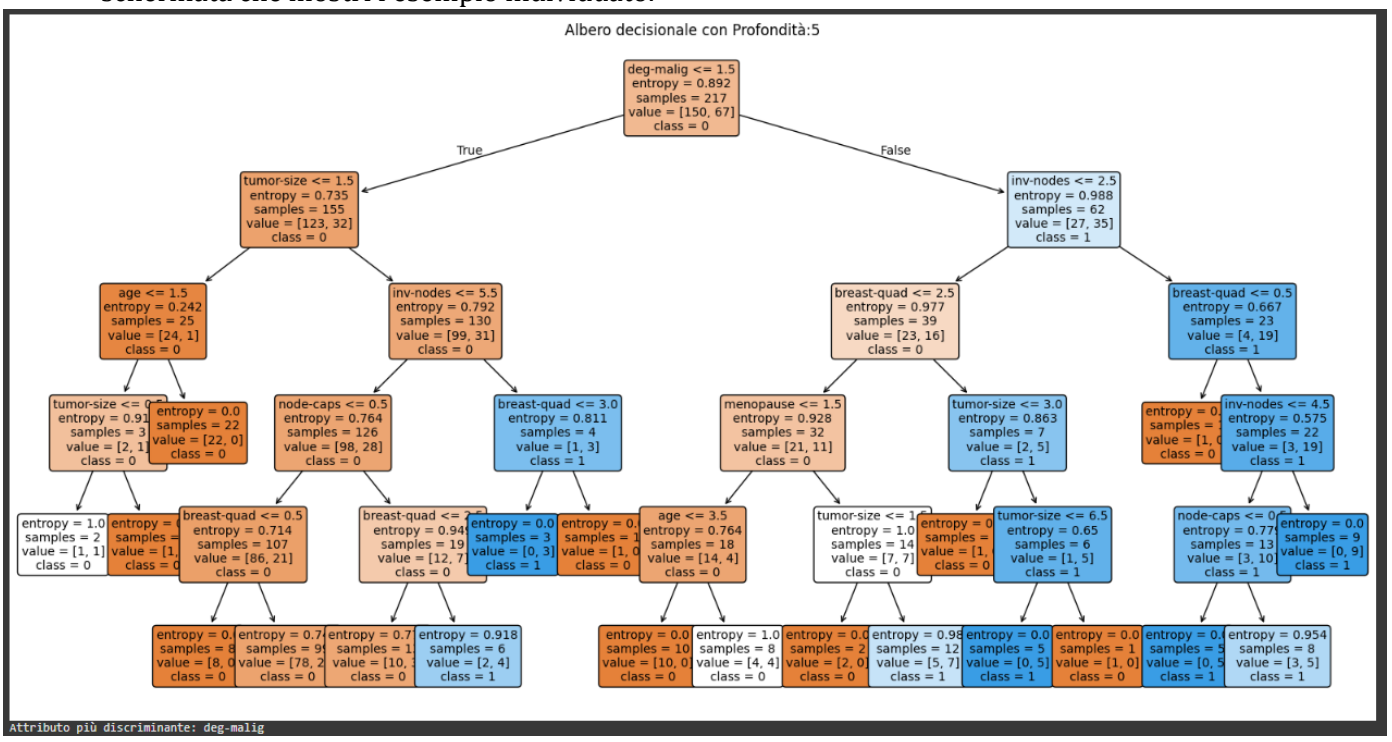
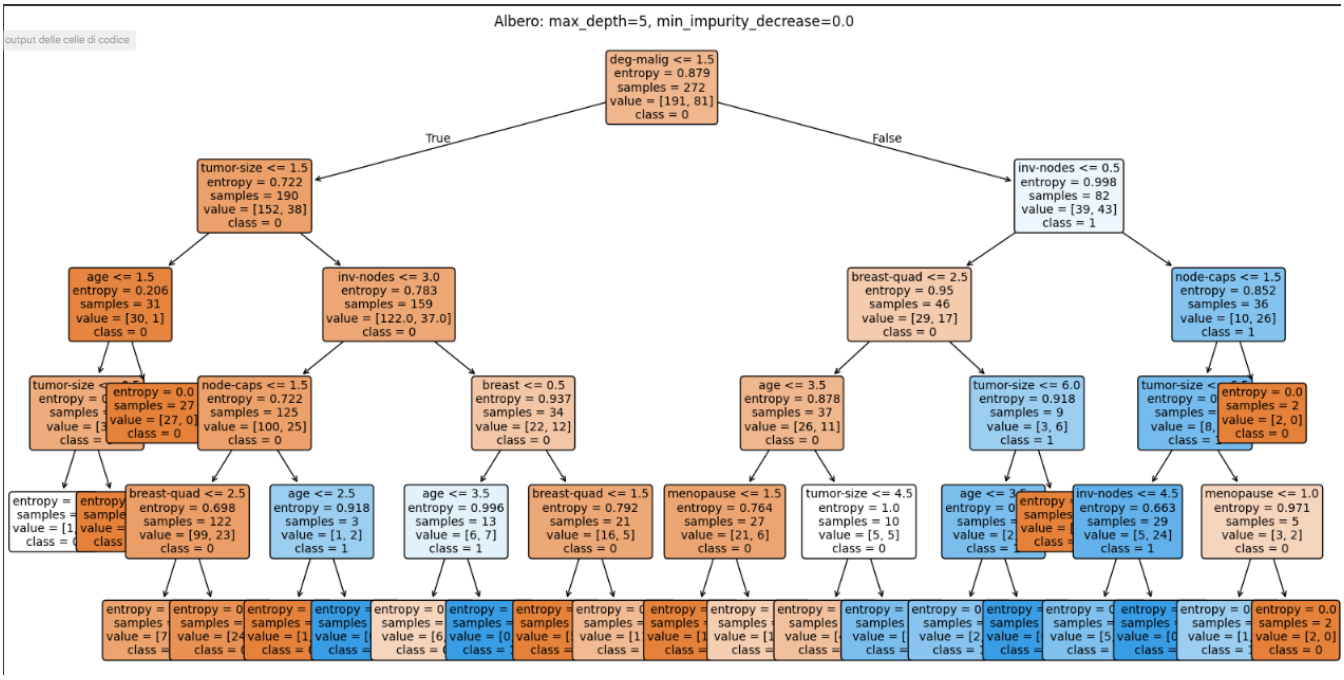
RELAZIONE HOMEWORK 2 – Domizio Mattia s341987

