

Homework 2 - Enrico Castelli s280124

Data Mining with RapidMiner

1.

(a)

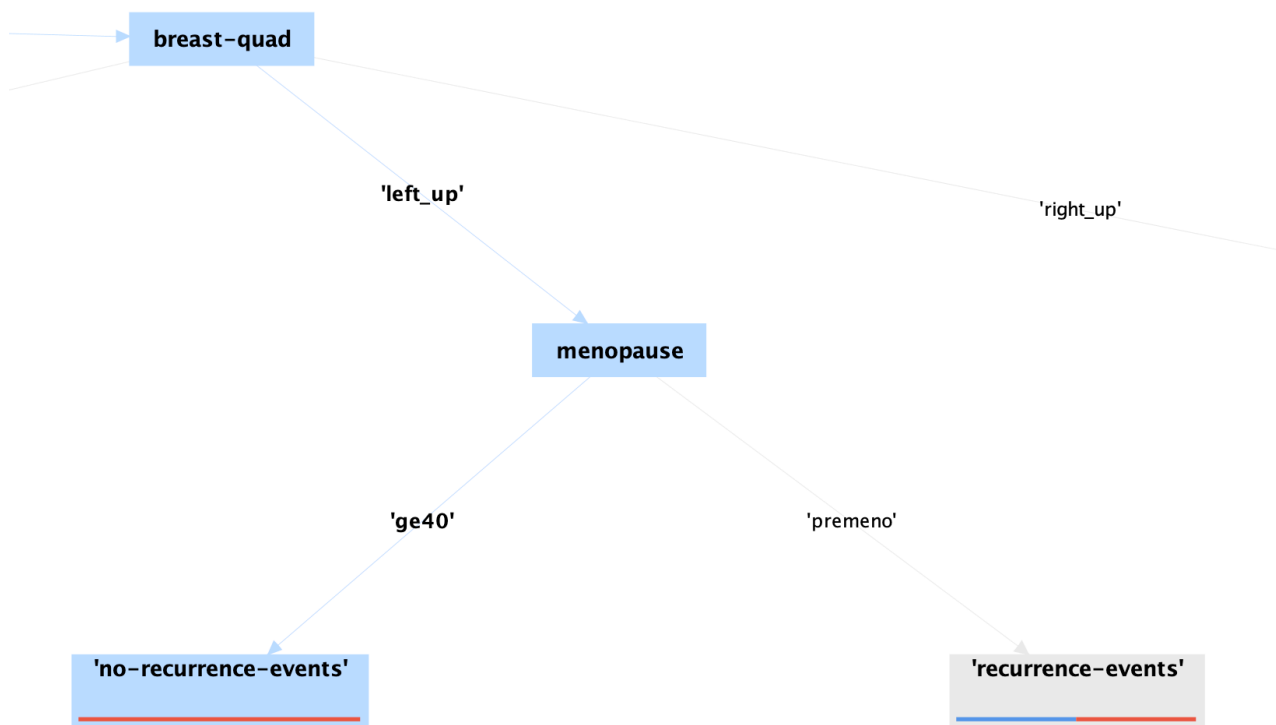
The most discriminative parameter, which is the root of the decision tree, is **node-caps**, meaning that if the cancer has infiltrated the lymph nodes (**node-caps = yes**) there is an higher probability that it will present itself again in the future.

node-caps	yes	no	?
probability of recurrence	55.36%	22.97%	37.5%

(b)

The height of the generated Decision Tree is 7, counting the root of the tree.

