 nodes of depth.
5. The difference between recall and accuracy becomes smaller with the depth due to the overfitting phenomenon.

## B. Exam 2021-02-05

Consider the problem of predicting if some patient will survive, through the use of a dataset with 165 medical records, described by 50 variables. From these the `class` variable has two possible values `survive` (102) and `die` (63). The tree on the left was learned through the C4.5 algorithm and the information gain criteria, when applied over 100 of the 165 records available, to learn the target variable `Class`, after applying some preparation techniques. The tree was printed through `sklearn.tree` package. Consider 5 new variables computed as follows: A=(A27<=0.5), B=(A32<=13.45), C=(A39<=57), D=(A39<=167.5) and E=(A24<=76.5).

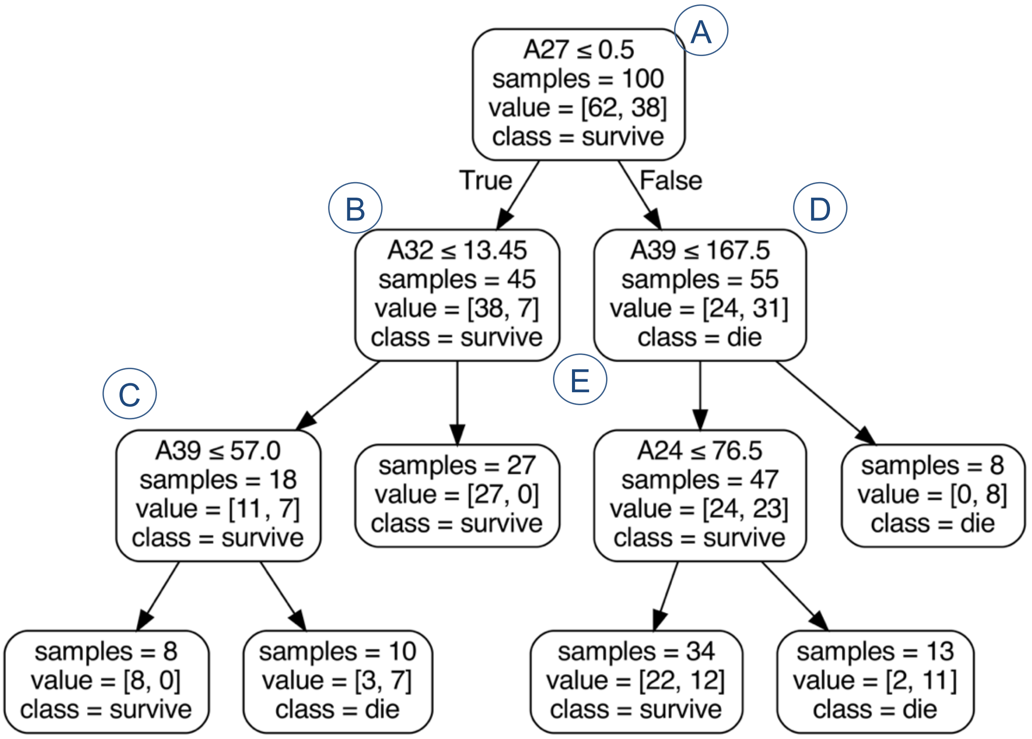

*Figure 3 Decision tree trained over 100 records*

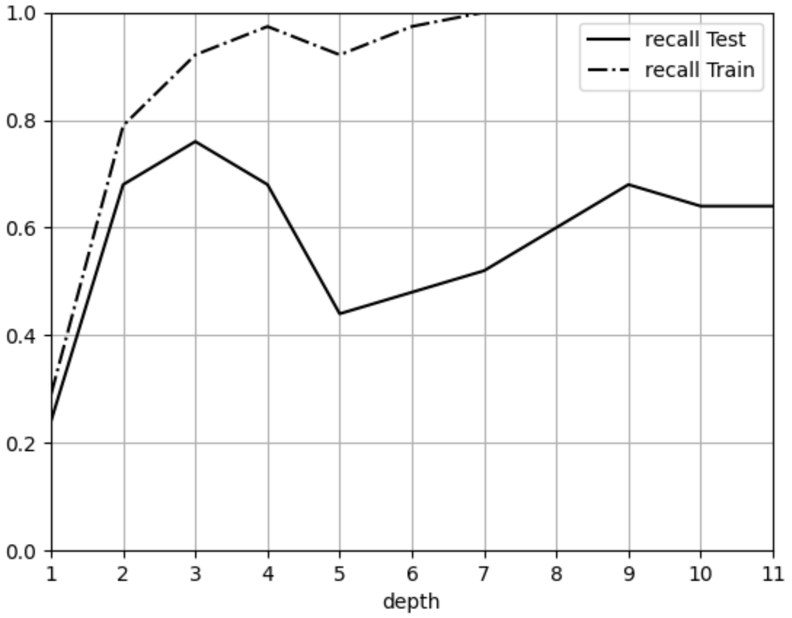

*Figure 4 Recall for different decision trees specializations*

1. The accuracy for the tree is 62%.
2. As reported in the tree, the number of False Positive is **smaller** than the number of False Negatives.
3. The recall for the tree is **less than** 75%.
4. The chart reporting the recall for different trees shows that the model **enters** in overfitting for models with **depth higher than 5**.
5. A smaller tree would be delivered if we would apply post-pruning, accepting an accuracy reduction of 5%.

## C. Exam 2022-02-10

Consider a classification task approached through the exploration of a dataset with 500 records, described by 12 variables. From these the `class` variable has two possible values `Pos` (100) and `Neg` (400). The tree below was **learned through** the C4.5 algorithm and the information gain criteria, when applied over **200** of the 500 records available, to learn the target variable `class`, after applying some preparation techniques.

