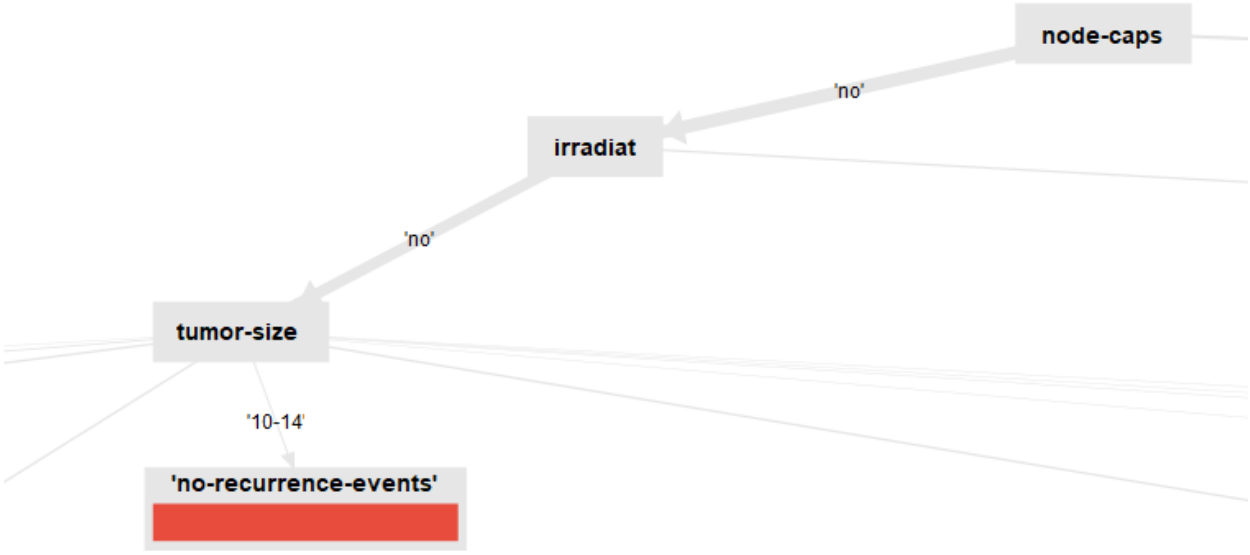
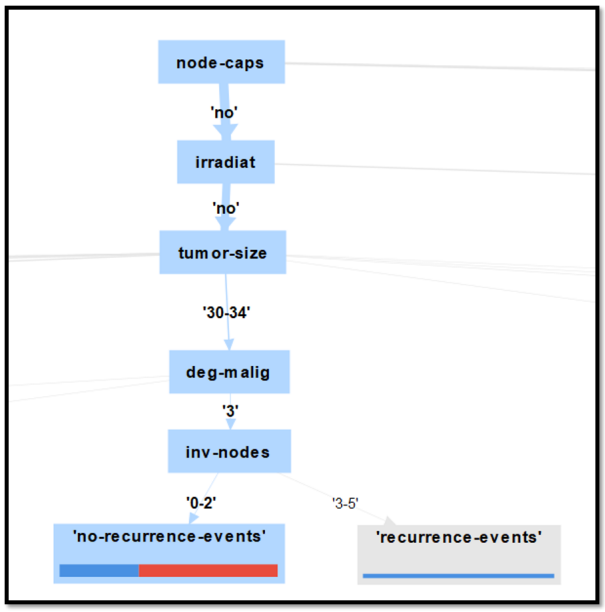
Report

