SCARANO NATALE (Matricola: 708842)
PUTIGNANO MARTINO (Matricola: 697513)

link repository: https://github.com/GruppoIcon/EsameIcon

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DOCUMENTAZIONE PROGETTO INGEGNERIA DELLA CONOSCENZA

SISTEMA INTELLIGENTE PER LA PREDIZIONE DEL CANCRO AL SENO ED EPATITE

Identificare le malattie è il primo passo nel trattamento dei pazienti. Se fatto in modo accurato e tempestivo, potrebbe modificare in modo significativo l'esito del trattamento. Quindi, l'apprendimento automatico può essere un ottimo strumento nella classificazione delle malattie. Questo studio indaga le prestazioni di due modelli di apprendimento automatico, K-Nearest Neighbor e Decision Trees, su due set di dati sanitari, vale a dire il cancro al seno e l'epatite. Dopo un'analisi dei due set di dati e la preelaborazione, ottimizziamo gli iperparametri, la distanza e la funzione di costo per entrambi i modelli utilizzando la convalida incrociata k-fold. Troviamo che i modelli sono simili in termini di accuratezza sul set di dati sul cancro al seno. Tuttavia, il DecisionTree supera il K-Nearest Neighbor sul set di dati dell'epatite a causa del set di dati che contiene pochi casi ma numerose funzionalità. Inevitabilmente, analizziamo l'effetto della selezione delle caratteristiche sul set di dati dell'epatite per superare lo svantaggio della dimensionalità. Concludiamo quindi il nostro studio analizzando i confini decisionali dei nostri modelli e parametri come la sensibilità e la specificità dei nostri modelli.

Introduzione

Negli ultimi anni, sono stati sviluppati modelli di apprendimento automatico per identificare le malattie principalmente in base ai loro sintomi.
In questo progetto, implementiamo due diverse tecniche di classificazione (K-Nearest Neighbor e Decision Tree) per il cancro al seno e l'epatite. I due classificatori sono stati formati e verificati utilizzando due dataset sanitari di riferimento, set di dati sul cancro (Breast Cancer Wisconsin) e sull'epatite (Hepatitis). La classificazione per entrambe le malattie include livelli benigni o maligni.
I due dataset sono stati preelaborati prima di essere pronti per l'addestramento, verifica e test.

- La fase di pre-elaborazione includeva la normalizzazione dei dati e la rimozione di caratteristiche irrilevanti (come l'ID del paziente), istanze duplicate e istanze con caratteristiche mancanti;
-E' stata condotta una fase di analisi sul set di dati per migliorare la comprensione.

Dopo la preelaborazione e l'analisi di entrambi i set di dati, i modelli K-Nearest Neighbor e Decision Tree sono stati implementati da zero. I modelli sono stati quindi sottoposti a convalida incrociata per ottimizzare iperparametri come K in K-Nearest Neighbor e profondità massima in Decision Tree e per trovare la funzione di distanza ideale per K-Nearest Neighbor e la funzione di costo per Decision Tree per ciascun set di dati.
I risultati della nostra analisi comparativa mostrano che il classificatore Decision Tree supera leggermente il classificatore K-Nearest Neighbor sul set di dati sul cancro al seno e con un'elevata differenza sul set di dati sull'epatite.
I due set di dati su cui conduciamo la nostra ricerca sono distinti e simili in molti modi. Ad esempio, entrambi contengono una distribuzione delle classi benigna rispetto a quella maligna molto disomogenea. Tuttavia, la differenza più significativa tra il cancro al seno e il set di dati sull'epatite è il numero di punti dati e le caratteristiche che contengono. Questo ci consente di identificare e confrontare i punti di forza e di debolezza dei nostri modelli. Dopo la preelaborazione, il set di dati sull'epatite, che contiene un gran numero di caratteristiche rispetto al set di dati sul cancro al seno, viene lasciato con un piccolo numero di punti dati.

