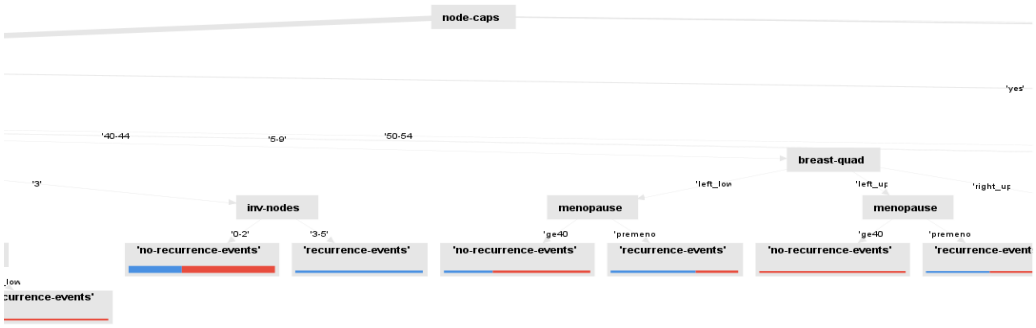
DOMANDA 1

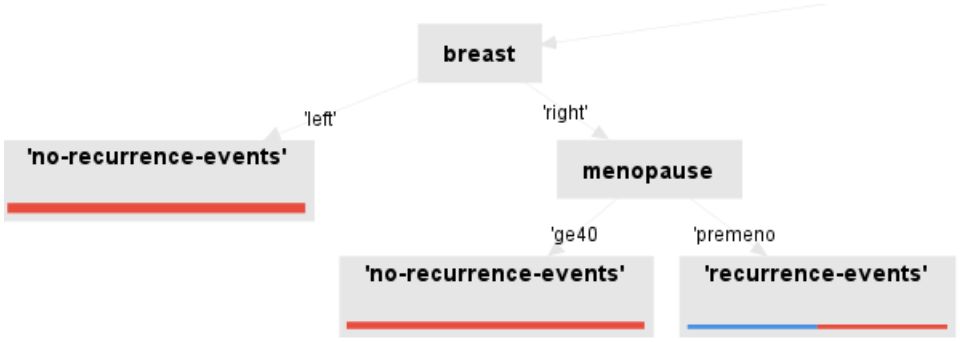
a)



L'algoritmo Decision Tree posiziona sempre l'attributo più selettivo nel nodo radice dell'albero, cioè nel nostro caso "node-caps"

b) L'altezza di un albero è determinata dalla lunghezza massima tra il nodo radice e un qualsiasi nodo foglia. L'albero generato da questo dataset è di altezza 6

c)



Un partizionamento puro si verifica quando uno split su un valore di un attributo dell'albero produce come risultato una partizione in cui tutti gli elementi appartengono alla stessa classe. Nel nostro caso possiamo vedere come la partizione ottenuta dal nodo "breast" scegliendo il ramo dello split "left" sia un partizionamento puro poiché tutti i suoi elementi appartengono solo alla classe rossa

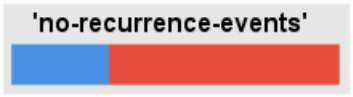
DOMANDA 2

Il valore del parametro "minimal-gain" stabilisce la soglia per cui l'albero deve continuare ad effettuare uno split nei nodi o meno. Uno split viene effettuato se il suo valore di "gain" è maggiore del valore del parametro "minimal-gain". Questo parametro determina sia il numero di foglie dell'albero sia la sua altezza poiché, se viene impostato un valore basso di "minimal-gain" allora l'albero tenderà a splittare più nodi e sarà più profondo e più fitto, d'altra parte con alti valori di "minimal-gain" l'albero avrà meno foglie e sarà più basso. Invece il valore del parametro "maximal-depth" stabilisce il valore massimo di altezza che l'albero può raggiungere, se è minore dell'altezza dell'albero esso verrà tagliato.

