CASO DI STUDIO ICON 2024

Cervical Cancer

AUTRICI

D'Imperio Francesca [741130], f.dimperio2@studenti.uniba.it
Franco Giorgia [738722], g.franco28@studenti.uniba.it

Link GitHub per codice sorgente

https://github.com/fradimp13/Caso-di-studio-ICON-D-Imperio-Franco.git

INDICE

INTRODUZIONE

INFORMAZIONI SUL PROGETTO

Il caso di studio mira a utilizzare i dati clinici raccolti da pazienti che hanno partecipato a programmi di screening per il cancro alla cervice con l'obiettivo di identificare pattern, correlazioni e caratteristiche che possano essere indicative di un potenziale sviluppo di cancro. Questo comprende l'uso di tecniche di apprendimento supervisionato, apprendimento non supervisionato per rilevare precocemente segnali di allarme per la malattia.

L'importanza di questo approccio risiede nel potenziale di individuare casi di cancro alla cervice in una fase più precoce, consentendo un intervento terapeutico tempestivo e migliorando così le prospettive di trattamento e sopravvivenza per le pazienti coinvolte. Inoltre, ciò potrebbe contribuire a ottimizzare le risorse sanitarie, dirigendo in modo più mirato gli interventi di screening e diagnostici verso le persone con maggior rischio di sviluppare la malattia.

INFORMAZIONI TECNICHE

Abbiamo impiegato nel nostro studio sia algoritmi di apprendimento supervisionato che non supervisionato. In seguito, ci siamo concentrati sulla costruzione di una rete bayesiana per condurre analisi e inferenze mirate.

DATASET DI RIFERIMENTO

Link: Cervical Cancer Dataset (kaggle.com)

ORGANIZZAZIONE DEL DATASET

ANALISI DEI DATI

Il dataset è composto da 835 righe e 36 colonne. Di seguito lo screen delle colonne presenti:

```
queste sono le colonne nel nostro dataset originale
Index(['Age', 'Number of sexual partners', 'First sexual intercourse',
       'Num of pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)',
       'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD',
       'IUD (years)', 'STDs', 'STDs (number)', 'STDs:condylomatosis',
       'STDs:cervical condylomatosis', 'STDs:vaginal condylomatosis',
       'STDs:vulvo-perineal condylomatosis', 'STDs:syphilis',
       'STDs:pelvic inflammatory disease', 'STDs:genital herpes',
       'STDs:molluscum contagiosum', 'STDs:AIDS', 'STDs:HIV',
       'STDs:Hepatitis B', 'STDs:HPV', 'STDs: Number of diagnosis',
       'STDs: Time since first diagnosis', 'STDs: Time since last diagnosis',
       'Dx:Cancer', 'Dx:CIN', 'Dx:HPV', 'Dx', 'Hinselmann', 'Schiller',
       'Citology', 'Biopsy'],
      dtype='object')
```

In seguito ad una ricerca su Internet abbiamo optato per la rimozione di colonne, non essenziali per la diagnosi finale. Le colonne da noi rimosse sono le seguenti:
-hormonal contraceptive
-hormonal contraceptive(year)
-IUD
-IUD (year)
-STDs cervical condylomatusis
-STD AIDS
-epatite B

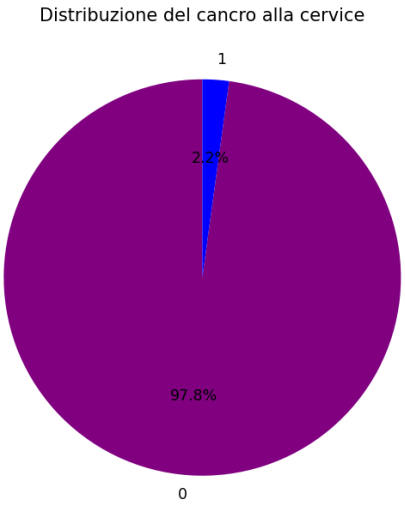
Abbiamo scelto di rimuovere le seguenti colonne poiché, su un totale di 835 valori disponibili per la colonna, solo un valore era pari a 1, i restanti avevano un valore comune di 0; di conseguenza, queste colonne sono sembrate poco influenti per lo scopo della predizione.
-STDs vaginal condylomatusis
-syphilis
-pelvic
-genital herpes
-molluscum contagiosum

Di seguito il dataset aggiornato:

```
queste sono le colonne nel nostro dataset aggiornato
Index(['Age', 'Number of sexual partners', 'First sexual intercourse',
       'Num of pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)',
       'STDs', 'STDs (number)', 'STDs:condylomatosis',
       'STDs:vulvo-perineal condylomatosis', 'STDs:HIV', 'STDs:HPV',
       'STDs: Number of diagnosis', 'STDs: Time since first diagnosis',
       'STDs: Time since last diagnosis', 'Dx:Cancer', 'Dx:CIN', 'Dx:HPV',
       'Dx', 'Hinselmann', 'Schiller', 'Citology', 'Biopsy'],
      dtype='object')
```

```
Numero di righe presenti nel Dataset: 835
Numero di colonne presenti nel Dataset: 24
```

È stato utilizzato un grafico a torta per comprendere meglio la distribuzione del dataset in relazione alla variabile target “Dx: Cancer”, ovvero la diagnosi finale, con due valori distinti:
0, casi senza cancro
1, casi con cancro



In seguito, abbiamo controllato quante colonne discrete e continue ci sono nel nostro dataset, e successivamente la verifica dei valori nulli presenti per ogni colonna, indicando inoltre il tipo di dati.

