# QUADERNO2 FEDERICO BUSSOLINO S317641

1.

a.L'attributo più selettivo è node-caps;

b.L'altezza massima dell' albero è 7 (maxdepth contando radice-foglia, 6 non contando radice o foglia);

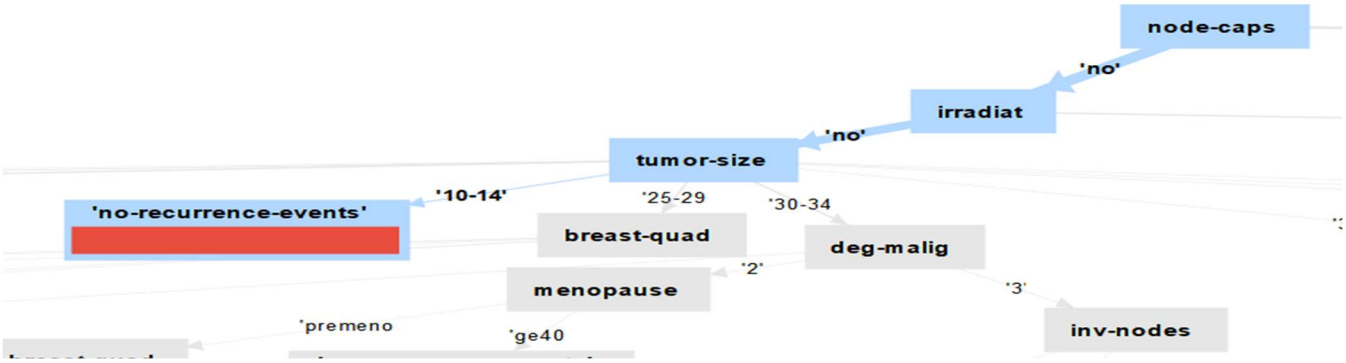
c.partizionamento puro:

node-caps: no

irradiat: no

tumor-size:10-14

→25 casi in cui il tumore non si è ripresentato




**Tree**

```
node-caps = 'no'
|   irradiat = 'no'
|   |   tumor-size = '0-4'
|   |   |   menopause = 'ge40': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=4}
|   |   |   menopause = 'premeno'
|   |   |   |   age = '30-39': 'recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=1}
|   |   |   |   age = '40-49': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=2}
|   |   tumor-size = '10-14': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=25}
```

2.

Osservazioni:

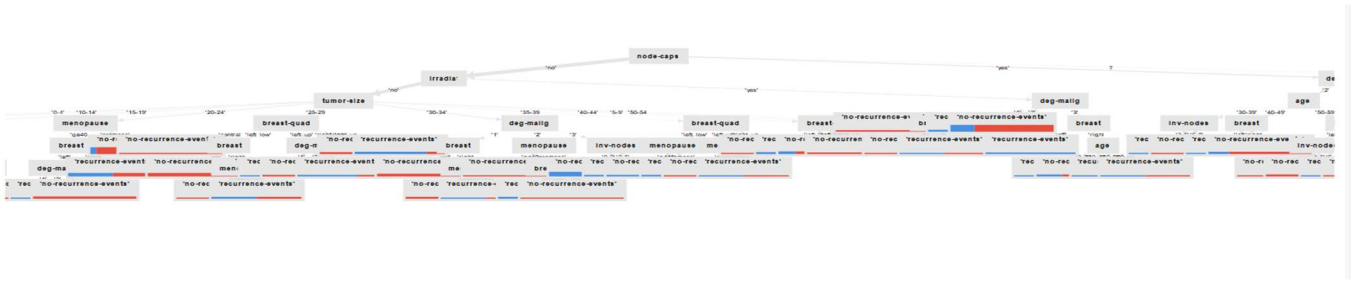
a)Maxilmal-depth può essere portato fino a 7 con min-gain 0.01 senza modificare l'albero

b)con  mingain=0.001 ottengo il massimo numero di ripartizioni tenendo invariato 2 come minimo numero di attributi per foglia, diminuendo ulteriormente mingain gli attributi non vengono ulteriormente suddivisi

e)Mingain=0.01 Maximal-depth=5 → ottengo un albero in cui vi sono ancora partizioni pure, anche se sono di meno, vi sono più attributi per foglia →casi + generali (si può settare il numero di attributi per foglia, ragionevolmente almeno 2 per evitare che un caso specifico di 1 tupla influenzi le decisioni su tutte le tuple aventi gli stessi valori di attributi →rischio overfitting da verificare in fase di test)

