Cazzola Michele s323270

Data Science e Tecnologie per le Basi di Dati

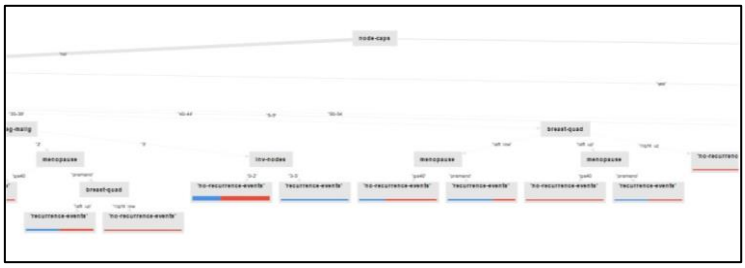
Quaderno #2 – Data mining

Esercizio 1

Generare un albero di decisione con l'algoritmo Decision Tree usando l'intero dataset per il training, settando il *minimal gain* a 0.01 e mantenendo la configurazione di default per gli altri parametri.

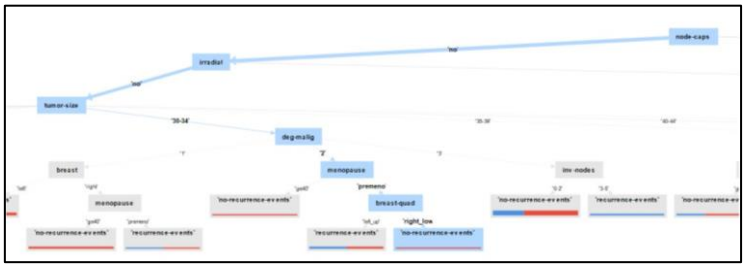
a) Quale attributo è considerato dall'algoritmo il più selettivo al fine di predire la classe di un nuovo dato di test?

L'attributo considerato più selettivo è ***node-caps***.



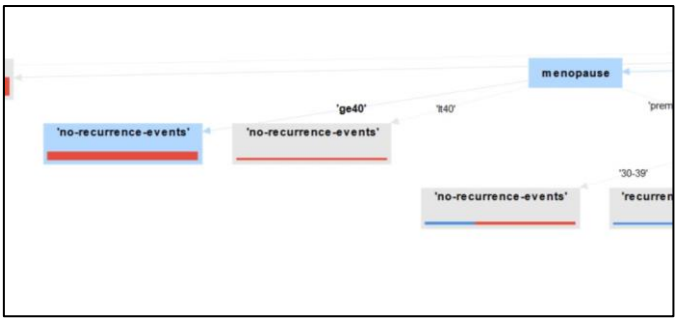
b) Qual è l'altezza dell'albero di decisione generato?

L'albero di decisione generato ha un'altezza pari a 7, includendo i nodi radice e foglia.



c) Trovare un esempio di partizionamento puro all'interno dell'albero di decisione generato e riportare un screenshot che mostri l'esempio trovato.
Il nodo consegue dal verificarsi delle condizioni seguenti:

- *node-caps* = 'no';
- *irradiat* = 'no';
- *tumor-size* = '15-19';
- *menopause* = 'ge40';

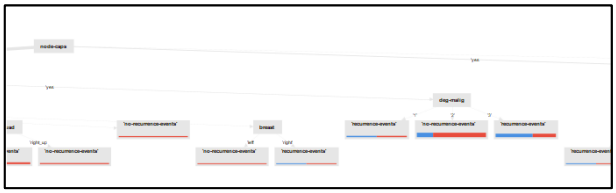


e genera la partizione seguente:

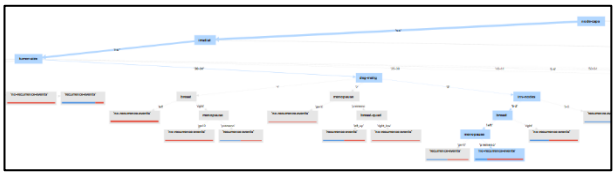
- *'recurrence-events'* = 0;
- *'no-recurrence-events'* = 11.

Esercizio 2

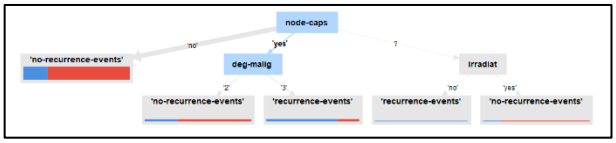
Analizzare l'impatto del minimal gain (considerando il gain ratio come criterio di splitting) e del maximal depth sulle caratteristiche dell'albero di decisione generato dall'intero dataset (mantenendo la configurazione di default per gli altri parametri di configurazione). Riportare almeno 5 screenshot differenti che mostrino gli alberi di decisione (o porzioni di essi) generati con differenti configurazioni.



