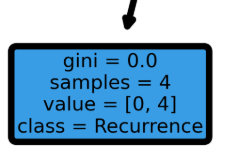
**Punto 1:**

Obbiettivo: costruire un albero decisionale con profondità massima 5 per determina se e il tumore sia un evento ricorrente o non ricorrente nella vita della paziente.

Risultati: Dopo un’attenta fase di pre-processing dei dati e encoding di è ottenuto che l’attributo più discriminante era “inv-nodes” e l’albero ha un’altezza di 5. Sono state individuate alcune partizioni pure.

Immagine che contiene testo, schermata, software, Software multimediale

Descrizione generata automaticamenteCodice pre- processing:



Partizione pura:

**Punto 2:**

Obbiettivo: Analizzare vari alberi decisionali ottenuti con il criterio di partizione basato sull’entropia usando parametri quali impurità minima, minimo di campioni per ogni foglia e profondità massima.

Risultati: Sono stati generati diversi alberi variando i parametri e si è visto quanto possano essere diversi l’uno dall’altro. Nelle immagini seguenti sono riportate configurazioni e viste degli alberi.

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamenteConfigurazioni:

Immagine che contiene testo, diagramma, Piano, linea

Descrizione generata automaticamenteAlbero 1:

Immagine che contiene diagramma, linea

Descrizione generata automaticamenteAlbero 2:

Immagine che contiene diagramma, mappa, Piano, testo

Descrizione generata automaticamenteAlbero 3:

Immagine che contiene testo, calligrafia, Carattere, diagramma

Descrizione generata automaticamenteAlbero 4:

Immagine che contiene testo, Carattere, calligrafia, schermata

Descrizione generata automaticamenteAlbero 5:

**Punto 3:**

Obbiettivo: valutare l’accuratezza dei 5 modelli precedenti e visualizzare le 5 matrici di confusione associate.

Risultati: Alcuni modelli performano meglio rispetto ad altri con percentuali di accuratezze che variano tra 67% e 73%. Accuratezze migliori potrebbero essere ottenute variando i parametri e andrebbero valutate su un dataset di validation.

Immagine che contiene testo, schermata, diagramma, numero

Descrizione generata automaticamenteImmagine che contiene testo, schermata, diagramma, numero

Descrizione generata automaticamenteImmagine che contiene testo, schermata, diagramma, numero

Descrizione generata automaticamenteImmagine che contiene testo, schermata, diagramma, numero

Descrizione generata automaticamenteMatrici di confusioni ordinato da albero 1 fino a 5:

Immagine che contiene testo, schermata, diagramma, numero

Descrizione generata automaticamente

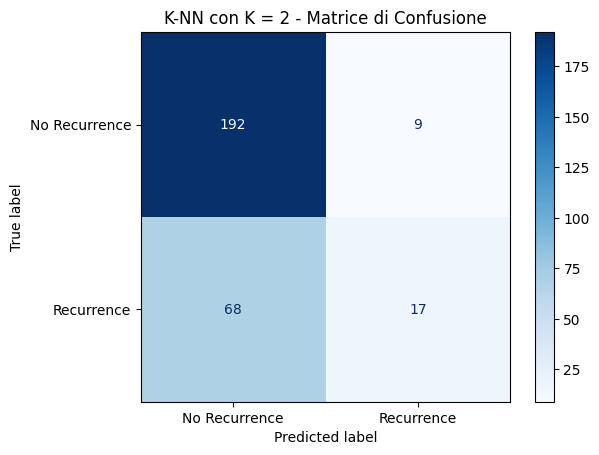
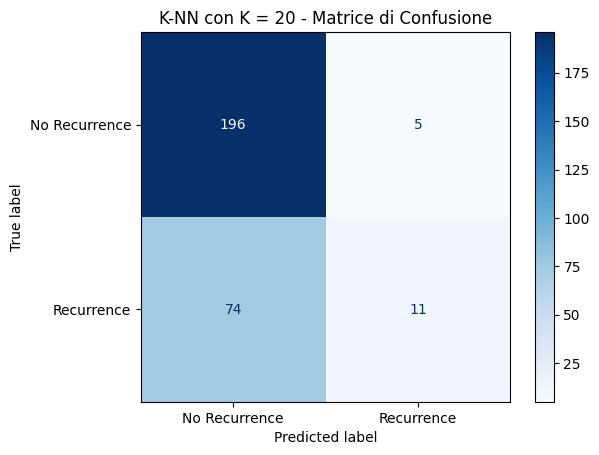
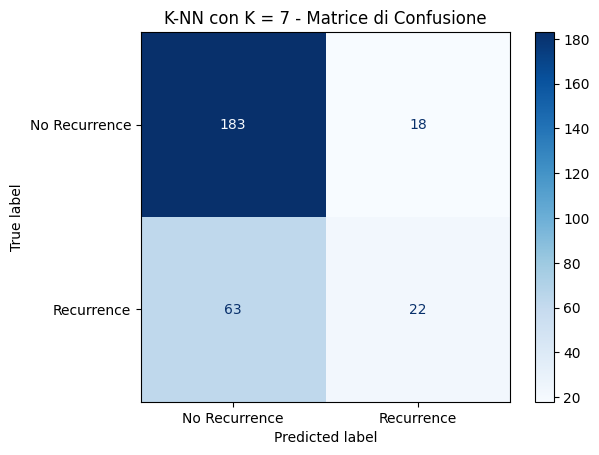
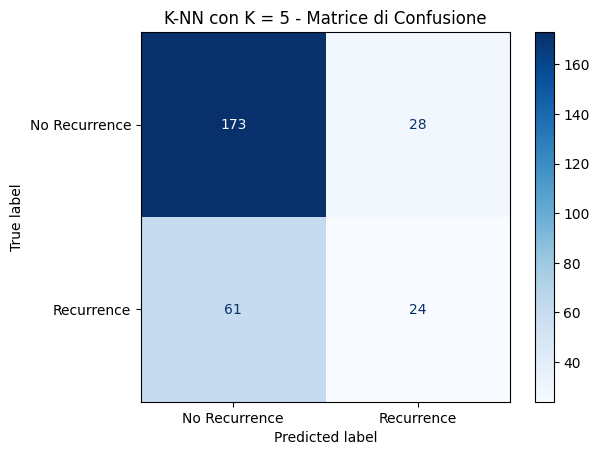
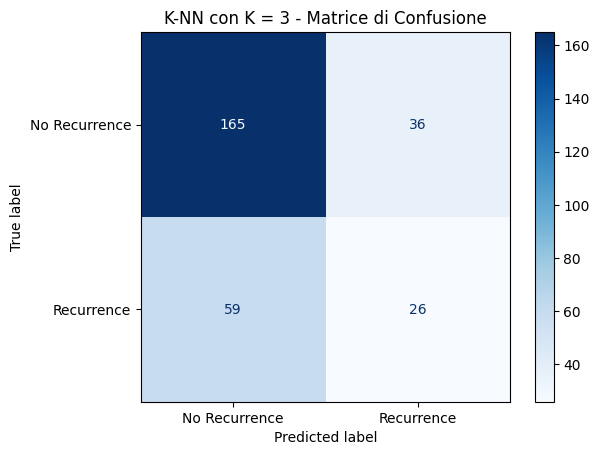
**Punto 4:**

Obbiettivo: Calcolare accuratezza e matrici di confusione di 5 modelli K-NN e confrontare l’accuratezza migliore con l’accuratezza di un classificatore Naïve Bayes usando una convalida incrociata stratificata.

Risultati: Si è ottenuto che l’accuratezza migliore di un modello K-NN (73.08%) è superiore rispetto a quella del classificatore Naïve Bayes (64.69%). Si è normalizzato i dati nel K-NN visto che alcuni attributi avevano valori numerici con ordini di grandezza diversi da altri e si è notato un incremento del 4% di accuratezza rispetto al miglior K-NN senza normalizzazione.

Matrici di confusione KNN e Naïve Bayes:

Immagine che contiene testo, schermata, diagramma, numero

Descrizione generata automaticamente 

**Punto 5:**

Obbiettivo: Analizzare la matrice di correlazione per scoprire le correlazioni a coppie tra gli attributi dei dati.

Risultati: Si è osservata una forte correlazione tra gli attributi “menopause\_premeno” (attributo derivato da “menopause” usando il Label Encoding) e “age” (-0.72). L'ipotesi di indipendenza Naïve non è valida per il set di dati Breast.

Immagine che contiene testo, schermata, numero, Parallelo

Descrizione generata automaticamente