# Homework #2

Matteo Corain S256654

Data Science and Database Technology – A.Y. 2018-19

## 1 Decision tree building

### 1.1 Question #1

After building the decision tree with the given parameters, we have the following results:

a) The attribute deemed to be the most discriminative for class prediction is `node-caps`, which is put in the root of the generated decision tree.
b) The maximum height of the generated decision tree is 7, along the path: `node-caps='no' → irradiat='no' → tumor-size='30-34' → deg-malig='2' → menopause='premeno' → breast-quad='left_up' → 'recurrence-events'` (or `breast-quad='right_low' → 'no-recurrence-events'`).
c) A pure partition can be identified in the generated decision tree, with 25 items labeled as `no-recurrence-events` and 0 items labeled as `recurrence-events`, by following the path: `node-caps='no' → irradiat='no' → tumor-size='10-14'`.

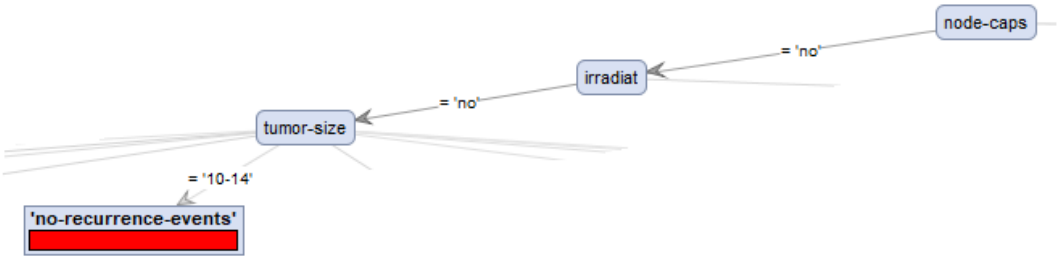


*Figure 1 – Pure partition in the generated decision tree*

### 1.2 Question #2

The minimal gain parameter identifies the minimum value for the gain ratio that the algorithm will consider while deciding to further split or not a node on the decision tree; the higher the minimal gain is, the less the tree will feature multiple paths.

The maximum height parameter identifies instead the maximum length of a path in the generated tree; the higher the minimum height is, the more the tree can potentially grow in height.

Both parameters, therefore, have an impact on the number of attributes considered for building the tree; attributes may in fact not be selected either if their associated gain is too low or if the maximum height has been reached. Decision trees with the following configurations have been considered in the analysis:

| Configuration # | Minimal gain | Maximum height |
|---|---|---|
| **0** | 0.01 | 20 |
| **1** | 0.01 | 5 |
| **2** | 0.04 | 20 |
| **3** | 0.04 | 5 |
| **4** | 0.08 | 20 |
| **5** | 0.08 | 5 |

*Table 1 – Configurations used in the analysis*

#### 1.2.1.1 *Configuration #1*

In configuration #1, the minimal gain value was left at the initial value of 0.01 while the maximum height value was reduced to 5. With respect to the initial configuration, in this case the tree has a lower height, with a number of subtrees condensed into single leaf nodes, due to the reaching of the maximum allowed height.

The following figures show the path `node-caps='no' → irradiat='no' → tumor-size='25-29'` in the original tree and in configuration #1.