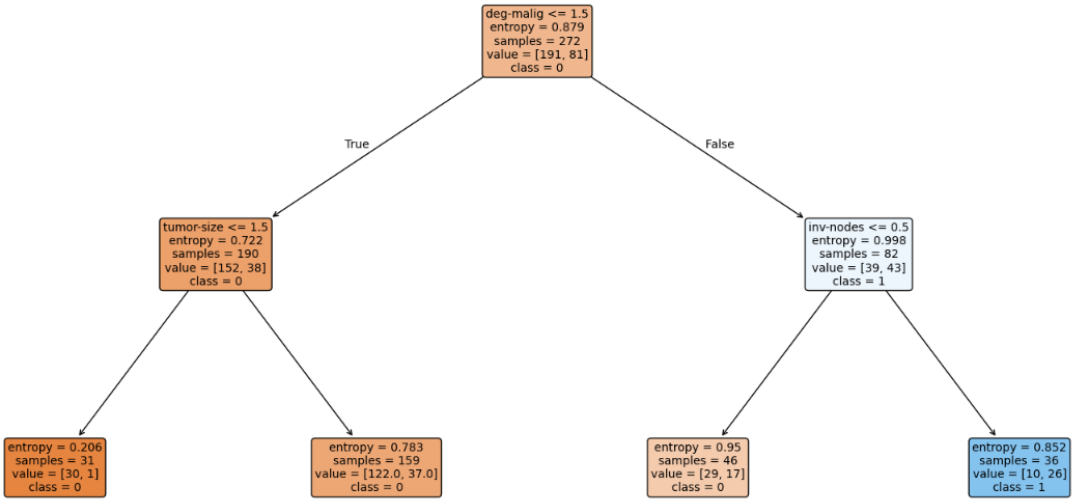
1. Addestrare un albero decisionale dall'intero dataset impostando la soglia di profondità massima a 5, mantenendo la configurazione predefinita per tutti gli altri parametri. (a) Quale attributo è ritenuto il più discriminante per la previsione della classe? (b) Qual è l'altezza dell'albero decisionale generato? (b) Individuare una partizione pura nell'albero decisionale e riportare una schermata che mostri l'esempio individuato.



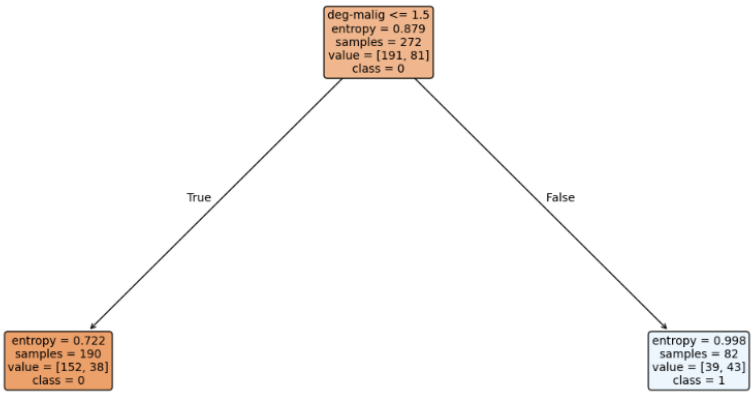
2. Analizzare l'impatto dei parametri impurità minima (utilizzando il criterio di suddivisione dell'entropia), numero minimo di campioni per ogni foglia e profondità massima sulle caratteristiche del modello di albero decisionale appreso dall'intero set di dati (mantenere la configurazione predefinita per tutti gli altri parametri). Riportare almeno 5 diverse schermate che mostrino gli Alberi decisionali (o porzioni di essi) generati con diverse impostazioni di configurazione.



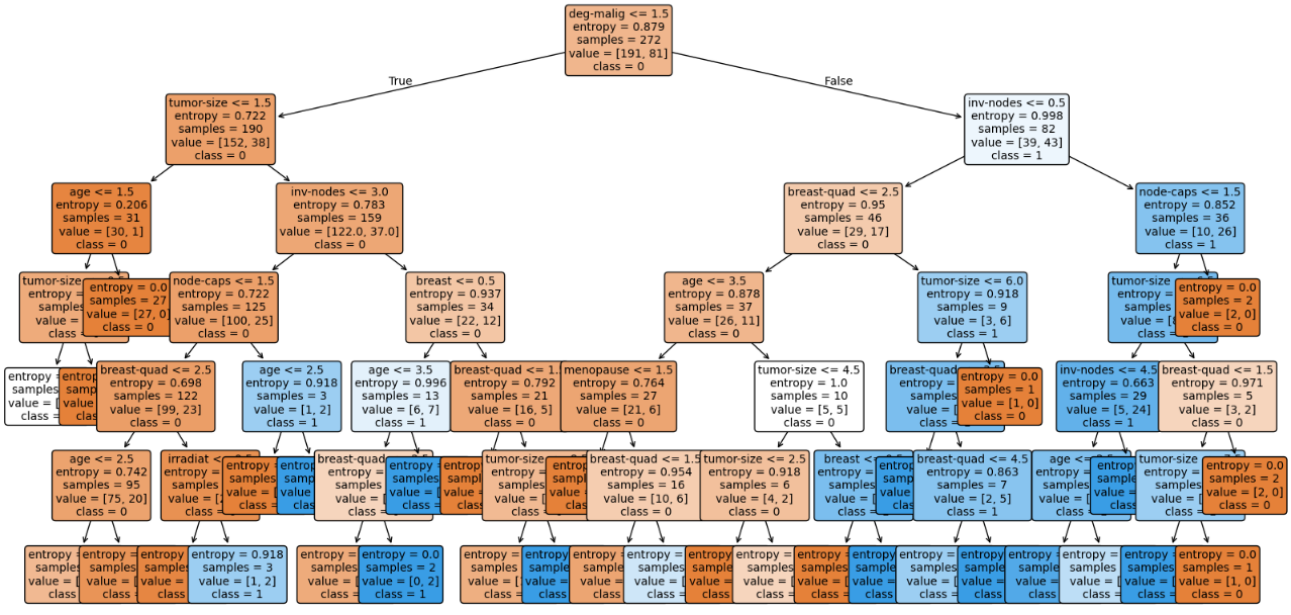
Albero: max_depth=4, min_impurity_decrease=0.02

