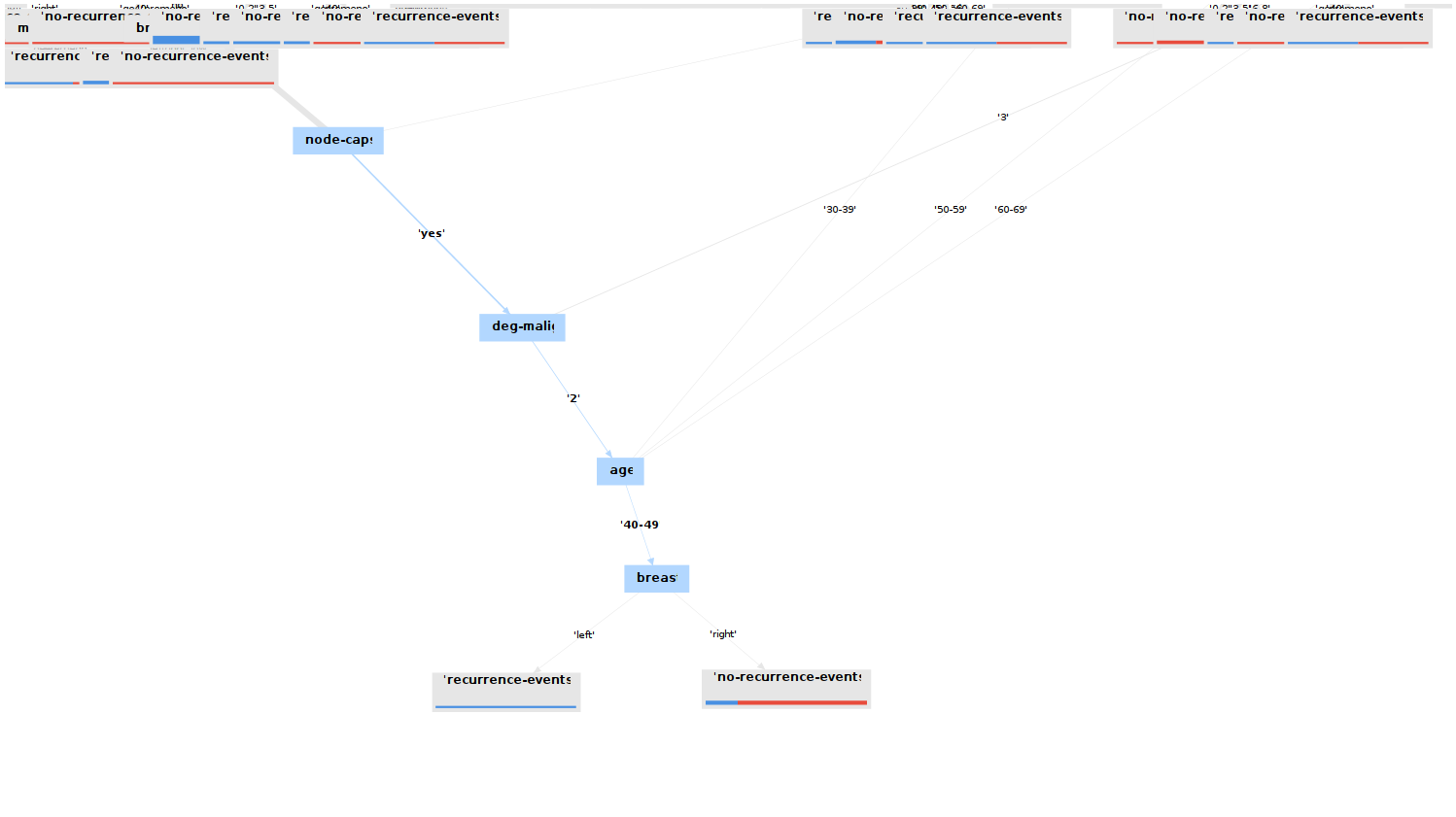
Homework 2

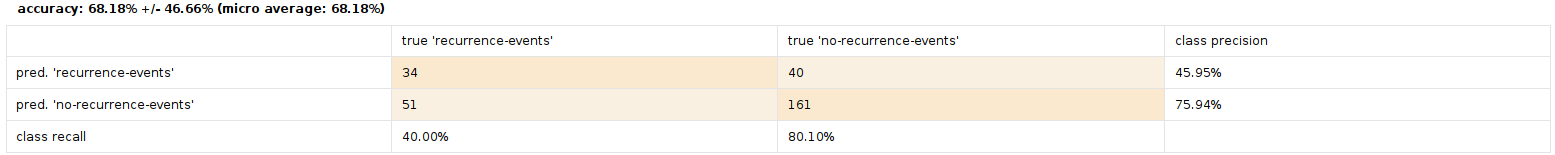
Sattolo Francesco S290158

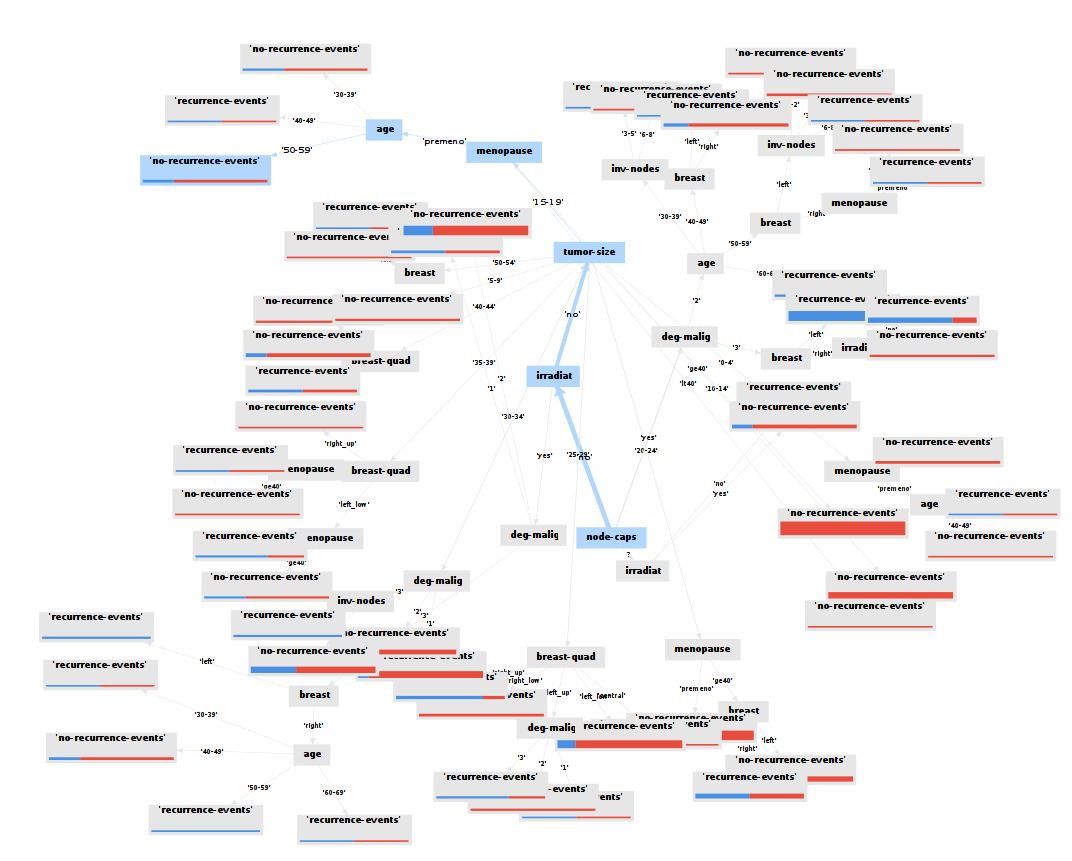
Domanda 1

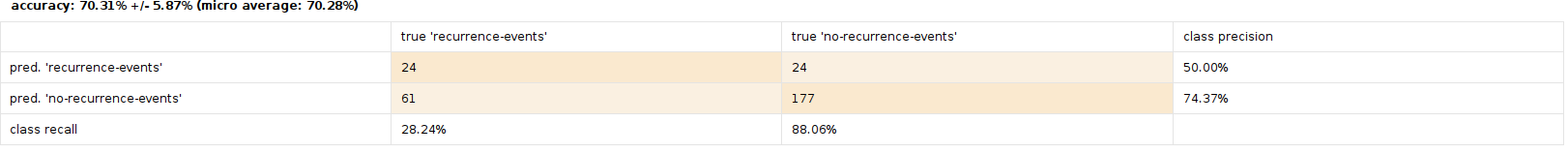
1. L'attributo più selettivo al fine di predire la classe di un nuovo dato di test è