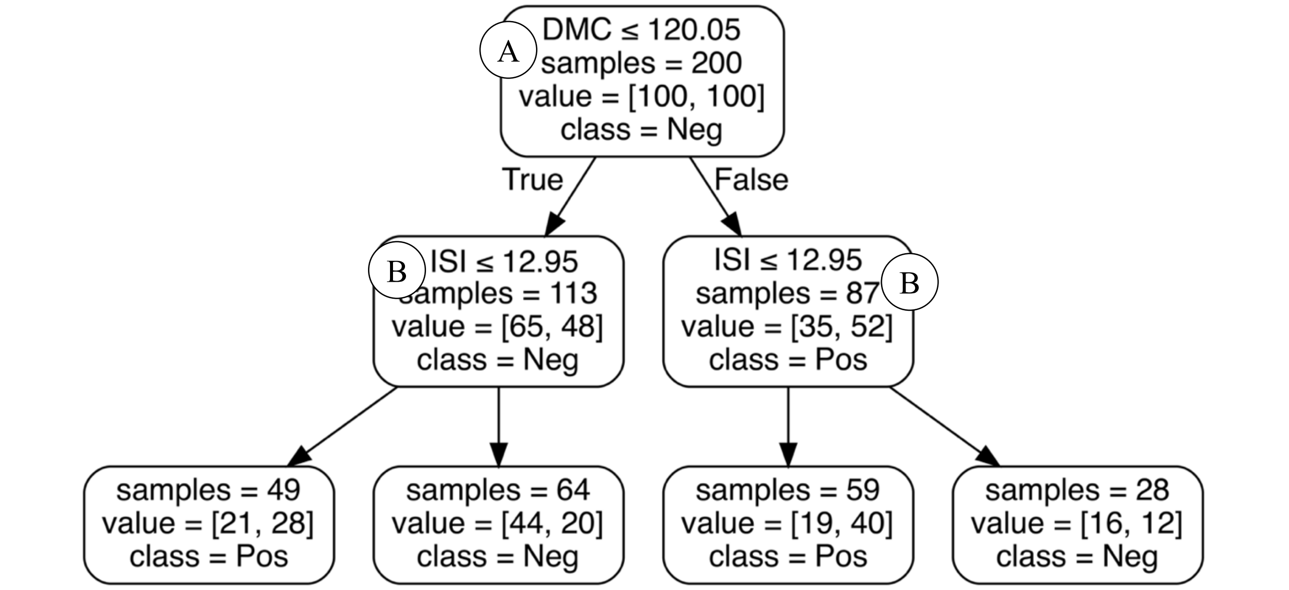

*Figure 5 Decision tree trained over 200 records*

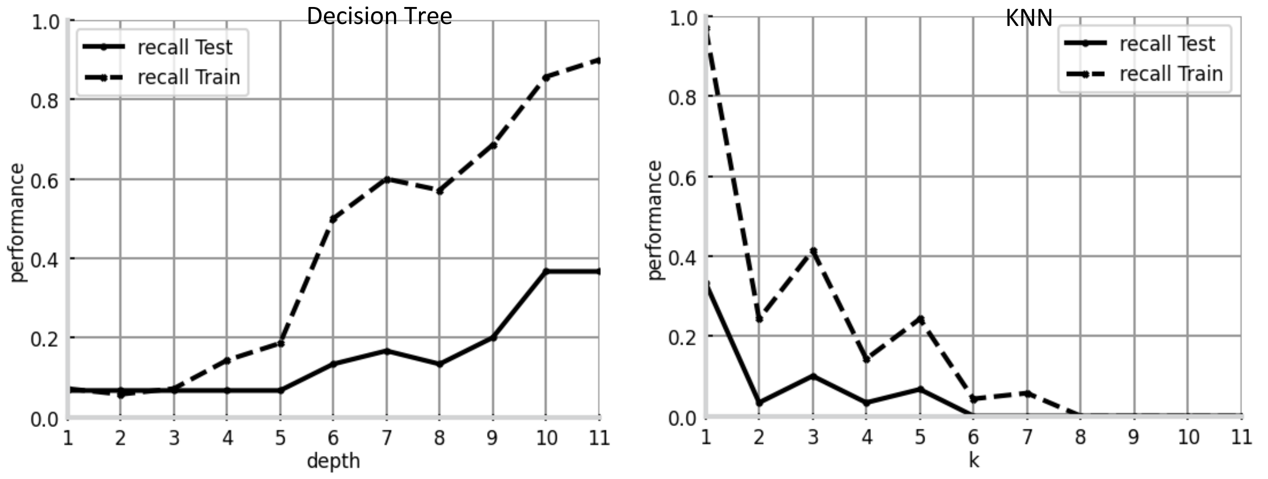

*Figure 6 Recall for different decision trees specializations (left) and different instantiations of KNN (right)*

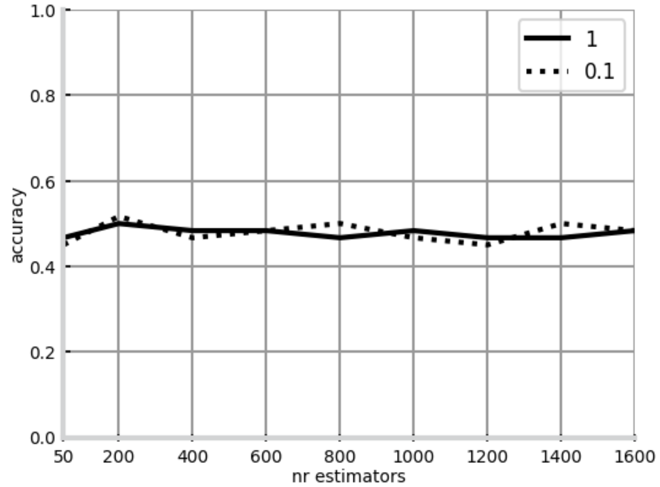

*Figure 7 Accuracy for random forests trained over the balanced dataset with 200 records, all under the same conditions, but the number of estimators and the number of features to consider when looking for the best split*

1. The accuracy for the presented tree is **lower** than 75%.
2. The precision for the presented tree is **higher** than its recall
3. According to the chart on the left, the tree with **6 nodes of depth** is in overfitting.
4. According to the charts, **KNN and Decision Trees present a similar behaviour**.
5. The random forests results shown can be explained by the lack of diversity resulting from the number of features considered.