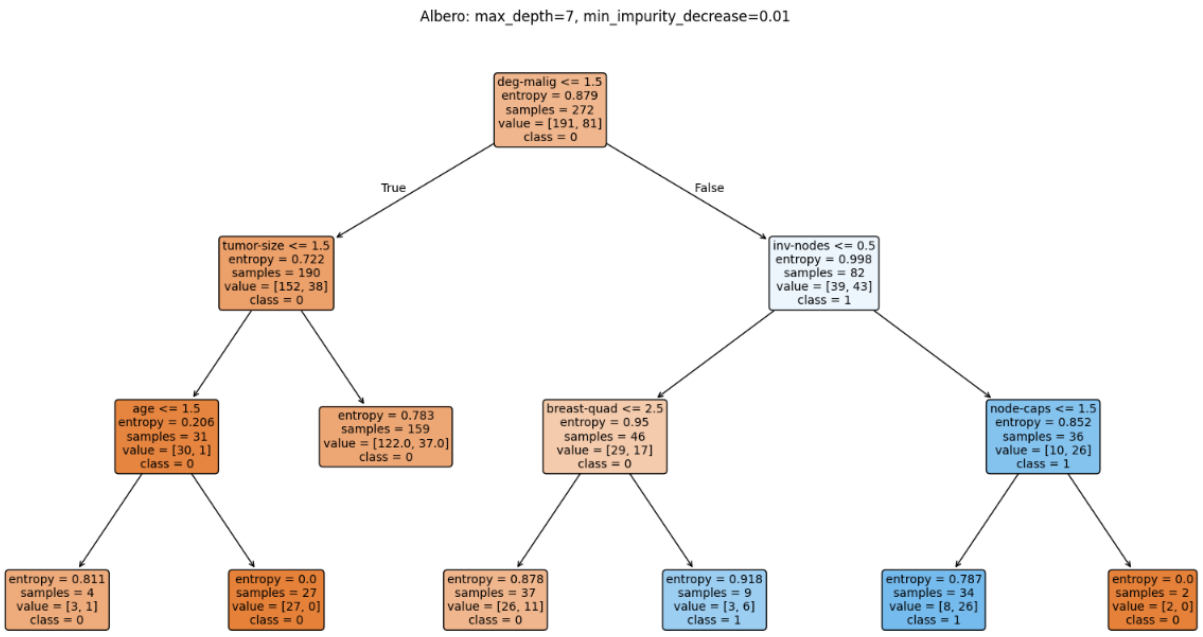


Albero: max_depth=3, min_impurity_decrease=0.05



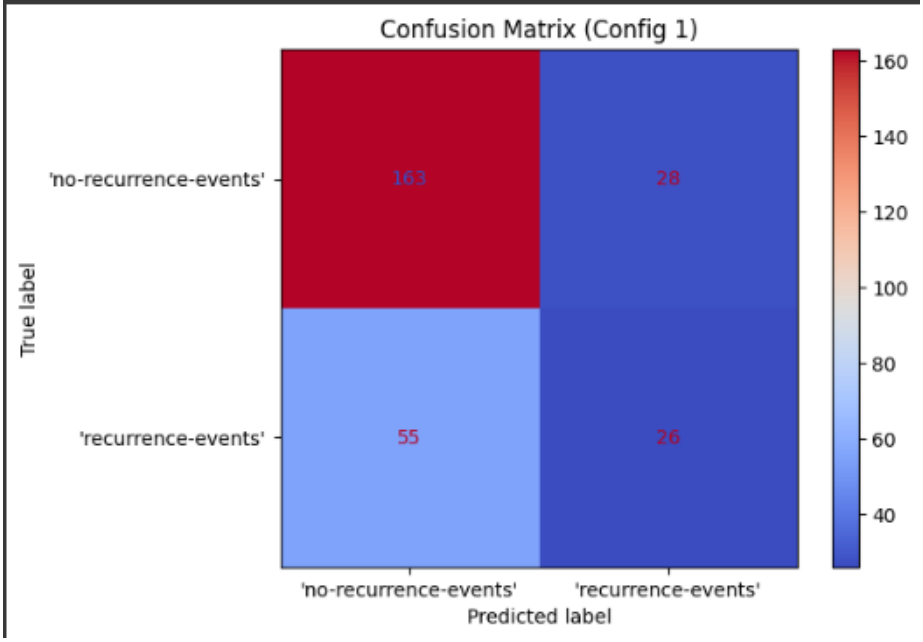
Albero: max_depth=6, min_impurity_decrease=0.0



