Tumore cervicale: fattori che influenzano il rischio di cancro

Chiaretti Giulia[1], Fiorentini Federica[2], Maganza Riccardo[3], Monaco Alberto[4], Rubino Massimiliano[5]

Abstract

Nonostante il numero di nuove manifestazioni della malattia sia diminuito costantemente negli ultimi decenni, ogni anno negli Stati Uniti si registrano circa 11.000 nuovi casi di cancro cervicale invasivo. Sebbene sia il tipo di tumore più prevenibile, annualmente negli Stati Uniti perdono la vita a causa di esso 4.000 donne e il numero cresce fino a circa 300.000 se si considera il resto del mondo. I tassi di mortalità per il cancro della cervice uterina sono crollati del 74% dal 1955 al 1992, grazie all'aumento dello screening e della diagnosi precoce attraverso il test di Papanicolaou, o meglio conosciuto come Pap Test. I metodi di trattamento del tumore variano in base allo stadio della malattia al momento della diagnosi, ma tengono in considerazione anche altri fattori come lo stato di salute generale del paziente, l'età e lo stile di vita. Risulta essenziale, quindi, uno studio più ampio della malattia al fine di determinare le cause che maggiormente influenzano l'insorgenza di essa. [1]

L'analisi riportata in questo report ha l'obiettivo non solo di ampliare, per quanto possibile, la conoscenza su questa tipologia di cancro ma anche capire quali e quanto sono importanti nella prevenzione le abitudini dell'individuo. Lo scopo ultimo è quello di sensibilizzare la popolazione femminile ad adottare misure di prevenzione mirate per ridurre al minimo il rischio di contrarre la malattia.

Per riuscire nell'obiettivo di evidenziare i fattori che favoriscono lo sviluppo del cancro cervicale sono stati utilizzati degli algoritmi di machine learning e sono stati esposti, applicati e discussi dei modelli di classificazione.

Keywords:
Machine Learning — Cancro — Fattori di rischio — Classificazione

[1]*Università degli Studi di Milano Bicocca, CdLM Data Science*
[2]*Università degli Studi di Milano Bicocca, CdLM Data Science*
[3]*Università degli Studi di Milano Bicocca, CdLM Data Science*
[4]*Università degli Studi di Milano Bicocca, CdLM Data Science*
[5]*Università degli Studi di Milano Bicocca, CdLM CLAMSES*

Contents

INTRODUZIONE

Il cancro alla cervice uterina è stato classificato dal World Health Organization come la quarta tipologia di cancro più frequente nelle donne, che nell'ultimo anno ha portato alla comparsa di circa 570.000 nuovi casi. Il 90% delle morti dovute al cancro cervicale si verificano nei paesi in via di sviluppo, paesi in cui scarseggiano gli esami di screening, le pratiche di vaccinazione e le misure di prevenzione [2]. Nel mondo occidentale, invece, il numero dei decessi continua a diminuire soprattutto grazie all'introduzione del *Pap Test*, un esame di diagnosi precoce molto efficace.

Uno dei principali fattori di rischio per il tumore della cervice è l'infezione da papilloma virus umano (HPV).
Fortunatamente, però, è stato ormai accertato che solo alcuni degli oltre 100 tipi di HPV sono pericolosi dal punto di vista oncologico, mentre la maggior parte rimane silente o si limita a dare origine a piccoli tumori benigni detti papillomi. [3]

Anche se l'infezione da papilloma virus è la causa principale del cancro cervicale, lo scopo di questa indagine è capire, per quanto possibile, se la presenza di determinati fattori in combinazione con l'infezione aumenterà significativamente il rischio di sviluppare il tumore.
Per eseguire l'analisi di seguito, ci siamo basati su informazioni raccolte nel reparto oncologico dell'Ospedale Universitario di Caracas in Venezuela.
La relazione è suddivisa in quattro paragrafi:

- Nel primo paragrafo è riportata la descrizione del dataset e delle singole variabili;
- nel secondo paragrafo è stata effettuata un'analisi preliminare sul dataset che ha portato a scegliere le tecniche opportune di selezione delle variabili, missing replacement, discretizzazione e data preparation;
- nel terzo paragrafo sono descritti tutti i modelli di classificazione utilizzati;
- nel quarto paragrafo sono stati confrontati i modelli ottenuti, tramite il confronto delle misure di Precision e Recall e tramite l'utilizzo della ROC curve.