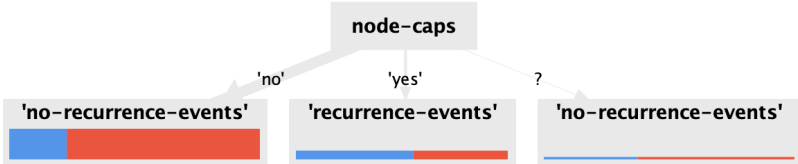
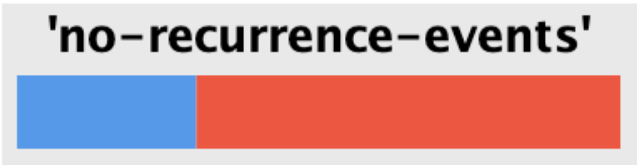
An example of a pure partition would be the following:



Where the decisions to reach it are:

node-caps[no] -> irradiat[no] -> tumor-size[35-39] -> breast-quad[left_up] -> menopause[ge40] -> no-recurrence-events

2.

minimal gain	maximal depth	screenshot
any number from 0 to 1	1	 <p>A screenshot of a decision tree with a single root node labeled 'no-recurrence-events'. Below the node is a horizontal bar divided into two segments: a blue segment on the left and a red segment on the right.</p>
0.01	2	 <p>A screenshot of a decision tree with root node 'node-caps'. It has three children: 'no-recurrence-events' (labeled 'no'), 'recurrence-events' (labeled 'yes'), and 'no-recurrence-events' (labeled '?'). Each child node has a corresponding horizontal bar with blue and red segments.</p>
0.06	3	 <p>A screenshot of a decision tree with root node 'node-caps'. It has three children: 'no-recurrence-events' (labeled 'no'), 'deg-malign' (labeled 'yes'), and 'irradiat' (labeled '?'). The 'deg-malign' node has two children: 'no-recurrence-events' (labeled '2') and 'recurrence-events' (labeled '3'). The 'irradiat' node has two children: 'recurrence-events' (labeled 'no') and 'no-recurrence-events' (labeled 'yes'). Each node has a corresponding horizontal bar with blue and red segments.</p>
0.07	3	 <p>A screenshot of a decision tree with a single root node labeled 'no-recurrence-events'. Below the node is a horizontal bar divided into two segments: a blue segment on the left and a red segment on the right.</p>
0.06	6	 <p>A screenshot of a complex decision tree with root node 'node-caps'. It has three children: 'no-recurrence-events' (labeled 'no'), 'deg-malign' (labeled 'yes'), and 'irradiat' (labeled '?'). The 'deg-malign' node has two children: 'no-recurrence-events' (labeled '2') and 'breast' (labeled '3'). The 'breast' node has two children: 'recurrence-events' (labeled 'left') and 'irradiat' (labeled 'right'). The 'irradiat' node has two children: 'recurrence-events' (labeled 'no') and 'no-recurrence-events' (labeled 'yes'). Each node has a corresponding horizontal bar with blue and red segments.</p>

From the table above, it emerges that the maximum value of the minimal gain that still yields a Decision Tree with nodes other than the root node is 0.06.

The maximal depth increases the number of decisions the Decision Tree can show. When it is 1, it is obvious that no node other than the root node can be shown. For higher values, it shows increasingly complex decision ramifications. When it is 2, only the first set of leaves is visible, hence the most discriminative parameter becomes apparent. For a value of 3, we can see how `deg-malign` and `irradiat` influence the decision; for example, a degree of 2 indicates a lower probability of reoccurring tumor instead of a degree of 3. When the maximal depth is 6, we can see more details in the ramification of the degree of 3.

3.

