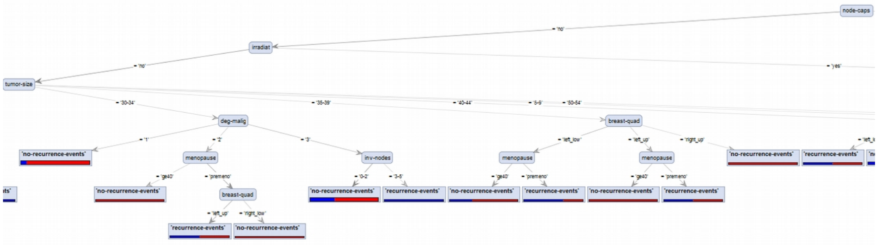
HOMEWORK 4

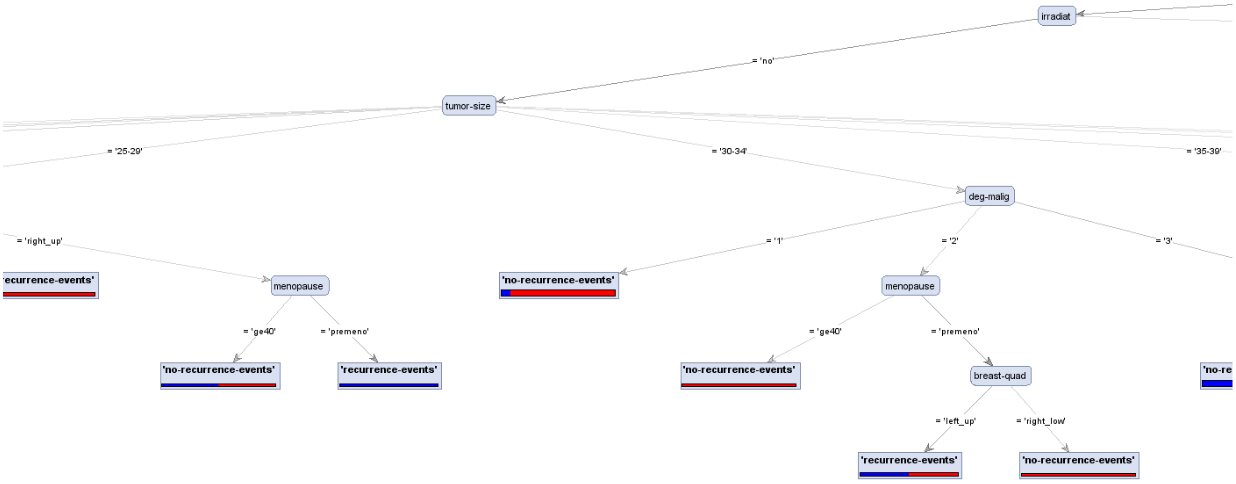
Learn a Decision Tree from the whole dataset by setting the minimum gain threshold to 0.01, while keeping the default configuration for all the other parameters.

(a) Which attribute is deemed to be the most discriminative one for class prediction?
The most discriminative attribute is “node-caps”.

(b) What is the height of the Decision Tree generated?
The height of the Decision Tree generated is equal to 6.

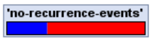


(c) Find a pure partition in the Decision Tree and report a screenshot that shows the example identified.
Some example of pure partition in the tree:

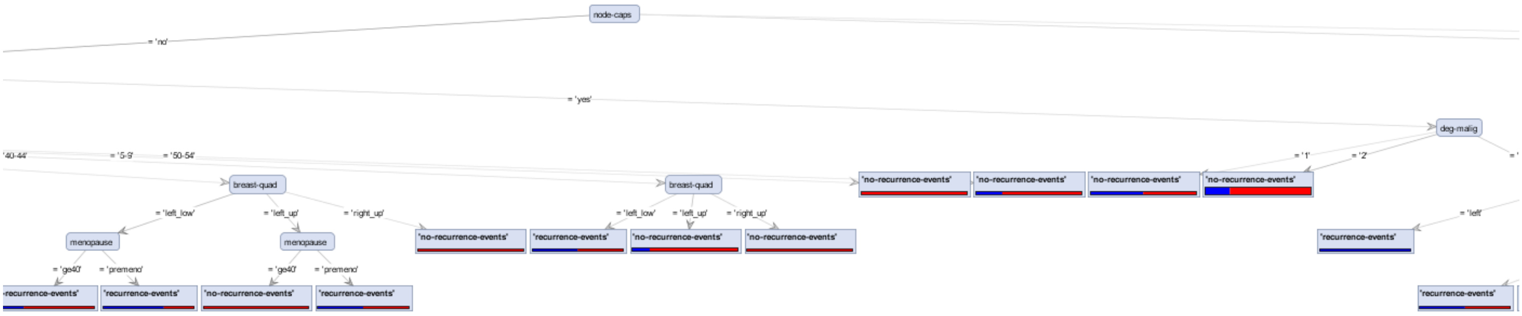


2. Analyze the impact of the minimal gain (using the gain ratio splitting criterion) and maximal depth parameters on the characteristics on the Decision Tree model learnt from the whole dataset (keep the default configuration for all the other parameters). Report at least 5 different screenshots showing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