## D. Exam 2022-02-26

Consider a classification task, whose goal is to determine a survival model. The task was approached through the exploration of a dataset with **1000 records**, described by **16 variables**. From these the `class` variable represents survival, and it has two possible values `Yes` (400) and `No` (600). The tree below was **learned through** the C4.5 algorithm and the information gain criteria, when applied over **all the 1000** records available, to learn the target variable `class`, after applying some preparation techniques. Consider the binarized dataset just described by *A* and *B*, where *A=True* ⇔ *ratio≤0.2* and *B=True* ⇔ *relapse≤0.5*.

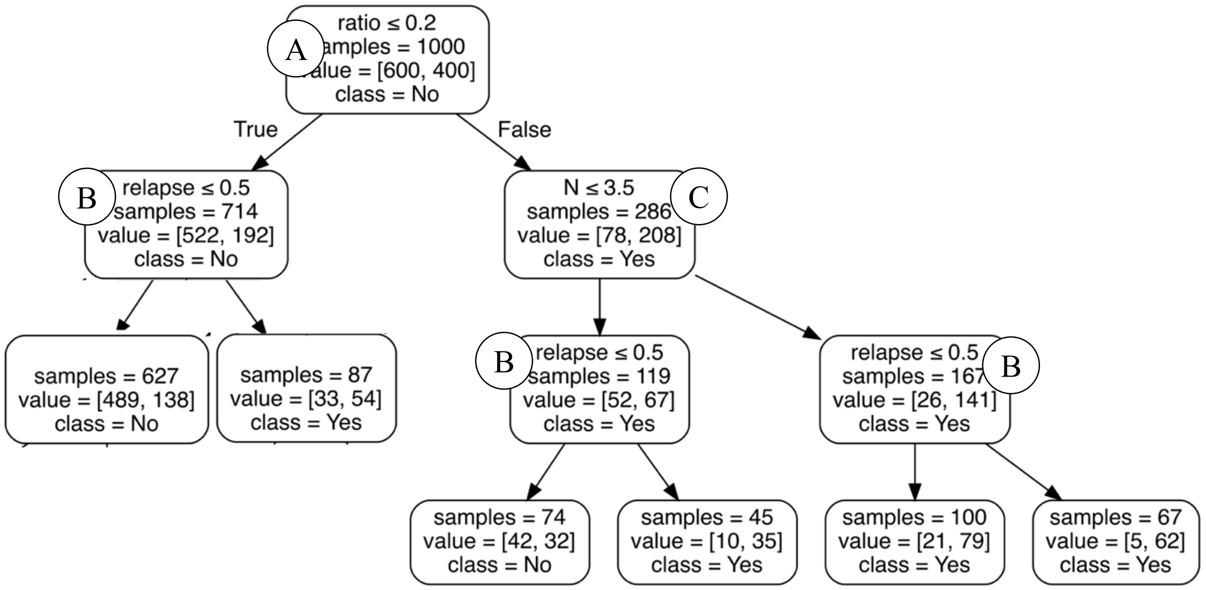