1.

```
Le colonne di tipologia discreta sono:
 []

Le colonne di tipologia continua sono:
 ['Age', 'Number of sexual partners', 'First sexual intercourse', 'Num of pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)', 'STDs', 'STDs (number)', 'STDs:condy
lomatosis', 'STDs:vulvo-perineal condylomatosis', 'STDs:HIV', 'STDs:HPV', 'STDs: Number of diagnosis', 'STDs: Time since first diagnosis', 'STDs: Time since last diagnosis', '
Dx:Cancer', 'Dx:CIN', 'Dx:HPV', 'Dx', 'Hinselmann', 'Schiller', 'Citology', 'Biopsy']
```

2.

```
Valori nulli per ogni colonna:
 Age                                   0
Number of sexual partners             25
First sexual intercourse               7
Num of pregnancies                    56
Smokes                                13
Smokes (years)                        13
Smokes (packs/year)                   13
STDs                                 100
STDs (number)                        100
STDs:condylomatosis                  100
STDs:vulvo-perineal condylomatosis    100
STDs:HIV                             100
STDs:HPV                             100
STDs: Number of diagnosis              0
STDs: Time since first diagnosis     764
STDs: Time since last diagnosis      764
Dx:Cancer                              0
Dx:CIN                                 0
Dx:HPV                                 0
Dx                                     0
Hinselmann                             0
Schiller                               0
Citology                               0
Biopsy                                 0
dtype: int64
```

È stata utilizzata un approccio di imputazione basato su KNN Imputer, che ha consentito di stimare e completare i valori mancanti considerando la somiglianza tra i casi circostanziali utilizzando una metrica di distanza appropriata.

Effettuando nuovamente la verifica dei valori nulli, tutti i campi sono stati popolati.

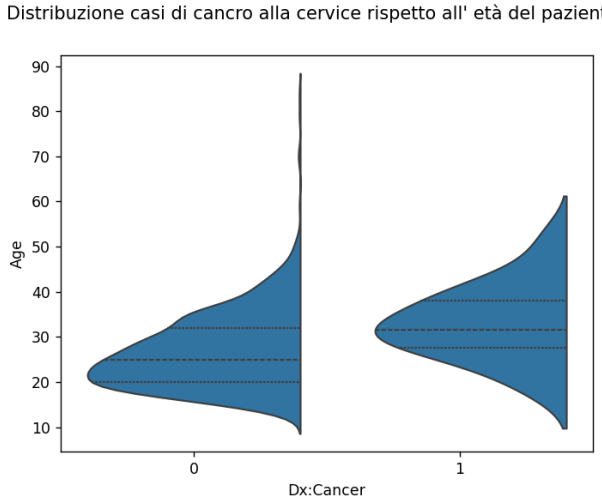
```
Age                                 0
Number of sexual partners           0
First sexual intercourse            0
Num of pregnancies                  0
Smokes                              0
Smokes (years)                      0
Smokes (packs/year)                 0
STDs                                0
STDs (number)                       0
STDs:condylomatosis                 0
STDs:vulvo-perineal condylomatosis  0
STDs:HIV                            0
STDs:HPV                            0
STDs: Number of diagnosis           0
STDs: Time since first diagnosis    0
STDs: Time since last diagnosis     0
Dx:Cancer                           0
Dx:CIN                              0
Dx:HPV                              0
Dx                                  0
Hinselmann                          0
Schiller                            0
Citology                            0
Biopsy                              0
```

OSSERVAZIONE GRAFICA DEI DATI

Siamo proceduti con l'analisi delle caratteristiche confrontate con il target al fine di identificare quali elementi hanno un impatto significativo sulla diagnosi finale. I risultati ottenuti sono stati successivamente supportati da approfondimenti condotti attraverso ricerche online.

Qui di seguito i grafici a violino da noi selezionati.

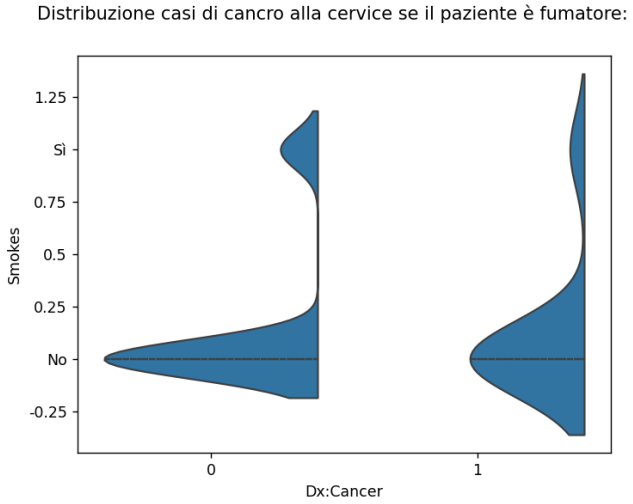
Grafico 1:



Il grafico mostra che le pazienti di età compresa tra i 20 e i 30 anni tendono a non essere positive alla diagnosi di cancro cervicale. Tuttavia, ci sono casi in cui anche le pazienti più giovani o più anziane possono risultare positive alla diagnosi del cancro.

In generale, si osserva un aumento del numero di casi positivi al cancro cervicale all'aumentare dell'età (tra i 30 e i 40 anni), come indicato dall'ampiezza del grafico, il che suggerisce una maggiore incidenza di questa malattia nelle fasce di età più avanzate.

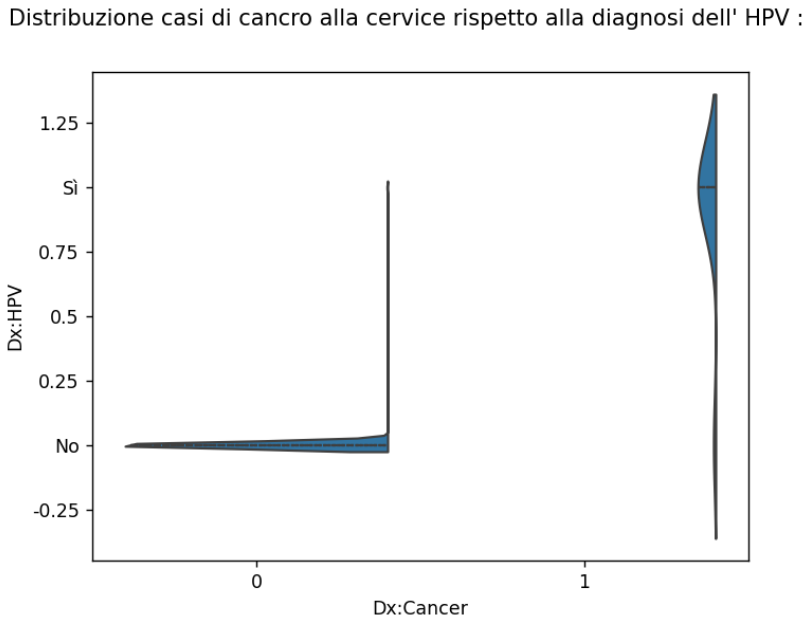
Grafico 2:



Il grafico suggerisce che il fumo può avere un impatto sulla diagnosi del cancro cervicale, tuttavia questo impatto sembra essere modesto. Vi sono casi di pazienti non fumatrici che non hanno contratto il cancro, così come ci sono pazienti non fumatrici che invece ne sono affette. Lo stesso si verifica anche per le pazienti fumatrici, sebbene in numero minore.