Dataset

I due set di dati, vale a dire il cancro al seno e l'epatite hanno dimensione 700 × 11 e 156 × 20 rispettivamente. Il cancro al seno contiene caratteristiche numeriche mentre l'epatite include caratteristiche sia numeriche che categoriali. Secondo la descrizione dei dati del cancro al seno, le etichette di classe sono associate a tumori benigni e maligni. Le etichette del

set di dati sull'epatite sono divise in due classi "die" e "live". Entrambi i set di dati contengono una distribuzione di classinon uniforme.

Per preelaborare entrambi i set di dati, viene definita una classe di preelaborazione, in cui vengono rimosse le informazioni mancanti. Quindi le probabili righe duplicate, vuote e malformate vengono omesse tramite una funzione dipulizia.

Viene rimossa una percentuale considerevole dal set di dati sull'epatite, circa il 48%, mentre solo il 3,46% dal set di datisul cancro al seno viene eliminato attraverso il processo di pulizia. Infine, entrambi i dataset vengono dichiarati in un formato unificato (l'etichetta della classe viene portata nella prima colonna, le funzionalità seguono nelle colonne successive). Le caratteristiche sono normalizzate nella funzione correlata tramite il ridimensionamento “minmax” implementato manualmente.

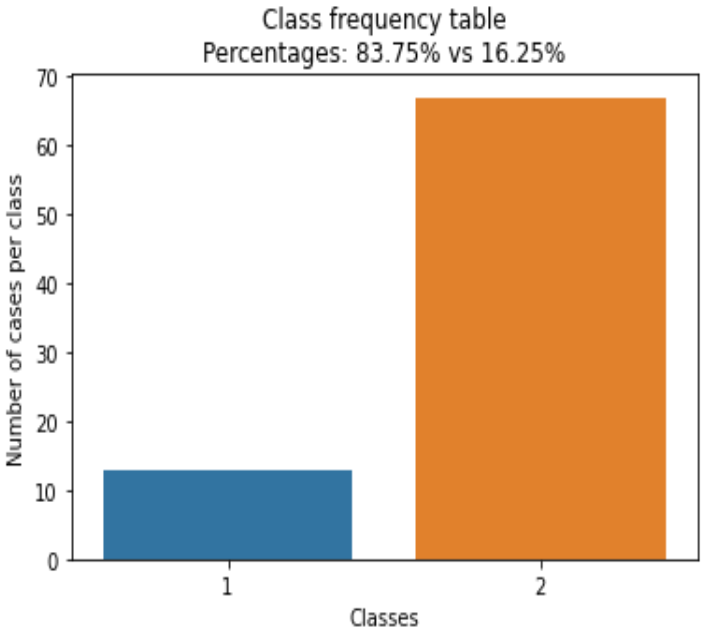
Per acquisire una migliore comprensione dei set di dati, viene distinto il numero di casi in ciascuna classe. Vienevisualizzata la proporzione di etichette maligne e benigne, oltre a quelle morte e vive.

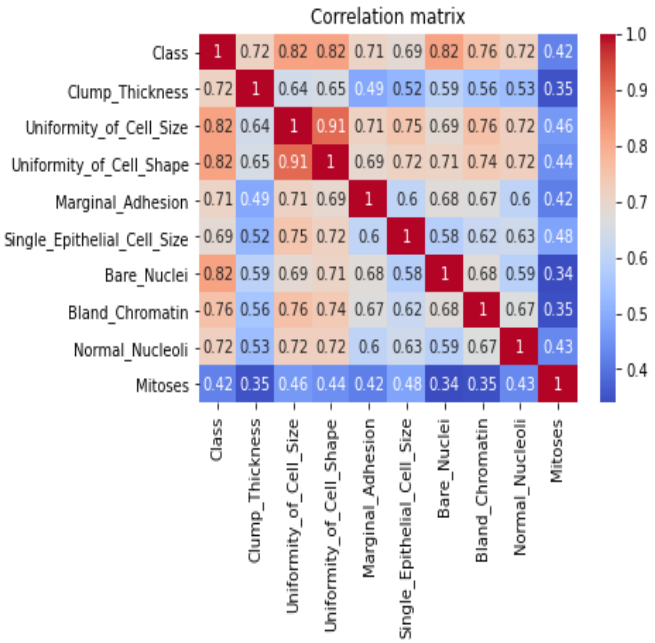
Un esempio può essere visto nella Figura 1 per il set di dati sul cancro al seno. L'analisi viene eseguita utilizzandoistogrammi e statistiche su set di dati.