(riporto rappresentazione testuale perché più comprensibile)

a)Mingain=0.01, maxdepth=7 →albero uguale a maxdepth=10



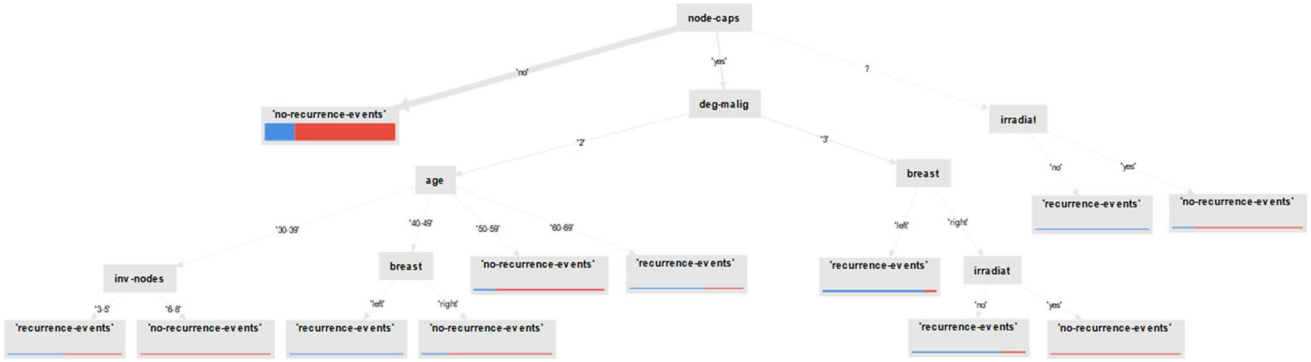
b)Mingain=0.001, maxdepth=10 →la foglia a profondità max è 7(maxdepth 8)

```
|   |   tumor-size = '30-34'
|   |   |   deg-malig = '1'
|   |   |   |   breast = 'left': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=6}
|   |   |   |   breast = 'right'
|   |   |   |   |   menopause = 'ge40': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=4}
|   |   |   |   |   menopause = 'premeno': 'recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=1}
|   |   |   deg-malig = '2'
|   |   |   |   menopause = 'ge40': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=2}
|   |   |   |   menopause = 'premeno'
|   |   |   |   |   breast-quad = 'left_up': 'recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=2}
|   |   |   |   |   breast-quad = 'right_low': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=
|   |   |   deg-malig = '3'
|   |   |   |   inv-nodes = '0-2'
|   |   |   |   |   breast = 'left'
|   |   |   |   |   |   menopause = 'ge40': 'recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=1}
|   |   |   |   |   |   menopause = 'premeno': 'no-recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=
```

c)Mingain=0.001, maxdepth=5 →la parte di albero precedente ha meno nodi (3)

```
node-caps = 'no'
|   irradiat = 'no'
|   |   tumor-size = '0-4': 'no-recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=7}
|   |   tumor-size = '10-14': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=25}
|   |   tumor-size = '15-19': 'no-recurrence-events' {'recurrence-events'=3, 'no-recurrence-events'=19}
|   |   tumor-size = '20-24': 'no-recurrence-events' {'recurrence-events'=6, 'no-recurrence-events'=28}
|   |   tumor-size = '25-29'
|   |   |   breast-quad = 'central': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=2}
|   |   |   breast-quad = 'left_low': 'no-recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=12}
|   |   |   breast-quad = 'left_up': 'no-recurrence-events' {'recurrence-events'=3, 'no-recurrence-events'=5}
|   |   |   breast-quad = 'right_low': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=4}
|   |   |   breast-quad = 'right_up': 'recurrence-events' {'recurrence-events'=4, 'no-recurrence-events'=1}
|   |   tumor-size = '30-34'
|   |   |   deg-malig = '1': 'no-recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=11}
|   |   |   deg-malig = '2': 'no-recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=6}
|   |   |   deg-malig = '3': 'recurrence-events' {'recurrence-events'=7, 'no-recurrence-events'=7}
|   |   tumor-size = '35-39'
|   |   |   breast-quad = 'left_low': 'recurrence-events' {'recurrence-events'=3, 'no-recurrence-events'=3}
|   |   |   breast-quad = 'left_up': 'no-recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=3}
|   |   |   breast-quad = 'right_up': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=2}
|   |   tumor-size = '40-44'
|   |   |   breast-quad = 'left_low': 'recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=2}
|   |   |   breast-quad = 'left_up': 'no-recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=5}
|   |   |   breast-quad = 'right_up': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=3}
|   |   tumor-size = '5-9': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=3}
|   |   tumor-size = '50-54'
|   |   |   breast = 'left': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=2}
|   |   |   breast = 'right': 'recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=1}
|   irradiat = 'yes'
|   |   deg-malig = '1': 'recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=2}
|   |   deg-malig = '2': 'no-recurrence-events' {'recurrence-events'=4, 'no-recurrence-events'=13}
|   |   deg-malig = '3': 'recurrence-events' {'recurrence-events'=8, 'no-recurrence-events'=5}
```