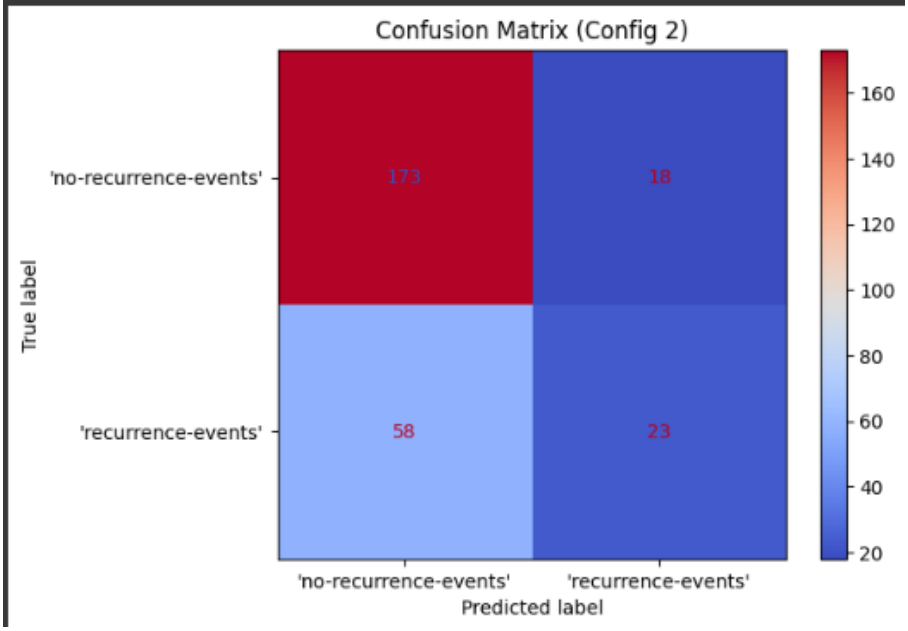


3. Eseguendo una 10-fold cross-validation stratificata, qual è l'impatto dei parametri impurità minima, numero minimo di campioni da dividere e profondità massima sull'accuratezza media ottenuta dall'albero decisionale? Riportare almeno 5 schermate che mostrino le matrici di confusione ottenute utilizzando diverse impostazioni dei parametri (considerare almeno tutte le configurazioni utilizzate per rispondere alla domanda 2). Mantenete la configurazione predefinita per tutti gli altri parametri.

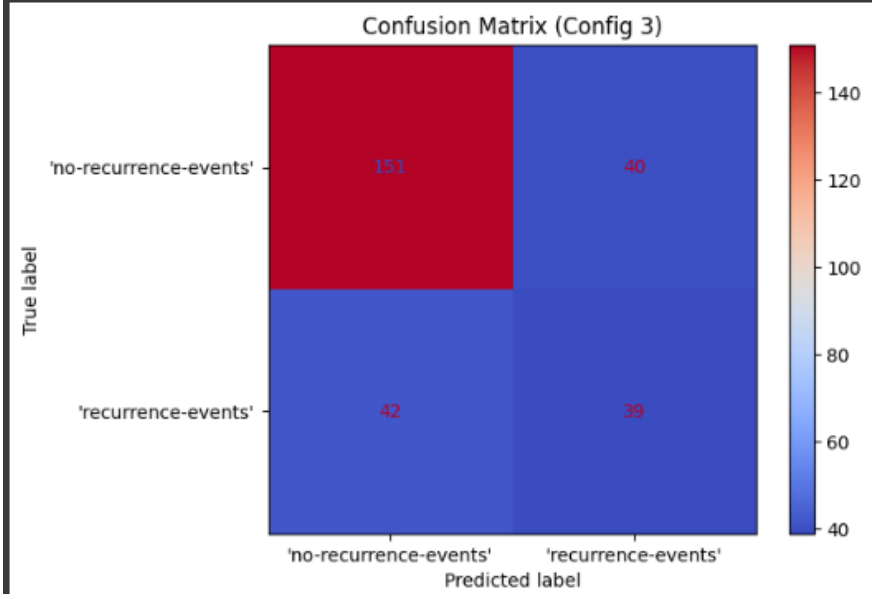
```
Configurazione 1: max_depth=5, min_impurity_decrease=0.0
Accuratezza: 0.6949
Matrice di confusione:
[[163  28]
 [ 55  26]]
              precision    recall  f1-score   support

           0       0.75      0.85      0.80       191
           1       0.48      0.32      0.39        81

    accuracy                           0.69       272
   macro avg       0.61      0.59      0.59       272
weighted avg       0.67      0.69      0.67       272
```



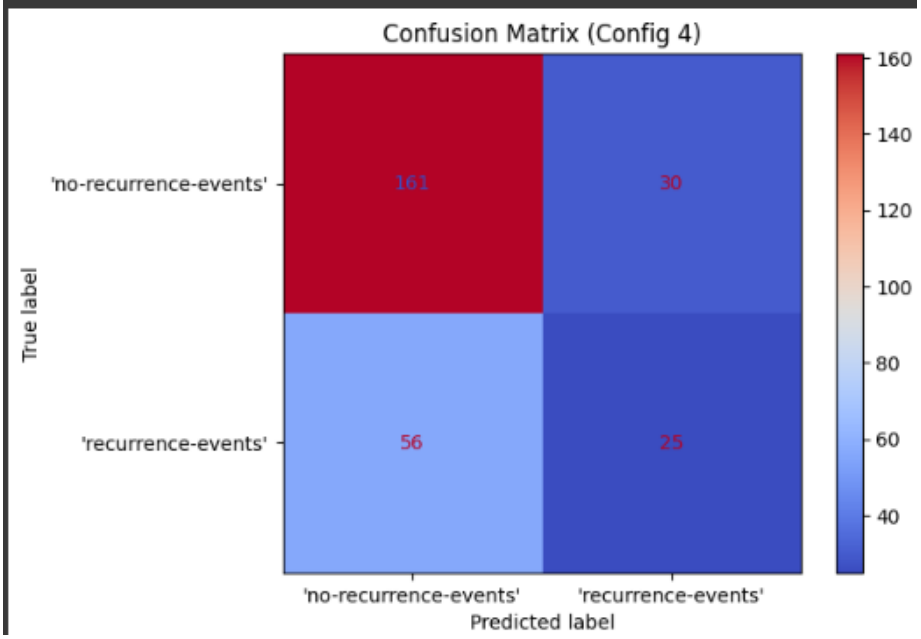
```
Configurazione 2: max_depth=4, min_impurity_decrease=0.02
Accuratezza: 0.7206
Matrice di confusione:
[[173  18]
 [ 58  23]]
              precision    recall  f1-score   support

           0       0.75      0.91      0.82       191
           1       0.56      0.28      0.38        81

    accuracy                           0.72       272
   macro avg       0.65      0.59      0.60       272
weighted avg       0.69      0.72      0.69       272
```