Gli istogrammi si trovano per illustrare la distribuzione della frequenza di classe e caratteristica all'interno di entrambi icampioni di set di dati.

La Figura 1 può rappresentare chiaramente caratteristiche fortemente correlate con l'obiettivo nel set di dati sul cancro al seno. Le caratteristiche con correlazione di alta classe sono state selezionate nella speranza di ottenere una maggiore precisione.

Figura 1. Tabella di frequenza delle classi (a sinistra) e matrice di correlazione (a destra) per il set di dati sul cancro al seno





Risultati e Analisi

La tabella 1 contiene l'accuratezza dei modelli, con iperparametri sintonizzati che utilizzano la convalida incrociata k-fold sui dataset. Si può vedere che entrambi i modelli K-Nearest Neighbor e Decision Tree hanno prestazioni migliori sul set di dati sul cancro al seno con un'accuratezza simile di circa il 96%. Tuttavia, il classificatore Decision Tree supera K-Nearest Neighbor sul set di dati sull'epatite con una precisione del 7% superiore (92,6% contro 85,2%).

Come accennato in precedenza, il set di dati sull'epatite contiene 19 caratteristiche e un gran numero di istanze vengono rimosse durante la pre-elaborazione a causa di caratteristiche mancanti, lasciando il set di dati sull'epatite preelaborato con solo 80 istanze.

D'altra parte, il set di dati sul cancro al seno comprende solo 9 caratteristiche, rispetto alle 19 caratteristiche del set di dati sull'epatite e contiene 675 istanze.

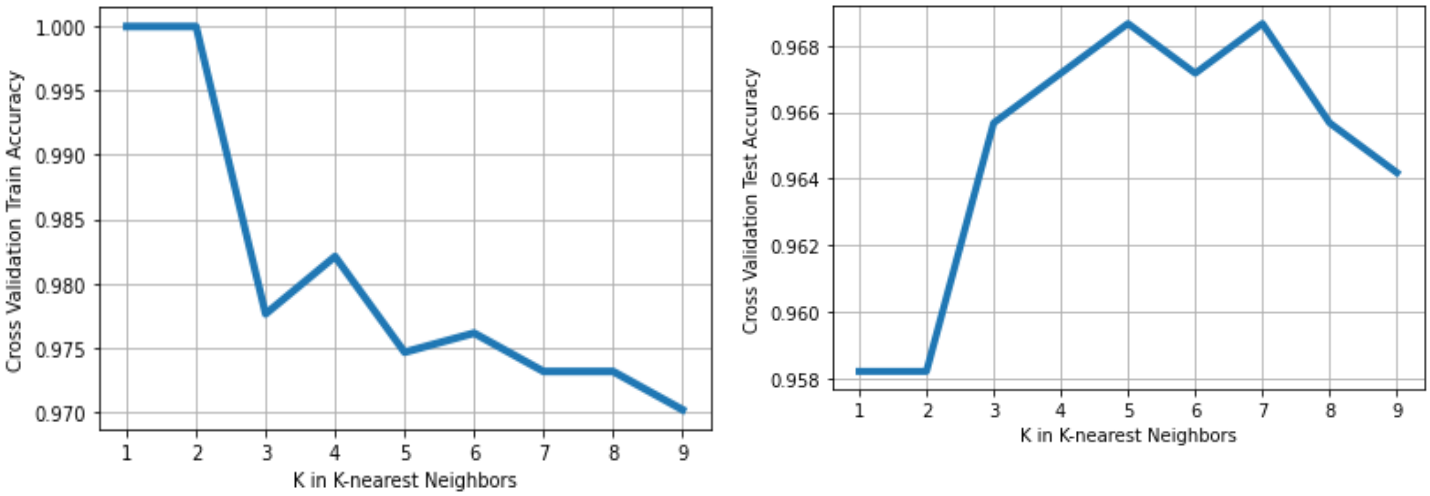
Poiché l'algoritmo K-Nearest Neighbor necessita di una grande quantità di dati per una buona prestazione, che il set di dati sull'epatite non fornisce, non è preciso rispetto all'albero decisionale. Inoltre, a causa della dimensionalità, l'apprendimento in 19 dimensioni con poche istanze diventa una sfida e l'accuratezza diminuisce per entrambi i modelli.