quello scelto come radice dell'albero, ossia "node-caps".
2. L’altezza dell’albero di decisione generato (intesa come numero massimo di archi di decisione attraversatidalla radice alle foglie) è pari a 6.
3. 

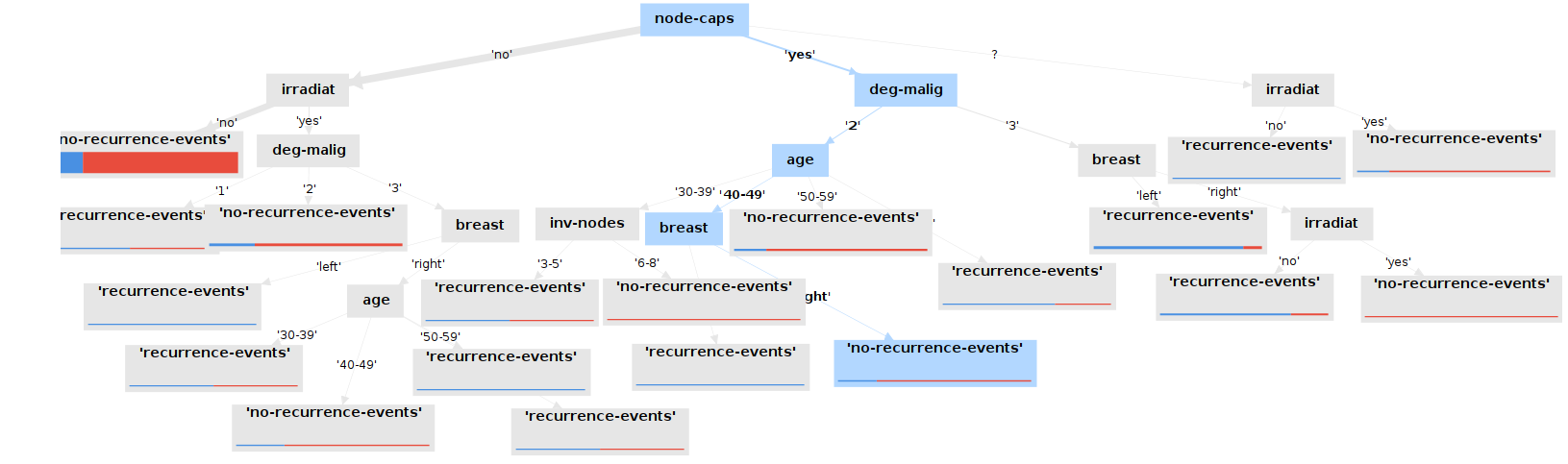
Un partizionamento puro identifica due partizioni pure, in cui ogni elemento della partizione appartenga alla stessa classe. In questo caso non sono presenti partizionamenti puri, ma il partizionamento sull'attributo "breast" in figura ci si avvicina molto, poiché genera una partizione pura ("left":2 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