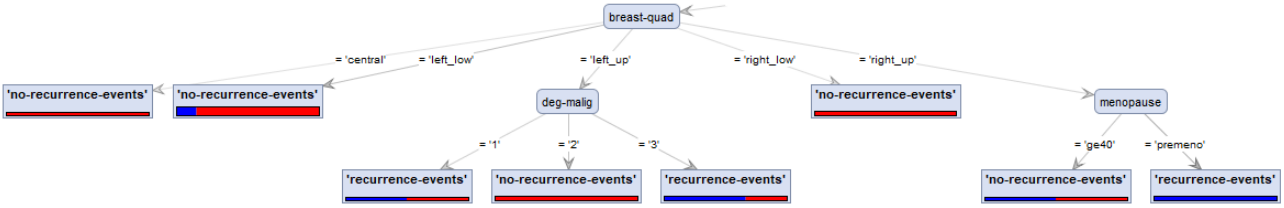


*Figure 2 – Path in the original tree*

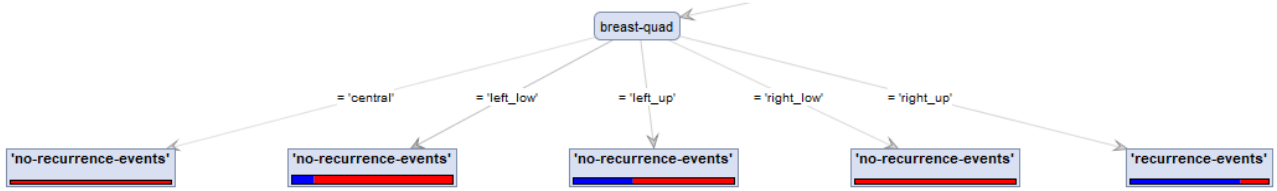


*Figure 3 – Path in the tree with configuration #1*

#### 1.2.1.2 *Configuration #2*

In configuration #2, the minimal gain value was increased to 0.04 while the maximum height value was left at the initial value of 20. With respect to the initial configuration, in this case the tree uses another set of attributes for performing the splits, selecting the ones for which the gain obtained by further splitting is more significant than the new minimal gain value.

The following figures show the uppermost splits in the left subtree of original and configuration #2 trees.

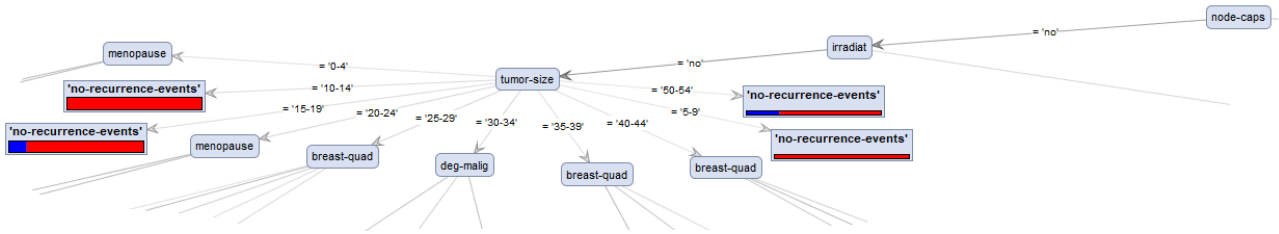


*Figure 4 – Path in the original tree*

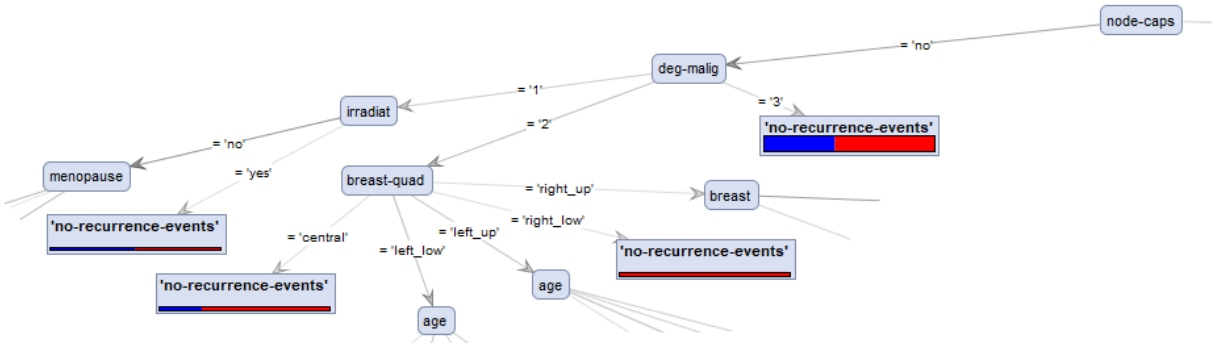


*Figure 5 – Path in the tree with configuration #2*

### *1.2.1.3 Configuration #3*

In configuration #3, the minimal gain value was increased to 0.04 while the maximum height value was reduced to 5. With respect to the previous configuration, a number of subtrees have been condensed into single leaf nodes due to the reaching of the maximum allowed height, thus reducing the overall height.

The following figures show the path `node-caps='no' → deg-malign='2' → breast-quad='left-low'` in configuration #2 and #3.