minimal gain	maximal depth	screenshot																
any number from 0 to 1	1	<div>accuracy: 70.30% +/- 1.43% (micro average: 70.28%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>0</td><td>0</td><td>0.00%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>85</td><td>201</td><td>70.28%</td></tr><tr><td>class recall</td><td>0.00%</td><td>100.00%</td><td></td></tr></table>		true 'recurrence-events'	true 'no-recurrence-events'	class precision	pred. 'recurrence-events'	0	0	0.00%	pred. 'no-recurrence-events'	85	201	70.28%	class recall	0.00%	100.00%	
	true 'recurrence-events'	true 'no-recurrence-events'	class precision															
pred. 'recurrence-events'	0	0	0.00%															
pred. 'no-recurrence-events'	85	201	70.28%															
class recall	0.00%	100.00%																
0.01	2	<div>accuracy: 68.90% +/- 6.96% (micro average: 68.88%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>28</td><td>32</td><td>46.67%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>57</td><td>169</td><td>74.78%</td></tr><tr><td>class recall</td><td>32.94%</td><td>84.08%</td><td></td></tr></table>		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%	
	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%																
0.01	10	<div>accuracy: 67.48% +/- 6.59% (micro average: 67.48%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>37</td><td>45</td><td>45.12%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>48</td><td>156</td><td>76.47%</td></tr><tr><td>class recall</td><td>43.53%</td><td>77.61%</td><td></td></tr></table>		true 'recurrence-events'	true 'no-recurrence-events'	class precision	pred. 'recurrence-events'	37	45	45.12%	pred. 'no-recurrence-events'	48	156	76.47%	class recall	43.53%	77.61%	
	true 'recurrence-events'	true 'no-recurrence-events'	class precision															
pred. 'recurrence-events'	37	45	45.12%															
pred. 'no-recurrence-events'	48	156	76.47%															
class recall	43.53%	77.61%																
0.06	10	<div>accuracy: 70.64% +/- 6.20% (micro average: 70.63%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>24</td><td>23</td><td>51.06%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>61</td><td>178</td><td>74.48%</td></tr><tr><td>class recall</td><td>28.24%</td><td>88.56%</td><td></td></tr></table>		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%	
	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%																
0.06	3	<div>accuracy: 70.97% +/- 4.76% (micro average: 70.98%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>12</td><td>10</td><td>54.55%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>73</td><td>191</td><td>72.35%</td></tr><tr><td>class recall</td><td>14.12%</td><td>95.02%</td><td></td></tr></table>		true 'recurrence-events'	true 'no-recurrence-events'	class precision	pred. 'recurrence-events'	12	10	54.55%	pred. 'no-recurrence-events'	73	191	72.35%	class recall	14.12%	95.02%	
	true 'recurrence-events'	true 'no-recurrence-events'	class precision															
pred. 'recurrence-events'	12	10	54.55%															
pred. 'no-recurrence-events'	73	191	72.35%															
class recall	14.12%	95.02%																
0.07	3	<div>accuracy: 70.30% +/- 1.43% (micro average: 70.28%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>3</td><td>3</td><td>50.00%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>82</td><td>198</td><td>70.71%</td></tr><tr><td>class recall</td><td>3.53%</td><td>98.51%</td><td></td></tr></table>		true 'recurrence-events'	true 'no-recurrence-events'	class precision	pred. 'recurrence-events'	3	3	50.00%	pred. 'no-recurrence-events'	82	198	70.71%	class recall	3.53%	98.51%	
	true 'recurrence-events'	true 'no-recurrence-events'	class precision															
pred. 'recurrence-events'	3	3	50.00%															
pred. 'no-recurrence-events'	82	198	70.71%															
class recall	3.53%	98.51%																
0.06	6	<div>accuracy: 69.21% +/- 4.11% (micro average: 69.23%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>11</td><td>14</td><td>44.00%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>74</td><td>187</td><td>71.65%</td></tr><tr><td>class recall</td><td>12.94%</td><td>93.03%</td><td></td></tr></table>		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%	
	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%																
from 0.01 to 0.05	3	<div>accuracy: 74.82% +/- 6.64% (micro average: 74.83%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>24</td><td>11</td><td>68.57%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>61</td><td>190</td><td>75.70%</td></tr><tr><td>class recall</td><td>28.24%</td><td>94.53%</td><td></td></tr></table>		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%	
	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%																

From the table above, we can see that with a maximal depth of 1 the tree does not predict any **recurrence** events, so the accuracy is simply the number of **no-recurrence** events divided by the total events multiplied by 100.

The minimal gain has no impact on the results in the range [0.01, 0.05]. It has a negative impact in the range [0.06, 1].

The maximal depth yields the best results when set with a value of 3. Any value below or above lowers the percentage of correct predictions of the model.