In generale, l'ampiezza del grafico è maggiore nel caso delle non fumatrici e minore nel caso delle fumatrici, indicando che il fumo potrebbe essere un fattore di rischio meno significativo rispetto ad altri fattori esaminati

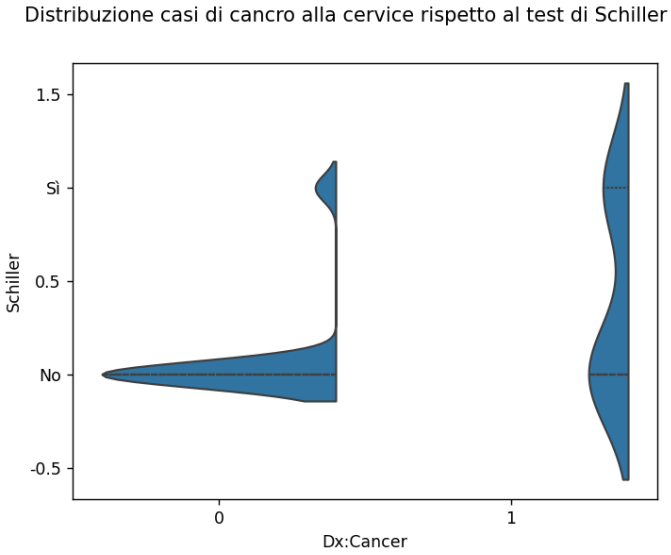
Grafico 3:



Dal grafico si deduce che le pazienti che risultano negative alla diagnosi dell'HPV generalmente non hanno il cancro, mentre quelle che risultano positive alla diagnosi dell'HPV spesso sono affette da questa malattia.

In generale, questo suggerisce che l'HPV è uno dei fattori più significativi nell'insorgenza del cancro cervicale.

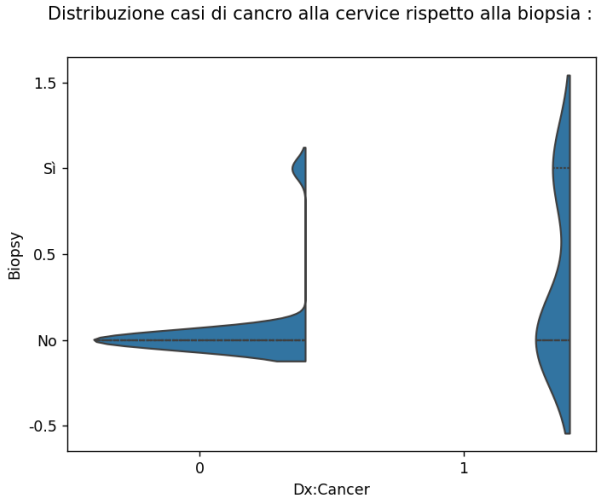
Grafico 4:



Dal grafico si evince che la maggior parte delle pazienti risulta negativa al test di Schiller e non ha contratto il cancro cervicale; invece, una piccola parte risulta positiva al test, ma non è affetta dalla malattia. Nel gruppo dei pazienti affetti da cancro, si nota un numero approssimativamente uguale di casi sia tra coloro che sono risultati positivi al test di Schiller sia tra coloro che sono risultati negativi.

In generale, il test di Schiller è un importante strumento medico utilizzato per valutare la salute della cervice uterina durante l'esame colposcopico. Pur non essendo direttamente correlato alla diagnosi dell'HPV o del cancro cervicale, il test fornisce informazioni cruciali sulla struttura e l'integrità della cervice uterina. Questo è fondamentale per individuare eventuali anomalie o lesioni precancerose che potrebbero essere presenti.

Grafico 5:



Dal grafico si deduce che la maggioranza delle pazienti risulta negativa alla biopsia e non ha contratto il cancro cervicale; mentre solo una piccola percentuale risulta positiva al test, ma non è affetta dalla malattia. Questi risultati riflettono la distribuzione dei casi nel nostro dataset, dove quasi il 98% delle pazienti non ha contratto il cancro, mentre solo il restante 2% ha

ricevuto una diagnosi positiva. Nel gruppo dei pazienti affetti da cancro invece, si osserva un numero approssimativamente uguale di casi sia tra coloro che sono risultati positivi alla biopsia sia tra coloro che sono risultati negativi.

Nonostante ciò, la biopsia rappresenta uno dei test fondamentali per individuare la presenza di cancro alla cervice uterina.