Infine, nella conclusione, si commentano i risultati ottenuti nell'analisi e si evidenziano gli aspetti più importanti emersi dall'indagine.

1. DATA INPUT

Si è scelto di utilizzare la versione 6 del dataset "*Cervical Cancer Risk Classification*", reso disponibile sulla piattaforma Kaggle [1]. Questo dataset contiene informazioni demografiche e mediche su 858 pazienti dell'Ospedale Universitario di Caracas, Venezuela. Il dataset contiene 36 variabili:

1. *Age*: l'età della paziente.
2. *Number of sexual partners*: il numero di partner con cui la paziente ha avuto uno o più rapporti sessuali.
3. *First sexual intercourse*: età in cui la paziente ha avuto il primo rapporto sessuale.
4. *Num of pregnancies*: numero di gravidanze della paziente.
5. *Smokes*: indica se la paziente è una fumatrice o meno.
6. *Smokes (years)*: indica da quanti anni la paziente è una fumatrice.
7. *Smokes (packs/years)*: per ogni paziente viene riportato il suo livello di pack-years, ovvero una misura clinica utilizzata per quantificare l'esposizione di una persona all'uso del tabacco nel corso del tempo. (1 pack-year equivale a fumare 20 sigarette (a pack) al giorno per un anno).
8. *Hormonal Contraceptives*: indica se la paziente utilizza contraccettivi ormonali o meno.
9. *Hormonal Contraceptives (years)*: indica da quanti anni la paziente utilizza contraccettivi ormonali;
10. *IUD*: indica se la paziente utilizza la spirale intrauterina (Intra Uterine Device) o meno.
11. *IUD (years)*: indica da quanti anni la paziente utilizza la spirale intrauterina (Intra Uterine Device).
12. *STDs*: indica se la paziente ha avuto malattie sessualmente trasmissibili (Sexually Transmitted Diseases) o meno. La variabile STDs è positiva se almeno una delle variabili della categoria "STDs:" è positiva.
13. *STDs (number)*: numero di malattie sessualmente trasmissibili (Sexually Transmitted Diseases) avute dalla paziente. Questa variabile è data dalla somma di tutte le variabili della categoria "STDs:".
14. *STDs:condylomatosis*: indica se la paziente è affetta da condilomatosi o meno.
15. *STDs:cervical condylomatosis*: indica se la paziente è stata affetta da condilomatosi cervicale.
16. *STDs:vaginal condylomatosis*: indica se la paziente è stata affetta da condilomatosi vaginale o meno.