4.

k	screenshot																
1	<div>accuracy: 66.44% +/- 7.28% (micro average: 66.43%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>30</td><td>41</td><td>42.25%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>55</td><td>160</td><td>74.42%</td></tr><tr><td>class recall</td><td>35.29%</td><td>79.60%</td><td></td></tr></table>		true 'recurrence-events'	true 'no-recurrence-events'	class precision	pred. 'recurrence-events'	30	41	42.25%	pred. 'no-recurrence-events'	55	160	74.42%	class recall	35.29%	79.60%	
		true 'recurrence-events'	true 'no-recurrence-events'	class precision													
	pred. 'recurrence-events'	30	41	42.25%													
	pred. 'no-recurrence-events'	55	160	74.42%													
	class recall	35.29%	79.60%														
2	<div>accuracy: 65.73% +/- 8.62% (micro average: 65.73%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>39</td><td>52</td><td>42.86%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>46</td><td>149</td><td>76.41%</td></tr><tr><td>class recall</td><td>45.88%</td><td>74.13%</td><td></td></tr></table>		true 'recurrence-events'	true 'no-recurrence-events'	class precision	pred. 'recurrence-events'	39	52	42.86%	pred. 'no-recurrence-events'	46	149	76.41%	class recall	45.88%	74.13%	
		true 'recurrence-events'	true 'no-recurrence-events'	class precision													
	pred. 'recurrence-events'	39	52	42.86%													
	pred. 'no-recurrence-events'	46	149	76.41%													
	class recall	45.88%	74.13%														
3	<div>accuracy: 70.26% +/- 7.23% (micro average: 70.28%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>27</td><td>27</td><td>50.00%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>58</td><td>174</td><td>75.00%</td></tr><tr><td>class recall</td><td>31.76%</td><td>86.57%</td><td></td></tr></table>		true 'recurrence-events'	true 'no-recurrence-events'	class precision	pred. 'recurrence-events'	27	27	50.00%	pred. 'no-recurrence-events'	58	174	75.00%	class recall	31.76%	86.57%	
		true 'recurrence-events'	true 'no-recurrence-events'	class precision													
	pred. 'recurrence-events'	27	27	50.00%													
	pred. 'no-recurrence-events'	58	174	75.00%													
	class recall	31.76%	86.57%														
4	<div>accuracy: 67.86% +/- 7.40% (micro average: 67.83%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>28</td><td>35</td><td>44.44%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>57</td><td>166</td><td>74.44%</td></tr><tr><td>class recall</td><td>32.94%</td><td>82.59%</td><td></td></tr></table>		true 'recurrence-events'	true 'no-recurrence-events'	class precision	pred. 'recurrence-events'	28	35	44.44%	pred. 'no-recurrence-events'	57	166	74.44%	class recall	32.94%	82.59%	
		true 'recurrence-events'	true 'no-recurrence-events'	class precision													
	pred. 'recurrence-events'	28	35	44.44%													
	pred. 'no-recurrence-events'	57	166	74.44%													
	class recall	32.94%	82.59%														
5	<div>accuracy: 73.77% +/- 5.98% (micro average: 73.78%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>26</td><td>16</td><td>61.90%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>59</td><td>185</td><td>75.82%</td></tr><tr><td>class recall</td><td>30.59%</td><td>92.04%</td><td></td></tr></table>		true 'recurrence-events'	true 'no-recurrence-events'	class precision	pred. 'recurrence-events'	26	16	61.90%	pred. 'no-recurrence-events'	59	185	75.82%	class recall	30.59%	92.04%	
		true 'recurrence-events'	true 'no-recurrence-events'	class precision													
	pred. 'recurrence-events'	26	16	61.90%													
	pred. 'no-recurrence-events'	59	185	75.82%													
	class recall	30.59%	92.04%														
6	<div>accuracy: 72.03% +/- 6.10% (micro average: 72.03%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>24</td><td>19</td><td>55.81%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>61</td><td>182</td><td>74.90%</td></tr><tr><td>class recall</td><td>28.24%</td><td>90.55%</td><td></td></tr></table>		true 'recurrence-events'	true 'no-recurrence-events'	class precision	pred. 'recurrence-events'	24	19	55.81%	pred. 'no-recurrence-events'	61	182	74.90%	class recall	28.24%	90.55%	
		true 'recurrence-events'	true 'no-recurrence-events'	class precision													
	pred. 'recurrence-events'	24	19	55.81%													
	pred. 'no-recurrence-events'	61	182	74.90%													
	class recall	28.24%	90.55%														
7	<div>accuracy: 74.84% +/- 6.23% (micro average: 74.83%)</div> <table><tr><td></td><td>true 'recurrence-events'</td><td>true 'no-recurrence-events'</td><td>class precision</td></tr><tr><td>pred. 'recurrence-events'</td><td>25</td><td>12</td><td>67.57%</td></tr><tr><td>pred. 'no-recurrence-events'</td><td>60</td><td>189</td><td>75.90%</td></tr><tr><td>class recall</td><td>29.41%</td><td>94.03%</td><td></td></tr></table>		true 'recurrence-events'	true 'no-recurrence-events'	class precision	pred. 'recurrence-events'	25	12	67.57%	pred. 'no-recurrence-events'	60	189	75.90%	class recall	29.41%	94.03%	
		true 'recurrence-events'	true 'no-recurrence-events'	class precision													
	pred. 'recurrence-events'	25	12	67.57%													
	pred. 'no-recurrence-events'	60	189	75.90%													
	class recall	29.41%	94.03%														