Di conseguenza, otteniamo una precisione inferiore rispetto al set di dati sul cancro al seno. La selezione delle caratteristiche viene utilizzata per superare la dimensionalità in set di dati come l'epatite.

Tabella 1. Accuratezza dei modelli sui dataset

	K-NN	Decision Tree
Tumore al seno	96,4%	96,8%
Epatite	85,2%	92,6%

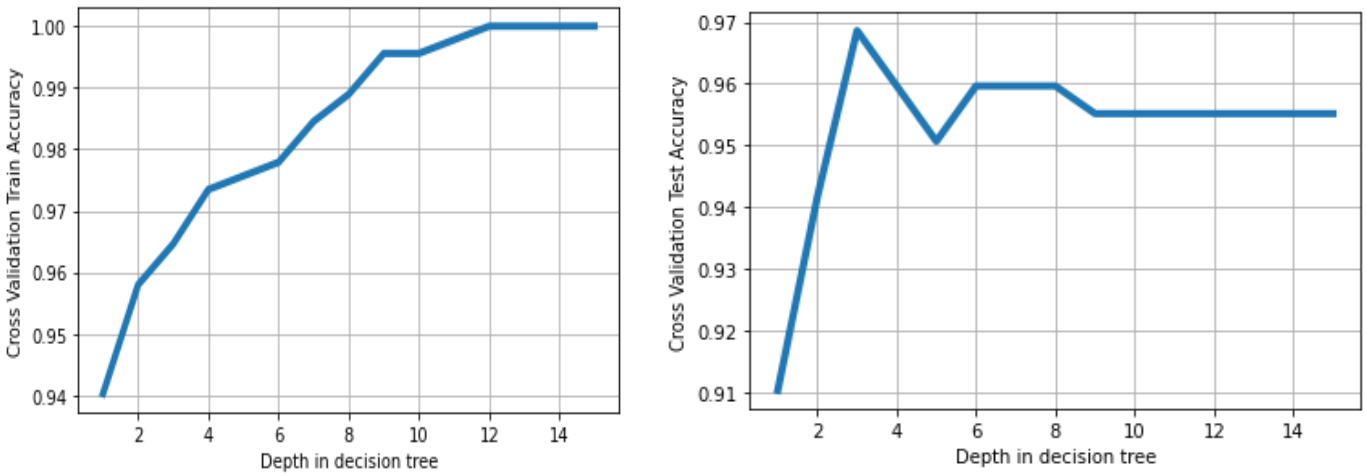
Figura 2. Accuratezza dei diversi valori K in KNN sul set di dati sul cancro al seno (training-sinistra, test-destra)



Valutiamo l'accuratezza del classificatore K-Nearest Neighbor sui dati di addestramento variando l'iperparametro K. Il risultato di questo esperimento sul set di dati sul cancro è mostrato nella Figura 2. Su entrambi i set di dati, l'algoritmo ha un'accuratezza del 100% quando K è uguale a 1. Questo risultato è previsto poiché stiamo valutando l'accuratezza in base agli esempi su cui è stato addestrato e il modello è overfitting. Man mano che l'iperparametro K cresce, il modello si adatta e la precisione diminuisce come risultato diretto.