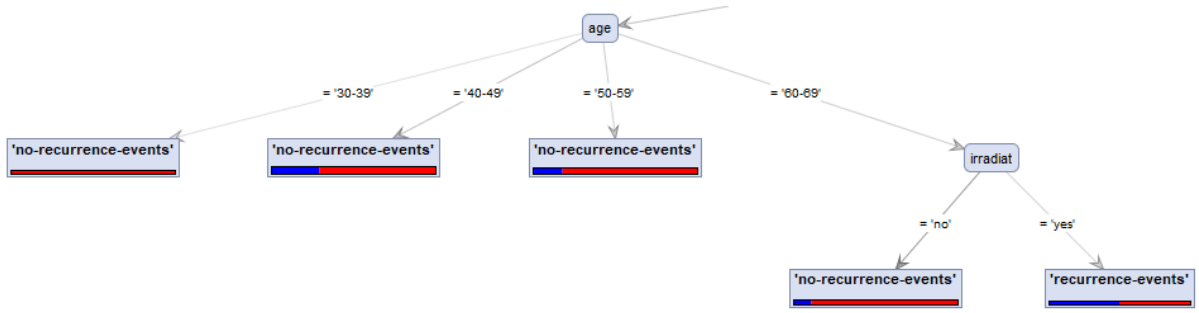


*Figure 6 – Path in the tree with configuration #2*

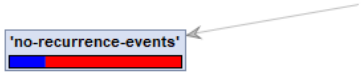


*Figure 7 – Path in the tree with configuration #3*

### *1.2.1.4 Configuration #4*

In configuration #4, the minimal gain value was increased again to 0.08 while the maximum height value was left at the initial value of 20. With respect to configuration #2, in this case the tree uses another set of attributes for performing the splits, selecting the ones for which the gain obtained by further splitting is more significant than the new minimal gain value.

The following figures show the uppermost splits in the left subtree of configuration #2 and #4 trees.