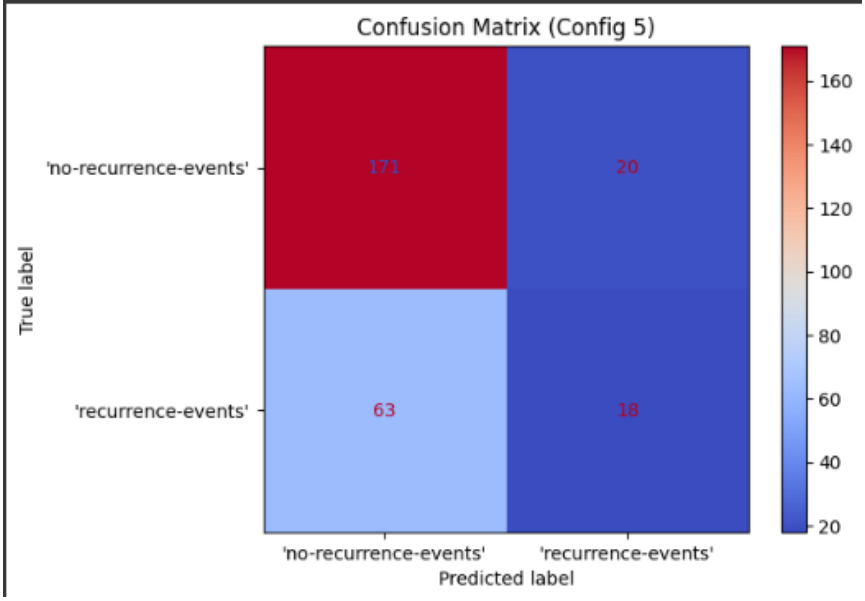

```
Configurazione 3: max_depth=3, min_impurity_decrease=0.05
Accuratezza: 0.6985
Matrice di confusione:
[[151  40]
 [ 42  39]]
              precision    recall  f1-score   support

           0       0.78      0.79      0.79       191
           1       0.49      0.48      0.49        81

    accuracy                           0.70       272
   macro avg       0.64      0.64      0.64       272
weighted avg       0.70      0.70      0.70       272
```



```
Configurazione 4: max_depth=6, min_impurity_decrease=0.0
Accuratezza: 0.6838
Matrice di confusione:
[[161  30]
 [ 56  25]]
              precision    recall  f1-score   support

           0       0.74      0.84      0.79       191
           1       0.45      0.31      0.37        81

    accuracy                           0.68       272
   macro avg       0.60      0.58      0.58       272
weighted avg       0.66      0.68      0.66       272
```

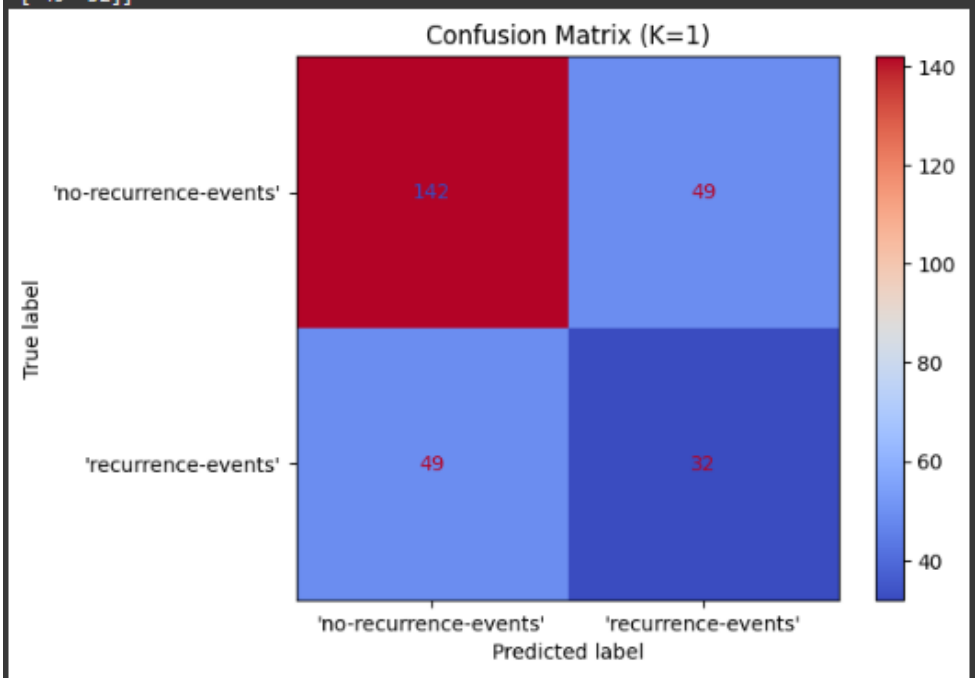



```
Configurazione 5: max_depth=7, min_impurity_decrease=0.01
Accuratezza: 0.6949
Matrice di confusione:
[[171  20]
 [ 63  18]]
              precision    recall  f1-score   support

           0       0.73      0.90      0.80       191
           1       0.47      0.22      0.30        81

    accuracy                           0.69       272
   macro avg       0.60      0.56      0.55       272
weighted avg       0.65      0.69      0.66       272
```

