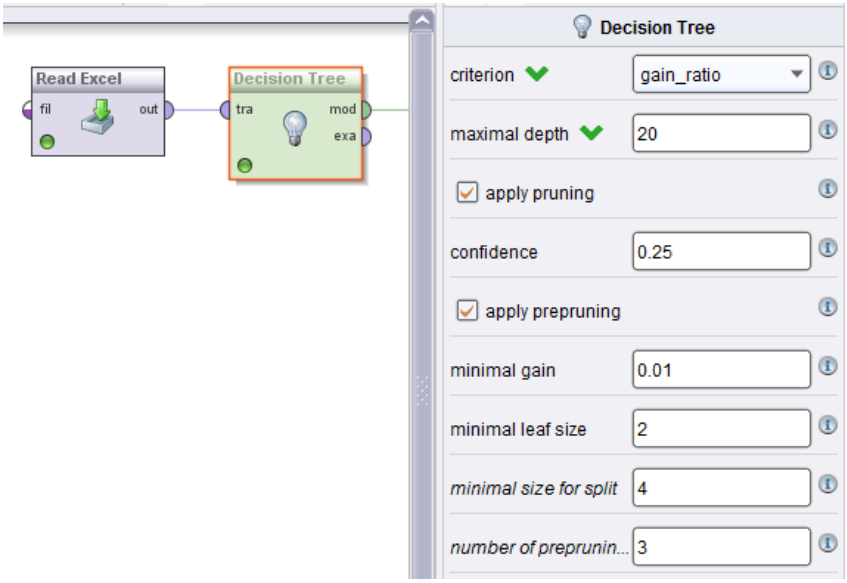
HomeWork 04 – Martino Mensio 191854/232297

1) Decision Tree

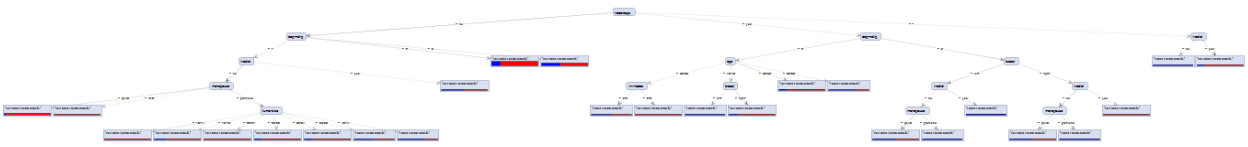


a) The most discriminative attribute for class prediction is node-caps.
b) The height of the tree is 6.
c) Pure partitions (circled in green)

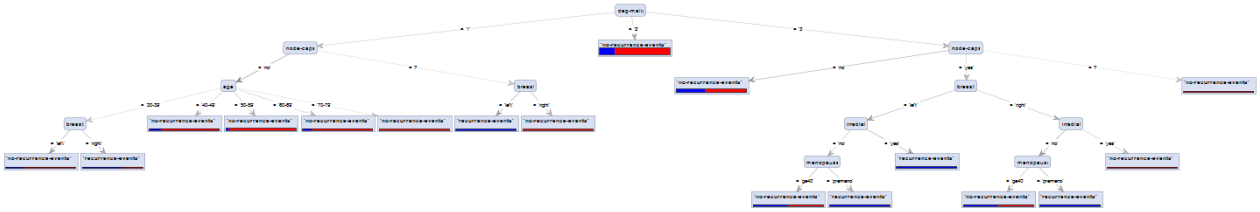
2) Impact of parameters of Decision Tree

Increasing the value of minimal gain:

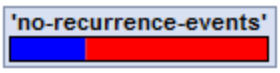
- Minimal gain: 0.05, max depth: 20



- Minimal gain: 0.08, ,max depth: 20



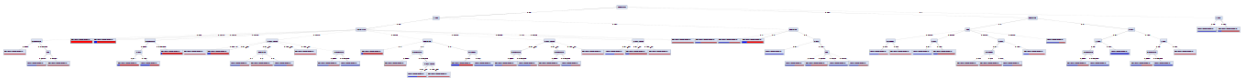
- Minimal gain: 0.1, max depth: 20



Conclusion: with higher values of minimal gain, a lower number of splits is done.