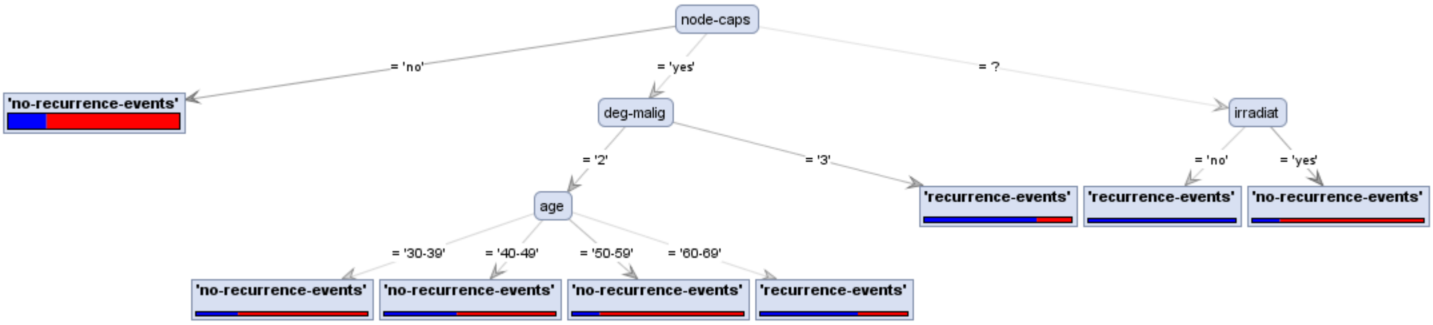
MIN GAIN: 0.1, DEPTH: 20



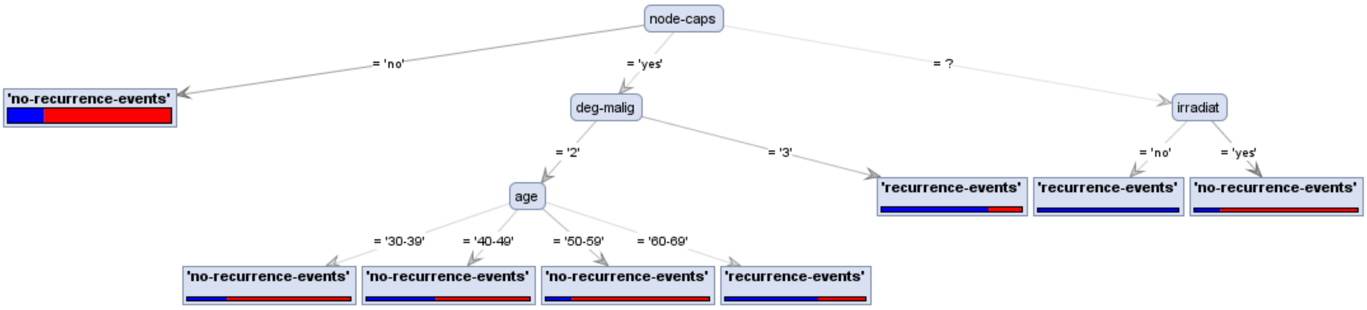
MIN GAIN: 0.01, DEPTH: 20



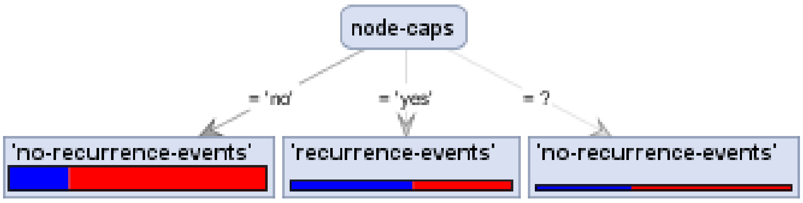
MIN GAIN: 0.02, DEPTH: 3



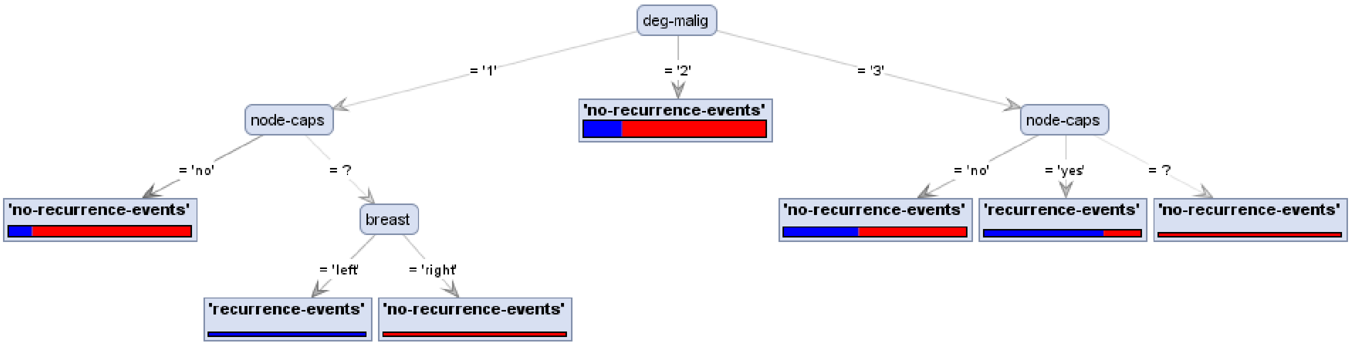
MIN GAIN: 0.05, DEPTH: 4



MIN GAIN: 0.05, DEPTH: 2

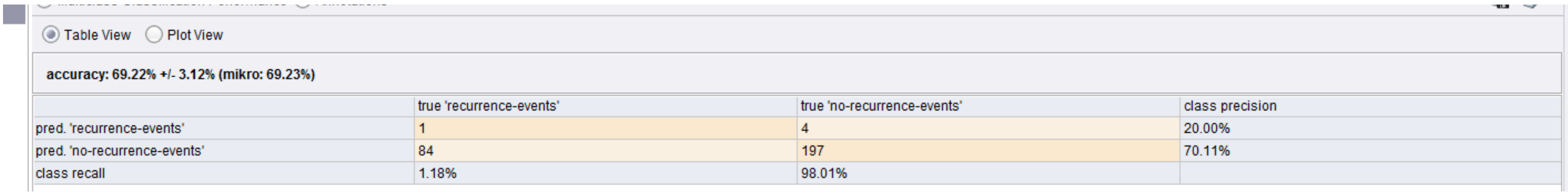


MIN GAIN: 0.08, DEPTH: 4



3. 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.

MIN GAIN: 0.1, DEPTH: 20

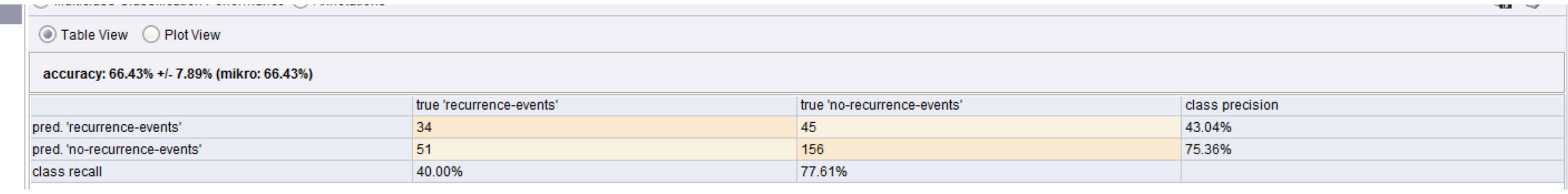


● Table View ○ Plot View

accuracy: 69.22% +/- 3.12% (mikro: 69.23%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	1	4	20.00%
pred. 'no-recurrence-events'	84	197	70.11%
class recall	1.18%	98.01%	