Valutiamo quindi l'accuratezza del classificatore K-Nearest Neighbor su dati non visti. Usiamo 2/3 dei dati per l'addestramento e manteniamo il restante 1/3 come invisibile per i test. Contrariamente all'accuratezza dell'addestramento, l'accuratezza del test non diminuisce dal 100% al diminuire di K perché stiamo testando il modello su dati non visti. Tuttavia, la Figura 2 mostra risultati interessanti. Su entrambi i set di dati, l'accuratezza aumenta all'aumentare di K fino a raggiungere un picco. Una volta raggiunta la precisione di picco, l'aumento di K riduce la precisione. Ancora una volta, questo può essere spiegato da overfitting e underfitting. Il modello è overfitting quando K è piccolo. Tuttavia, poiché stiamo utilizzando dati non visti per testare il modello, l'overfitting risulta in una bassa precisione. Allo stesso modo, l'aumento di K riduce la precisione. Tuttavia, esiste un valore K che bilancia il modello che si traduce nella massima precisione.
Per quanto riguarda l'implementazione dell'algoritmo dell'albero delle decisioni, la precisione aumenta per l'addestramento, poiché la profondità massima aumenta, può raggiungere il punto in cui può rimanere quasi costante. Si può dedurre che il modello potrebbe essere overfit in una certa profondità massima e la precisione è massima per l'addestramento.
Questo risultato è diverso da K-Nearest Neighbor che si sovrappone quando K è piccolo e fornisce una precisione del 100% sui dati di addestramento. Per valori inferiori impostati alla profondità massima, la precisione del test è bassa come previsto. All'aumentare della profondità massima, l'accuratezza del test aumenta notevolmente e a valori di profondità massima molto più elevati si sovrappone, provocando a sua volta un calo drastico. Nel mezzo, c'è un'area ottimale in cui l'accuratezza del test è intorno al suo picco. In poche parole, l'overfitting del modello si traduce in una maggiore precisione di addestramento mentre una minore precisione di test.

Figura 3. Precisione dei diversi valori di profondità massima in Decision Tree sul set di dati sul cancro al seno (training-sinistra, test-destra)



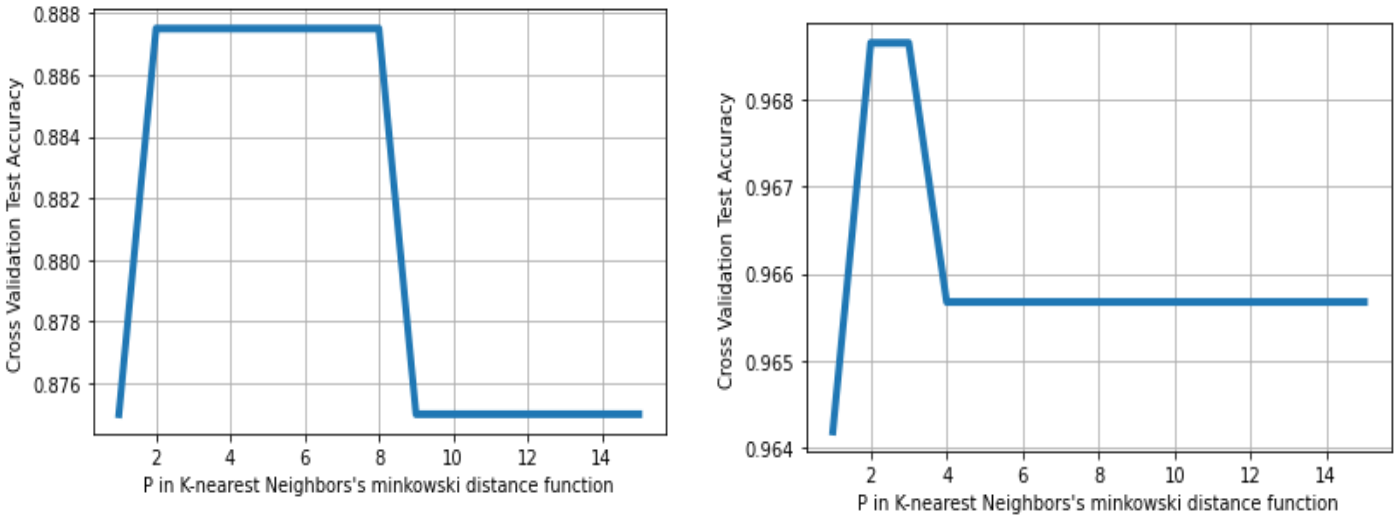
Nell'albero decisionale, per entrambi i set di dati, l'accuratezza viene calcolata per diversi metodi della funzione di costoin base al tasso di classificazione errata, all'entropia e all'indice di Gini. Per il set di dati sul cancro al seno, l'errata classificazione con un'accuratezza del 96,86% è la più alta tra le diverse funzioni di costo applicate, mentre l'entropia con un'accuratezza del 94,17% è la peggiore. Allo stesso modo, per il set di dati sull'epatite, l'accuratezza dell'errata classificazione

