Data Science e Tecnologie per le Basi di Dati

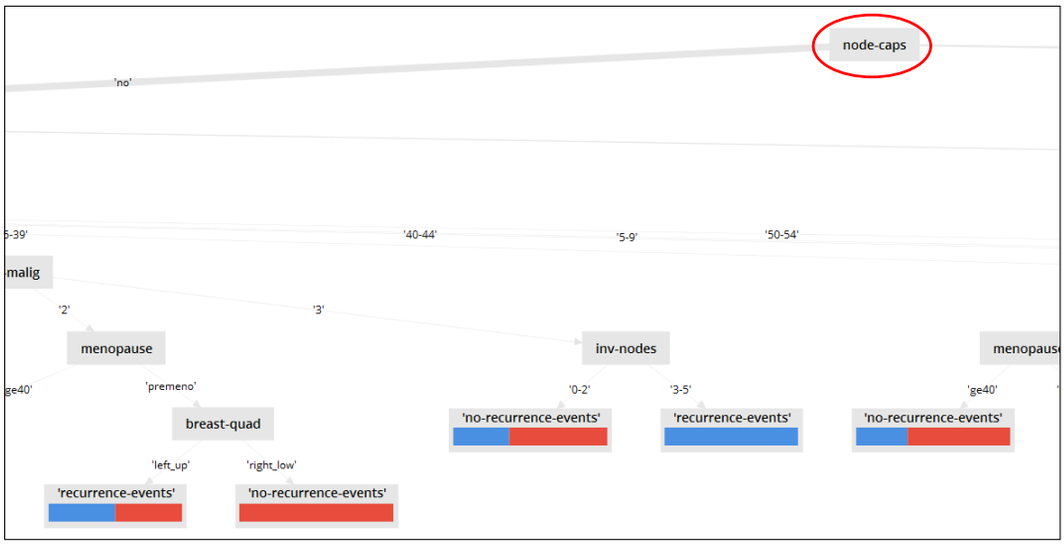
Quaderno #2 – Data mining

1)

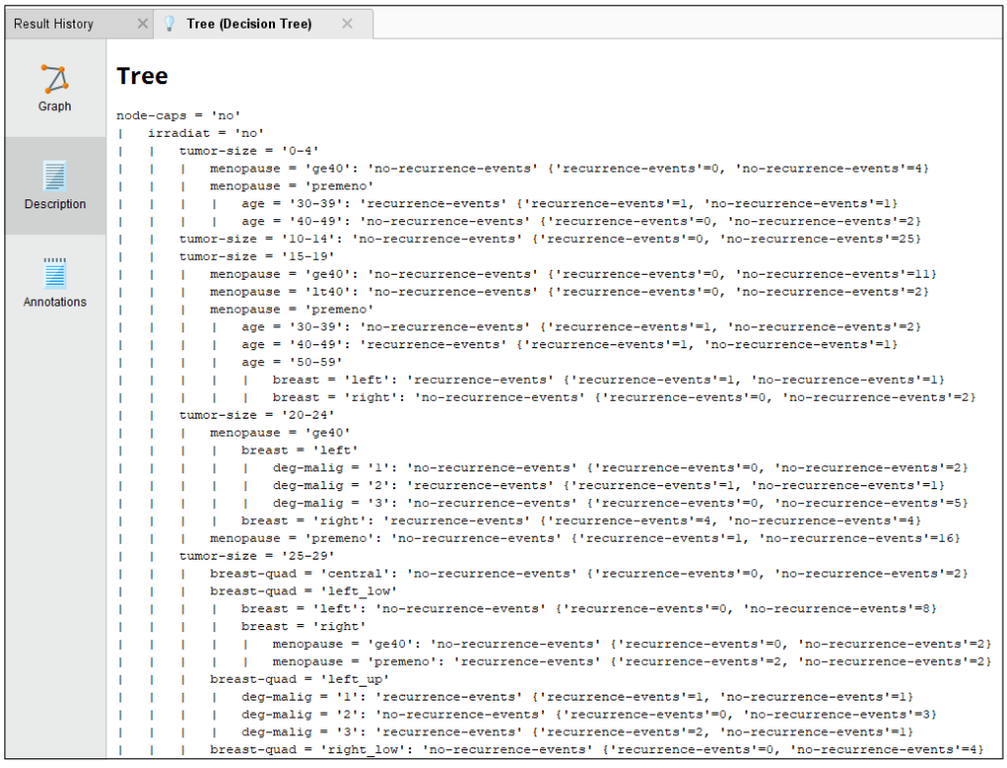
a. L'attributo considerato dall'algoritmo come il più selettivo al fine di predire la classe di un nuovo dato di test è 'node-caps', essendo il nodo radice dell'albero di decisione.