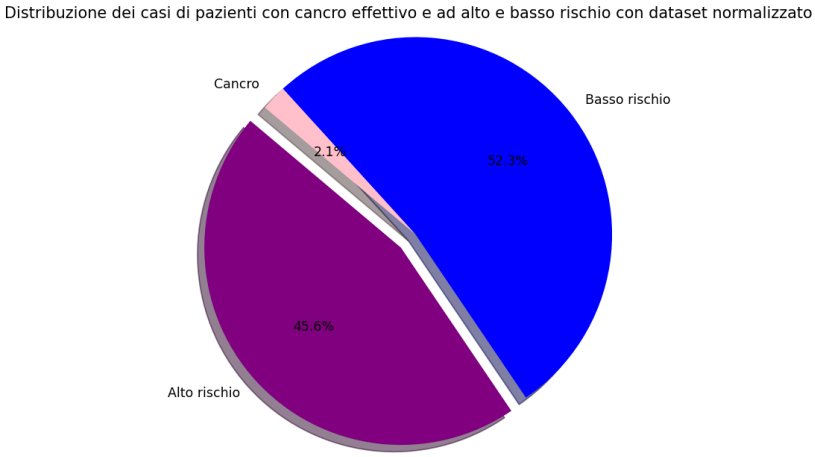
APPRENDIMENTO NON SUPERVISIONATO

CLUSTERING CON K-MEANS

Nel corso dello studio del nostro dataset, ci siamo resi conto di una distribuzione non equilibrata tra i casi positivi e negativi al cancro. Al fine di affrontare questa disparità e ottenere una suddivisione più bilanciata, abbiamo deciso di concentrarci sul 97.8% dei casi negativi al cancro e suddividerli ulteriormente in due sotto-categorie distinte: pazienti a basso rischio di cancro e pazienti ad alto rischio di cancro. Questa suddivisione è stata realizzata utilizzando l'algoritmo K-means per il clustering dei dati. Successivamente, abbiamo stabilito una soglia basata sulla distanza euclidea dei punti dal centroide del cluster al fine di distinguere tra pazienti ad alto e basso rischio di cancro.

In sintesi, questa strategia ci ha permesso di organizzare in modo più efficace lo studio del dataset, consentendo una migliore gestione e interpretazione dei dati relativi ai casi negativi al cancro.

Qui di seguito il grafico a torta aggiornato.

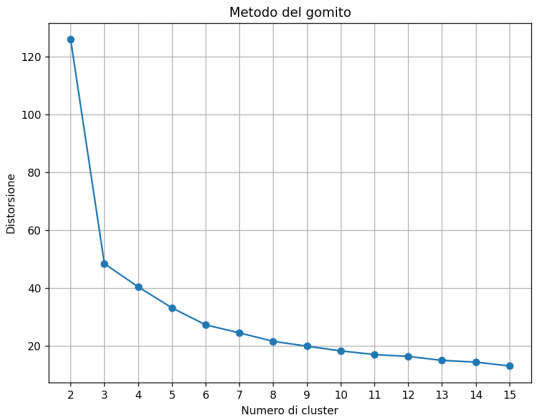


METODO DEL GOMITO E VALORE DI SILHOUETTE

Per garantire un processo decisionale ottimale nella suddivisione del 97.8% dei casi negativi, abbiamo adottato una strategia basata sul metodo del gomito e sul calcolo del valore di silhouette per ogni numero di cluster. Il metodo del gomito ci ha aiutato a individuare il numero ottimale di cluster, consentendoci di determinare un punto in cui l'aggiunta di ulteriori cluster non portava a miglioramenti significativi nella suddivisione dei dati. Inoltre, il valore di silhouette ci ha fornito un'indicazione sulla qualità della separazione dei cluster.

Integrando queste due tecniche, siamo stati in grado di prendere una decisione informata sulla suddivisione dei casi negativi, garantendo una suddivisione efficace e ben definita dei dati

Qui di seguito il grafico.



Il grafico del metodo del gomito ha indicato che il punto di svolta si verifica a 3 cluster. Questo ci conferma quindi che la suddivisione del nostro cluster in basso rischio e alto rischio è stata una scelta opportuna, il che è stato ulteriormente confermato dai valori di silhouette.

```
Con n_clusters=2, il valore di silhouette è 0.6948784895476883
Con n_clusters=3, il valore di silhouette è 0.7592594396069726
Con n_clusters=4, il valore di silhouette è 0.7643528907826049
Con n_clusters=5, il valore di silhouette è 0.40708975347148446
Con n_clusters=6, il valore di silhouette è 0.44855354684731347
Con n_clusters=7, il valore di silhouette è 0.37668913873663257
Con n_clusters=8, il valore di silhouette è 0.38029571441703314
Con n_clusters=9, il valore di silhouette è 0.3799215740434899
Con n_clusters=10, il valore di silhouette è 0.413220815953364
Con n_clusters=11, il valore di silhouette è 0.40099932553171536
Con n_clusters=12, il valore di silhouette è 0.38155606146751847
Con n_clusters=13, il valore di silhouette è 0.39501895081758154
Con n_clusters=14, il valore di silhouette è 0.40918046023409155
Con n_clusters=15, il valore di silhouette è 0.4017294753028466
```

Dopo aver esaminato i risultati ottenuti con il metodo del gomito e il valore di silhouette applicati al dataset originale e standardizzato, abbiamo constatato che non soddisfacevano i nostri criteri di qualità. Di conseguenza, abbiamo deciso di utilizzare il dataset normalizzato per l'analisi del clustering.

Qui di seguito riportiamo i valori di silhouette con dataset originale e standardizzato.

```
Con n_clusters=2, il valore di silhouette è 0.7111786782292054
Con n_clusters=3, il valore di silhouette è 0.43880435934243706
Con n_clusters=4, il valore di silhouette è 0.38210500655118923
Con n_clusters=5, il valore di silhouette è 0.3771880643825805
Con n_clusters=6, il valore di silhouette è 0.38259592564514977
Con n_clusters=7, il valore di silhouette è 0.3927435323717938
Con n_clusters=8, il valore di silhouette è 0.3194237085561624
Con n_clusters=9, il valore di silhouette è 0.3754597469569658
Con n_clusters=10, il valore di silhouette è 0.3609468978572043
Con n_clusters=11, il valore di silhouette è 0.3639182358491263
Con n_clusters=12, il valore di silhouette è 0.34745162106903615
Con n_clusters=13, il valore di silhouette è 0.33713253150954353
Con n_clusters=14, il valore di silhouette è 0.3375446394231825
Con n_clusters=15, il valore di silhouette è 0.34383698030270893
```

dataset originale

```
Con n_clusters=2, il valore di silhouette è 0.6646625327908515
Con n_clusters=3, il valore di silhouette è 0.5532408060150013
Con n_clusters=4, il valore di silhouette è 0.5541154195845184
Con n_clusters=5, il valore di silhouette è 0.30511339351330785
Con n_clusters=6, il valore di silhouette è 0.3882163940436535
Con n_clusters=7, il valore di silhouette è 0.3992657155650432
Con n_clusters=8, il valore di silhouette è 0.42153673761984406
Con n_clusters=9, il valore di silhouette è 0.39354883944117786
Con n_clusters=10, il valore di silhouette è 0.3853234030745797
Con n_clusters=11, il valore di silhouette è 0.36666886363264717
Con n_clusters=12, il valore di silhouette è 0.3892401667248759
Con n_clusters=13, il valore di silhouette è 0.3719263731879343
Con n_clusters=14, il valore di silhouette è 0.4026008029588794
Con n_clusters=15, il valore di silhouette è 0.36780153401506227
```