(92,29%) si trova in cima piuttosto che l'indice di Gini (88,88%) e l'accuratezza dell'entropia (77,77%). In termini di costocomputazionale, l'entropia è

il più costoso rispetto a errata classificazione e Gini. Ci sono due ragioni dietro l'alto costo computazionale dell'entropia.Non solo esegue la funzione logaritmica, ma impone una condizione per saltare il calcolo del logaritmo zero in ogni iterazione. La funzione tasso di classificazione errata sembra avere la spesa computazionale più bassa poiché non esegue operazioni matematiche così considerevoli all'interno.

Poiché il set di dati sull'epatite conteneva caratteristiche sia numeriche che categoriche, il nostro approccio a una funzione di distanza per l'algoritmo KNN è stato quello di prendere una somma ponderata della distanza di hamming e Minkowski calcolata rispettivamente per le caratteristiche categoriali e numeriche. Tuttavia, essendo il set di dati sull'epatite molto piccolo, non abbiamo ottenuto alcuna differenza nell'accuratezza utilizzando questo metodo. Abbiamoscoperto che KNN ha ottenuto una maggiore precisione quando p nella funzione Minkowski era uguale a 2 (distanza euclidea) per entrambi i set di dati. Come si vede nella Figura 4, la precisione diminuisce quando p è piccolo o molto grande.

Figura 4. Precisione dei diversi valori di p nella funzione distanza Minkowski (epatite-sinistra, cancro al seno-destra)



I limiti delle decisioni sono illustrati nella Figura 5. Come si può chiaramente vedere, i confini sono più lisci in KNN rispetto a quelli in Decision Tree. Le linee di separazione nei limiti di Decision Tree comunicano la discrezionalità delle condizioni applicate sulle soglie delle funzionalità. I confini uniformi di KNN, d'altra parte, derivano da calcoli continui della distanza nelle regioni di separazione. Inoltre dai limiti decisionali possiamo osservare che non tutti i casi sono stati classificati correttamente. Questo può essere visto dai punti gialli nelle regioni viola e dai punti viola nelle regioni gialle. Possiamo anche dedurre dai limiti decisionali che i nostri modelli non sono né overfitting né underfitting.

