Polytechnic University of Turin

Master of Science in Computer Engineering

Database Management Systems' fourth homework

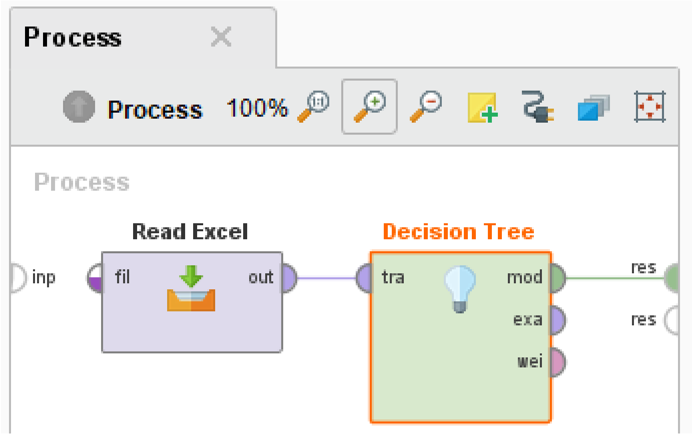
Marco Micera

Academic Year 2017-2018

1 First question

"Learn a Decision Tree from the whole dataset by setting the minimum gain threshold to 0.01, while keeping the default configuration for all the other parameters."

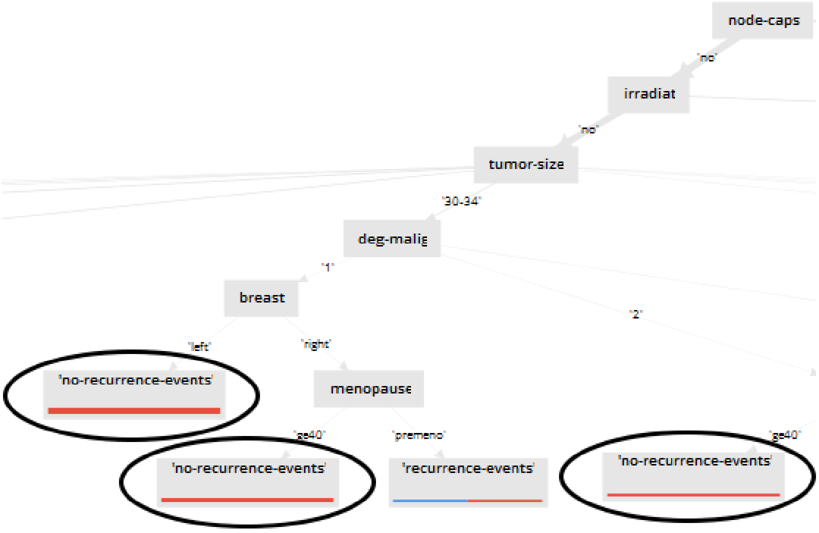
Building the decision tree:



a. **Q:** *Which attribute is deemed to be the most discriminative one for class prediction?*
A: The `node-caps` attribute is deemed to be the most discriminative one as it is the decision tree's root node.

b. **Q:** *What is the height of the Decision Tree generated?*
A: The height of the generated Decision Tree is 6.

c. **Q:** *Find a pure partition in the Decision Tree and report a screenshot that shows the example identified.*
A: An example of pure partitions (circled nodes) is shown below:

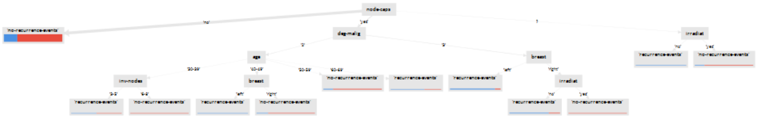


2 Second question

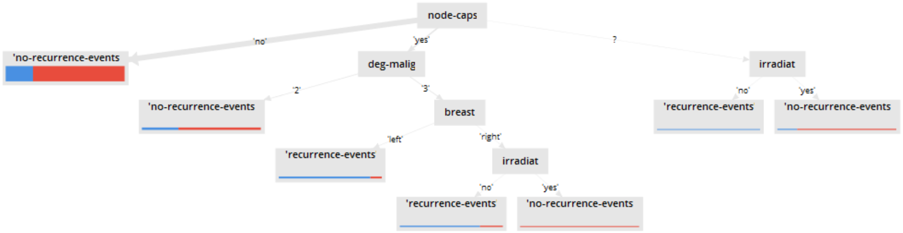
Analyze the impact of the minimal gain (using the gain ratio splitting criterion) and maximal depth parameters on the characteristics on the Decision Tree model learnt from the whole dataset (keep the default configuration for all the other parameters). Report at least 5 different screenshots showing Decision Trees (or portions of them) generated with different configuration settings.

