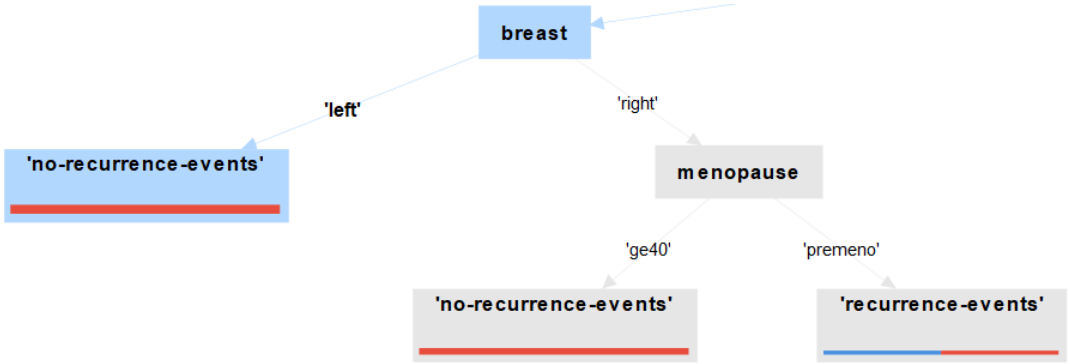
S330591 Lorenzo Stefanoni Homework 2

Question 1

a) The most discriminative attribute is the root of the decision tree, so in this scenario is “node caps”.
b) The height of a graph is equal to the longest path between the root and one of the leaves, and in this case is equal to seven.
c) In this image there are two pure partitions: one highlighted and a second one under “ge40”. All the leaves that have the same color under their name are pure partitions.



Question 2

The minimal gain threshold determines if a node should split or not. If when the threshold is compared to the gain of a leaf this last one is higher the split happens. This means that using a too high minimal gain can result in a single node tree.

On the other hand, the maximal depth is a constraint on the maximum height of the tree.