8	accuracy: 74.51% +/- 5.02% (micro average: 74.48%)			
		true 'recurrence-events'	true 'no-recurrence-events'	class precision
	pred. 'recurrence-events'	24	12	66.67%
	pred. 'no-recurrence-events'	61	189	75.60%
	class recall	28.24%	94.03%	
9	accuracy: 75.20% +/- 5.18% (micro average: 75.17%)			
		true 'recurrence-events'	true 'no-recurrence-events'	class precision
	pred. 'recurrence-events'	23	9	71.88%
	pred. 'no-recurrence-events'	62	192	75.59%
	class recall	27.06%	95.52%	
10	accuracy: 75.20% +/- 5.43% (micro average: 75.17%)			
		true 'recurrence-events'	true 'no-recurrence-events'	class precision
	pred. 'recurrence-events'	25	11	69.44%
	pred. 'no-recurrence-events'	60	190	76.00%
	class recall	29.41%	94.53%	

We can see from the table above that the K-Nearest Neighbor classifier with 10-fold cross-validation yields a correct prediction in at least 70% of the cases for a minimum value of 3. It performs best with a k value of 9, after that the performance starts dropping.

Note: the even values of k have been included for completeness, but it is known that K-NN is facilitated in taking a decision when the number of considered neighbors k is odd.

Using the Naïve Bayes classifier with 10-fold cross-validation:

accuracy: 72.45% +/- 7.70% (micro average: 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%	

Considering only the accuracy of the K-NN classifier for odd values of k, its average performance is:

$$acc_{avg,k} = \frac{\sum_{k \in \{1,3,5,7,9\}} acc_k}{5} = \frac{66.44+70.26+73.77+74.84+75.20}{5} = \frac{360.47}{5} = 72.09 \%$$

The accuracy of the Naïve Bayes classifier is 72.45%, hence it performs better, in average and with the constraints considered above.

5.

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

(a)

Seeing the correlation matrix above, we cannot state that the Naïve Bayes independence assumption holds true for the features of the Breast dataset. Even if most of them have a very low correlation or anti-correlation, meaning correlation values near zero, not one of the features is completely independent of all of the others.

This makes sense, since the characteristics of an entry in this dataset only refer to a specific tumor of a specific person.

(b)

The pair of most correlated attributes is `inv-nodes` - `irradiat` (0.399).

The pair of most anti-correlated attributes is `inv-nodes` - `node-caps` (-0.465).