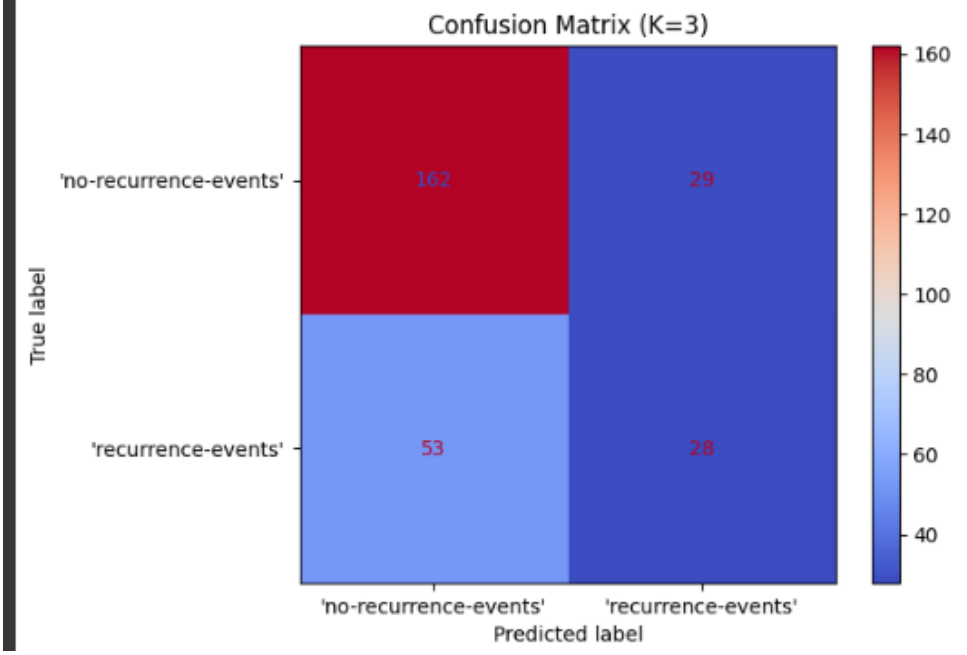


4. Considerando il classificatore K-Nearest Neighbor (K-NN) ed eseguendo una 10 fold CrossValidation stratificata, qual è l'impatto del parametro K sull'accuratezza media del classificatore? Riportate almeno 5 schermate che mostrino le matrici di confusione ottenute utilizzando diversi valori del parametro K. Eseguite una convalida incrociata stratificata a 10 volte con il classificatore Naïve Bayes. K-NN si comporta in media meglio o peggio del classificatore Naïve Bayes sui dati analizzati? Riportare una schermata che mostri la matrice di confusione ottenuta da Naïve Bayes sul set di dati analizzato.

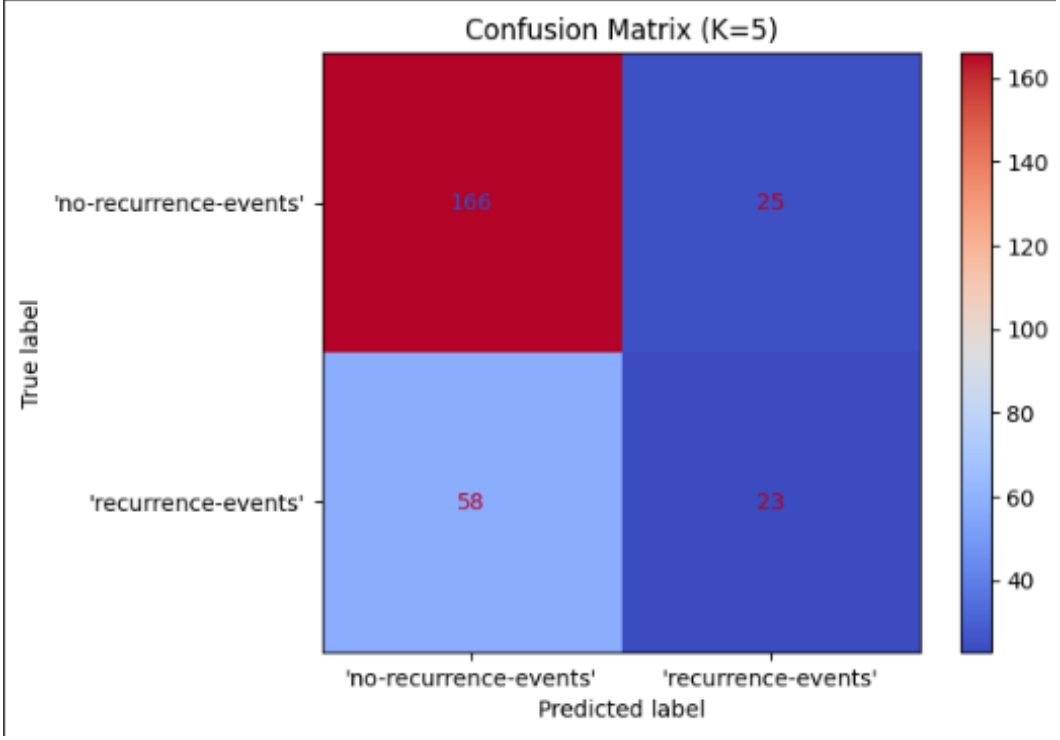
```
K=1
Accuratezza media: 0.6397
Matrice di confusione:
[[142  49]
 [ 49  32]]
```



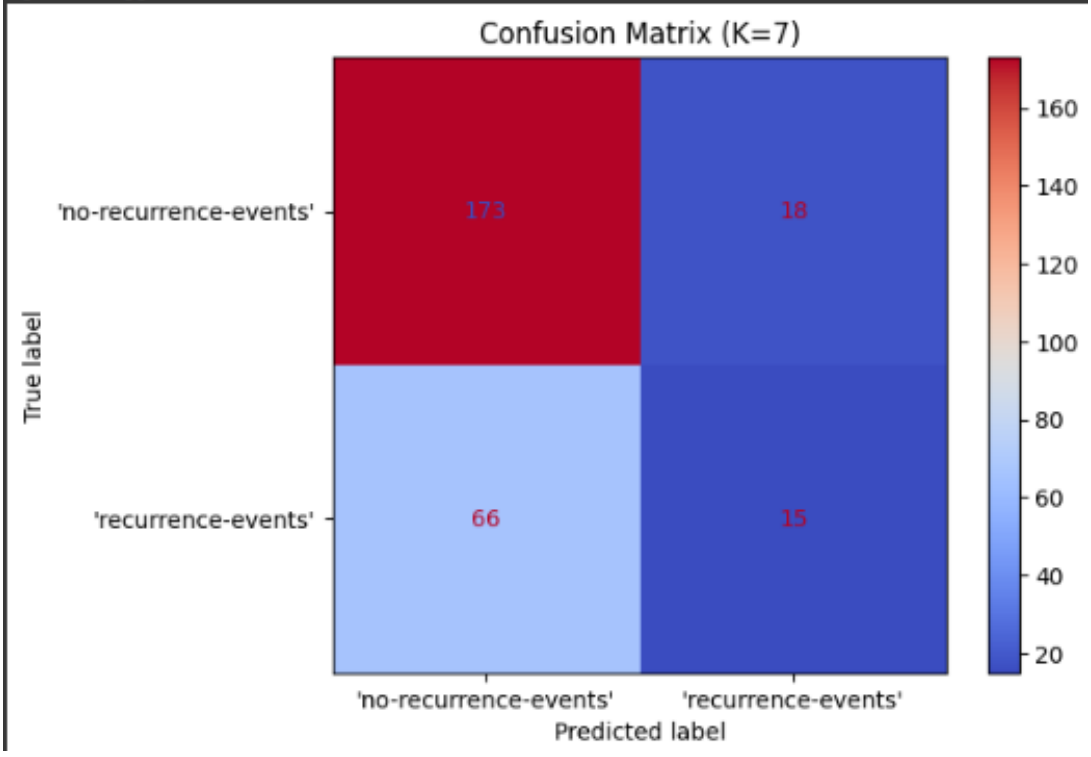

```
K=3
Accuratezza media: 0.6985
Matrice di confusione:
[[162  29]
 [ 53  28]]
```