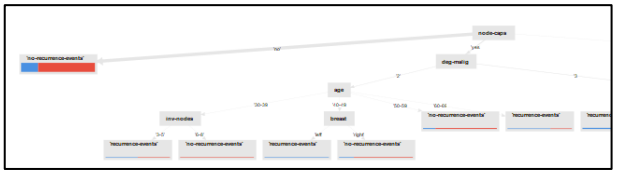
1: max_depth = 5, min_gain = 0.01



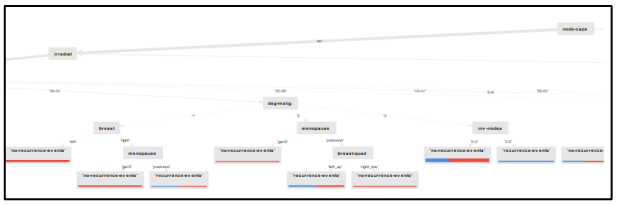
4: max_depth = 10, min_gain = 0.001



2: max_depth = 3, min_gain = 0.01



5: max_depth = 10, min_gain = 0.05



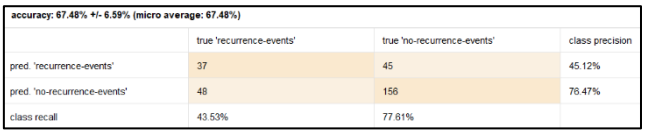
3: max_depth = 8, min_gain = 0.01

Si nota che:

- Nelle configurazioni 1 e 2, i partizionamenti sembrano essere meno puri, essendo l'altezza dell'albero inferiore;
- La configurazione 3 è analoga a quella di default, essendo variato il parametro *max depth*, ma rimanendo superiore all'altezza dell'albero nella configurazione di default (pari a 7);
- La configurazione 4 ha un'altezza dell'albero superiore a quello della configurazione di default, avendo diminuito il *minimal gain* e di conseguenza effettuato più split;
- La configurazione 5 ha un'altezza dell'albero inferiore a quello della configurazione di default, poiché il *minimal gain* è superiore e vengono effettuati meno split.

Esercizio 3

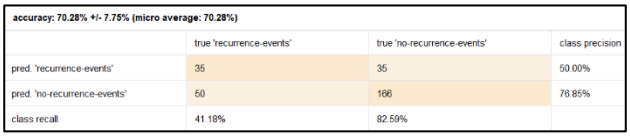
Applicando un 10-fold Stratified Cross-Validation, qual è l'effetto del *minimal gain* e del *maximal depth* sull'accuratezza media ottenuta da Decision Tree? Riportare almeno 5 sceenshot che mostrino le matrici di confusione ottenute usando diverse configurazioni per i parametri sopra citati (considerare almeno le 5 configurazioni usate per rispondere alla domanda 2). Mantenere la configurazione di default per tutti gli altri parametri.



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%	

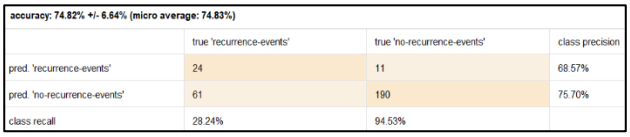
Default: max_depth = 10, min_gain = 0.01



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%	

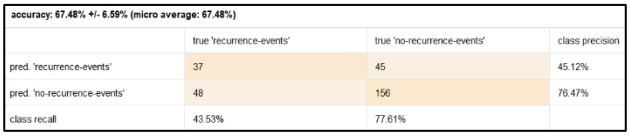
1: max_depth = 5, min_gain = 0.01



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%	

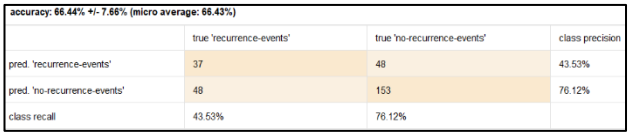
2: max_depth = 3, min_gain = 0.01



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%	

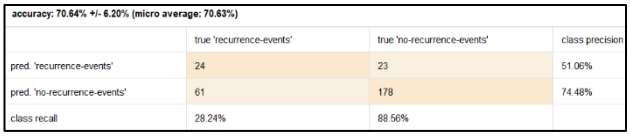
3: max_depth = 8, min_gain = 0.01



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%	

4: max_depth = 10, min_gain = 0.001



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%	

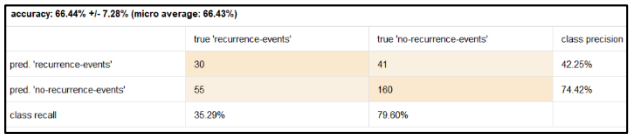
5: max_depth = 10, min_gain = 0.05

Si nota che:

- a parità di *minimal gain*, la diminuzione di *max depth* porta ad un'accuratezza maggiore, ovvero gli attributi usati per effettuare i primi due split sono probabilmente più adatti ad effettuare la predizione rispetto ai successivi;
- a parità di *max depth,* una diminuzione di *minimal gain* porta ad una diminuzione di accuratezza, mentre un aumento porta ad un aumento di accuratezza, in quanto permette di effettuare più split, rendendo i partizionamenti più puri.