The node is split if its gain is greater than the `minimal gain`.
A higher value of minimal gain results in fewer splits and thus a smaller tree: here is an example of tree reduction due to a `minimal gain` increase.

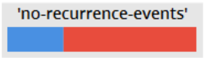
1. `minimal gain = 0.05, maximum depth = 20;`



2. `minimal gain = 0.06, maximum depth = 20;`

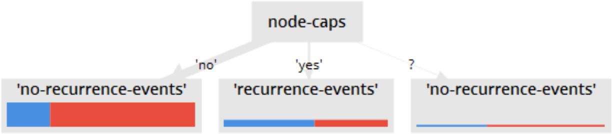


3. `minimal gain = 0.07, maximum depth = 20;`

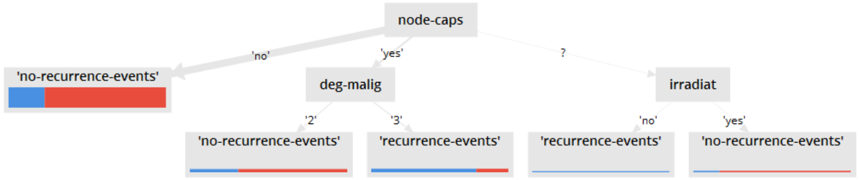


The `maximum depth` setting's aim is just to allow the user to restrict the depth of the decision tree.

4. `minimal gain = 0.01, maximum depth = 2;`



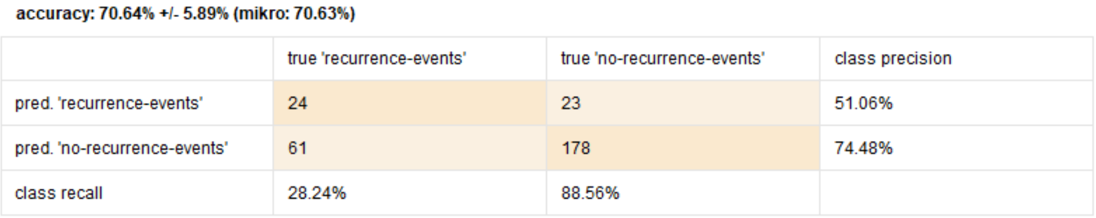
5. `minimal gain = 0.01, maximum depth = 3;`



3 Third question

Performing a 10-fold Stratified Cross-Validation, what is the impact the maximal gain and maximal depth parameters on the average accuracy achieved by Decision Tree? Report at least 5 screenshots showing the confusion matrices achieved using different parameter settings (consider at least all the configurations used to answer Question 2). Keep the default configuration for all the other parameters.

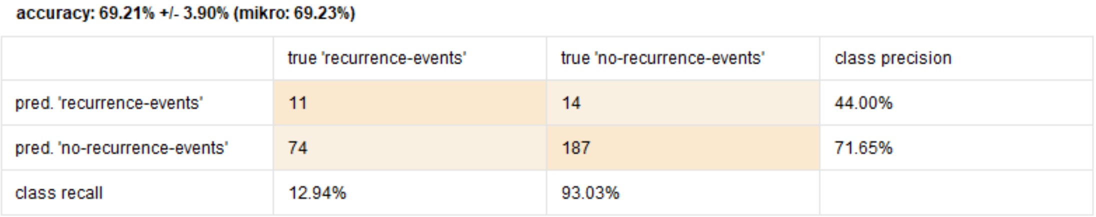
1. `minimal gain = 0.05, maximum depth = 20;`



accuracy: 70.64% +/- 5.89% (mikro: 70.63%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	23	51.06%
pred. 'no-recurrence-events'	61	178	74.48%
class recall	28.24%	88.56%	