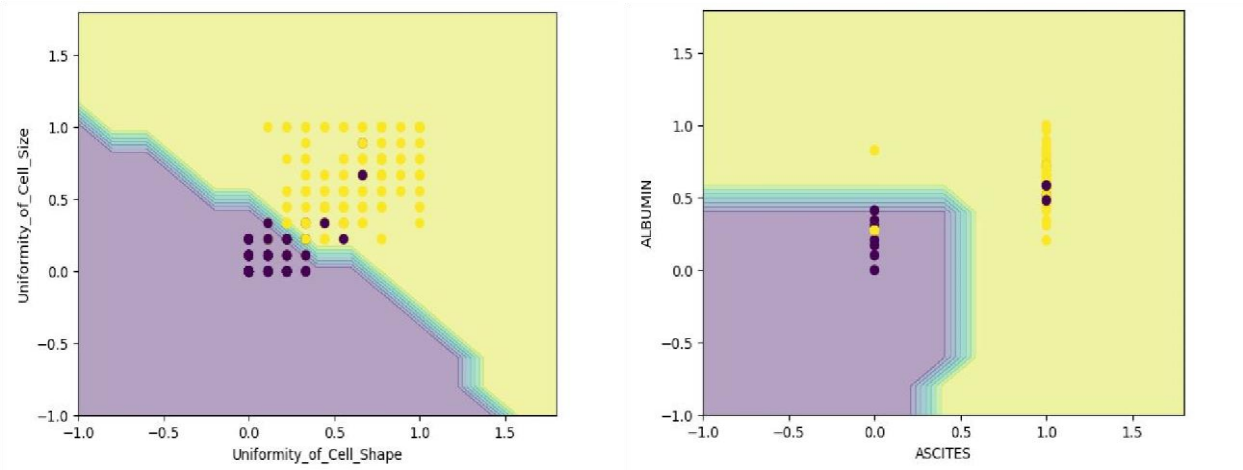
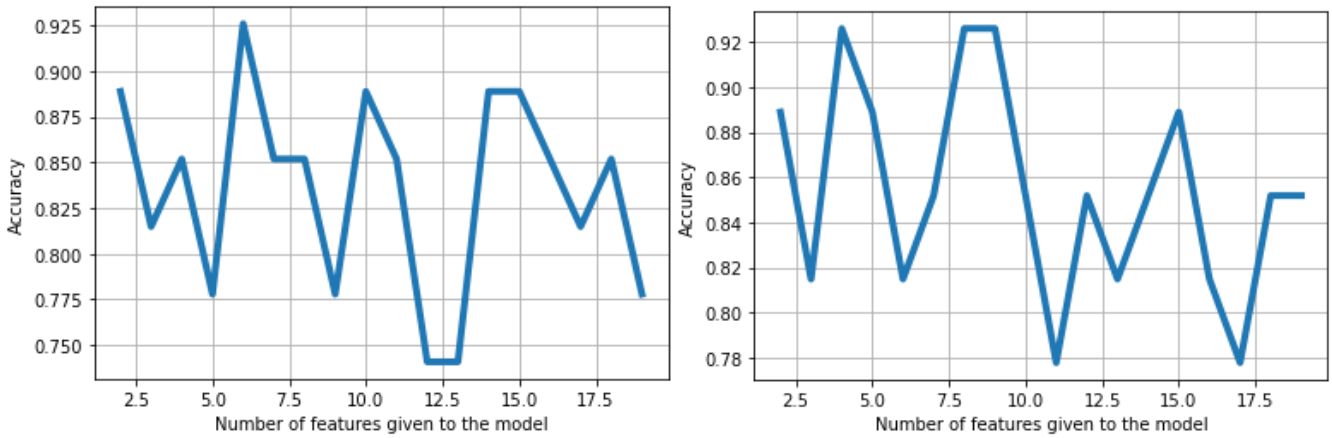


Figura 5. Numero di caratteristiche dell'epatite fornite ai modelli (KNN-sinistra Decision Tree-right)



In tali malattie potenzialmente letali come il cancro, è fondamentale rilevare tutti i casi positivi (tumori maligni). In questasituazione, la sensibilità è una metrica decisiva che afferma quanto il nostro modello rileverebbe con precisione i casi positivi. La tabella 2 è completata dopo aver tracciato la Confusion Matrix. Sia la sensibilità che le metriche delle prestazioni di precisione possono essere idealmente derivate per quanto riguarda la matrice di confusione tracciata.