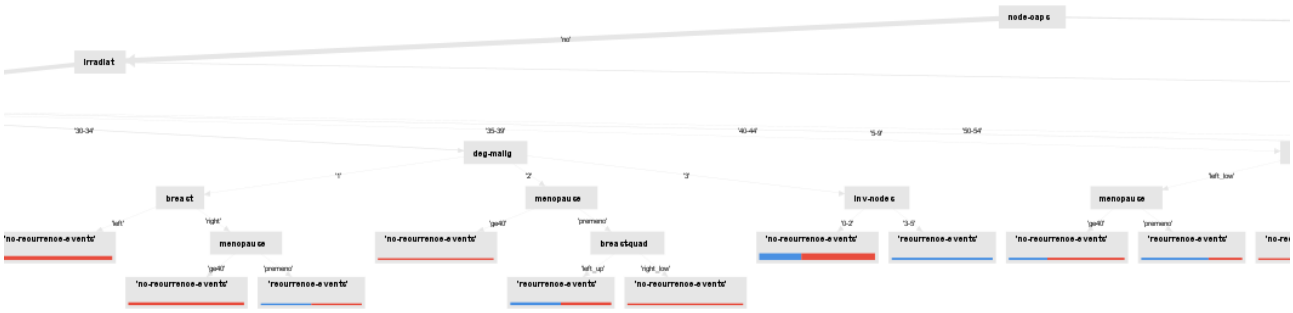
Proviamo a variare i valori di "minimal-gain" e "maximal-depth" per vedere il risultato.

-Prima modifichiamo solo il valore di "minimal-gain" che di default è impostato ad 0.01. Per valori di poco più alti rispetto al default, es. 0.04, otterremo un albero meno fitto e alto rispetto a quello di default

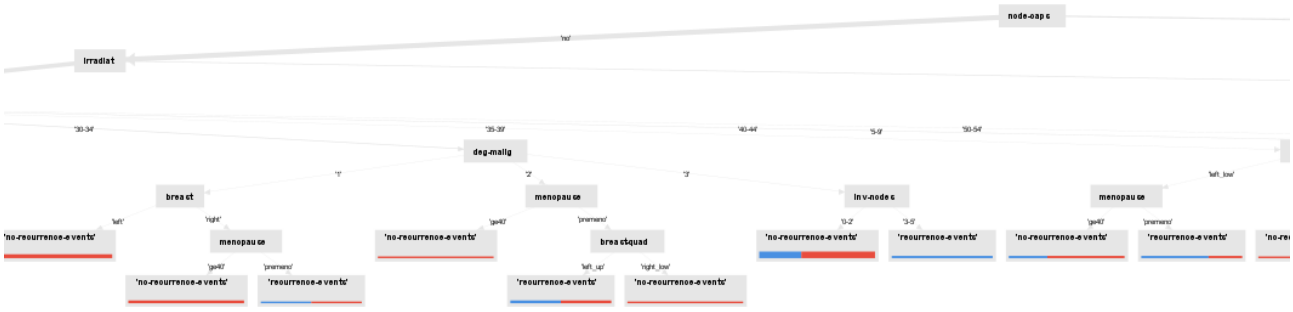
- Mentre per valori molto più alti, es. 0.1, l'albero verrà compresso tutto in una foglia dato che non riesce a calcolare il gain necessario per splittarsi



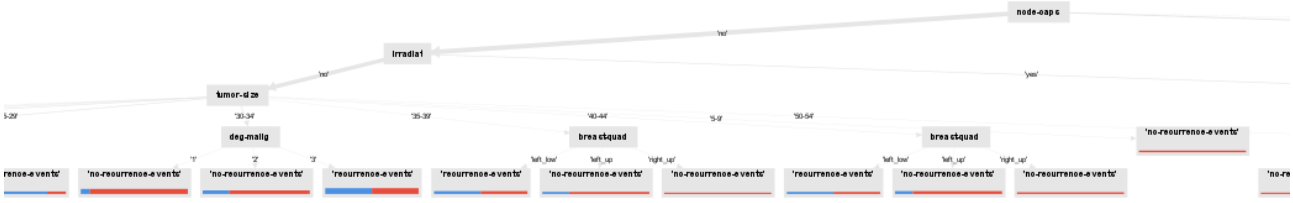
-Se invece impostiamo un valore minore di 0.01 non noteremo alcun cambiamento nell'albero generato