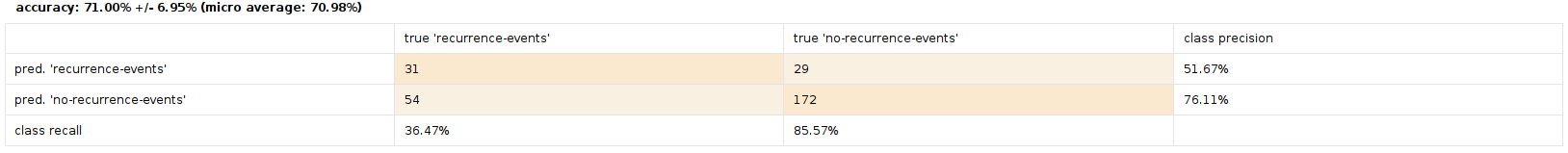




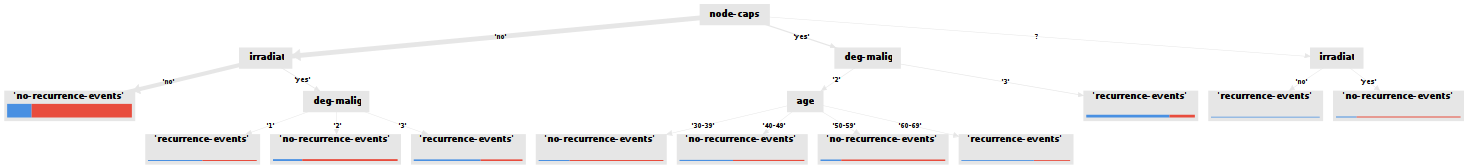
2. Maximal depth=7 | minimal gain=0.03



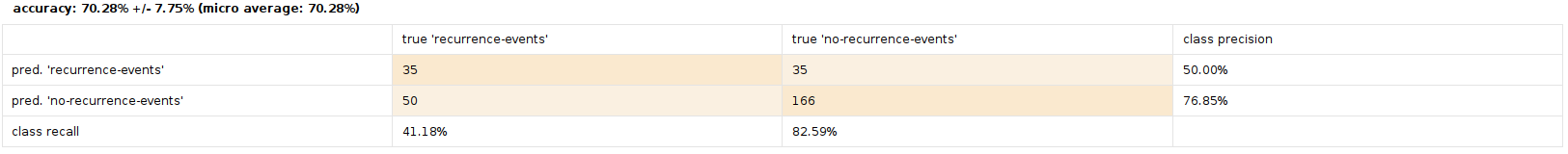
3. Maximal depth=4 | minimal gain=0.01

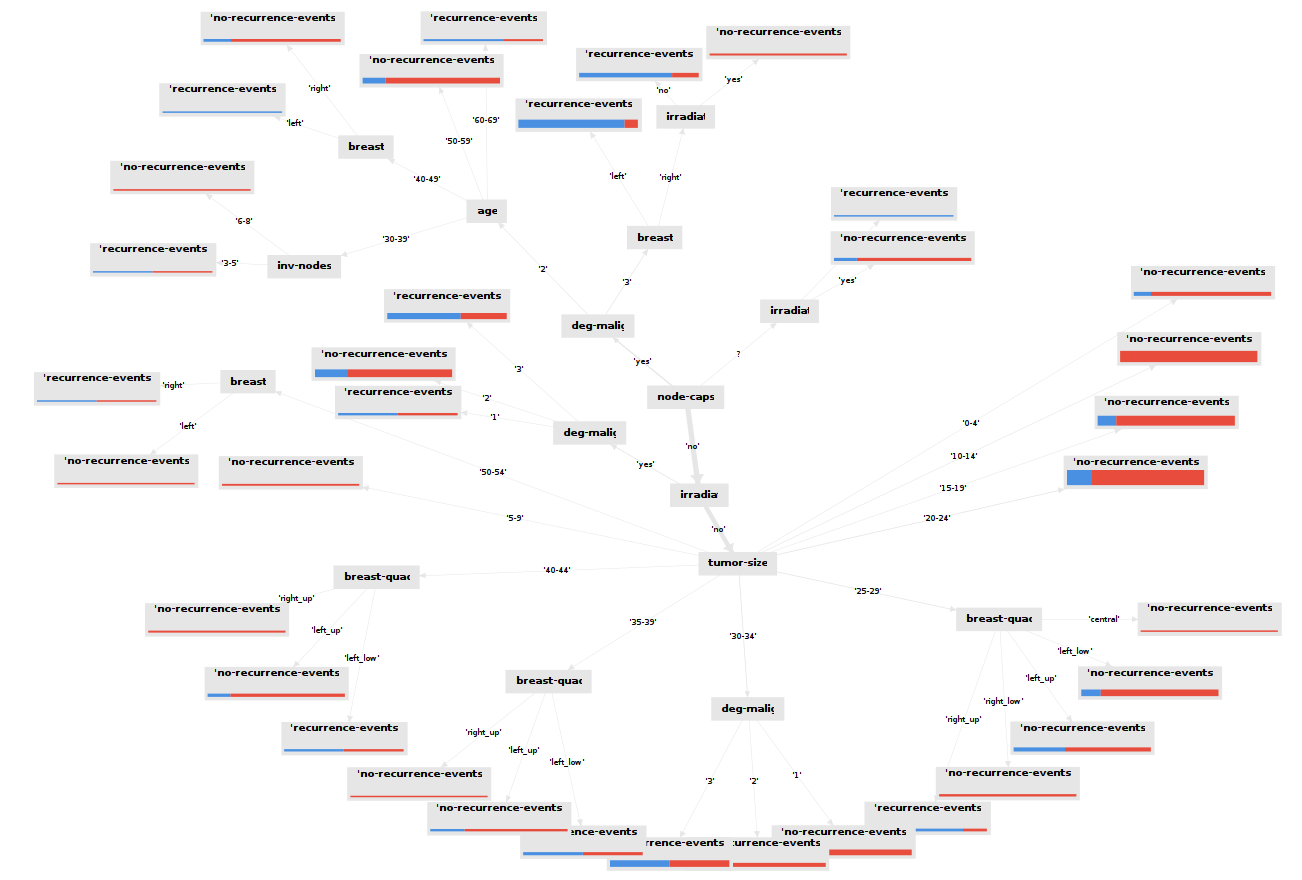