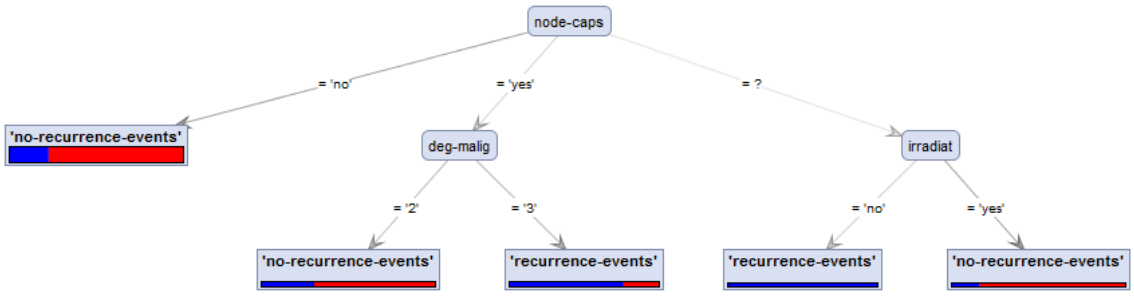
Decreasing the value of max depth:

- minimal gain: 0.01, max depth: 10



- minimal gain: 0.01, max depth: 5

- minimal gain: 0.01, max depth: 3



Conclusion: the height of the tree decreases if we limit the max depth of leaf nodes.

3) 10-fold stratified Cross-Validation

- minimal gain: 0.01, max depth: 20

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%	

- minimal gain: 0.05, max depth: 20

accuracy: 64.66% +/- 7.90% (mikro: 64.69%)			
	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	28	44	38.89%
pred. 'no-recurrence-events'	57	157	73.36%
class recall	32.94%	78.11%	

- minimal gain: 0.08, max depth: 20

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%	

- minimal gain: 0.1, max depth: 20

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%	

- minimal gain: 0.5, max depth: 20

accuracy: 70.30% +/- 1.36% (mikro: 70.28%)			
	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%	

- minimal gain: 0.01, max depth: 10

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%	

- minimal gain: 0.01, max depth: 5

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%	

- minimal gain: 0.01, max 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%	

4) K-NN

Impact of K

- 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 = 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%	

- K = 5

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%	

- K = 6

accuracy: 73.09% +/- 5.72% (mikro: 73.08%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	34	26	56.67%
pred. 'no-recurrence-events'	51	175	77.43%
class recall	40.00%	87.06%	

- K = 7

accuracy: 74.84% +/- 5.91% (mikro: 74.83%)

	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%	

- 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%	

- K = 9

accuracy: 75.20% +/- 4.91% (mikro: 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%	

- 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 = 15

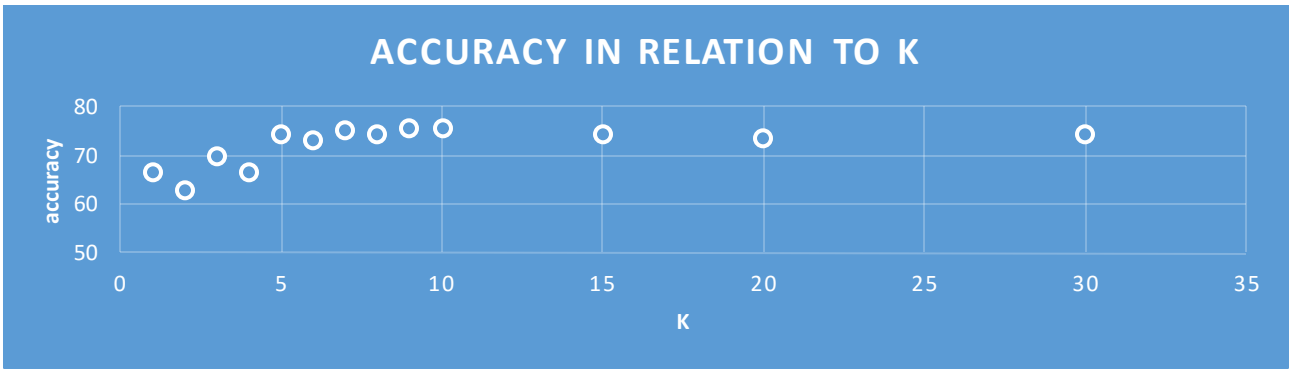
accuracy: 74.13% +/- 5.38% (mikro: 74.13%)			
	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	18	7	72.00%
pred. 'no-recurrence-events'	67	194	74.33%
class recall	21.18%	96.52%	

- 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%	

- K = 30

accuracy: 74.11% +/- 3.64% (mikro: 74.13%)			
	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	15	4	78.95%
pred. 'no-recurrence-events'	70	197	73.78%
class recall	17.65%	98.01%	



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%	

Naive Bayes classifier works better than K-NN only for low N (<=4).

5) Correlation Matrix

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

The highest value is 0.399, while the minimum is -0.465, so the maximum correlation in absolute value is less than 0.5. This implies that the attributes are not strongly correlated, so the Bayesian assumption is not so wrong.

The pair of most correlated attributes (absolute value) is

inv-nodes	node-caps	-0.465

Considering only positive correlation:

inv-nodes	irradiat	0.399