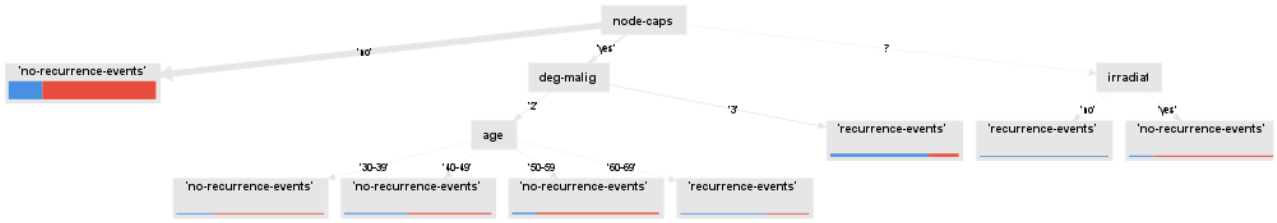
-Modifichiamo ora solo il valore di "maximal-depth". Impostiamo prima il suo valore a 10, in modo tale che sia maggiore dell'altezza effettiva dell'albero e quindi possiamo osservare alcun cambiamento all'albero



-Mentre se si imposta un valore minore, es. 5, otterremo un taglio dell'albero che andrà ad eliminare tutti i nodi con altezza maggiore di 5



-Infine, modifichiamo entrambi i valori di “minimal-gain” e “maximal-depth”, impostandoli rispettivamente a: minimal-gain:0.04 e maximal-depth:4, in cui l’albero potrebbe avere più nodi in profondità poiché il minimal-gain lo permetterebbe ma il maximal-depth dli impedisce di avere un’altezza superiore a 4



DOMANDA 3

In generale, riducendo il valore del minimal gain e aumentando la maximal depth si genera un modello di classificazione più dettagliato e quindi più accurato. Tuttavia, sulla base dei risultati ottenuti in precedenza otteniamo in alcuni casi l’effetto denominato “overfitting”, ovvero il modello risulta troppo “focalizzato” sui dati di train per classificare in modo accurato nuovi dati di test.

accuracy: 70.30% +/- 1.43% (micro average: 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:0.1, maximal depth 10

accuracy: 69.95% +/- 6.37% (micro average: 69.93%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	25	48.98%
pred. 'no-recurrence-events'	61	176	74.26%
class recall	28.24%	87.56%	

Minimal gain:0.04, maximal depth:10

accuracy: 70.28% +/- 7.75% (micro average: 70.28%)

	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%	

Minimal gain:0.01, maximal depth:5

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%	

Minimal gain:0.01, maximal depth:7

accuracy: 70.30% +/- 7.18% (micro average: 70.28%)

	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%	

Minimal gain:0.04, maximal depth: 4

DOMANDA 4

Incrementando il valore di K, nell'algoritmo K-NN, il classificatore considera un numero maggiore di dati di train "vicini" al dato di test e quindi l'accuratezza media cresce: 73.77% con K=5, 74.51% con K=8, 75.20% con K=10, 70.26% con K=13, 73.79% con K=20. Considerando un numero molto elevato di record di train "vicini", nel nostro caso quando K>10, la presenza di dati rumorosi comincia ad inficiare le performance di classificazione e dunque l'accuratezza media di classificazione diminuisce leggermente. L'accuratezza di Naïve Bayes risulta essere inferiore al K-NN con K=5, quindi per valori di K superiori a 5 conviene utilizzare l'algoritmo K-NN per ottenere un'accuratezza maggiore

K=5

accuracy: 73.77% +/- 5.98% (micro average: 73.78%)

	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%	

K=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%	

K=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%	

K=13

accuracy: 70.26% +/- 7.23% (micro average: 70.28%)

	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=20

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%	

Mentre utilizzando Naïve Bayes

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%	

DOMANDA 5

La figura sottostante mostra la matrice di correlazione ottenuta dal dataset analizzato. Essa riporta la correlazione mutua e simmetrica tra coppie di attributi. La coppia di attributi “irradiant” e “inv-nodes” è la coppia col più alto valore di correlazione. L’ipotesi d’indipendenza Naïve risulta essere irrealistica per il dataset analizzato, in quanto i dati del dataset presentano una dipendenza tra di loro non trascurabile, e ciò non si può conciliare con l’ipotesi Naïve.

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