17. *STDs:vulvo-perineal condylomatosis*: indica se la paziente è stata affetta da condilomatosi vulvo perineale o meno.
18. *STDs:syphilis*: indica se la paziente è stata affetta da sifilide o meno.
19. *STDs:pelvic inflammatory disease*: indica se la paziente è stata affetta da malattia infiammatoria pelvica o meno.
20. *STDs:genital herpes*: indica se la paziente è stata affetta da herpes genitale o meno.
21. *STDs:molluscum contagiosum*: indica se la paziente è stata affetta dall'infezione virale mollusco contagioso o meno.
22. *STDs:AIDS*: indica se la paziente ha contratto l'AIDS o meno.
23. *STDs:HIV*: indica se la paziente ha contratto l'HIV o meno.
24. *STDs:Hepatitis B*: indica se la paziente è stata affetta da epatite B o meno.
25. *STDs:HPV*: indica se la paziente è stata contagiata dal Papilloma virus o meno (Human Papilloma Virus).
26. *STDs: Number of diagnosis*: numero di volte a cui alla paziente è stata diagnosticata una malattia sessuale.
27. *STDs: Time since first diagnosis*: tempo intercorso dalla prima diagnosi.
28. *STDs: Time since last diagnosis*: tempo intercorso dall'ultima diagnosi.
29. *Dx:Cancer*: indica se la paziente ha già avuto il cancro o meno.
30. *Dx:CIN*: rivela se la paziente è stata affetta da neoplasia intraepiteliale cervicale (CIN) o meno.
31. *Dx:HPV*: segnala se la paziente aveva già contratto il Papilloma virus o meno (Human Papilloma Virus).
32. *Dx*: è una variabile binaria che è positiva se almeno una delle variabili della categoria "Dx: è positiva.
33. *Hinselmann*: indica se la colposcopia (esame ideato dal ginecologo tedesco Hans Hinselmann) ha rilevato o meno la presenza del tumore.
34. *Schiller*: suggerisce se il Test di Schiller ha rilevato o meno la presenza del tumore.
35. *Citology*: indica se l'esame citologico ha rilevato o meno la presenza del tumore.
36. *Biopsy*: è stata utilizzata come variabile risposta. Indica se l'esame istologico ha confermato o meno la presenta del tumore alla cervice uterina (Biopsy=1) o se si esclude la presenza del tumore (Biopsy=0).

Si pone l'attenzione sul fatto che questo dataset è caratterizzato da **classi sbilanciate**. Il numero di records appartenenti alla classe positiva (*Biopsy*=1), infatti, è pari a 55 su 858, ovvero il 6.4% del totale.

2. PREPROCESSING

Inizialmente è stato ritenuto necessario eliminare alcuni attributi. In particolare, sono state omesse diverse variabili riferite a malattie sessualmente trasmissibili. Le cause dell'eliminazione risiedono nella massiccia presenza di missing values rilevata in questa classe di attributi e nell'esistenza di rari valori positivi (la quasi totalità dei record presentava modalità pari a 0 che si traduce in un debole potere discriminante). Inoltre, tra le variabili eliminate rientrano anche quelle ritenute tipologie della variabile risposta.

Nello specifico, sono state omesse le variabili:

- *Hinselmann, Schiller e Citology*; eliminate in quanto tipologie della variabile risposta Biopsy.
- *STD: number of diagnosis*; eliminata perché somma di tutte le seguenti variabili STDs.
- *STDs: time since first diagnosis* e *STDs: time since last diagnosis*, rimosse per massiva quantità di missing (ogni qual volta che la variabile STD risultava essere pari a 0).
- *STDs: AIDS, STDs: condylomatosis, STDs: cervical condylomatosis, STDs: molluscum contagiosus, STDs: HPV, STDs: hepatites B, STDs: genital herpes, STDs: pelvic infiammatory disease, STDs: vaginal condylomatosis*; eliminate per scarsa presenza generale di valori diversi da 0, non utili a discriminare le osservazioni.
- *Dx*; rimossa perché somma delle restanti variabili della tipologia Dx.

Conclusa l'eliminazione, gli attributi binari rimanenti sono stati trasformati da numerici a stringhe.

2.2 Discretizzazione

La discretizzazione, in generale, è utile per ridurre la varianza delle stime e consentire l'utilizzo di algoritmi che non trattano, per loro natura, attributi continui.

Nello studio, è stata utilizzata una discretizzazione *unsupervised* che non tiene conto dei valori dell'attributo di classe (*Biopsy*, in questo caso).

È stato discretizzata la variabile *Age* in 5 classi, ognuna avente al proprio interno ugual numero di record (*Equal frequency unsupervised discretization*). In questo modo, anche il *missing replacement*, descritto nel paragrafo successivo, può essere effettuato con una maggior precisione.

2.3 Missing Replacement

Pur avendo operato un'eliminazione degli attributi con forte presenza di valori mancanti, 200 records risultavano ancora caratterizzati da almeno un *missing value*. Dato che l'eliminazione di suddetti records avrebbe comportato una perdita d'informazione troppo onerosa, è stato ritenuto doveroso procedere con una sostituzione di tali valori.