Maximal depth: 20

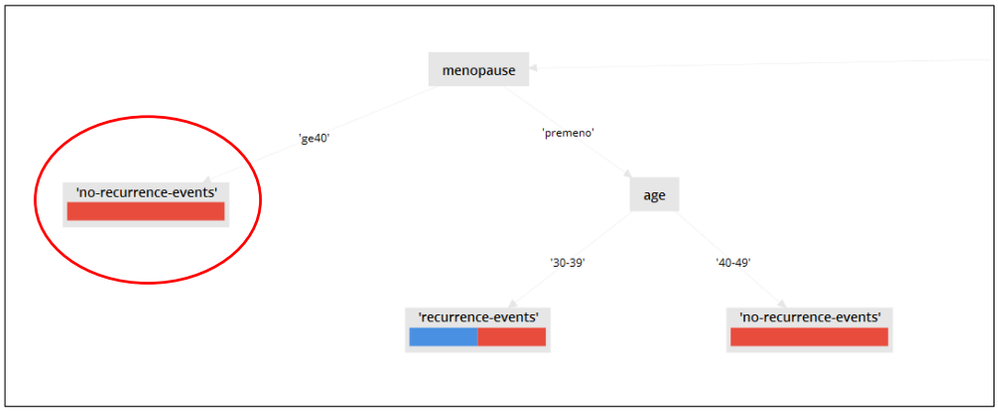
Minimal gain: 0.01



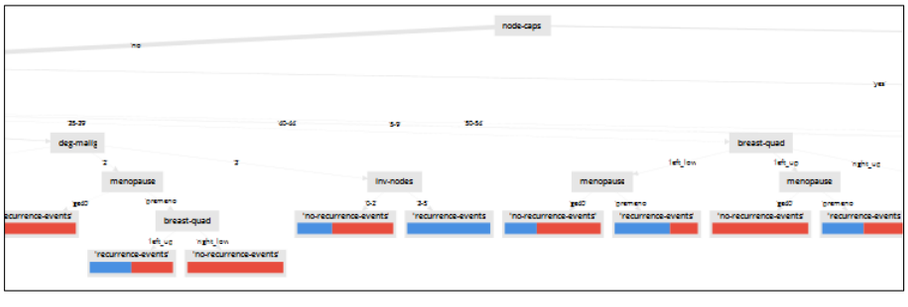
b. L'altezza dell'albero, ovvero la lunghezza massima di un percorso che collega la radice ad una foglia dell'albero è 6.
Si può anche calcolare dalla sezione 'Description' di 'Result', una volta runnato il processo, contando manualmente il 'numero di split'.



c. Un esempio di partizionamento puro all'interno dell'albero di decisione generato è visibile in corrispondenza dell'attributo 'menopause', dove la prima partizione è proprio un partizionamento puro, dato che tutte le relative istanze appartengono alla classe 'no-recurrence-events', mentre la seconda copre record di entrambe le classi.

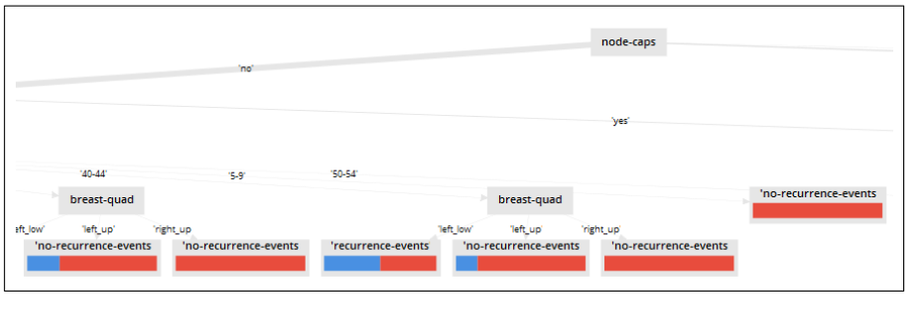


2) Modificando i parametri 'maximal depth' e 'minimal gain' si va a modificare rispettivamente l'altezza massima dell'albero di decisione e la soglia minima di split di un nodo (un nodo viene splittato se il suo gain è superiore alla soglia minima indicata).
Detto ciò, modificare il parametro 'maximal depth', il cui valore di default è pari a 20, non causerà nessun cambiamento all'albero di decisione finché non si scenderà sotto all'altezza massima registrata al momento.



