Homework 2

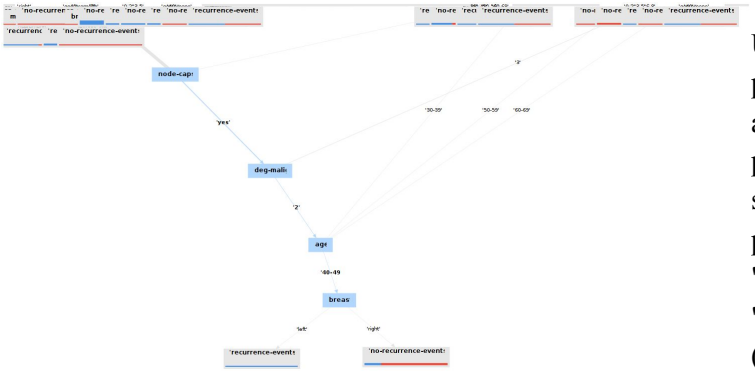
Sattolo Francesco S290158

Domanda 1

a) L'attributo più selettivo al fine di predire la classe di un nuovo dato di test è quello scelto come radice dell'albero, ossia "node-caps".

b) L’altezza dell’albero di decisione generato (intesa come numero massimo di archi di decisione attraversatidalla radice alle foglie) è pari a 6.

c)



Un partizionamento puro identifica due partizioni pure, in cui ogni elemento della partizione appartenga alla stessa classe. In questo caso non sono presenti partizionamenti puri, ma il partizionamento sull'attributo "breast" in figura ci si avvicina molto, poiché genera una partizione pura ("left":2 elementi "recurrence-event" e nessun elemento "no-recurrence-event") e una partizione quasi pura ("right":1 elemento "recurrence-event" e 4 elementi "no-recurrence-events").

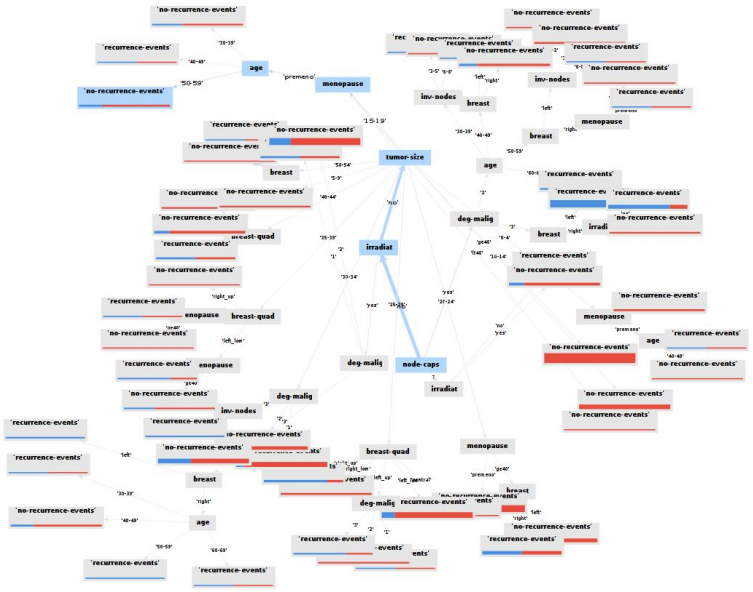
Domanda 2+3

1.Default:Maximal depth=7 o più | minimal gain=0.01

accuracy: 68.18% +/- 46.66% (micro average: 68.18%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	34	40	45.95%
pred. 'no-recurrence-events'	51	161	75.94%
class recall	40.00%	80.10%	

precision: 76.51% +/- 3.64% (micro average: 76.47%) **recall: 77.62% +/- 9.76% (micro average: 77.61%)**

