

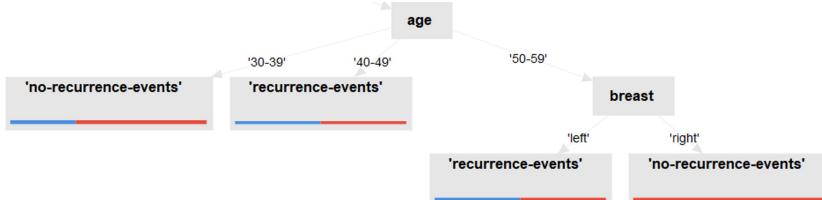
# QUADERNO\_2

giovedì 1 dicembre 2022 11:17

1.

- (a) L'attributo più selettivo è 'Node-caps'
- (b) Altezza dell'albero di decisione generato: 6 (considerando come altezza la lunghezza massima di un percorso che collega la radice ad una foglia)
- (c) Esempio di un partizionamento puro

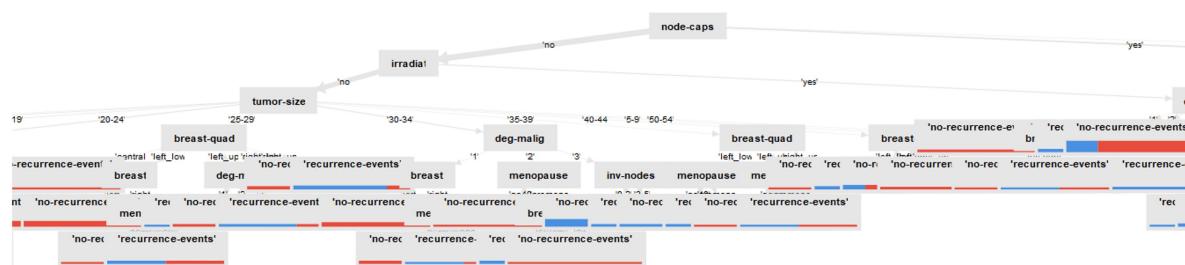
Un partizionamento puro è uno split sui valori di un attributo tale per cui i record corrispondenti appartengono tutti alla medesima classe. Per esempio, nella figura sottostante è possibile vedere che i valori dell'attributo 'age' sono suddivisi in tre gruppi: '30-39', '40-49', '50-59'. Mentre la terza partizione è impura perché copre record etichettati sia con la classe 'recurrence-events' che 'no-recurrence-events', la prima e la seconda sono pure perché tutte le relative istanze appartengono rispettivamente alle classi 'no-recurrence-events' e 'recurrence-events'.



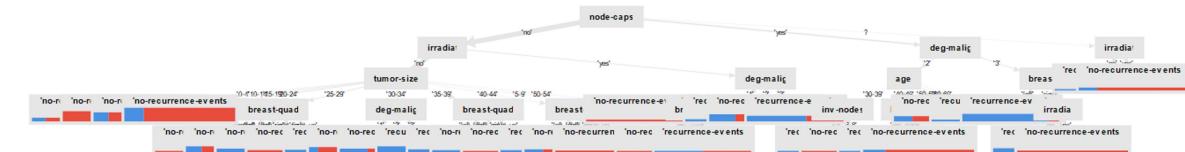
2.

Il parametro "maximal depth" permette di specificare l'altezza massima dell'albero di decisione generato. Usando la configurazione di default (con maximal depth = 10) l'altezza dell'albero risulta essere 6 e quindi il processo ricorsivo di learning dell'albero viene completato. Al contrario, settando un valore di maximal depth inferiore a 6 (ad es. 5) la ricorsione viene interrotta e quindi la qualità del modello generato potenzialmente decresce. Il parametro "minimal gain" permette di scegliere se dividere ulteriormente un nodo dell'albero oppure no. In particolare, un nodo viene diviso se il suo gain è superiore alla soglia minima (minimal gain). Valori elevati di minimal gain producono un numero limitato di partizionamenti e, di conseguenza, alberi di decisione più piccoli. Valori troppo elevati di minimal gain (ad es., 0.9) impediscono completamente lo split dei valori degli attributi e quindi l'albero risultante conterrà un singolo nodo. Dato che il valore di default di minimal gain è moderatamente basso (ad es., 0.1), esso produce generalmente uno splitting degli attributi abbastanza fitto.