*Figure 8 Decision tree trained over all records*

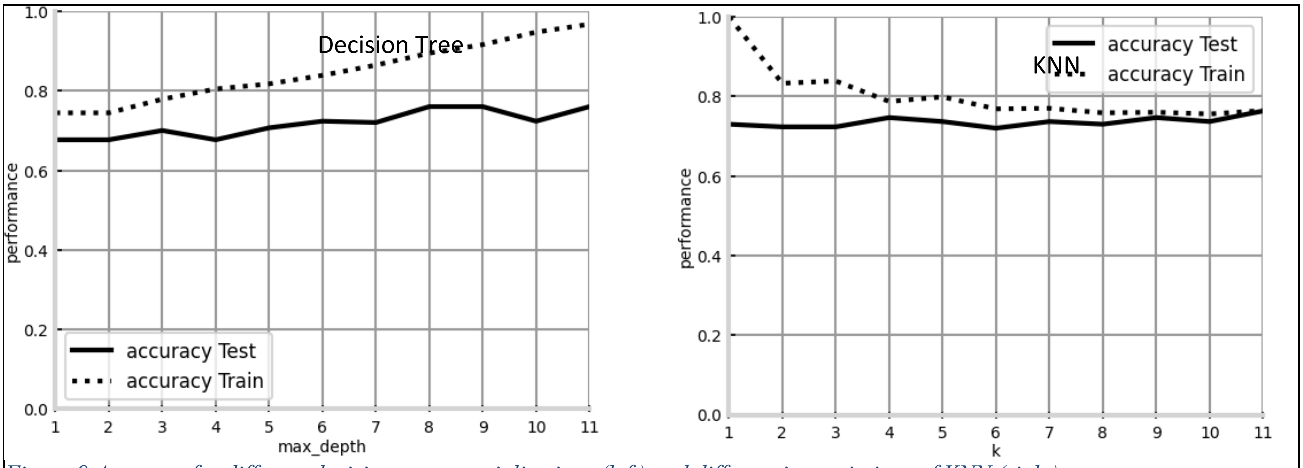

*Figure 9 Accuracy for different decision trees specializations (left) and different instantiations of KNN (right)*

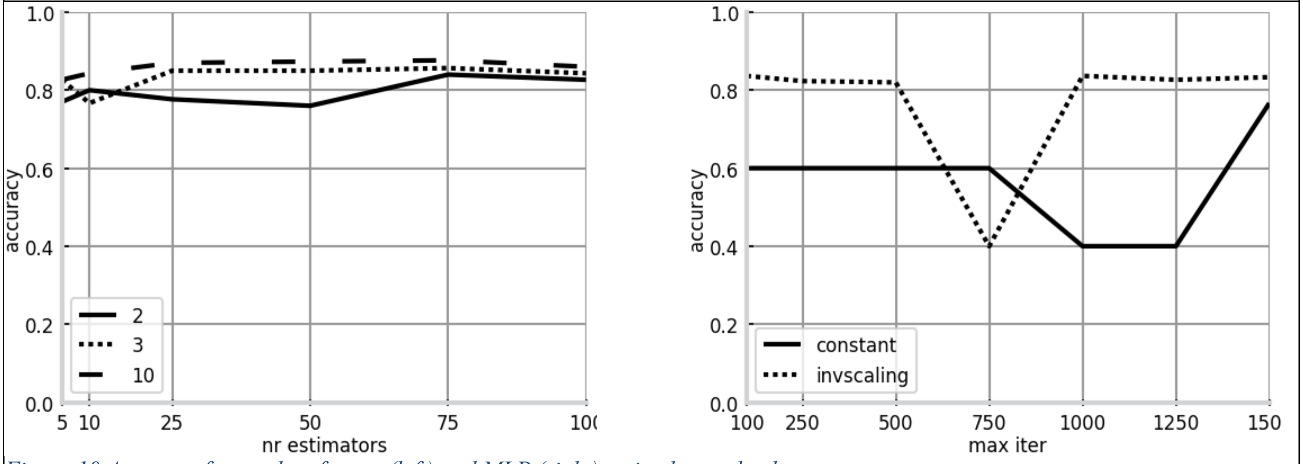

*Figure 10 Accuracy for random forests (left) and MLP (right) trained over the dataset*

1. The accuracy for the presented tree is **lower** than 75%.
2. The recall for the presented tree is **higher** than its precision.
3. **KNN** with **5 neighbours** is in overfitting.
4. **KNN** and **Decision Trees** show a similar trend.
5. Results for **Random Forests** identified as **2**, may be explained by its estimators being in *underfitting*.