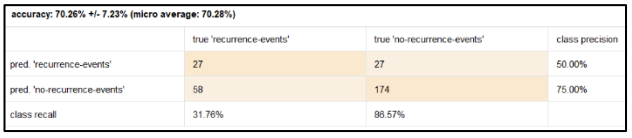
Esercizio 4

Considerando il classificatore K-Nearest Neighbor (K-NN) e applicando un 10-fold Stratified Cross-Validation, qual è l'effetto del parametro K sull'accuratezza media del classificatore? Riportare almeno 5 sceenshot che mostrino le matrici di confusione ottenute usando diversi valori di K. Applicare un 10-fold Stratified Cross-Validation con il classificatore Naïve Bayes. K-NN ottiene mediamente prestazioni superiori o inferiori a Naïve Bayes classifier sul dataset analizzato? Riportare uno screenshot che mostri la matrice di confusione ottenuta con Naive Bayes sul dataset analizzato.



accuracy: 66.44% +/- 7.28% (micro average: 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%	

1: k-NN, k = 1

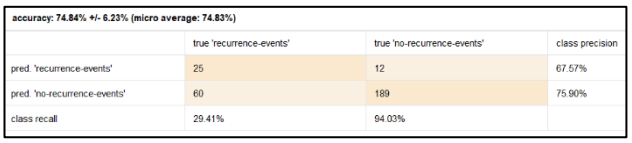


accuracy: 70.26% +/- 7.23% (micro average: 70.28%)			
	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	27	27	50.00%
pred. 'no-recurrence-events'	58	174	75.00%
class recall	31.76%	86.57%	

2: k-NN, k = 3

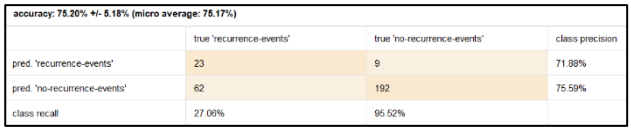
accuracy: 73.77% +/- 5.98% (micro average: 73.78%)			
	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	16	61.90%
pred. 'no-recurrence-events'	59	185	75.82%
class recall	30.59%	92.04%	

3: k-NN, k = 5



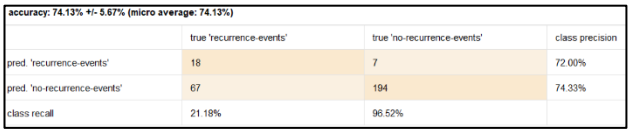
accuracy: 74.84% +/- 6.23% (micro average: 74.83%)			
	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	25	12	67.57%
pred. 'no-recurrence-events'	60	189	75.90%
class recall	29.41%	94.03%	

4: k-NN, k = 7



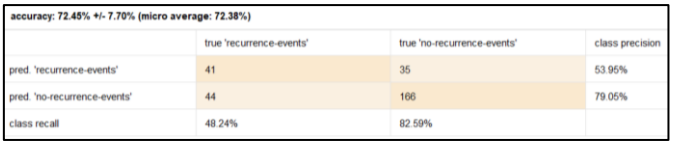
accuracy: 75.20% +/- 5.18% (micro average: 75.17%)			
	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	23	9	71.88%
pred. 'no-recurrence-events'	62	192	75.59%
class recall	27.06%	95.52%	

5: k-NN, k = 9



accuracy: 74.13% +/- 5.67% (micro average: 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%	

6: k-NN, k = 15



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%	

Naive Bayes

Si nota che:

- la classificazione con k-NN ha il suo massimo per k = 9 e in tale configurazione l'accuratezza è simile a quella ottenuta utilizzando un decision tree, nella configurazione 2 (*max depth* = 3, *minimal gain* = 0.01);
- il classificatore bayesiano ha accuratezza di poco inferiore al miglior valore ottenuto con classificatore k-NN, ma vicino al suo valore medio: i due metodi sembrano quindi avere prestazioni equivalenti.

Esercizio 5

Analizzare la matrice di correlazione per valutare la correlazione tra coppie di attributi del dataset. Riportare uno screenshot che mostri la matrice di correlazione ottenuta. Alla luce dei risultati ottenuti, l'ipotesi d'indipendenza Naïve risulta valida per il dataset Breast? Qual è la coppia di attributi maggiormente correlati?

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

Le uniche coppie di attributi (debolmente) correlati sono:

- *irradiat, inv-nodes*: correlazione positiva debole, con coefficiente 0.399;
- *node-caps, inv-nodes*: correlazione negativa debole, con coefficiente -0.465.

L'ipotesi Naïve per lo stimatore bayesiano risulta quindi valida, seppur parzialmente, in quanto è presente correlazione debole tra due sole coppie di attributi.