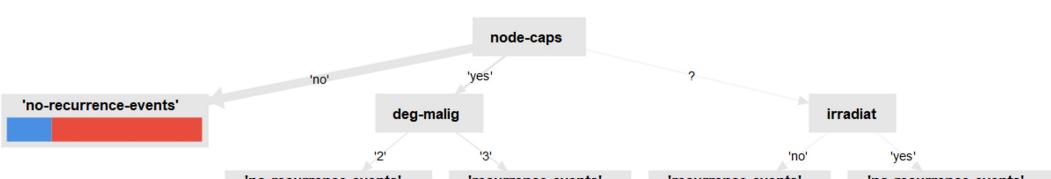
Minimal gain = 0.01  
Maximal depth = 10

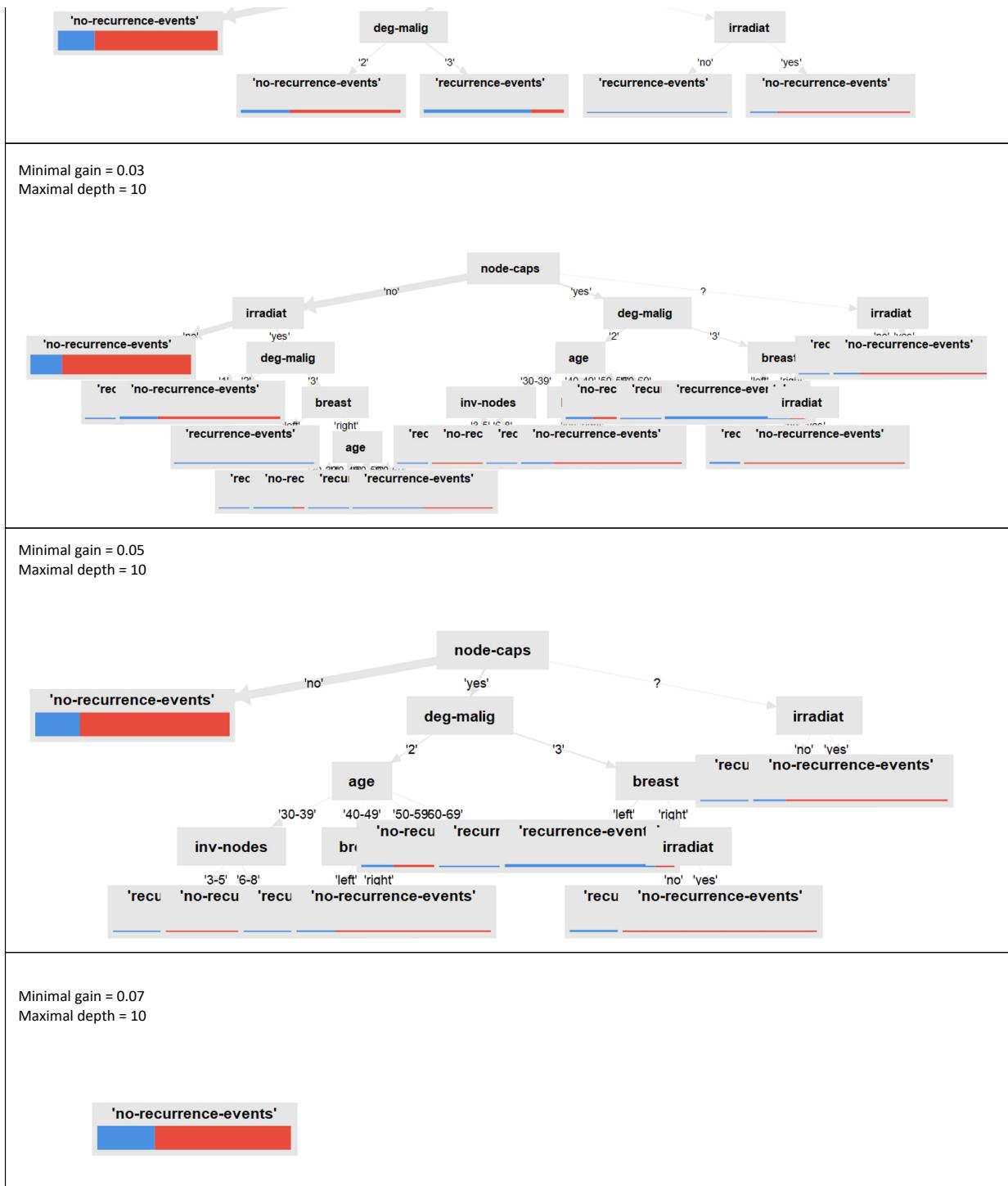


Minimal gain = 0.01  
Maximal depth= 5



Minimal gain = 0.01  
Maximal depth = 3





### 3. 10-fold Stratified Cross-Validation

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 = 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%

	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: 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%	

Minimal gain = 0.01  
Maximal depth = 3

accuracy: 70.31% +/- 5.87% (micro average: 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%	

Minimal gain = 0.03  
Maximal depth = 10

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%	

Minimal gain = 0.05  
Maximal depth = 10

accuracy: 69.61% +/- 1.89% (micro average: 69.58%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	2	4	33.33%
pred. 'no-recurrence-events'	83	197	70.36%
class recall	2.35%	98.01%	

Minimal gain = 0.07  
Maximal depth = 10

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, impostando valori di maximal depth superiori a 5 e minimal gain inferiori a 0.05 si produce l'effetto denominato "overfitting", ovvero il modello risulta troppo "focalizzato" sui dati di train per classificare in modo accurato nuovi dati di test.

4.

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%	

K-NN. Matrice di confusione. K=1

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

K-NN. Matrice di confusione. 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%	

K-NN. Matrice di confusione. 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%	