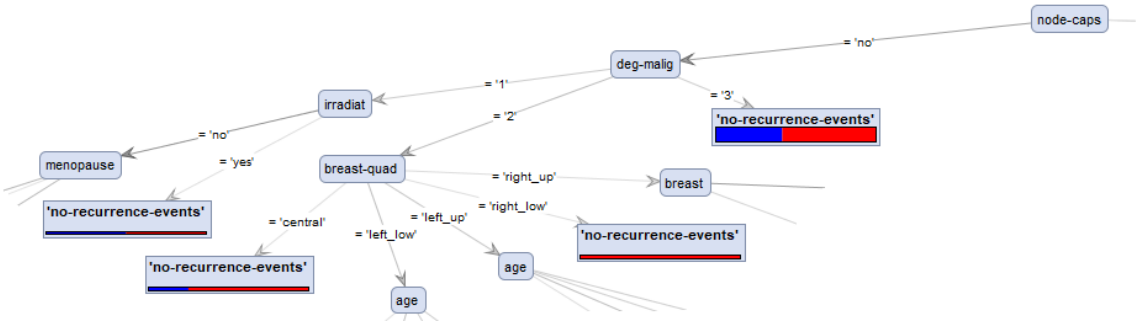


*Figure 8 – Path in the tree with configuration #2*

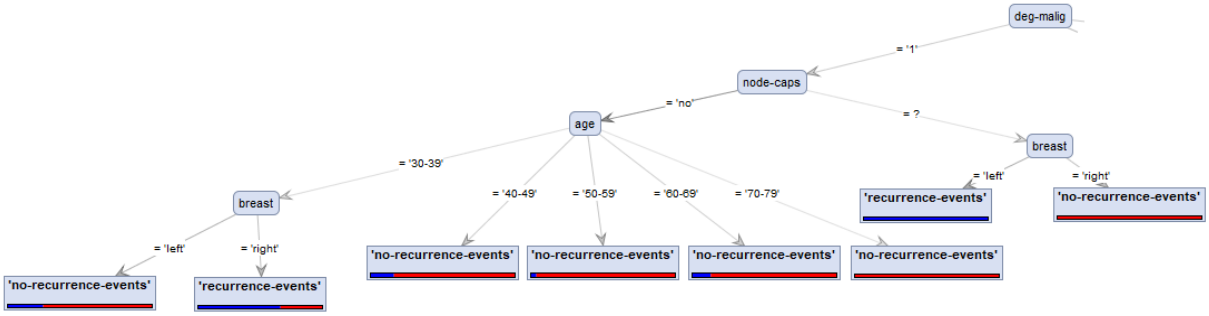


*Figure 9 – Path in the tree with configuration #4*

#### *1.2.1.5 Configuration #5*

In configuration #5, the minimal gain value was increased to 0.08 while the maximum height value was reduced to 5. With respect to the previous configuration, a number of subtrees have been condensed into single leaf nodes due to the reaching of the maximum allowed height, thus reducing the overall height.

The following figures show the path `deg-malig='3' → node-caps='yes'` in configuration #4 and #5.

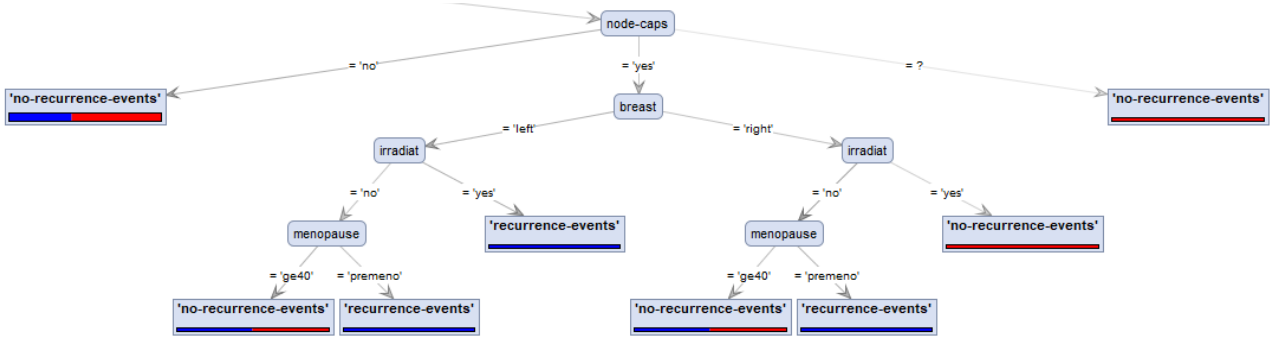


*Figure 10 – Path in the tree with configuration #4*

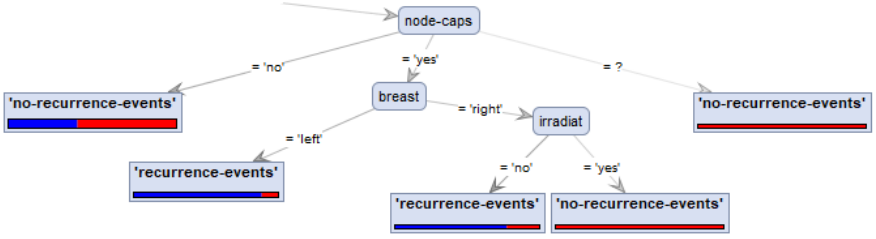


*Figure 11 – Path in the tree with configuration #5*

# 2 Validation and correlation

## 2.1 Question #3

In order to analyze the dependency of the produced decision tree from the chosen values for the minimal gain and the maximum height, the following graphs have been produced; the first one plots the accuracy of a tree with a fixed maximal height (20) and a variable minimal gain in the interval $(0.01, 0.1)$, while the second one plots the accuracy of a tree with a fixed minimal gain (0.01) and a variable maximum height in the interval $(1, 20)$. The data for the graphs have been obtained with the following procedure:

- The input data set was read from the Excel file;
- The input data was is passed to a loop subprocess, containing three operators:
  - A 10-folds cross-validation operator, using as a model a decision tree whose parameters are dynamically set via macros depending on the loop iteration; the random seed was fixed at the same value (2001) for all the validation operation, for the sake of consistency;
  - A performance-to-data operator, transforming the performance vector obtained by the cross-validation into a data set;
  - A filter operator, to select only the entries relative to the accuracy measure.
- The results of the different iterations were joined via an append operator;
- The appended results were written to an Excel file.