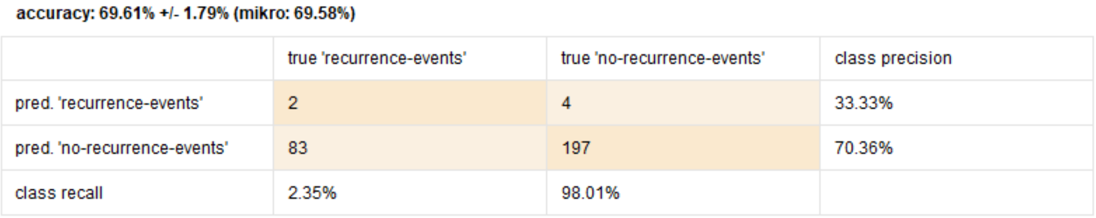
2. `minimal gain = 0.06, maximum depth = 20;`



accuracy: 69.21% +/- 3.90% (mikro: 69.23%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	11	14	44.00%
pred. 'no-recurrence-events'	74	187	71.65%
class recall	12.94%	93.03%	

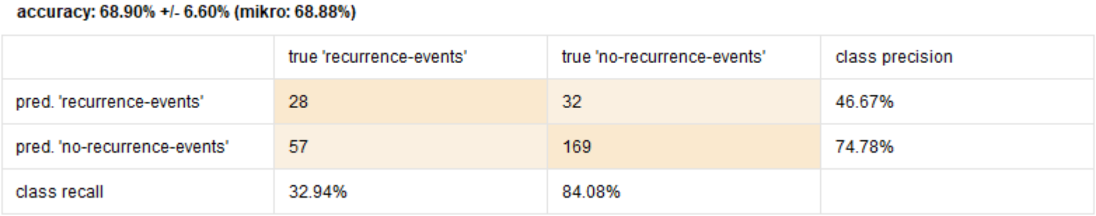
3. `minimal gain = 0.07, maximum depth = 20;`



accuracy: 69.61% +/- 1.79% (mikro: 69.58%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	2	4	33.33%
pred. 'no-recurrence-events'	83	197	70.36%
class recall	2.35%	98.01%	

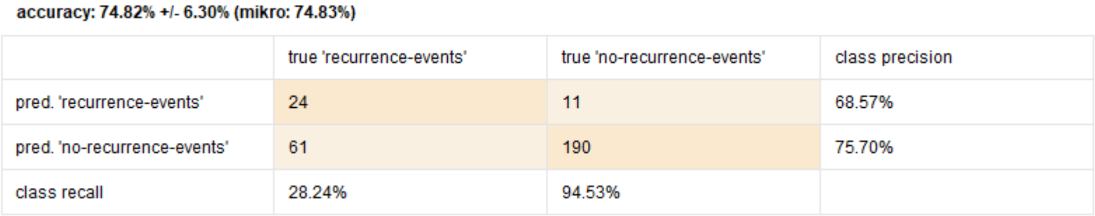
4. `minimal gain = 0.01, maximum depth = 2;`



accuracy: 68.90% +/- 6.60% (mikro: 68.88%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	28	32	46.67%
pred. 'no-recurrence-events'	57	169	74.78%
class recall	32.94%	84.08%	

5. `minimal gain = 0.01, maximum depth = 3;`



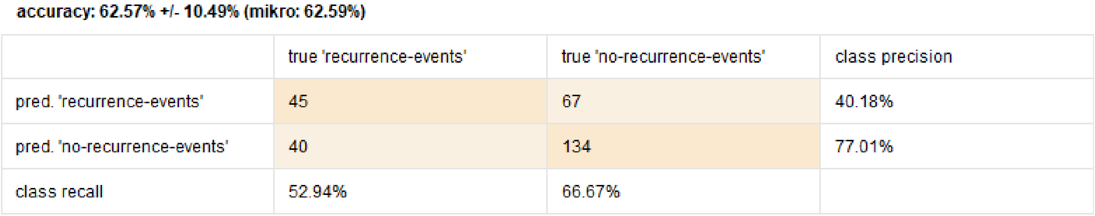
accuracy: 74.82% +/- 6.30% (mikro: 74.83%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	11	68.57%
pred. 'no-recurrence-events'	61	190	75.70%
class recall	28.24%	94.53%	

4 Fourth question

Considering the K-Nearest Neighbor (K-NN) classifier and performing a 10-fold Stratified Cross-Validation, what is the impact of parameter K on the average classifier accuracy? Report at least 5 screenshots showing the confusion matrices achieved using different K parameter values. Perform a 10-fold Stratified Cross-Validation with classifier Naïve Bayes.