K-NN. Matrice di confusione. 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%	

K-NN. Matrice di confusione. K=7

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-NN. Matrice di confusione. K=10

Come visibile nelle precedenti immagini, all'aumentare del parametro K aumenta anche l'accuratezza del classificatore K-Nearest Neighbour.

Infatti, incrementando il valore di K, il classificatore considera un numero maggiore di dati di train "vicini" al dato di test e quindi l'accuratezza media cresce. Tuttavia considerando un numero molto elevato di record di train "vicini" (ad es., K>15) la presenza di dati rumorosi comincia ad inficiare le performance di classificazione e dunque l'accuratezza media di classificazione diminuisce leggermente come mostrato nell'immagine seguente.

accuracy: 74.85% +/- 7.88% (micro average: 74.83%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	21	8	72.41%
pred. 'no-recurrence-events'	64	193	75.10%
class recall	24.71%	96.02%	

K-NN. Matrice di confusione. K=15

accuracy: 72.75% +/- 11.32% (micro average: 72.73%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	40	33	54.79%
pred. 'no-recurrence-events'	45	168	78.87%
class recall	47.06%	83.58%	

Naive-Bayes

Dovendo effettuare un confronto tra i due classificatori si può dire che, per K≤3, il K-NN ha prestazioni, in termini di accuratezza, inferiori rispetto a Naive-Bayes. Inoltre, facendo una media delle accuratezze ottenute con i diversi valori del parametro K, l'accuratezza media del classificatore K-NN risulta essere 72.102 e pertanto Naive-Bayes sembra avere prestazioni superiori a K-NN per questo dataset, quantomeno con bassi valori del parametro K.