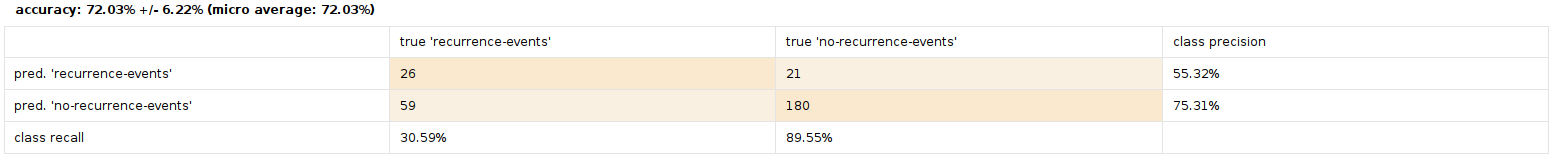


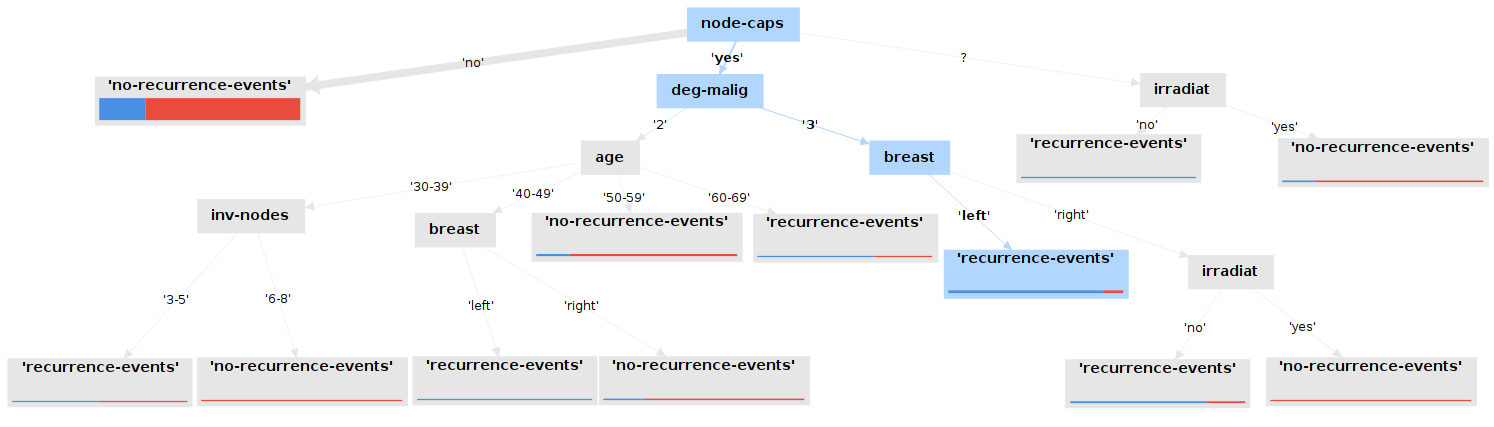
4. Maximal depth=5 | minimal gain=0.01





5. Maximal depth=5 | minimal gain=0.05





Con la configurazione di default l'albero completa il processo ricorsivo e raggiunge altezza 6.

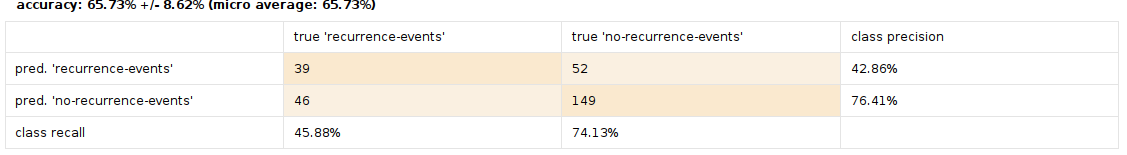
A parità di minimal gain, al diminuire della maximal depth l'accuratezza e il richiamo aumentano, ma rischiando di rendere l'albero troppo generico.

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.