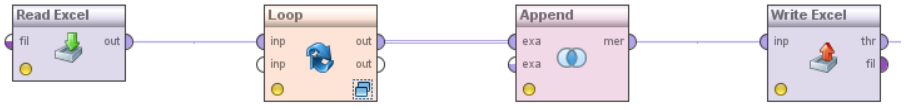


*Figure 12 – Process used for the extraction of the accuracy data*

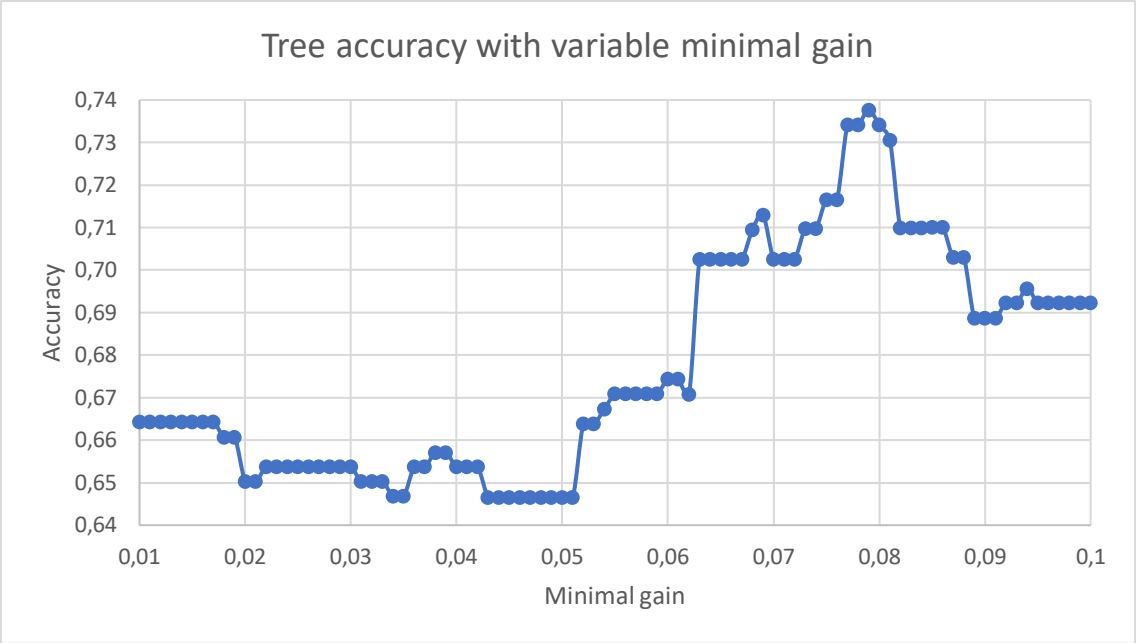


*Figure 13 – Tree accuracy with variable minimal gain (maximal height: 20)*

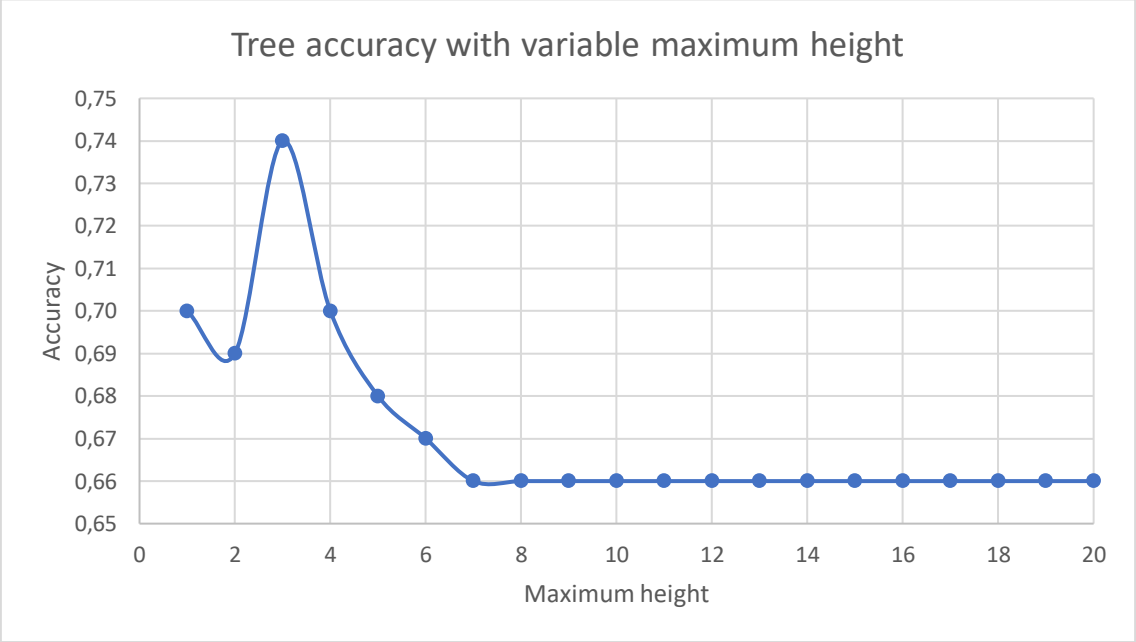


*Figure 14 – Tree accuracy with variable maximum height (minimal gain: 0.01)*

It can be noted that the choice of the values for the parameters has a strong impact on the overall accuracy of the model. In particular, the overall accuracy of the tree is lower when:

- The minimal gain is too low or the maximum height is too high: in this case, the decision tree is too precise and lacks the possibility to generalize based on the training data (overfitting);
- The minimal gain is too high or the maximum height is too low: in this case, the decision tree is too small and does not represent well the features of the training data (underfitting).

In between the two zones, a peak in the accuracy value can be found in both graphs, representing the optimal choices for the values of the parameters. As an example, the confusion matrices for the configurations considered in the previous part are reported in the following figures.

#### 2.1.1.1 Original tree

| accuracy: 66.43% +/- 7.89% (mikro: 66.43%) | | | |
|---|---|---|---|
| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
| pred. 'recurrence-events' | 34 | 45 | 43.04% |
| pred. 'no-recurrence-events' | 51 | 156 | 75.36% |
| class recall | 40.00% | 77.61% | |

*Figure 15 – Confusion matrix for the original tree*

#### 2.1.1.2 Configuration #1

| accuracy: 65.71% +/- 7.40% (mikro: 65.73%) | | | |
|---|---|---|---|
| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
| pred. 'recurrence-events' | 28 | 41 | 40.58% |
| pred. 'no-recurrence-events' | 57 | 160 | 73.73% |
| class recall | 32.94% | 79.60% | |