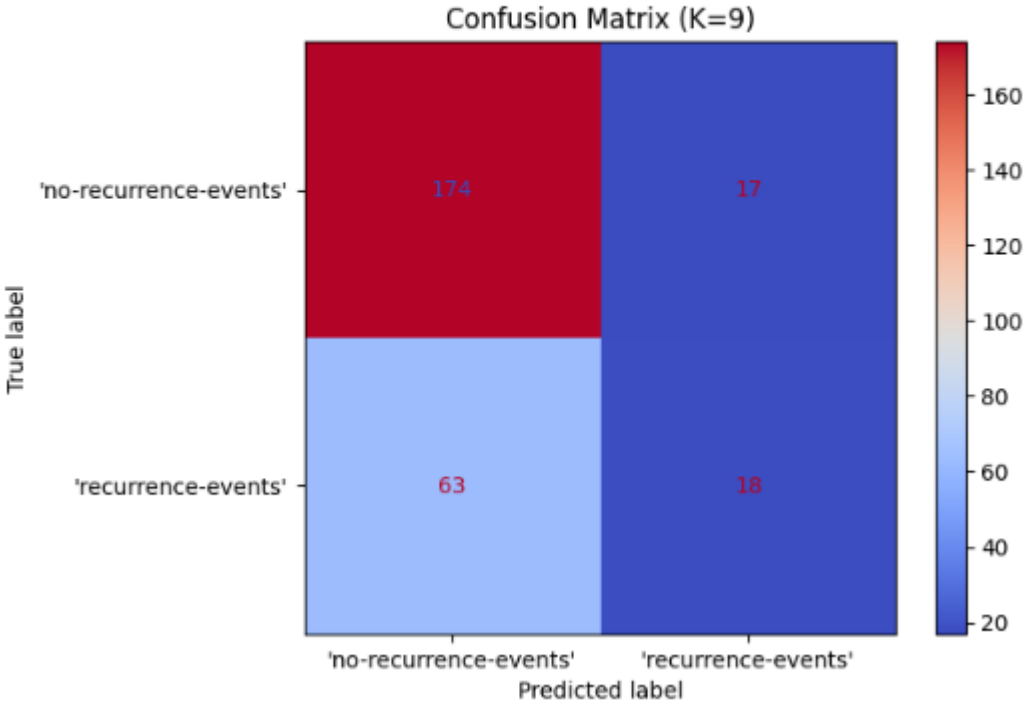
```
K=5
Accuratezza media: 0.6949
Matrice di confusione:
[[166  25]
 [ 58  23]]
```

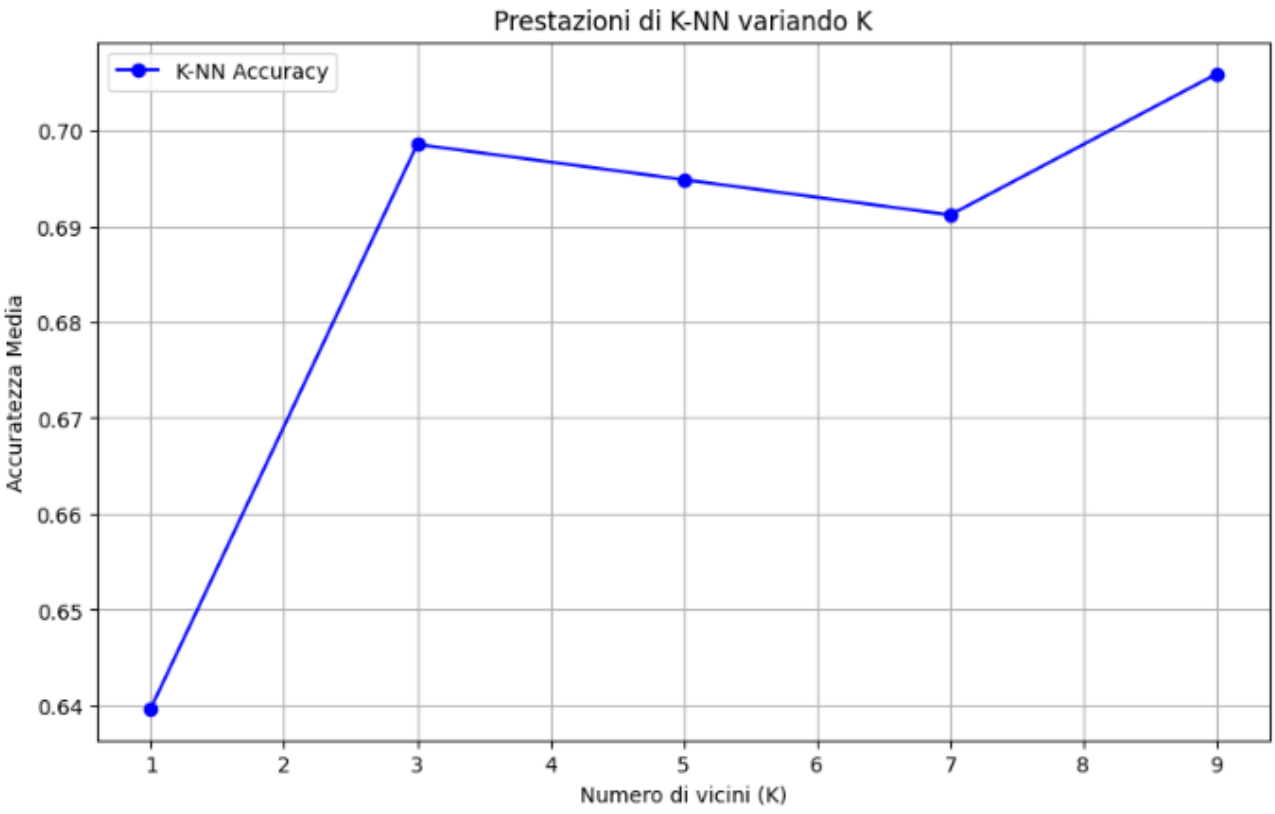



```
K=7
Accuratezza media: 0.6912
Matrice di confusione:
[[173  18]
 [ 66  15]]
```