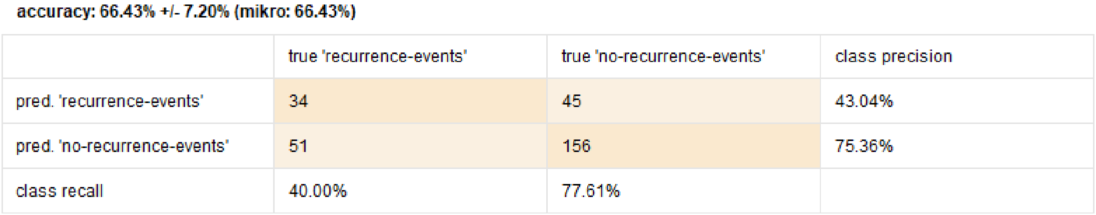
1. K-Nearest Neighbor, with `k = 2`



accuracy: 62.57% +/- 10.49% (mikro: 62.59%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	45	67	40.18%
pred. 'no-recurrence-events'	40	134	77.01%
class recall	52.94%	66.67%	

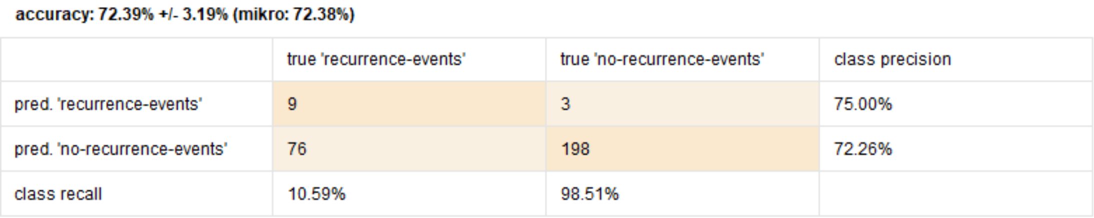
2. K-Nearest Neighbor, with `k = 4`



accuracy: 66.43% +/- 7.20% (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%	

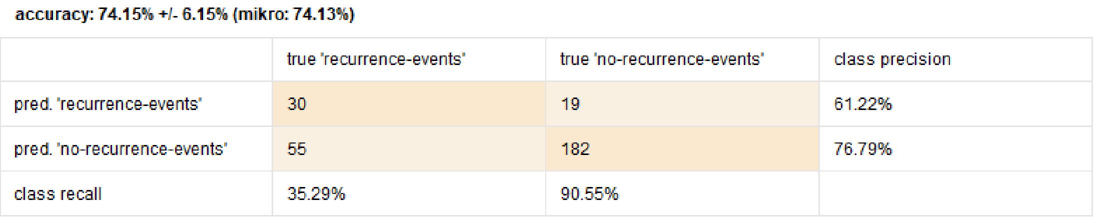
3. K-Nearest Neighbor, with `k = 5`



accuracy: 72.39% +/- 3.19% (mikro: 72.38%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	9	3	75.00%
pred. 'no-recurrence-events'	76	198	72.26%
class recall	10.59%	98.51%	

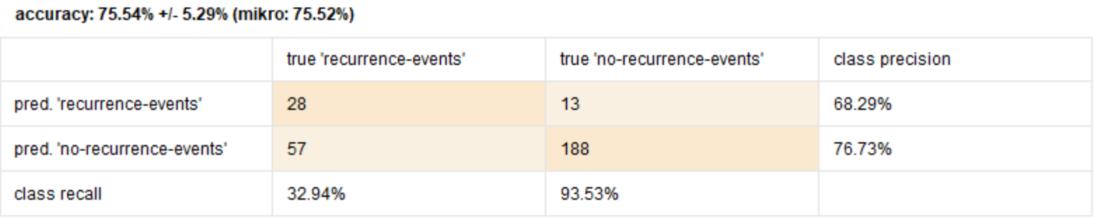
4. K-Nearest Neighbor, with `k = 8`



accuracy: 74.15% +/- 6.15% (mikro: 74.13%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	30	19	61.22%
pred. 'no-recurrence-events'	55	182	76.79%
class recall	35.29%	90.55%	