dataset standardizzato

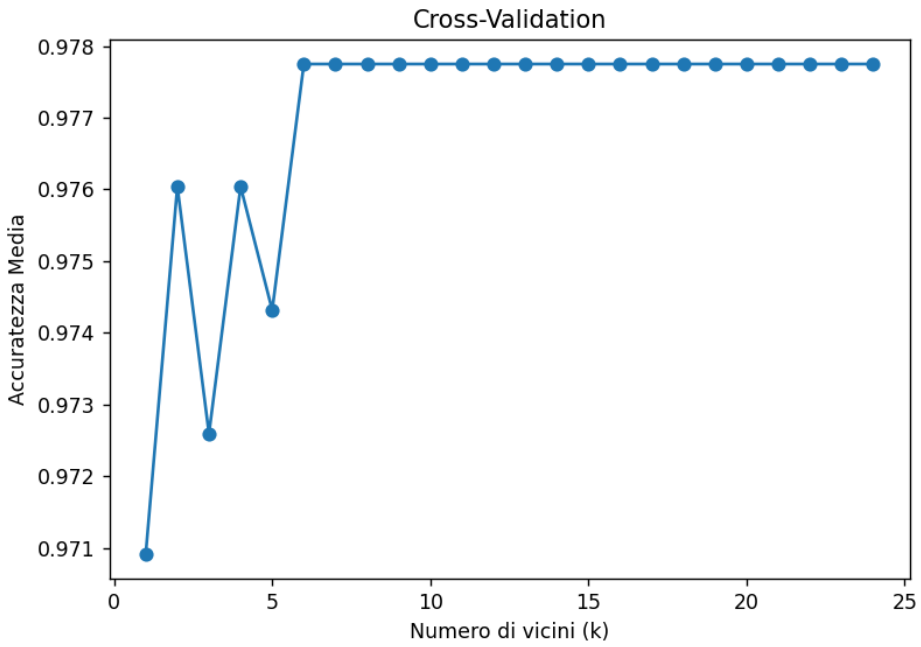
APPRENDIMENTO SUPERVISIONATO

KNN E CROSS VALIDATION

Nell'ambito della classificazione utilizzando il classificatore KNN (K-Nearest Neighbors), viene adottata la tecnica della cross-validation per determinare il valore ottimale del parametro 'k'.

A tal fine, è stato eseguito un esperimento che ha testato un range di valori per 'k', da 1 a 25. Questa scelta è stata guidata dalla considerazione del numero di colonne nel nostro dataset, che è pari a 24.
Nel corso dell'analisi dei risultati, è emerso che l'accuratezza del classificatore KNN aumenta significativamente quando il valore di 'k' supera il limite di 6 nel range testato da 1 a 25 e rimane costante. Questo fenomeno potrebbe indicare una migliore capacità del modello di generalizzare i dati e di ridurre l'overfitting.

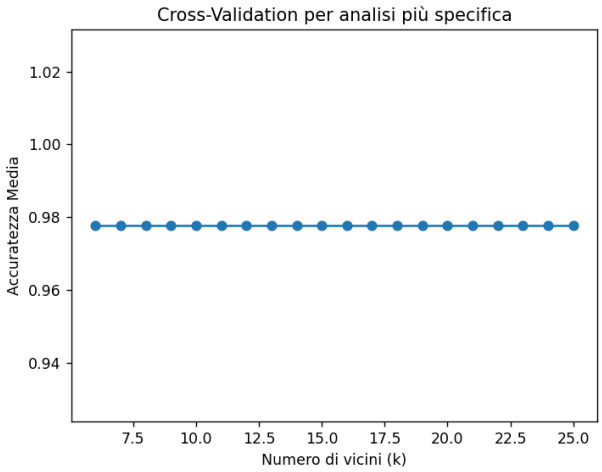
Qui di seguito il grafico.



Tuttavia, è fondamentale eseguire ulteriori analisi e valutazioni approfondite per accertare la coerenza di questo pattern. Queste analisi più dettagliate verranno esaminate in seguito nella documentazione.

Abbiamo eseguito un ulteriore controllo mediante cross-validation, limitando l'analisi ai valori di k compresi tra 6 e 24, al fine di valutare eventuali cambiamenti e individuare il valore ottimale di k.

Qui di seguito il grafico.



Dai risultati del grafico, abbiamo ottenuto ulteriore conferma che l'accuratezza media rimane costante per i valori di k compresi tra 6 e 24.

SMOTE E KNEIGHBOURS CLASSIFER

Davanti allo sbilanciamento dei dati nel dataset, abbiamo scelto di impiegare la tecnica SMOTE al fine di correggere questa discrepanza. Questo ci ha permesso di ottenere stime più precise dell'accuratezza, nonché della matrice di confusione e del report di classificazione.
Inoltre, abbiamo condotto test supplementari senza l'uso di SMOTE per valutare eventuali cambiamenti sostanziali con esito negativo.

Dopo aver confermato la stabilità nell'analisi del k ottimale, abbiamo proceduto calcolando l'accuratezza, la matrice di confusione e il report di classificazione per i primi tre valori di k (k=6, k=7 e k=8) e gli ultimi tre valori di k (k=22, k=23 e k=24). Attraverso la valutazione di precision, recall e f1-score, e l'identificazione degli iperparametri ottimali abbiamo individuato il valore di k=6 come il più ottimale per il nostro modello e la distanza di Manhattan pari a p=1.