d)Mingain=0.05, maxdepth=10 →tutto l'albero è rappresentabile in un immagine (pochi split)



e)Mingain=0.01, maxdepth=5 →buon albero 7 nodi foglia con 2 soli attributi su 35 foglie

```
node-caps = 'no'
|   irradiat = 'no'
|   |   tumor-size = '0-4': 'no-recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=7}
|   |   tumor-size = '10-14': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=25}
|   |   tumor-size = '15-19': 'no-recurrence-events' {'recurrence-events'=3, 'no-recurrence-events'=19}
|   |   tumor-size = '20-24': 'no-recurrence-events' {'recurrence-events'=6, 'no-recurrence-events'=28}
|   |   tumor-size = '25-29'
|   |   |   breast-quad = 'central': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=2}
|   |   |   breast-quad = 'left_low': 'no-recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=12}
|   |   |   breast-quad = 'left_up': 'no-recurrence-events' {'recurrence-events'=3, 'no-recurrence-events'=5}
|   |   |   breast-quad = 'right_low': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=4}
|   |   |   breast-quad = 'right_up': 'recurrence-events' {'recurrence-events'=4, 'no-recurrence-events'=1}
|   |   tumor-size = '30-34'
|   |   |   deg-malig = '1': 'no-recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=11}
|   |   |   deg-malig = '2': 'no-recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=6}
|   |   |   deg-malig = '3': 'recurrence-events' {'recurrence-events'=7, 'no-recurrence-events'=7}
|   |   tumor-size = '35-39'
|   |   |   breast-quad = 'left_low': 'recurrence-events' {'recurrence-events'=3, 'no-recurrence-events'=3}
|   |   |   breast-quad = 'left_up': 'no-recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=3}
|   |   |   breast-quad = 'right_up': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=2}
|   |   tumor-size = '40-44'
|   |   |   breast-quad = 'left_low': 'recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=2}
|   |   |   breast-quad = 'left_up': 'no-recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=5}
|   |   |   breast-quad = 'right_up': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=3}
|   |   tumor-size = '5-9': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=3}
|   |   tumor-size = '50-54'
|   |   |   breast = 'left': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=2}
|   |   |   breast = 'right': 'recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=1}
|   irradiat = 'yes'
|   |   deg-malig = '1': 'recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=2}
|   |   deg-malig = '2': 'no-recurrence-events' {'recurrence-events'=4, 'no-recurrence-events'=13}
|   |   deg-malig = '3': 'recurrence-events' {'recurrence-events'=8, 'no-recurrence-events'=5}
node-caps = 'yes'
|   deg-malig = '2'
|   |   age = '30-39'
|   |   |   inv-nodes = '3-5': 'recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=1}
|   |   |   inv-nodes = '6-8': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=2}
|   |   age = '40-49'
|   |   |   breast = 'left': 'recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=0}
|   |   |   breast = 'right': 'no-recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=4}
|   |   age = '50-59': 'no-recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=10}
|   |   age = '60-69': 'recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=1}
|   deg-malig = '3'
|   |   breast = 'left': 'recurrence-events' {'recurrence-events'=16, 'no-recurrence-events'=2}
|   |   breast = 'right'
|   |   |   irradiat = 'no': 'recurrence-events' {'recurrence-events'=7, 'no-recurrence-events'=2}
|   |   |   irradiat = 'yes': 'no-recurrence-events' {'recurrence-events'=0, 'no-recurrence-events'=3}
node-caps = ?
|   irradiat = 'no': 'recurrence-events' {'recurrence-events'=2, 'no-recurrence-events'=0}
|   irradiat = 'yes': 'no-recurrence-events' {'recurrence-events'=1, 'no-recurrence-events'=5}
```