Otteniamo una sensibilità e specificità zero per il set di dati sull'epatite utilizzando l'algoritmo KNN. Ciò è dovuto principalmente al set di dati contenente pochi numeri di casi di “die” che limita la precisione della nostra ricerca. A parte questo, siamo in grado di raggiungere questi parametri per il cancro al seno. Una maggiore sensibilità si trova utilizzando il modello Decision Tree. Pertanto, questo modello può essere migliore per rilevare malattie potenzialmente letali.

Tabella 2. Metriche delle prestazioni per i modelli di entrambi i set di dati

	Algoritmo distribuito	sensibilità	Specificità	Precisione
Set di dati sul cancro al seno	Algoritmo dell'albero decisionale	0,944	0,980	0,957
	Algoritmo KNN	0,941	0,955	0,930
Set di dati sull'epatite	Algoritmo dell'albero decisionale	0,625	0,844	0,625
	Algoritmo KNN	0	0,888	0

L'algoritmo KNN ha dato luogo a scarse prestazioni sul set di dati sull'epatite. Di conseguenza, abbiamo eseguito la selezione delle caratteristiche addestrando il modello con le caratteristiche con la correlazione di classe più elevata. Abbiamo calcolato l'accuratezza del modello addestrato con N caratteristiche più correlate, i cui risultati possono essere visti sempre nella Figura 5. L'accuratezza dei nostri modelli rimane per lo più la stessa con un numero ridotto di caratteristiche. Alla luce di questi risultati, possiamo concludere che il dataset dell'epatite, una volta preelaborato, perde una notevole quantità di dati e, poiché ha dimensioni elevate, l'apprendimento diventa una sfida per il nostro algoritmo. Pertanto, la riduzione delle caratteristiche deve essere utilizzata per affrontare questo problema su set di dati simili al setdi dati sull'epatite.

Conclusioni

In questo progetto, abbiamo addestrato ed esaminato due tecniche di classificazione, vale a dire K-Nearest Neighbor e Decision tree su due dataset sanitari. Il set di dati sul cancro al seno conteneva meno caratteristiche rispetto al set di datisull'epatite, pur contenendo più punti dati. Questa differenza tra i nostri due set di dati ci ha permesso di estrarre somiglianze e differenze tra i due modelli. Abbiamo osservato che entrambi i modelli potevano raggiungere il 100% di precisione, adattandosi così ai dati di addestramento. Questo overfitting si è verificato quando K era piccolo nel modello KNN e la profondità massima era grande nell'albero decisionale. Entrambi i modelli hanno ottenuto la massima precisione sui dati di test mentre gli iperparametri erano bilanciati. Inoltre, utilizzando la distanza euclidea e la funzione dicosto, la massima precisione è stata ottenuta rispettivamente dal KNN e dagli alberi decisionali. Dopo aver messo a punto gli iperparametri, abbiamo scoperto che entrambi i modelli hanno funzionato in modo simile sul set di dati sul cancro al seno mentre il modello Decision Tree ha superato KNN nel set di dati sull'epatite. Abbiamo scoperto che ciò era dovuto alla dimensionalità poiché il set di dati sull'epatite conteneva un numero sbilanciato di caratteristiche rispetto ai punti dati. Per superare questo problema, abbiamo provato la selezione delle caratteristiche utilizzando la correlazionedi classe. La riduzione delle funzionalità sul set di dati sull'epatite non ha ridotto la precisione. Pertanto, abbiamo concluso che il set di dati sull'epatite non contiene dati sufficienti per prestazioni adeguate utilizzando l'algoritmo KNN. Una direzione interessante per le indagini future sarebbe migliorare l'accuratezza del modello KNN su set di dati con elevata dimensionalità e quantità limitata di dati attraverso tecniche di selezione delle caratteristiche.