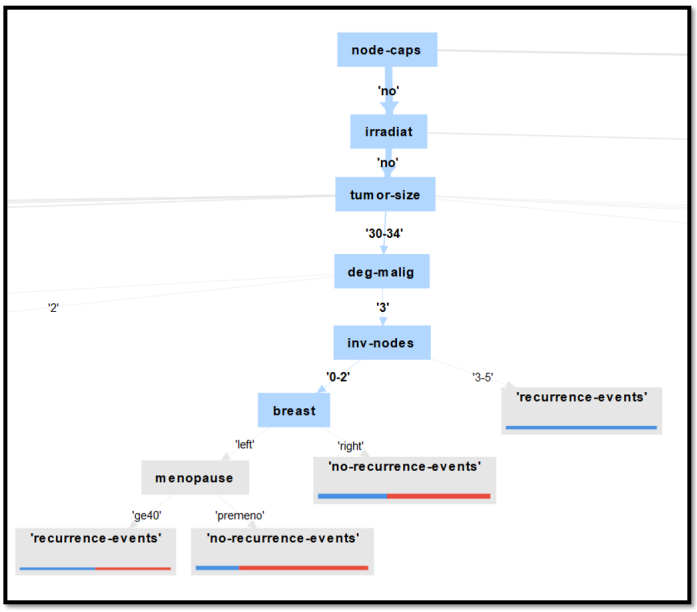
Answers Question 1

1. The most discriminative attribute is "node-caps", the root node. This means that the algorithm which builds the decision tree consider "Node-caps" as the best attribute to split the dataset.
2. The height of the decision tree generated is seven.
3. *Find a pure partition in the Decision Tree and report a screenshot that shows the example identified.*



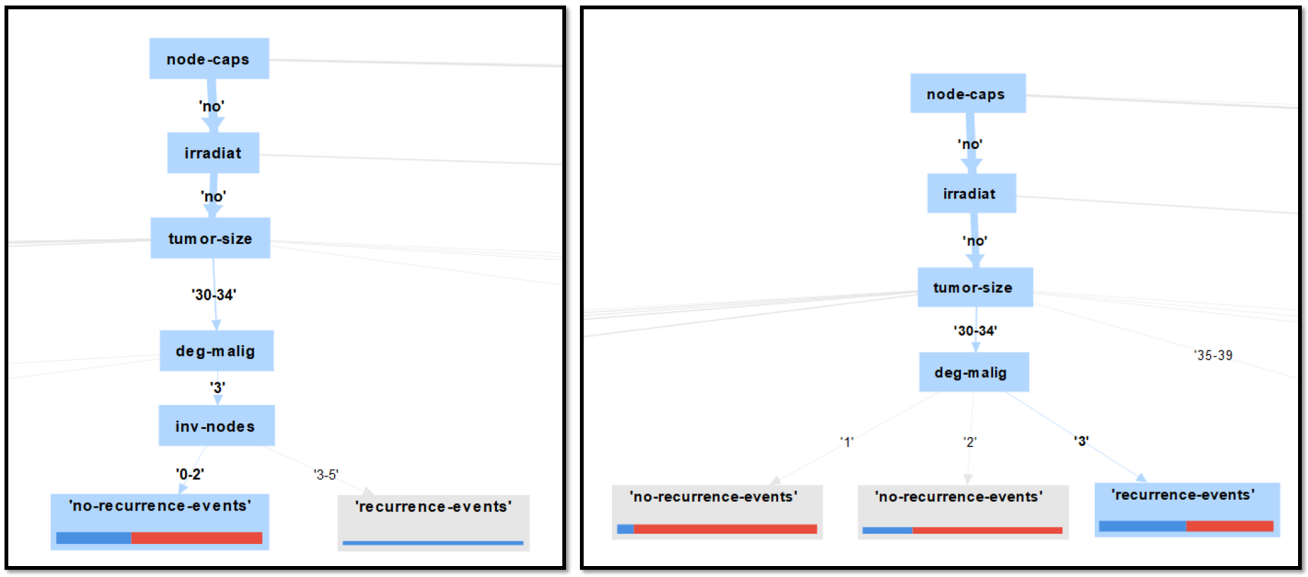
Answer Question 2



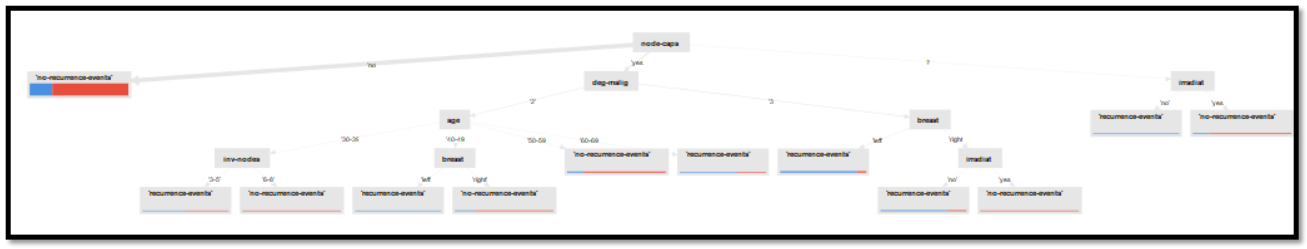


Both trees have 10 as maximal depth, but the first one has 0.01 as minimal gain (default