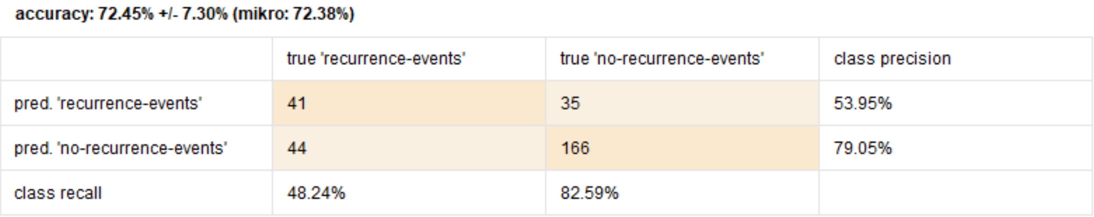
5. K-Nearest Neighbor, with `k = 10`



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%	

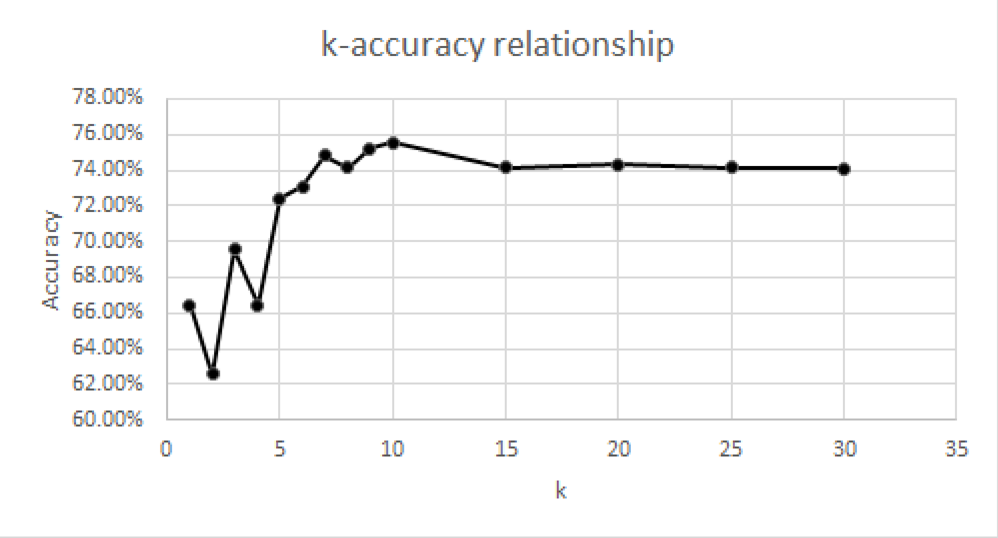
6. Naïve Bayes



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%	

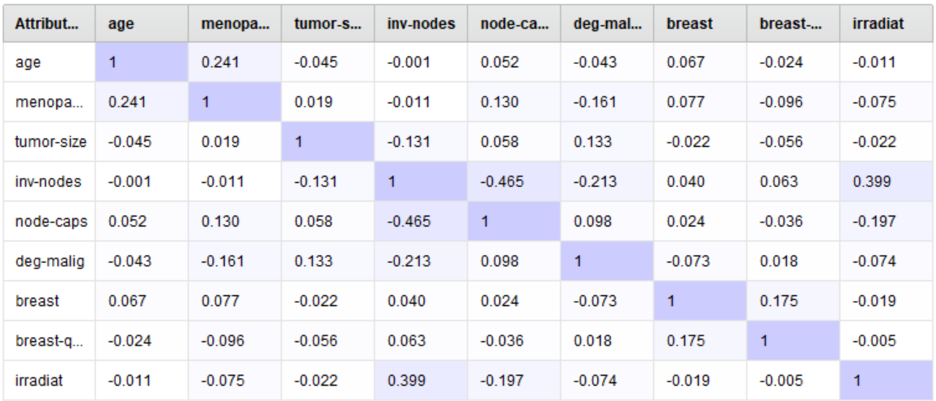
Q: *Does K-NN perform on average better or worse than the Naïve Bayes classifier on the analyzed data?*
A: The Naïve Bayes classifier outperforms the K-Nearest Neighbor classifier when `K < 5`.



5 Fifth question

Analyze the Correlation Matrix to discover pairwise correlations between data attributes. Report a screenshot showing the correlation matrix achieved.

The Correlation Matrix is reported below:



Attribut...	age	menopa...	tumor-s...	inv-nodes	node-ca...	deg-mal...	breast	breast-...	irradiat
age	1	0.241	-0.045	-0.001	0.052	-0.043	0.067	-0.024	-0.011
menopa...	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-q...	-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

a. **Q:** *Does the Naïve independence assumption actually hold for the Breast dataset?*
A: Almost for every attribute pair, the correlation (always discussed in its absolute value from now on) is very low. In just two cases the correlation reaches 0.399 and 0.465, but they could still be considered low values as they are less than 0.5.

b. **Q:** *Which is the pair of most correlated attributes?*
A: The most correlated attributes, with a correlation of -0.465, are `node-caps` and `inv-nodes`.