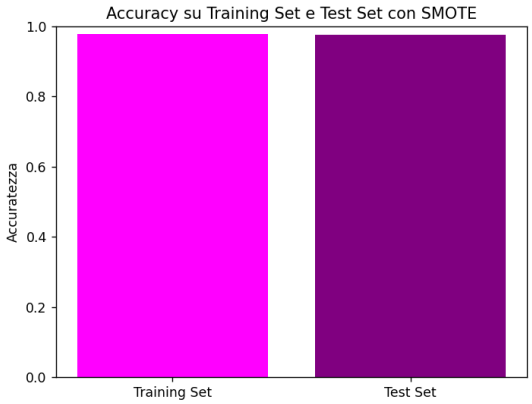
Qui di seguito i risultati ottenuti.

```
Accuratezza con SMOTE: 0.9761

Classification Report per K-Nearest Neighbors (k=6) con SMOTE:

              precision    recall  f1-score   support

   Cancro No       0.98      0.99      0.99       246
   Cancro Si       0.33      0.20      0.25         5

    accuracy                           0.98       251
   macro avg       0.66      0.60      0.62       251
weighted avg       0.97      0.98      0.97       251
```

Dai risultati ottenuti abbiamo ritenuto necessario adottare la tecnica SMOTE nonostante vi sia una leggera riduzione dell'accuratezza complessiva del modello, poiché, senza l'utilizzo di tale tecnica, abbiamo constatato che il bilanciamento del dataset risulta compromesso, portando a valori di precision, recall e F1-score nulli.

Questa situazione è dovuta alla netta prevalenza di una classe rispetto all'altra nel dataset. Pertanto, abbiamo deciso di utilizzare SMOTE al fine di correggere questa discrepanza e migliorare le prestazioni del modello.

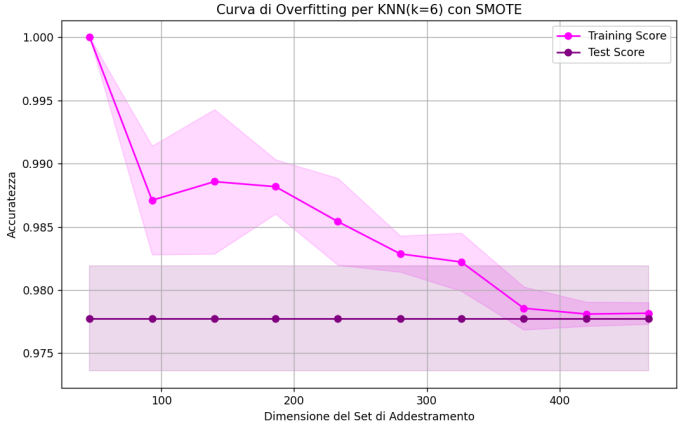
ACCURATEZZA SU TRAINING SET E TEST SET E CURVA DI OVERFITTING



Dall'analisi del grafico emerge che la barra fucsia, rappresentante il set di addestramento, ha un valore pari a 0.97, mentre la barra viola, corrispondente al set di test, mostra un valore di 0.95. Questo indica che il modello ha raggiunto ottime performance su entrambi gli insiemi, con una differenza minima di soli 0.02 punti percentuali tra di essi.

Se i valori di accuratezza sul set di test fossero stati bassi, questo avrebbe potuto indicare che il modello abbia imparato bene solo dai dati di addestramento e non sarebbe stato in grado di generalizzare bene su nuovi dati, suggerendo la presenza di overfitting.

Tuttavia, come possiamo osservare dal grafico a barre, l'accuratezza sul set di test non è bassa rispetto a quella sul set di addestramento, di conseguenza non sembra esserci un evidente caso di overfitting. Questo potrebbe indicare che il modello è in grado di generalizzare bene su dati non visti in precedenza, senza compromettere l'accuratezza sul set di addestramento.

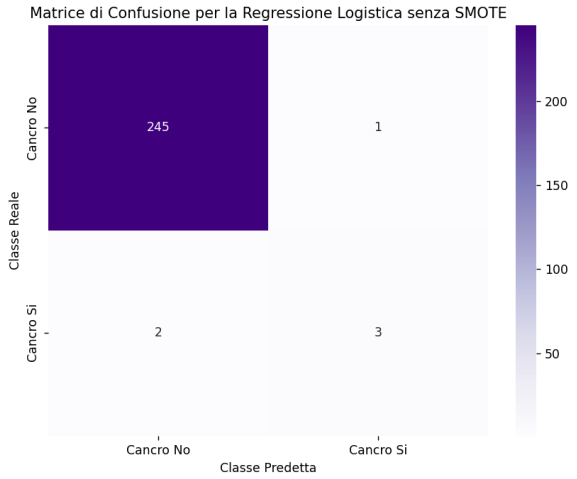


Il grafico sopra indicato conferma ciò che è stato detto in precedenza, ovvero che l'overfitting non sembra essere presente.

REGRESSIONE LOGISTICA

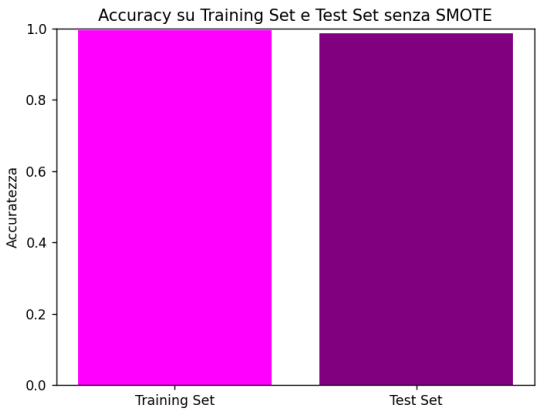
Per quanto riguarda la regressione logistica, abbiamo scelto di non utilizzare SMOTE poiché abbiamo dato maggiore peso alla prevenzione dell'overfitting. Questo perché l'overfitting potrebbe compromettere la capacità del modello di generalizzare su nuovi dati e di adattarsi a variazioni nei dati di addestramento.