MIN GAIN: 0.01, DEPTH: 20

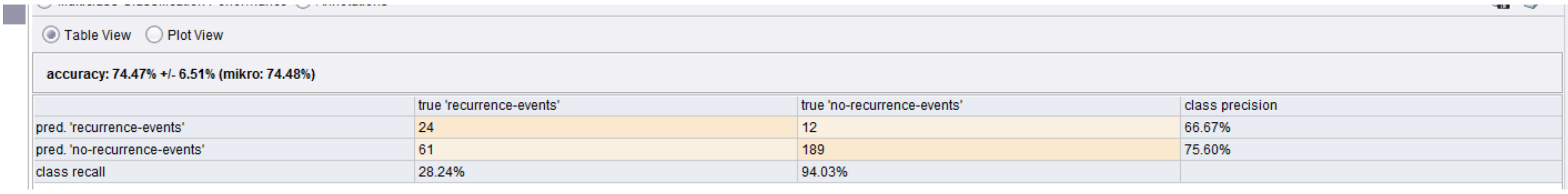


● Table View ○ Plot View

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%	

MIN GAIN: 0.02, DEPTH: 3

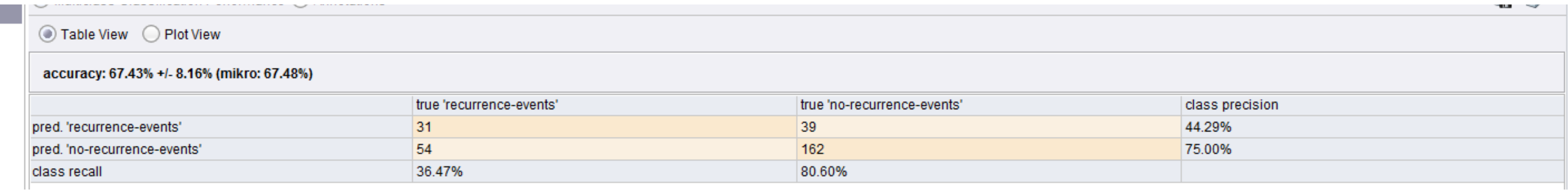


● Table View ○ Plot View

accuracy: 74.47% +/- 6.51% (mikro: 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%	

MIN GAIN: 0.05, DEPTH: 4

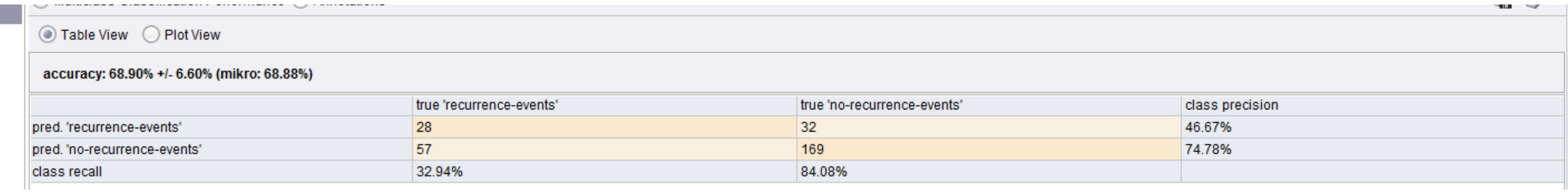


● Table View ○ Plot View

accuracy: 67.43% +/- 8.16% (mikro: 67.48%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	31	39	44.29%
pred. 'no-recurrence-events'	54	162	75.00%
class recall	36.47%	80.60%	