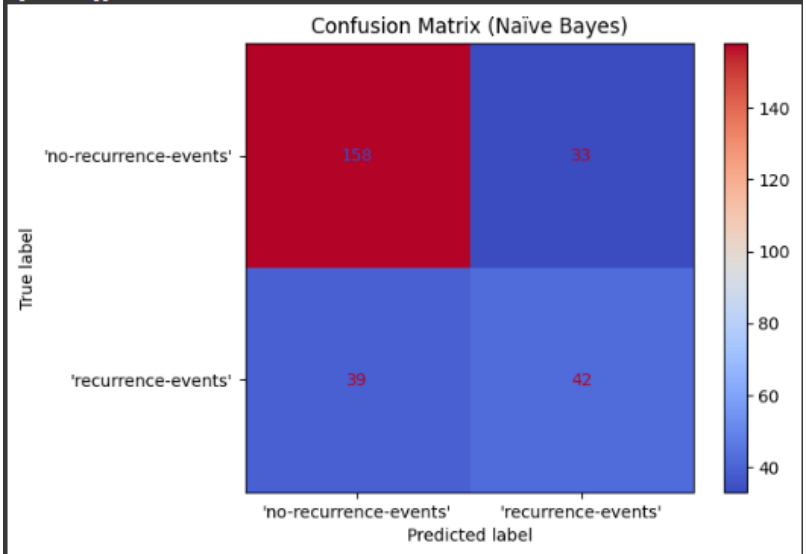


```
K=9
Accuratezza media: 0.7059
Matrice di confusione:
[[174  17]
 [ 63  18]]
```



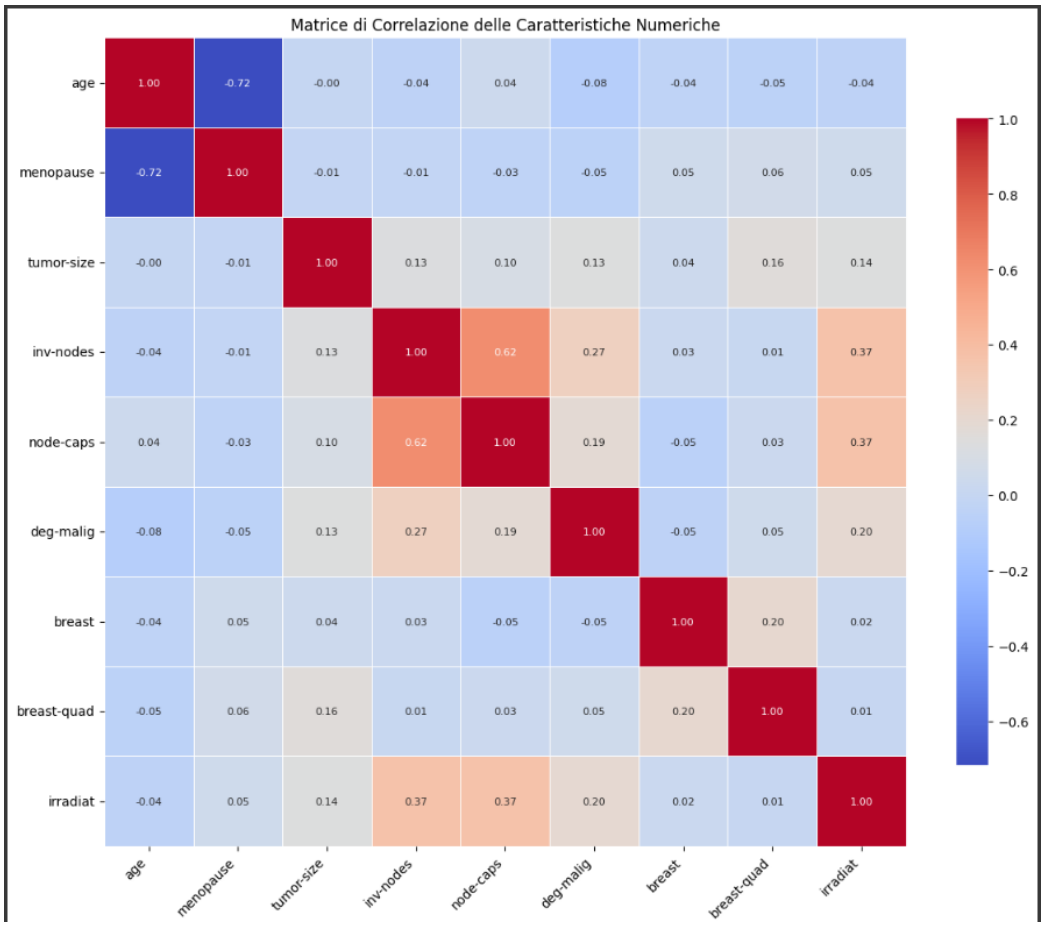


```
Analisi Naïve Bayes:

Naïve Bayes Classifier
Accuratezza media: 0.7353
Matrice di confusione:
[[158  33]
 [ 39  42]]
```




```
Confronto delle prestazioni:
Accuratezza media migliore di K-NN: 0.7059
Accuratezza media di Naïve Bayes: 0.7353
Naïve Bayes si comporta meglio di K-NN.
```

5. Analizzare la matrice di correlazione per scoprire le correlazioni a coppie tra gli attributi dei dati. Riportare una schermata che mostri la matrice di correlazione ottenuta. (a) L'ipotesi di indipendenza Naïve è effettivamente valida per il set di dati Breast? (b) Qual è la coppia di attributi più correlata?



L'ipotesi di indipendenza Naïve afferma che, nel classificatore Naïve Bayes, gli attributi sono indipendenti tra loro, ossia non c'è correlazione tra essi. Non ci sono forti correlazioni in questo dataset (la maggiore in valore assoluto è -0.72), quindi l'ipotesi può essere considerata valida. La coppia con la correlazione più alta è tra menopause e age, con una correlazione di -0.72.