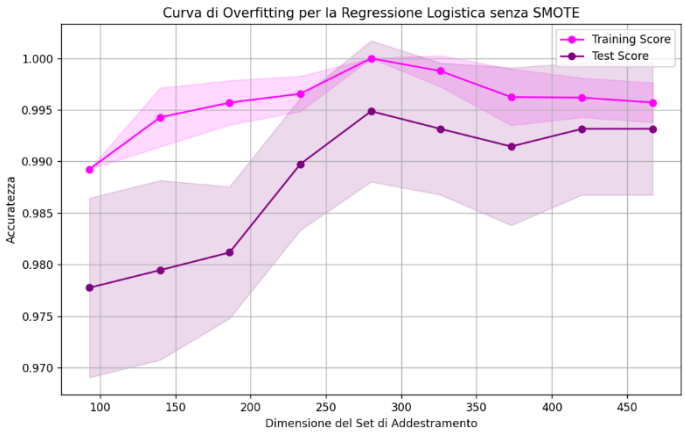
Abbiamo ritenuto prioritario preservare la capacità del modello di adattarsi a nuovi contesti e dati piuttosto che concentrarci sull'individuazione precisa dei casi rari, anche a costo di un aumento del tasso di falsi positivi.



Dall'analisi del report emerge che, senza l'utilizzo di SMOTE, si riscontra un valore di recall leggermente basso; tuttavia, nonostante questo risultato abbiamo scelto di confermare la decisione presa in precedenza.

```
Accuratezza senza SMOTE: 0.9880

Classification Report per la Regressione Logistica senza SMOTE:

              precision    recall  f1-score   support

   Cancro No       0.99      1.00      0.99       246
   Cancro Si       0.75      0.60      0.67         5

    accuracy                           0.99       251
   macro avg       0.87      0.80      0.83       251
weighted avg       0.99      0.99      0.99       251
```

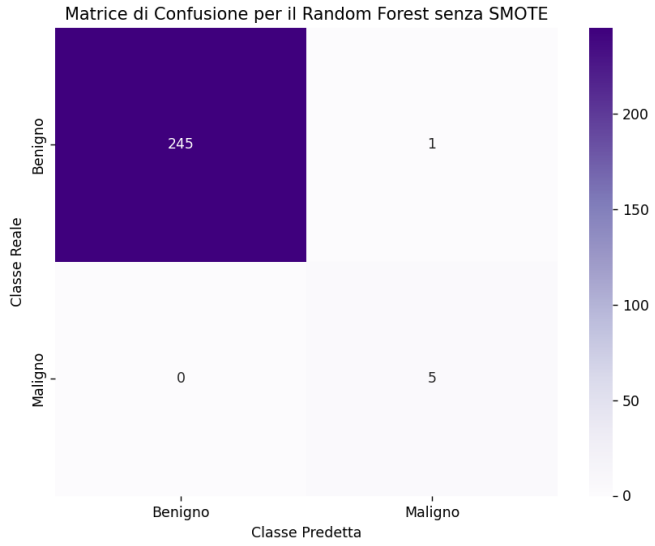




Anche osservando il grafico a barre e la curva di overfitting, si nota che la scelta di non utilizzare SMOTE permette di ottenere risultati ottimali e ciò è confermato dalla differenza tra l'accuratezza del set di addestramento (0.99) e quella del set di test (0.98) la quale è di soli 0.01 punti percentuali.

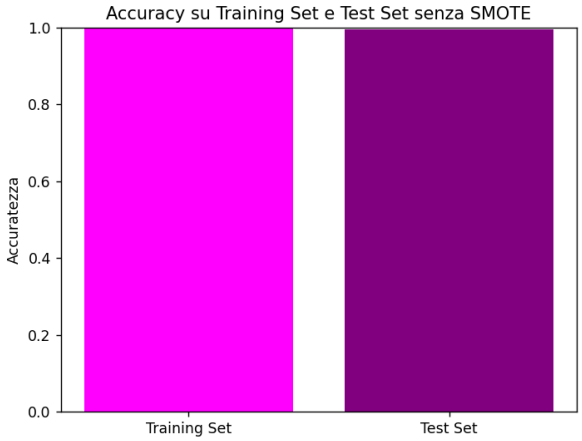
RANDOM FOREST

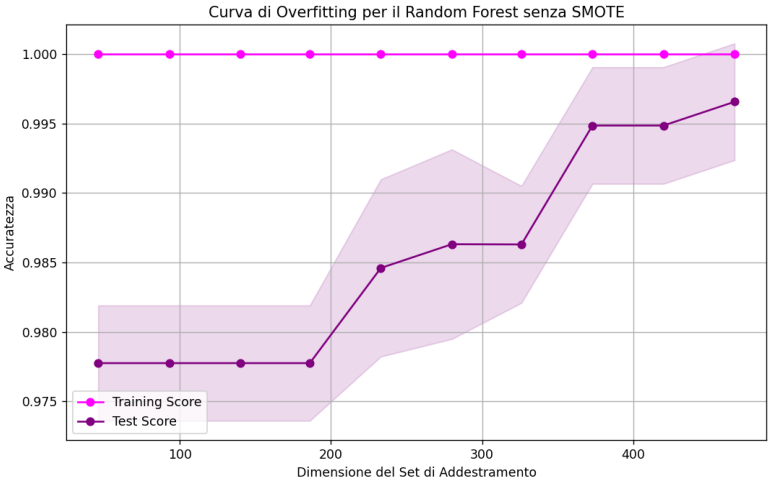
Per la tecnica del Random Forest, abbiamo seguito lo stesso approccio adottato per la regressione logistica, dove abbiamo dato la priorità alla gestione dell'overfitting. Di conseguenza, non abbiamo utilizzato la tecnica SMOTE.




```
Accuratezza senza SMOTE: 0.9960

Classification Report per il Random Forest senza SMOTE:

              precision    recall  f1-score   support

     Benigno       1.00      1.00      1.00       246
     Maligno       0.83      1.00      0.91         5

    accuracy                           1.00       251
   macro avg       0.92      1.00      0.95       251
weighted avg       1.00      1.00      1.00       251
```

L'accuratezza estremamente elevata del modello indica che l'applicazione di SMOTE potrebbe non essere cruciale in questo contesto. Questa conclusione è ulteriormente supportata dai valori del report, i quali suggeriscono che il modello ottiene prestazioni valide anche senza l'utilizzo di SMOTE.





Anche in questo caso, l'omissione di SMOTE ha portato a risultati eccellenti, come evidenziato dal grafico a barre e dalla curva di overfitting. La differenza minima tra l'accuratezza del set di addestramento (1.00) e quella del set di test (0.99), con soli 0.01 punti percentuali, conferma la solidità del modello senza l'aggiunta di dati sintetici tramite SMOTE.