MIN GAIN: 0.05, DEPTH: 2

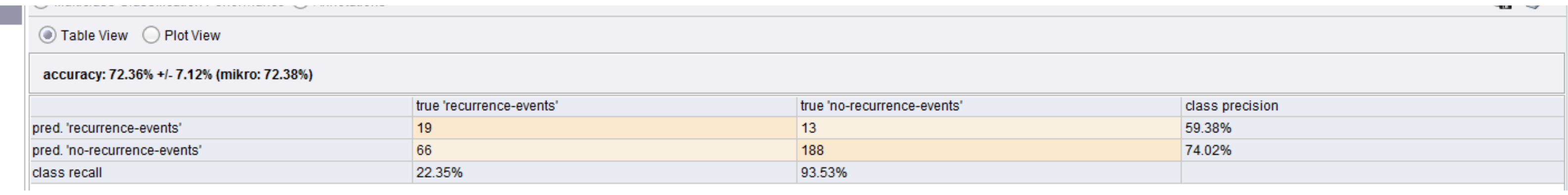


● Table View ○ Plot View

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%	

MIN GAIN: 0.08, DEPTH: 4



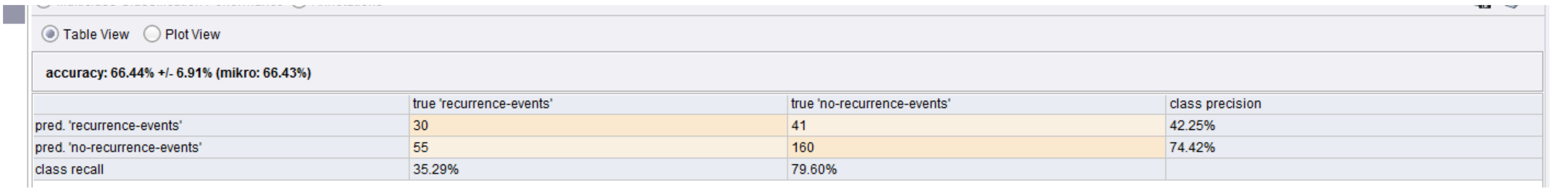
● Table View ○ Plot View

accuracy: 72.36% +/- 7.12% (mikro: 72.38%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	19	13	59.38%
pred. 'no-recurrence-events'	66	188	74.02%
class recall	22.35%	93.53%	

4. Considering the K-Nearest Neighbor (K-NN) classifier and performing a 10-fold Stratified CrossValidation, 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. Does K-NN perform on average better or worse than the Naïve Bayes classifier on the analyzed data? Report a screenshot showing the confusion matrix achieved by Naïve Bayes on the analyzed dataset.

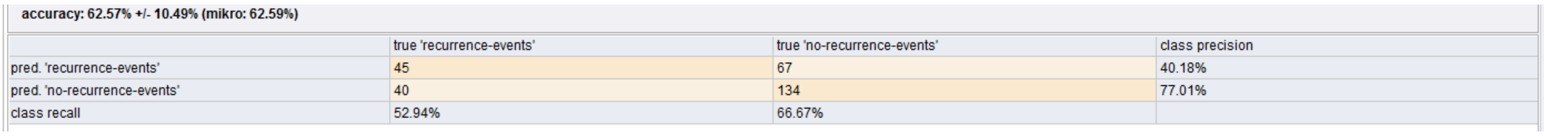
K=1



accuracy: 66.44% +/- 6.91% (mikro: 66.43%)

	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%	

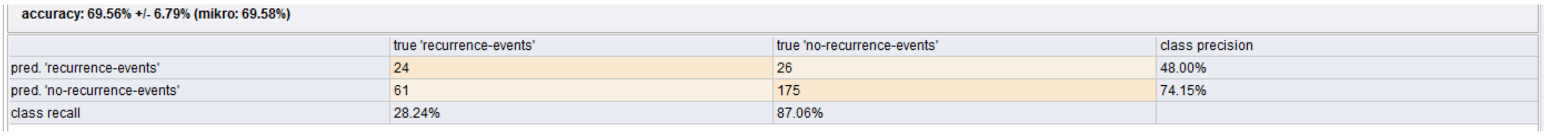
K=2



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%	

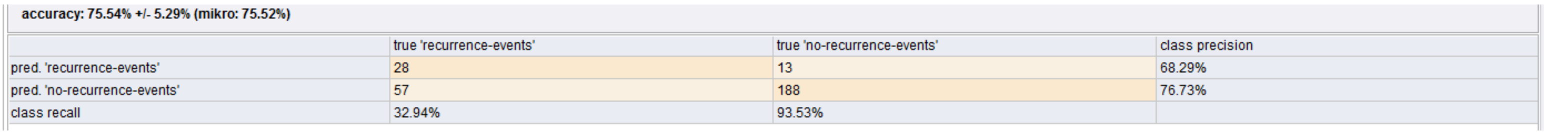
K=3



accuracy: 69.56% +/- 6.79% (mikro: 69.58%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	26	48.00%
pred. 'no-recurrence-events'	61	175	74.15%
class recall	28.24%	87.06%	