*Figure 16 – Confusion matrix for configuration #1*

#### 2.1.1.3 Configuration #2

| accuracy: 68.18% +/- 8.20% (mikro: 68.18%) | | | |
|---|---|---|---|
| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
| pred. 'recurrence-events' | 32 | 38 | 45.71% |
| pred. 'no-recurrence-events' | 53 | 163 | 75.46% |
| class recall | 37.65% | 81.09% | |

*Figure 17 – Confusion matrix for configuration #2*

#### 2.1.1.4 Configuration #3

| accuracy: 68.15% +/- 6.69% (mikro: 68.18%) | | | |
|---|---|---|---|
| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
| pred. 'recurrence-events' | 30 | 36 | 45.45% |
| pred. 'no-recurrence-events' | 55 | 165 | 75.00% |
| class recall | 35.29% | 82.09% | |

*Figure 18 – Confusion matrix for configuration #3*

#### 2.1.1.5 Configuration #4

| accuracy: 73.77% +/- 5.30% (mikro: 73.78%) | | | |
|---|---|---|---|
| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
| pred. 'recurrence-events' | 22 | 12 | 64.71% |
| pred. 'no-recurrence-events' | 63 | 189 | 75.00% |
| class recall | 25.88% | 94.03% | |

*Figure 19 – Confusion matrix for configuration #4*

#### 2.1.1.6 Configuration #5

| accuracy: 73.78% +/- 6.83% (mikro: 73.78%) | | | |
|---|---|---|---|
| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
| pred. 'recurrence-events' | 19 | 9 | 67.86% |
| pred. 'no-recurrence-events' | 66 | 192 | 74.42% |
| class recall | 22.35% | 95.52% | |

*Figure 20 – Confusion matrix for configuration #5*

## 2.2 Question #4

For performing an analysis by means of a $K$-Nearest Neighbors classifier, first of all the dependency of the model accuracy from the value of $K$ has been analyzed. The following plot, obtained with a procedure similar to the one described for the accuracy measures of the trees, shows the impact of the choice of the value of $K$ in the interval $(1, 100)$ on the accuracy of the model.