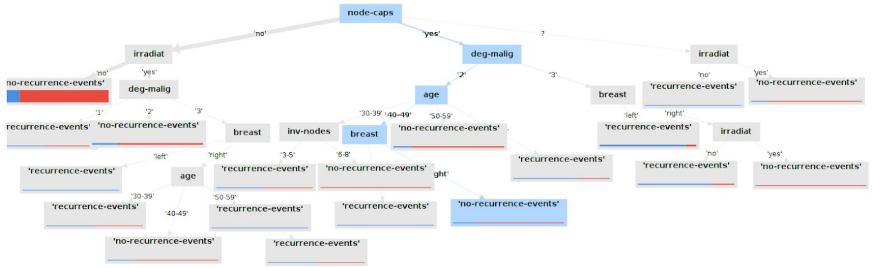


2. Maximal depth=7 | minimal gain=0.03

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%	

precision: 74.50% +/- 3.25% (micro average: 74.37%) **recall: 88.07% +/- 9.17% (micro average: 88.06%)**

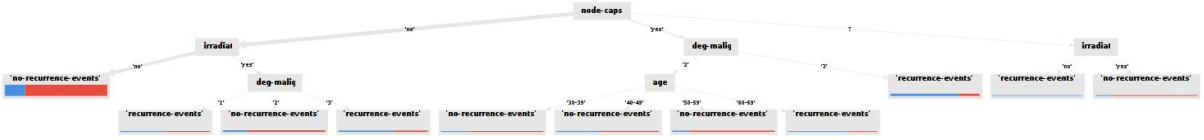


3. Maximal depth=4 | minimal gain=0.01

accuracy: 71.00% +/- 6.95% (micro average: 70.98%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	31	29	51.67%
pred. 'no-recurrence-events'	54	172	76.11%
class recall	36.47%	85.57%	

precision: 76.26% +/- 4.26% (micro average: 76.11%) recall: 85.64% +/- 9.31% (micro average: 85.57%)

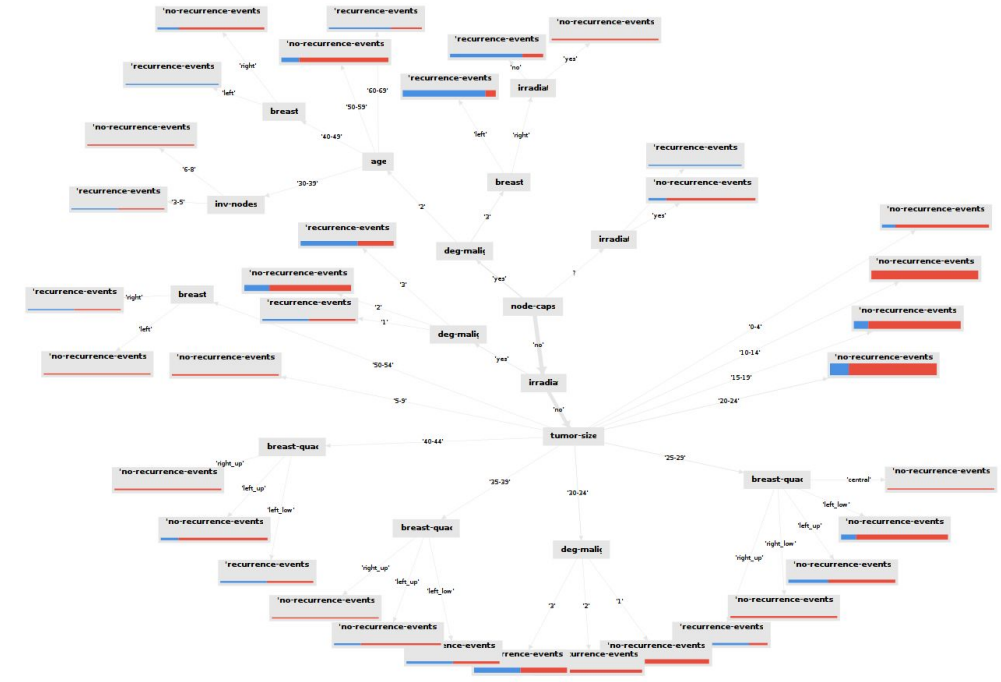


4. Maximal depth=5 | minimal gain=0.01

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%	

precision: 76.80% +/- 3.84% (micro average: 76.85%) recall: 82.60% +/- 10.32% (micro average: 82.59%)