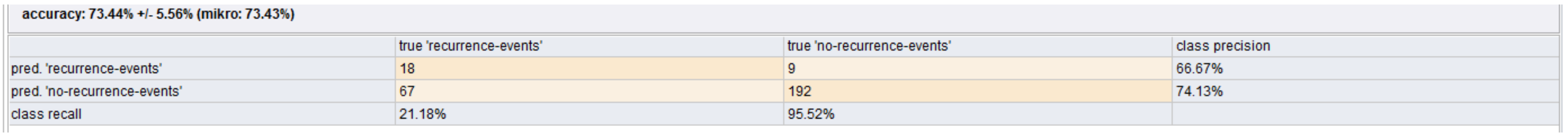
K=10



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%	

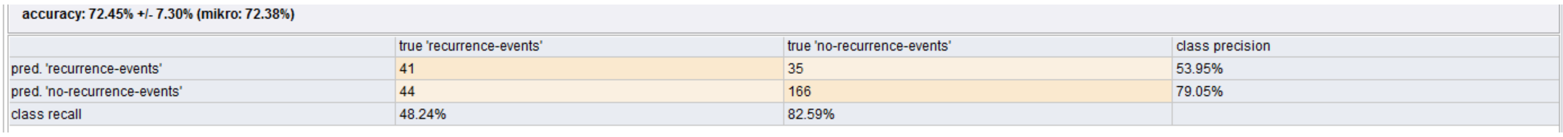
K=20



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%	

NAIVE BAYES

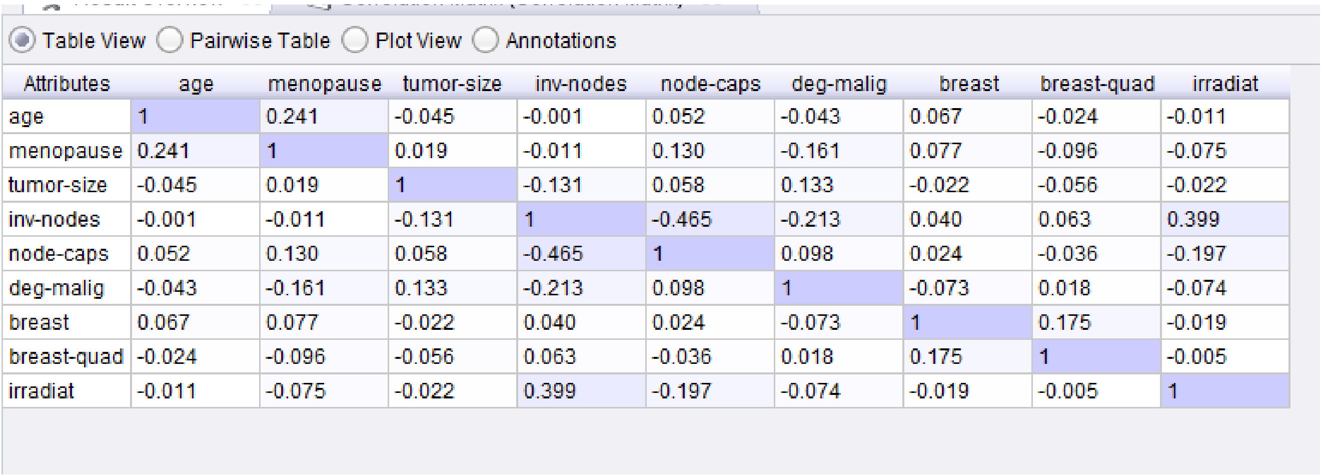


accuracy: 72.45% +/- 7.30% (mikro: 72.38%)

	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%	

On the analyzed dataset Naive Bayes performs on average better than the K-NN:

5. Analyze the Correlation Matrix to discover pairwise correlations between data attributes. Report a screenshot showing the correlation matrix achieved. (a) Does the Naïve independence assumption actually hold for the Breast dataset? (b) Which is the pair of most correlated attributes?



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) Does the Naive independence assumption actually hold for the Breast dataset?
No, because there are some high values in absolute value.

(b) Which is the pair of most correlated attributes?
The pair of most correlated attributes is (node-caps, inv-nodes).