RETE BAYESIANA

È stata implementata una rete bayesiana per determinare la probabilità che un individuo possa risultare positivo al cancro alla cervice, considerando tutte le caratteristiche specificate nel dataset del soggetto.

Raccolta dati:

È stato utilizzato un dataset contenente le caratteristiche di diversi soggetti, inclusi fattori come età, storia familiare di cancro alla cervice, risultati dei test di Pap smear, presenza di HPV (Human Papillomavirus), e altri attributi rilevanti.

Pre-processamento dei dati:

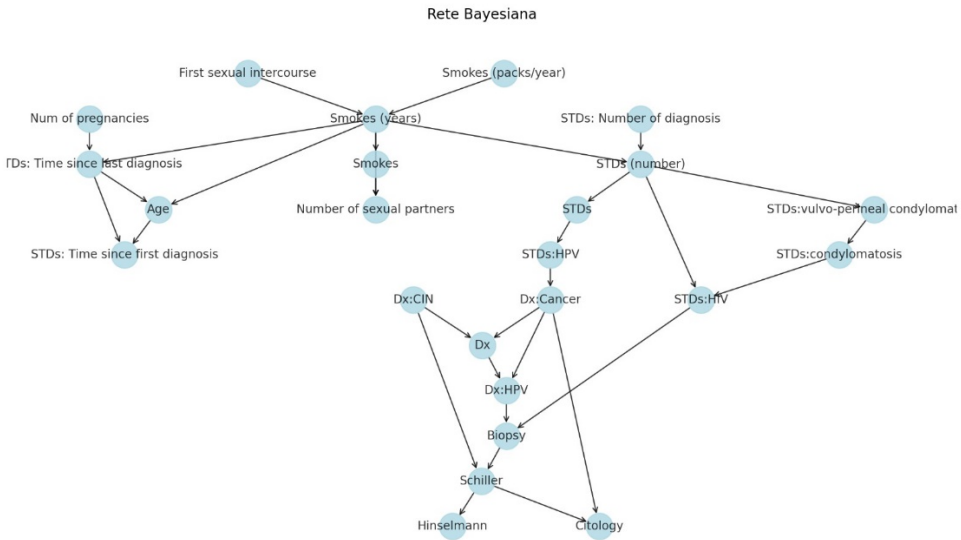
I dati sono stati esaminati per identificare valori mancanti o inconsistenti, che sono stati successivamente trattati tramite tecniche di imputazione o eliminazione delle righe/colonne.

```
Nodi della rete bayesiana:
Age
STDs: Time since first diagnosis
First sexual intercourse
Smokes (years)
Num of pregnancies
STDs: Time since last diagnosis
Smokes
Number of sexual partners
STDs (number)
Smokes (packs/year)
STDs
STDs:HPV
STDs:vulvo-perineal condylomatosis
STDs:HIV
STDs:condylomatosis
Biopsy
Dx:HPV
STDs: Number of diagnosis
Dx:Cancer
Dx
Citology
Dx:CIN
Schiller
Hinselmann
```

```
Archi nella rete bayesiana:
('Age', 'STDs: Time since first diagnosis')
('First sexual intercourse', 'Smokes (years)')
('Smokes (years)', 'Age')
('Smokes (years)', 'Smokes')
('Smokes (years)', 'STDs (number)')
('Smokes (years)', 'STDs: Time since last diagnosis')
('Smokes (years)', 'Number of sexual partners')
('Num of pregnancies', 'STDs: Time since last diagnosis')
('STDs: Time since last diagnosis', 'STDs: Time since first diagnosis')
('STDs: Time since last diagnosis', 'Age')
('Smokes', 'Number of sexual partners')
('STDs (number)', 'STDs')
('STDs (number)', 'STDs:vulvo-perineal condylomatosis')
('STDs (number)', 'STDs:HIV')
('Smokes (packs/year)', 'Smokes (years)')
('STDs', 'STDs:HPV')
('STDs:HPV', 'Dx:HPV')
('STDs:vulvo-perineal condylomatosis', 'STDs:condylomatosis')
('STDs:HIV', 'Biopsy')
('STDs:condylomatosis', 'STDs:HIV')
('Biopsy', 'Schiller')
('Biopsy', 'Dx:CIN')
('Dx:HPV', 'Dx:Cancer')
('STDs: Number of diagnosis', 'STDs (number)')
('Dx:Cancer', 'Dx')
('Dx:Cancer', 'Biopsy')
('Dx:Cancer', 'Citology')
('Dx:CIN', 'Dx')
('Dx:CIN', 'Schiller')
('Schiller', 'Hinselmann')
('Schiller', 'Citology')
```

Costruzione della rete bayesiana:

Utilizzando gli attributi raccolti, è stata costruita una rete bayesiana che modella le dipendenze probabilistiche tra di essi.



Stima delle probabilità condizionate:

Utilizzando il modello di rete bayesiana, è stata effettuata un'analisi inferenziale per stimare la probabilità che un individuo sia positivo al cancro alla cervice, data la sua combinazione di attributi.

```
Probabilità per una donna di non avere il cancro alla cervice:
+--------------+------------------+
| Dx:Cancer    |   phi(Dx:Cancer) |
+==============+==================+
| Dx:Cancer(0) |           0.9994 |
+--------------+------------------+
| Dx:Cancer(1) |           0.0006 |
+--------------+------------------+
```

```
Probabilità per una donna di avere il cancro alla cervice:
+--------------+------------------+
| Dx:Cancer    |   phi(Dx:Cancer) |
+==============+==================+
| Dx:Cancer(0) |           0.0002 |
+--------------+------------------+
| Dx:Cancer(1) |           0.9998 |
+--------------+------------------+
```

Valutazione del modello:

Sono stati valutati i parametri di accuratezza, precision, recall e f1 score della rete bayesiana. Di seguito i risultati ottenuti.

```
Valutazione della rete Bayesiana:
Accuracy: 0.9298245614035087
Precision: 0.9119257703081234
Recall: 0.909523809523806
F1-score: 0.903815569042117
```