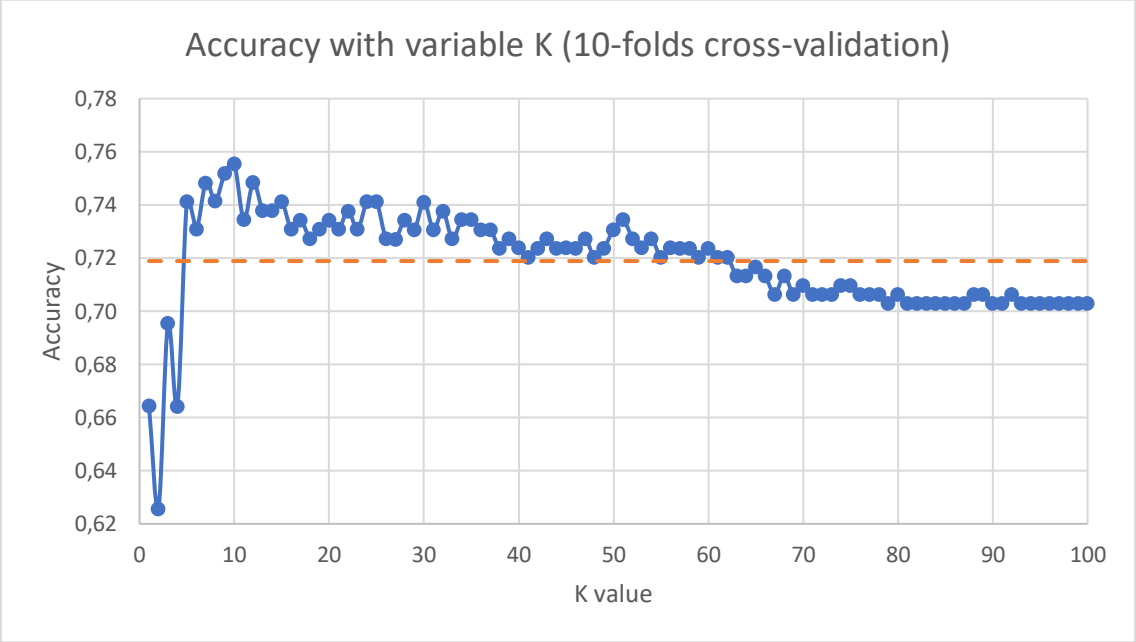


*Figure 21 – Overall accuracy after a 10-folds cross-validation*

In this case, the accuracy of the model is higher for values of $K$ in the range $(5, 60)$; the absolute peak is reached for $K = 10$ (accuracy: 75.54%), while for higher values of $K$ the accuracy value is stable at 70.30%. The average accuracy value for $K$ in the given range is 71.89%.

The cross-validation task was also carried out on a Naïve Bayes classifier, performed by using the same random seed (2001) used for the $K$-NN classifier for consistency. Its overall accuracy (72.45%) is just slightly higher than the average of the values obtained with the $K$-NN classifier, with $K$ in the range $(1, 100)$.

As an example, the confusion matrices of the following configurations for the $K$-NN classifier and of the Naïve Bayes classifier are shown in the figures below.

| Configuration # | $K$ value | Overall accuracy |
|---|---|---|
| **1** | 5 | 74.13% |
| **2** | 10 | 75.54% |
| **3** | 15 | 74.13% |
| **4** | 20 | 73.44% |
| **5** | 25 | 74.13% |

*Table 2 – Configurations used for the K-NN analysis*

### *2.2.1.1 Configuration #1*

**accuracy: 74.13% +/- 5.62% (mikro: 74.13%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 26 | 15 | 63.41% |
| pred. 'no-recurrence-events' | 59 | 186 | 75.92% |
| class recall | 30.59% | 92.54% | |