# E. Exam 2023-01-23

Consider a classification task, whose goal is to determine if some patient will make an insurance claim above a threshold (`class` variable). The task was approached through the exploration of a dataset with **1200 records**, described by **9 variables**. There are 490 records `Yes` (41%) and 710 records `No` (59%) regarding the `class` variable. The tree below was **learned through** the C4.5 algorithm and the information gain criteria, when applied over **only 1000** records from the available, to learn the target variable `class`, after applying some preparation techniques. Consider the binarized dataset just described by *A* and *B*, where $A=True \Leftrightarrow bloodpressure \leq 95.5$ and $B=True \Leftrightarrow bmi \leq 25.05$.

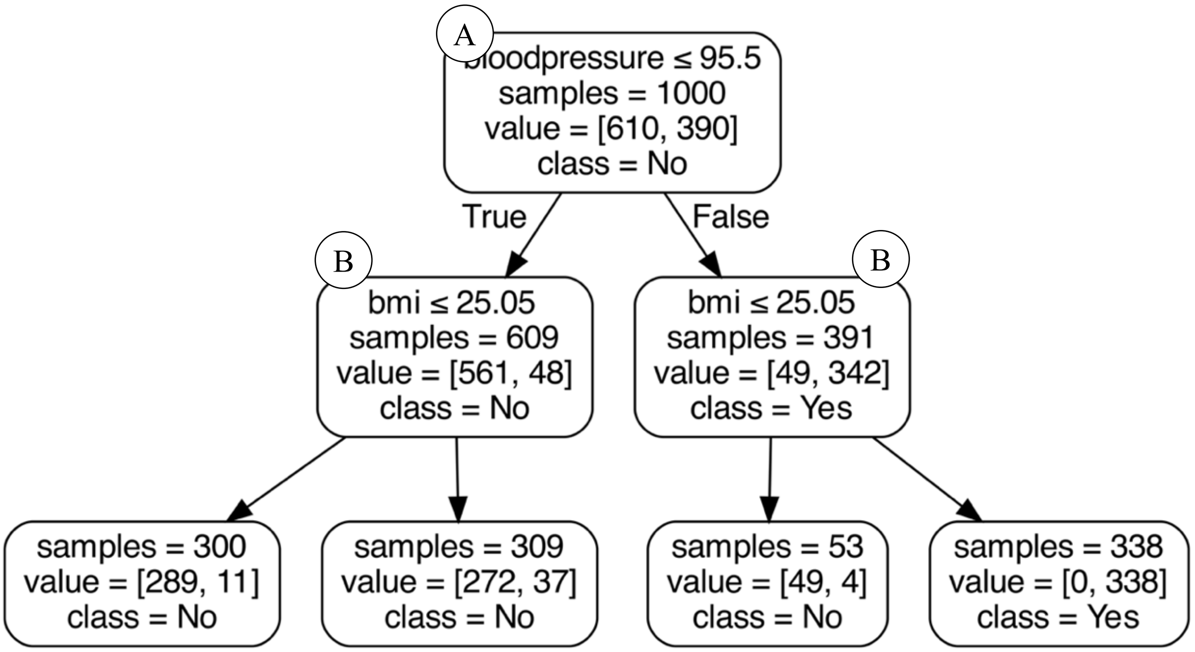

*Figure 11 Decision tree trained over all records*

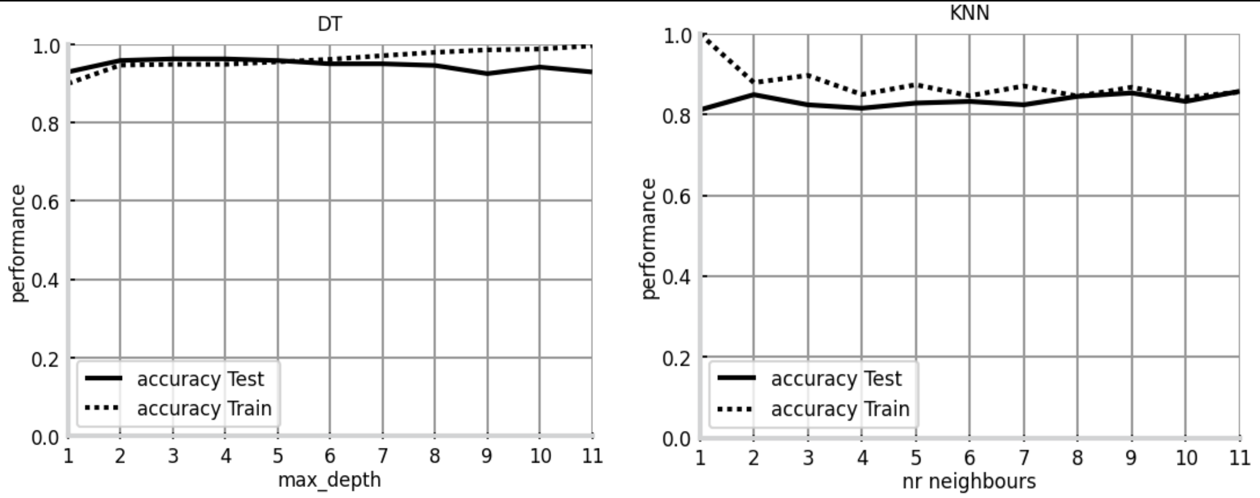

*Figure 12 Accuracy for different decision trees specializations (left) and different instantiations of KNN (right)*

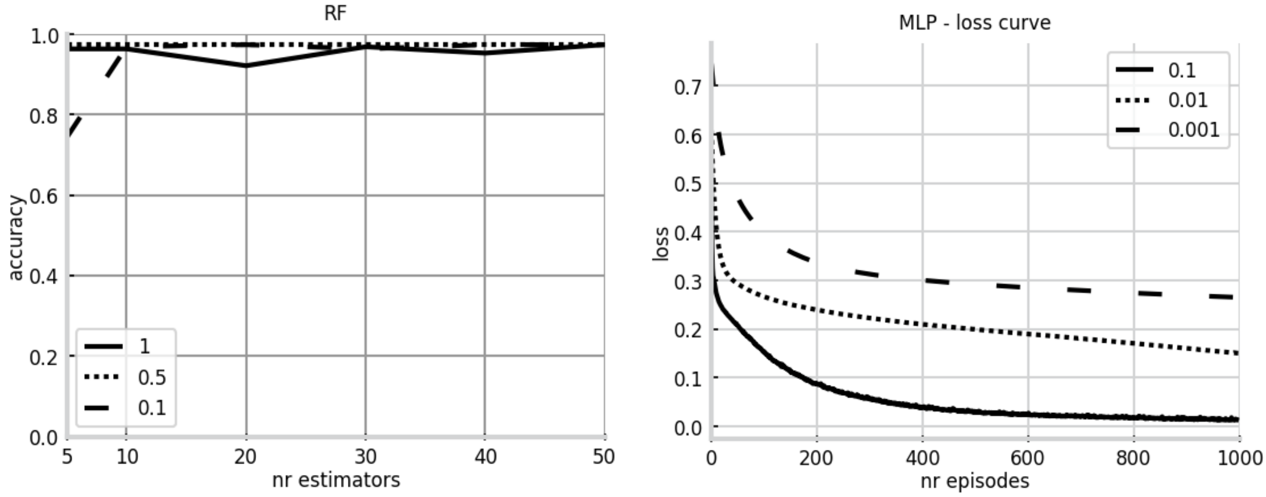

*Figure 13 Accuracy for random forests for multiple features (left) and loss curves for MLP multiple learning rates (right)*

1. The accuracy for the presented tree is **lower** than 75%.
2. The recall for the presented tree is **higher** than its precision.
3. KNN with **1 neighbour** is in **overfitting**.
4. KNN and Decision Trees show a **different** trend in the majority of hyperparameters tested.
5. The best MLP model learnt is **at least as good as** the best Random Forest model learnt.