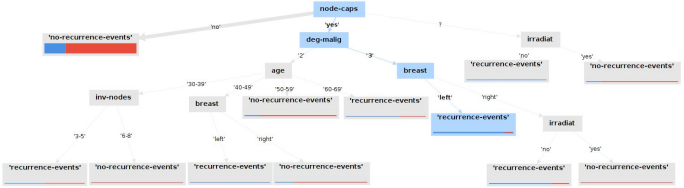


5. Maximal depth=5 | minimal gain=0.05

accuracy: 72.03% +/- 6.22% (micro average: 72.03%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	21	55.32%
pred. 'no-recurrence-events'	59	180	75.31%
class recall	30.59%	89.55%	

precision: 75.43% +/- 3.58% (micro average: 75.31%) recall: 89.57% +/- 8.61% (micro average: 89.55%)



Con la configurazione di default l'albero completa il processo ricorsivo e raggiunge altezza 6.
A parità di minimal gain, al diminuire della maximal depth l'accuratezza e il richiamo aumentano, ma rischiando di rendere l'albero troppo generico.
A parità di maximal depth, all'aumentare del minimal gain l'albero si "sfoltisce": si decide di non partizionare l'albero se il guadagno di informazione che si avrebbe non è abbastanza elevato. Può essere utile per evitare l'overfitting del modello. Accuratezza e richiamo aumentano, mentre la precisione diminuisce.

Domanda 4

K-NN

K=2

accuracy: 65.73% +/- 8.62% (micro average: 65.73%)

	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

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=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=6

accuracy: 72.03% +/- 6.10% (micro average: 72.03%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	19	55.81%
pred. 'no-recurrence-events'	61	182	74.90%
class recall	28.24%	90.55%	

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=25

accuracy: 74.13% +/- 4.66% (micro average: 74.13%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	17	6	73.91%
pred. 'no-recurrence-events'	68	195	74.14%
class recall	20.00%	97.01%	

Bayes

precision: 79.49% +/- 6.06% (micro average: 79.05%) **recall: 82.64% +/- 9.00% (micro average: 82.59%)**

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%	

In questo problema l'accuratezza da sola non basta a determinare la bontà del classificatore, poiché si ha una distribuzione delle classi sbilanciata (70% non-recurrence-events, 30% recurrence-events). Per questo motivo al crescere di k (numero di vicini da considerare), aumentano gli elementi assegnati alla classe non-recurrence-event, anche andando a considerare vicini non significativi. Questo fa aumentare l'accuratezza (finchè non inizia a entrare in gioco il rumore ad esempio con k=25), ma a scapito del richiamo della classe minoritaria (recurrence-events), che è anche la più critica in ambito medico, perché l'obiettivo è quello di individuare precocemente proprio i possibili eventi ricorrenti. Per questo motivo, sebbene per alcuni valori di k l'accuratezza del modello sia superiore, con il Naive Bayes Classifier si ottengono prestazioni migliori.

Domanda 5

Attributes	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat
age	1	?	?	?	?	?	?	?	?
menopause	?	1	?	?	?	?	?	?	?
tumor-size	?	?	1	?	?	?	?	?	?
inv-nodes	?	?	?	1	?	?	?	?	?
node-caps	?	?	?	?	1	?	?	?	?
deg-malig	?	?	?	?	?	1	?	?	?
breast	?	?	?	?	?	?	1	?	-0.019
breast-quad	?	?	?	?	?	?	?	1	?
irradiat	?	?	?	?	?	?	-0.019	?	1

La matrice di correlazione è formata dal coefficiente di correlazione di Pearson, calcolato tra tutte le coppie di attributi numerici. Per attributi categorici non ordinabili il coefficiente perde di significato, in quanto non si è in grado di calcolare se essi crescano o decrescano in modo correlato. Poiché quasi tutti gli attributi sono proprio di questo tipo, non si è in grado di stabilire con certezza se l'ipotesi d'indipendenza naive risulti valida per il dataset.