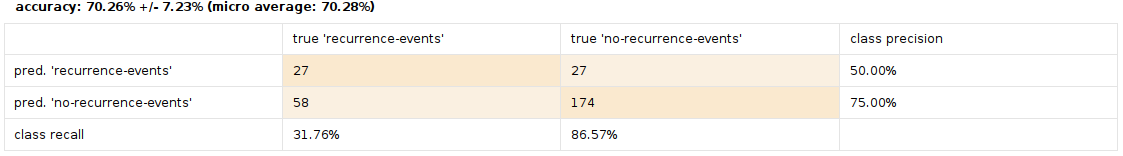
Domanda 4

K-NN

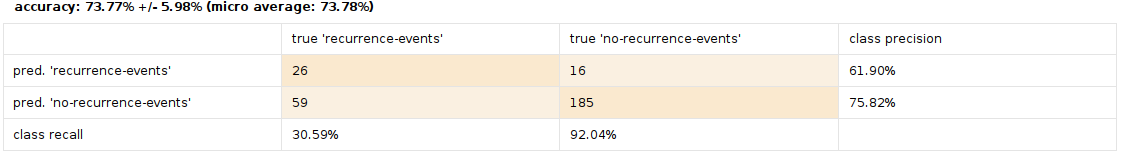
K=2



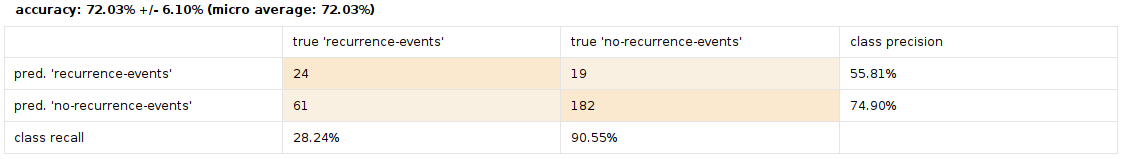
K=3



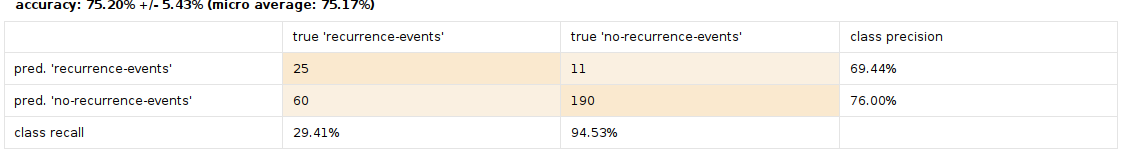
K=5



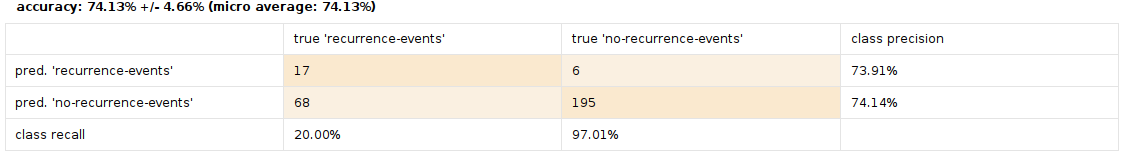
K=6



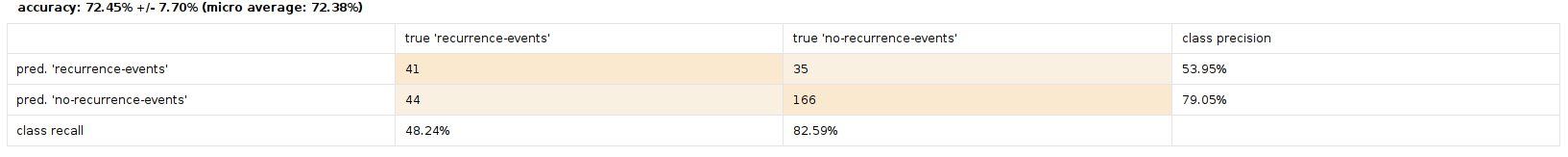
K=10



K=25

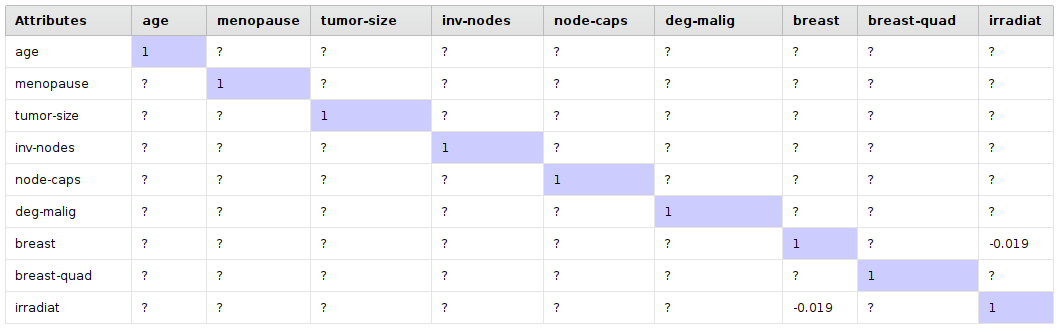


Bayes



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



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.