## F. Exam 2023-02-10

Consider a classification task, whose goal is to determine if the driver in a tesla car accident will die in the accident (class=`driver_death`). The task was approached through the exploration of a dataset curated from **235 records**, described by **11 variables** plus the class, where there were 96 records `Yes` (41%) and 139 records `No` (59%) regarding the `driver_death` variable. One of the eleven variables available contained a description of the accident, "*car collides with tesla, both drivers die*" is one of the 200 descriptions provided. The tree below was **learned through** the C4.5 algorithm and the information gain criteria, when applied over **the curated dataset**, to learn the target variable `driver_death`. Consider the binarized dataset just described by *A* and *B*, where *A=True ⇔ fire ≤ 0.5* and *B=True ⇔ model ≤ 3.5*.

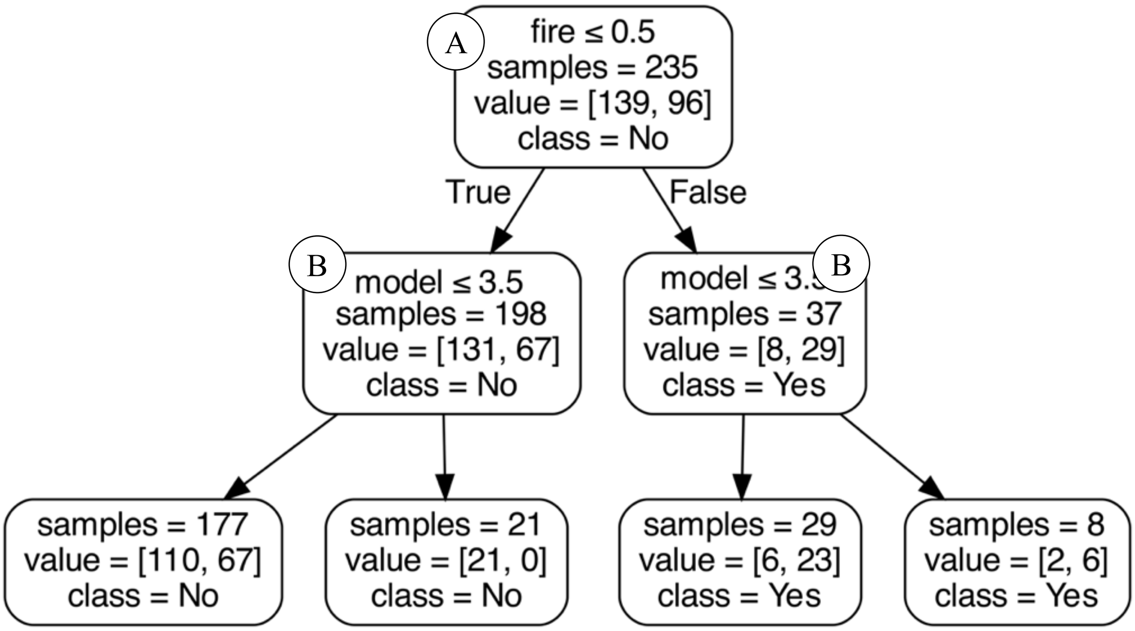

*Figure 14 Decision tree trained over all records*

1. The accuracy for the presented tree is **higher** than 75%.
2. The recall for the presented tree is **higher** than its precision.
3. The decision tree is in **overfitting** for depths above 4.
4. KNN is in **overfitting** for k larger than 5.
5. Decision trees and KNN show similar behaviours according to **Error! Reference source not found.**.