*Figure 22 – Confusion matrix for K-NN configuration #1*

#### 2.2.1.2 Configuration #2

| accuracy: 75.54% +/- 5.29% (mikro: 75.52%) | | | |
|---|---|---|---|
| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
| pred. 'recurrence-events' | 28 | 13 | 68.29% |
| pred. 'no-recurrence-events' | 57 | 188 | 76.73% |
| class recall | 32.94% | 93.53% | |

*Figure 23 – Confusion matrix for K-NN configuration #2*

#### 2.2.1.3 Configuration #3

| accuracy: 74.13% +/- 5.38% (mikro: 74.13%) | | | |
|---|---|---|---|
| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
| pred. 'recurrence-events' | 18 | 7 | 72.00% |
| pred. 'no-recurrence-events' | 67 | 194 | 74.33% |
| class recall | 21.18% | 96.52% | |

*Figure 24 – Confusion matrix for K-NN configuration #3*

#### 2.2.1.4 Configuration #4

| accuracy: 73.44% +/- 5.56% (mikro: 73.43%) | | | |
|---|---|---|---|
| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
| pred. 'recurrence-events' | 18 | 9 | 66.67% |
| pred. 'no-recurrence-events' | 67 | 192 | 74.13% |
| class recall | 21.18% | 95.52% | |

*Figure 25 – Confusion matrix for K-NN configuration #4*

#### 2.2.1.5 Configuration #5

| accuracy: 74.13% +/- 4.42% (mikro: 74.13%) | | | |
|---|---|---|---|
| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
| pred. 'recurrence-events' | 17 | 6 | 73.91% |
| pred. 'no-recurrence-events' | 68 | 195 | 74.14% |
| class recall | 20.00% | 97.01% | |

*Figure 26 – Confusion matrix for K-NN configuration #5*

#### 2.2.1.6 Naïve Bayes classifier

| accuracy: 72.45% +/- 7.30% (mikro: 72.38%) | | | |
|---|---|---|---|
| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
| pred. 'recurrence-events' | 41 | 35 | 53.95% |
| pred. 'no-recurrence-events' | 44 | 166 | 79.05% |
| class recall | 48.24% | 82.59% | |

*Figure 27 – Confusion matrix for the Naïve Bayes classifier*

## 2.3 Question #5

By analyzing the correlation matrix obtained from the input data set, it is possible to state that:

a) The majority of the attributes describing the objects in the data set have a very little value for the correlation measure, except for some cases. For this reason, it is possible to say that the naïve independence assumption holds for the considered data set.
b) The most significant correlations can be identified between:
   - The attribute pair `node-caps` and `inv-nodes` (negative correlation, -0.465);
   - The attribute pair `inv-nodes` and `irradiat` (positive correlation, 0.399).

The correlation matrix obtained from the input data set is shown in the figure below.

| Attributes | age | menopause | tumor-size | inv-nodes | node-caps | deg-malig | breast | breast-quad | irradiat |
|---|---|---|---|---|---|---|---|---|---|
| age | 1 | 0.241 | -0.045 | -0.001 | 0.052 | -0.043 | 0.067 | -0.024 | -0.011 |
| menopause | 0.241 | 1 | 0.019 | -0.011 | 0.130 | -0.161 | 0.077 | -0.096 | -0.075 |
| tumor-size | -0.045 | 0.019 | 1 | -0.131 | 0.058 | 0.133 | -0.022 | -0.056 | -0.022 |
| inv-nodes | -0.001 | -0.011 | -0.131 | 1 | -0.465 | -0.213 | 0.040 | 0.063 | 0.399 |
| node-caps | 0.052 | 0.130 | 0.058 | -0.465 | 1 | 0.098 | 0.024 | -0.036 | -0.197 |
| deg-malig | -0.043 | -0.161 | 0.133 | -0.213 | 0.098 | 1 | -0.073 | 0.018 | -0.074 |
| breast | 0.067 | 0.077 | -0.022 | 0.040 | 0.024 | -0.073 | 1 | 0.175 | -0.019 |
| breast-quad | -0.024 | -0.096 | -0.056 | 0.063 | -0.036 | 0.018 | 0.175 | 1 | -0.005 |
| irradiat | -0.011 | -0.075 | -0.022 | 0.399 | -0.197 | -0.074 | -0.019 | -0.005 | 1 |

*Figure 28 – Correlation matrix in the data set*