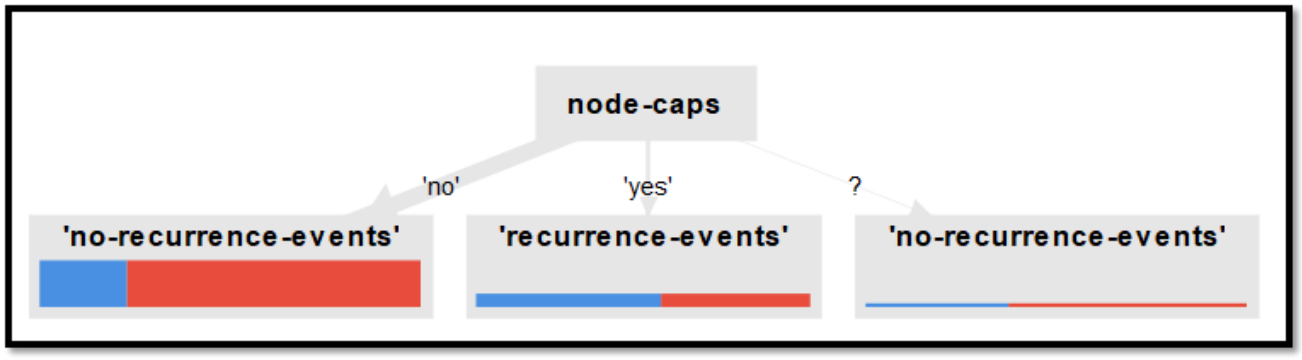
value), while the second one is 0.001. As we can see, when the miminal gain is lower, more splits are generated.



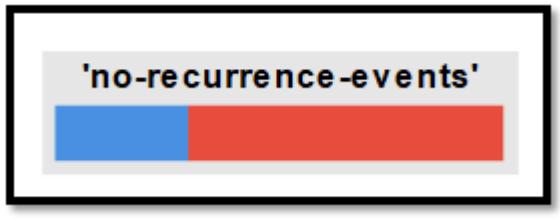
In this example, the maximal depth of the tree is decreased to 5. The same path of the last example now has 5 levels and all the aspects related to inv-nodes are collected directly in the classification node.



In this example, the increased minimal gain is 0.05. This has as a consequence the height's decrease, as the values has to have a more significant impact, so the small number and path are merged.



Here, the maximal depth is set to 2.



Finally, this is the result of setting the minimal gain to 0.1.

Answer Question 3

Decreasing these parameters has effects on the precision of classification and the risk of having overfitted or underfitted results. For example, increasing maximal depth can create more detailed trees, but have a higher risk of overfitting and so they wouldn't be able to generalize the information.

Maximal depth = 10; Minimal Gain = 0.001

	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%	

Maximal depth = 5; Minimal Gain = 0.01

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	35	35	50.00%
pred. 'no-recurrence-events'	50	166	76.85%
class recall	41.18%	82.59%	

Maximal depth = 10; Minimal Gain = 0.05

	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%	

Maximal depth = 2; Minimal Gain = 0.01

	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%	

Maximal depth = 10; Minimal Gain = 0.1

	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%	

Answer Question 4

The value of K in the K-NN classifier represent the error rate of the model, so changing the value of k has effects on the precision of the validation. In particular, we can wee that if k is

too small, it is influenced by noise points and be overfitted, while if it's too large, it could be inserted in other classes and be underfitted.

k=5 (Default case)

	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%	

k=2

	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%	

k=3

	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%	

k=1

	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=9

	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%	

Naïve Bayes

	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%	

The result of comparing the K-NN performance with k = 5 and the Naïve Bayes is that the

perfomance of the second one is slightly better of the K-NN.

Answer Question 5

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

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

First Attribute	Second Attribute	Correlation ↓
inv-nodes	irradiat	0.399
age	menopause	0.241
breast	breast-quad	0.175
tumor-size	deg-malig	0.133
menopause	node-caps	0.130
node-caps	deg-malig	0.098
menopause	breast	0.077
age	breast	0.067
inv-nodes	breast-quad	0.063
tumor-size	node-caps	0.058
age	node-caps	0.052
inv-nodes	breast	0.040
node-caps	breast	0.024
menopause	tumor-size	0.019
deg-malig	breast-quad	0.018
age	inv-nodes	-0.001
breast-quad	irradiat	-0.005
menopause	inv-nodes	-0.011

We can see that some attributes are dipentent from each others, so the assumption of indipendence of Naive Classification is not correct.

The most correlated attributes are inv-nodes and irradiant, followed by age and menopause.