Maximal depth: 15

Minimal gain: 0.01

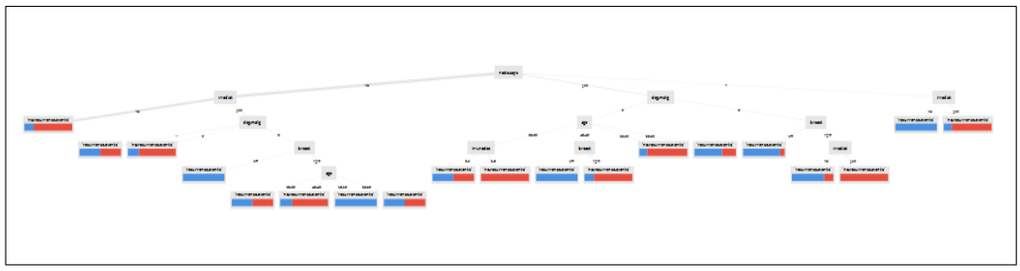


Maximal depth: 5

Minimal gain: 0.01

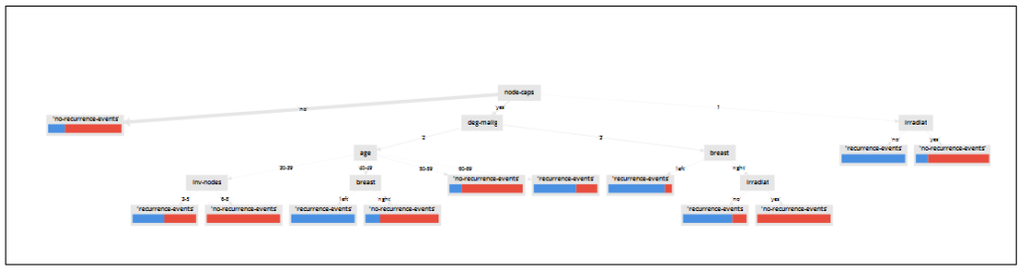
Maximal depth: 20

Minimal gain: 0.03



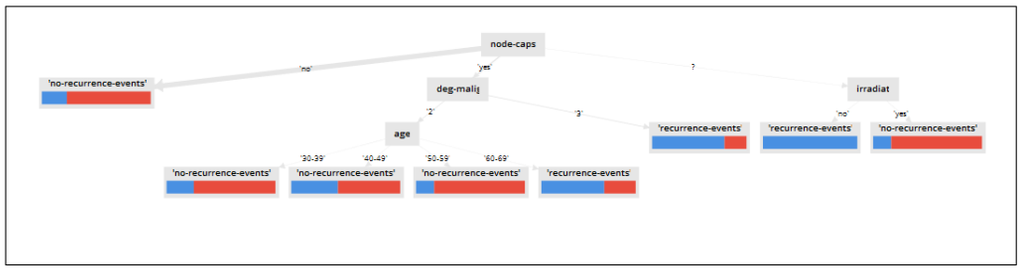
Maximal depth: 20

Minimal gain: 0.05



Maximal depth: 4

Minimal gain: 0.05



3)

Matrici di confusione:

accuracy: 67.48% +/- 6.25% (mikro: 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%	

Maximal depth: 15

Minimal gain: 0.01

accuracy: 70.28% +/- 7.35% (mikro: 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%	

Maximal depth: 5

Minimal gain: 0.01

accuracy: 70.31% +/- 5.57% (mikro: 70.28%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	24	50.00%
pred. 'no-recurrence-events'	61	177	74.37%
class recall	28.24%	88.06%	

Maximal depth: 20

Minimal gain: 0.03

accuracy: 70.64% +/- 5.89% (mikro: 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%	

Maximal depth: 20

Minimal gain: 0.05

accuracy: 71.33% +/- 6.24% (mikro: 71.33%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	23	53.06%
pred. 'no-recurrence-events'	59	178	75.11%
class recall	30.59%	88.56%	

Maximal depth: 4

Minimal gain: 0.05

Da queste figure si può dunque dedurre che:

Riducendo il valore di 'minimal gain' e aumentando 'maximal depth' si genera un modello di classificazione più dettagliato e quindi più accurato. Tuttavia, sulla base dei risultati riportati nelle figure precedenti, impostando valori di maximal depth superiori a 5 e minimal gain inferiori a 0.05 si produce l'effetto denominato *overfitting*, ovvero il fenomeno per cui il modello si adatta troppo ai dati di training per classificare in modo accurato nuovi dati.

4)

Matrici di confusione:

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-Nearest Neighbor, K = 1

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-Nearest Neighbor, K = 2

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-Nearest Neighbor, K = 3

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-Nearest Neighbor, K = 4

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-Nearest Neighbor, K = 5

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%	

Naïve Bayes (Default settings)

Come si può dedurre dagli screenshot sopra riportati, Naïve Bayes ottiene mediamente prestazioni superiori rispetto al K-NN, avendo una accuracy media maggiore.
Si parla infatti del 72.45% (N-B) contro il 67.8% (K-NN, valore medio tra le simulazioni riportate).

5)

Matrice di correlazione:

Attribut...	age	menopa...	tumor-s...	inv-nodes	node-ca...	deg-mal...	breast	breast-...	irradiat
age	1	0.241	-0.045	-0.001	0.052	-0.043	0.067	-0.024	-0.011
menopa...	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-q...	-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

Alla luce dei risultati ottenuti, l'ipotesi di indipendenza Naïve non risulta del tutto valida per il dataset *Breast* utilizzato, visto che la correlazione tra gli attributi è diversa da 0. Tuttavia si parla pur sempre di valori molto bassi, tendenti allo 0, motivo per cui il classificatore riesce comunque a produrre un risultato accettabile.

First Att...	Second Attribute	Corr... ↑
inv-nodes	node-caps	-0.465
inv-nodes	deg-malig	-0.213
node-caps	irradiat	-0.197
menopa...	deg-malig	-0.161
tumor-size	inv-nodes	-0.131
menopa...	breast-quad	-0.096

Come si evince dalla figura sopra riportata, la coppia di attributi maggiormente correlati in valore assoluto (sarebbero correlati negativamente, in questo caso) è la coppia ('inv-nodes'-'node-caps').