3.

a)

**accuracy: 66.35% +/- 4.40% (micro average: 66.35%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 135 | 193 | 41.16% |
| pred. 'no-recurrence-events' | 161 | 563 | 77.76% |
| class recall | 45.61% | 74.47% | |

b)

**accuracy: 65.78% +/- 4.07% (micro average: 65.78%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 136 | 200 | 40.48% |
| pred. 'no-recurrence-events' | 160 | 556 | 77.65% |
| class recall | 45.95% | 73.54% | |

c)

**accuracy: 67.87% +/- 3.95% (micro average: 67.87%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 125 | 167 | 42.81% |
| pred. 'no-recurrence-events' | 171 | 589 | 77.50% |
| class recall | 42.23% | 77.91% | |

d)

**accuracy: 72.65% +/- 4.06% (micro average: 72.72%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 78 | 69 | 53.06% |
| pred. 'no-recurrence-events' | 218 | 687 | 75.91% |
| class recall | 26.35% | 90.87% | |

e)

**accuracy: 67.96% +/- 3.92% (micro average: 67.97%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 125 | 166 | 42.96% |
| pred. 'no-recurrence-events' | 171 | 590 | 77.53% |
| class recall | 42.23% | 78.04% | |

4.

a) k=2 → considerazioni: se si vuole intervenire in ottica preventiva rispetto al ripresentarsi dei tumori conviene prendere un classificatore con alto recall per recurrence-events

**accuracy: 67.22% +/- 4.20% (micro average: 67.21%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 134 | 183 | 42.27% |
| pred. 'no-recurrence-events' | 162 | 573 | 77.96% |
| class recall | 45.27% | 75.79% | |

b) k=5 → per non allarmare inutilmente le persone coniene scegliere un classificatore con + accuracy

**accuracy: 72.63% +/- 3.76% (micro average: 72.62%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 112 | 104 | 51.85% |
| pred. 'no-recurrence-events' | 184 | 652 | 77.99% |
| class recall | 37.84% | 86.24% | |

c) k=10 → accuracy elevata, buon class recall per recurrence-events

**accuracy: 75.04% +/- 2.31% (micro average: 75.00%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 107 | 74 | 59.12% |
| pred. 'no-recurrence-events' | 189 | 682 | 78.30% |
| class recall | 36.15% | 90.21% | |

d) k=15 → perdita di class recall per recurrence-events

**accuracy: 74.39% +/- 2.84% (micro average: 74.43%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 86 | 59 | 59.31% |
| pred. 'no-recurrence-events' | 210 | 697 | 76.85% |
| class recall | 29.05% | 92.20% | |

e) k=8 → alternativa migliore di k=5: buon compromesso accuracy e class recall recurrence-events

**accuracy: 73.55% +/- 3.32% (micro average: 73.48%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 113 | 96 | 54.07% |
| pred. 'no-recurrence-events' | 183 | 660 | 78.29% |
| class recall | 38.18% | 87.30% | |

naïve-bayes ottiene prestazioni migliori: ottimo recall per recurrence-events e buona precisione

**accuracy: 73.27% +/- 3.87% (micro average: 73.19%)**

| | true 'recurrence-events' | true 'no-recurrence-events' | class precision |
|---|---|---|---|
| pred. 'recurrence-events' | 143 | 129 | 52.57% |
| pred. 'no-recurrence-events' | 153 | 627 | 80.38% |
| class recall | 48.31% | 82.94% | |

5. MATRICE DI CORRELAZIONE

| 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 |

Gli attributi sono incorrelati, alcuni sono debolmente correlati→ si può applicare l'ipotesi naïve

La coppia di attributi maggiormente correlata è: inv-nodes e node-caps