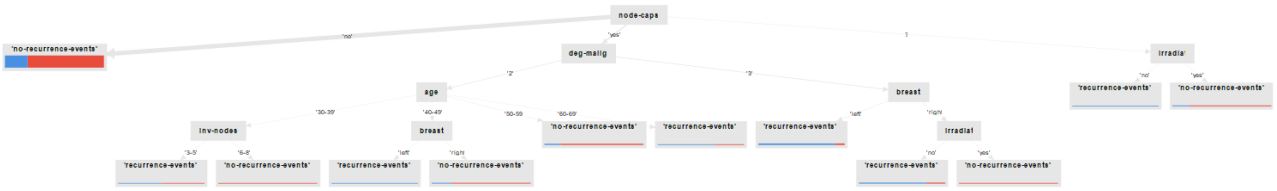


Figure 1: Minimal gain = 0.05, Maximal depth = 10

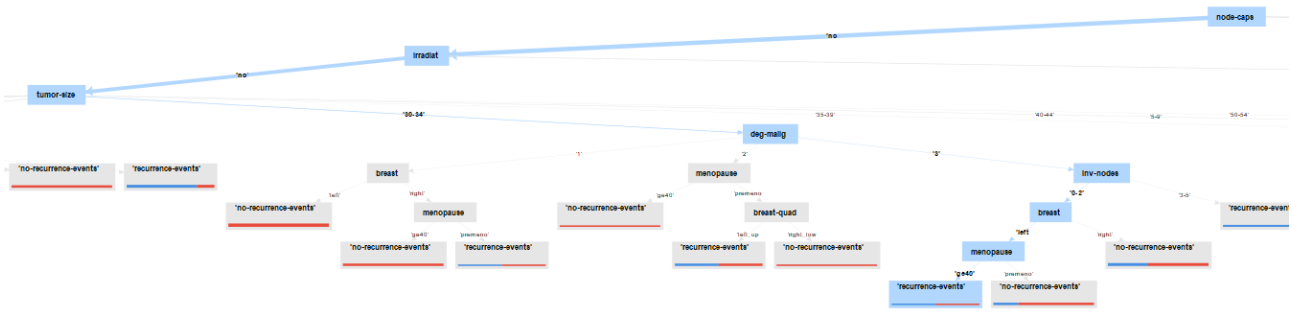


Figure 2: Minimal gain = 0.001, Maximal depth = 10

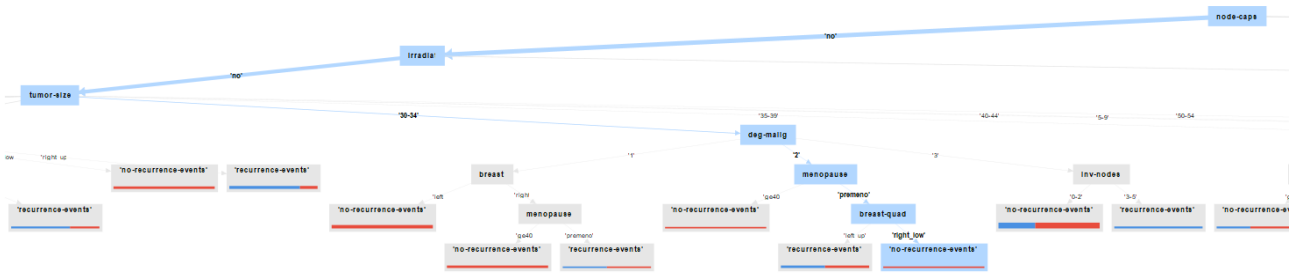


Figure 3: Minimal gain = 0.005, Maximal depth = 10

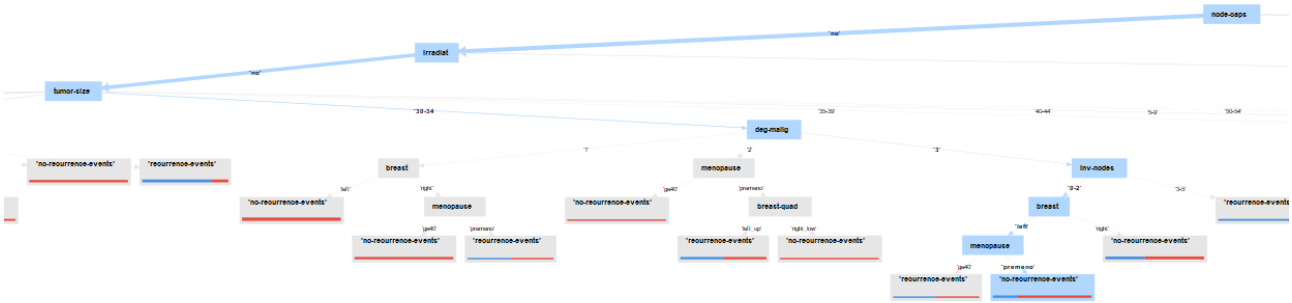


Figure 4: Minimal gain = 1e-100, Maximal depth = 10

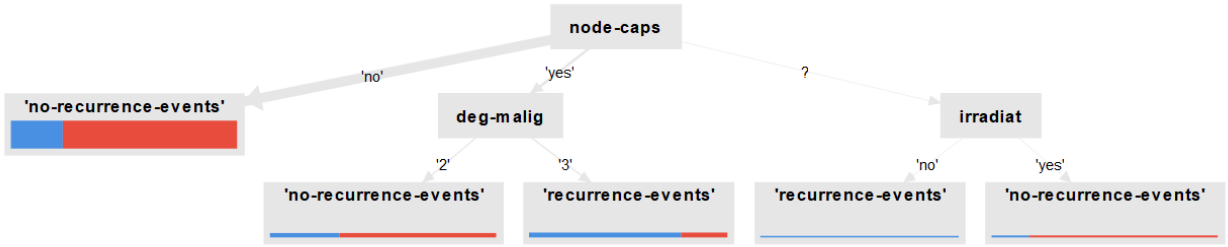


Figure 5: Minimal gain = 1e-100, Maximal depth = 3

I personally wanted to see how far I could take the values to the limit; figures 3 and 4 show the same tree although they have a different gain value; all the freedom Rapid Miner had in these two graphs was lost with the last one, where the maximal depth has limited all the job.

Question 3

accuracy: 70.64% +/- 6.20% (micro average: 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%	

Figure 6: Minimal gain = 0.05, Maximal depth = 10

accuracy: 67.48% +/- 6.59% (micro average: 67.48%)

	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%	

Figure 7: Minimal gain = 0.001, Maximal depth = 10

accuracy: 66.44% +/- 7.66% (micro average: 66.43%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	37	48	43.53%
pred. 'no-recurrence-events'	48	153	76.12%
class recall	43.53%	76.12%	

Figure 8: Minimal gain = 0.005, Maximal depth = 10

accuracy: 66.44% +/- 7.66% (micro average: 66.43%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	37	48	43.53%
pred. 'no-recurrence-events'	48	153	76.12%
class recall	43.53%	76.12%	

Figure 9: Minimal gain = 1e-100, Maximal depth = 10 (the same as the previous one, as a confirm that is the same tree)

accuracy: 74.82% +/- 6.64% (micro average: 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%	

Figure 10: Minimal gain = 1e-100, Maximal depth = 3

accuracy: 72.03% +/- 6.22% (micro average: 72.03%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	21	55.32%
pred. 'no-recurrence-events'	59	180	75.31%
class recall	30.59%	89.55%	

Figure 11: Minimal gain = 0.05, Maximal depth = 5

The best values for the minimal gain and maximal depth are the ones in the middle. Talking about the minimal gain, the lower the value is, the more accurate it is, but going too low there is the overfitting

problem which occurs when the data are too tightly coupled to the current situation and therefore do not allow for prediction of other input values.

Question 4

accuracy: 65.73% +/- 8.62% (micro average: 65.73%)

	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%	

Figure 12: K = 2

accuracy: 67.86% +/- 7.40% (micro average: 67.83%)

	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%	

Figure 13: K = 4

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%	

Figure 14: K = 8

accuracy: 73.79% +/- 5.61% (micro average: 73.78%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	17	7	70.83%
pred. 'no-recurrence-events'	68	194	74.05%
class recall	20.00%	96.52%	

Figure 15: K = 20

accuracy: 73.44% +/- 3.67% (micro average: 73.43%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	12	3	80.00%
pred. 'no-recurrence-events'	73	198	73.06%
class recall	14.12%	98.51%	

Figure 16: K = 50

Looking at the different values of accuracy we can see that in a first moment as K increases the accuracy increases too, and this because K-NN is a technique that establish the membership of an item by its k

neighbors, so the more are them and the more accurate is the decision. At a certain point the accuracy starts to fall because of the overfitting; the result became too much dependent from the input set of data.

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%	

Figure 17: Confusion matrix of Naive – Bayes

It depends by the value of K; as we can see looking at the screens above if K is at least equal to 8 the K-NN is better than the Naïve-Bayes

Question 5

Attributes	age	menopa...	tumor-s...	inv-nod...	node-ca...	deg-malig	breast	breast-...	irradiat
age	1	0.241	-0.045	-0.001	0.052	-0.043	0.067	-0.024	-0.011
menopau...	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-qu...	-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

In this scenario the Naïve hypothesis allows us to obtain meaningful data. As we can see looking at the table, the most correlated attributes are inv-nodes with irradiant, reaching a value of 0.399 (normalized).