MEDIA = (66.44+70.26+73.77+74.84+75.20)/5= 72.102

## 5. MATRICE DI CORRELAZIONE

Attribut...	age = '4...	age = '5...	age = '6...	age = '3...	age = '7...	age = '2...	menop...	menop...	menop...	tumor-s...	tumor-s...	tumor-s...	tumor-s...	tumo...
age = '4...	1	-0.482	-0.338	-0.257	-0.099	-0.040	0.510	-0.478	-0.107	-0.109	0.031	0.021	0.019	-0.054
age = '5...	-0.482	1	-0.355	-0.270	-0.104	-0.042	-0.257	0.234	0.079	-0.002	0.019	-0.003	0.054	0.017
age = '6...	-0.338	-0.355	1	-0.189	-0.073	-0.030	-0.524	0.515	0.034	0.086	-0.098	0.022	-0.039	-0.045
age = '3...	-0.257	-0.270	-0.189	1	-0.056	-0.022	0.340	-0.344	0.008	0.042	0.026	-0.014	-0.021	0.049
age = '7...	-0.099	-0.104	-0.073	-0.056	1	-0.009	-0.154	0.161	-0.023	0.030	-0.039	-0.075	-0.071	0.141
age = '2...	-0.040	-0.042	-0.030	-0.022	-0.009	1	0.056	-0.054	-0.009	-0.020	0.222	-0.031	-0.029	-0.017
menopa...	0.510	-0.257	-0.524	0.340	-0.154	0.056	1	-0.952	-0.166	-0.062	0.085	-0.025	0.119	-0.065
menopa...	-0.478	0.234	0.515	-0.344	0.161	-0.054	-0.952	1	-0.144	0.034	-0.073	0.016	-0.096	0.081
menopa...	-0.107	0.079	0.034	0.008	-0.023	-0.009	-0.166	-0.144	1	0.093	-0.042	0.030	-0.076	-0.046
tumor-siz...	-0.109	-0.002	0.086	0.042	0.030	-0.020	-0.062	0.034	0.093	1	-0.091	-0.176	-0.165	-0.095
tumor-siz...	0.031	0.019	-0.098	0.026	-0.039	0.222	0.085	-0.073	-0.042	-0.091	1	-0.137	-0.129	-0.071
tumor-siz...	0.021	-0.003	0.022	-0.014	-0.075	-0.031	-0.025	0.016	0.030	-0.176	-0.137	1	-0.249	-0.145
tumor-siz...	0.019	0.054	-0.039	-0.021	-0.071	-0.029	0.119	-0.096	-0.076	-0.165	-0.129	-0.249	1	-0.135

<                    |||                    >

First Attribute	Second Attribute	Correla...
age = '40-49'	age = '50-59'	-0.482
age = '40-49'	age = '60-69'	-0.338
age = '40-49'	age = '30-39'	-0.257
age = '40-49'	age = '70-79'	-0.099
age = '40-49'	age = '20-29'	-0.040
age = '40-49'	menopause = 'premeno'	0.510
age = '40-49'	menopause = 'ge40'	-0.478
age = '40-49'	menopause = 'lt40'	-0.107
age = '40-49'	tumor-size = '15-19'	-0.109
age = '40-49'	tumor-size = '35-39'	0.031
age = '40-49'	tumor-size = '30-34'	0.021
age = '40-49'	tumor-size = '25-29'	0.019
age = '40-49'	tumor-size = '40-44'	-0.054

First Attribute	Second Attribute	Cor... ↓
inv-nodes = '0-2'	node-caps = 'no'	0.686
age = '60-69'	menopause = 'ge40'	0.515
age = '40-49'	menopause = 'premeno'	0.510
inv-nodes = '6-8'	node-caps = 'yes'	0.398
inv-nodes = '0-2'	irradiat = 'no'	0.389
node-caps = 'no'	irradiat = 'no'	0.370
tumor-size = '0-4'	breast-quad = 'central'	0.359
age = '30-39'	menopause = 'premeno'	0.340
inv-nodes = '3-5'	node-caps = 'yes'	0.317
inv-nodes = '9-11'	node-caps = ?	0.314
node-caps = 'yes'	irradiat = 'yes'	0.304
breast = 'left'	breast-quad = 'left_low'	0.281
inv-nodes = '0-2'	deg-malig = '1'	0.262

Le figure sopra mostrano la matrice di correlazione ottenuta dal dataset analizzato. Essa riporta la correlazione mutua (e simmetrica) tra coppie di attributi. Per esempio, l'attributo "inv-nodes" risulta essere molto correlato con l'attributo "node-caps" (correlazione pari a 0.686). Pertanto l'ipotesi di indipendenza statistica Naïve risulta essere irrealistica per il dataset analizzato. Tuttavia, le performance di Naïve Bayes risultano essere mediamente buone (vedi riposta alla domanda precedente).