Si è scelto di effettuare un replacement degli attributi condizionandolo alla variabile *Age*. Questa decisione è dovuta ad un duplice motivo: molte delle variabili (numero di partner sessuali, gli anni da fumatrice, gli anni di utilizzo di contraccettivi ormonali etc.) sono presumibilmente legate all'età della paziente. L'altra motivazione, secondaria ma non trascurabile, è data dal fatto che l'attributo *Age* è privo di missing values; questo ha evitato l'insorgere di errori e problematiche nell'effettuare il replacement condizionato.

Quindi, i *missing values* riferiti ad attributi quantitativi sono stati sostituiti con la mediana dell'attributo in questione, calcolata all'interno della classe di età a cui apparteneva il record d'interesse.

Per quanto riguarda gli attributi qualitativi, invece, i valori mancanti sono stati rimpiazzati prendendo in considerazione la moda della variabile d'interesse nella classe di età di riferimento.

2.4 Data exploration

Per verificare la relazione tra variabili categoriche e la variabile target è stato applicato il test Chi-Quadro.

Tale test è stato svolto su tutte le variabili qualitative, ognuna di esse viene sottoposta al test in coppia con la variabile dipendente "Biopsy", in modo tale da verificarne l'ipotesi nulla d'indipendenza.

Si riportano in Figura 2.4.1 i risultati del test.

Variabile	Chi-Quadro	DF	P-value
Smoke	0.7079	1	0.04
IUD	3.01	1	0.08
Hormonal contraceptives	0.0152	1	0.9
Vulvo perineal condylomatosis	7.348	1	0.006
STDs	11.179	1	0.0008
STDs: syphillis	0.1054	1	0.75
STDs: HIV	13.99	1	0.0002
Dx:Cancer	22.21	1	2.44e^-6
Dx: CIN	10.98	1	0.0009
Dx: HPV	22.21	1	2.44^-6

Figura 2.4.1: Risultati del test Chi-Quadro.

Dalla tabella si evince che molti degli attributi influenzano la variabile risposta ad un livello di significatività del 95%.
Fanno eccezione IUD (che influenza, però, Biopsy a livello di significativà del 90%), Hormonal contraceptives e STDs: syphillis.

2.5 Partitioning and oversampling

Giunti a questo punto, sono state applicate due principali tecniche di partizionamento dei dati.

Nel primo metodo i dati disponibili sono stati suddivisi in due subset: il 67% dei record è stato attribuito alla partizione A (training set) ed i restanti 33% alla partizione B (test set). Nella divisione del dataset è stato utilizzato il campionamento stratificato affinché fosse mantenuta la proporzione delle classi della variabile dipendente (Biopsy = 0 e Biopsy = 1) nelle due partizioni.

Inoltre, è stata applicata una seconda tecnica di partizionamento durante la fase di apprendimento dei modelli: la *K-fold cross validation*, tecnica statistica che consiste nella divisione del dataset in *k* parti di uguale numerosità di cui si approfondisce nel paragrafo 4.1.

Per procedere ad uno sviluppo significativo dei modelli di classificazione, si è deciso di ricampionare adeguatamente il dataset, in quanto i dati a disposizione erano caratterizzati da un forte sbilanciamento delle classi della variabile dipendente. (Soltanto il 6.4% del totale, infatti, presentava modalità positiva dell'attributo Biopsy).

La tecnica scelta a tal scopo è stata l'oversampling, la quale consiste nel campionare casualmente (con reinserimento) dalla classe minoritaria un numero di osservazioni tale da pareggiare la cardinalità della classe maggioritaria.

Nello specifico, è stata utilizzata la tecnica SMOTE (Synthetic Minority Over-sampling Technique), che consiste nell'estrarre data point con reinserimento dalla minority class e nel creare nuove unità artificiali considerando i suoi K-nearest-neighbors, prendendo il vettore fra l'unità estratta e uno dei suoi K vicini e moltiplicandolo per un numero casuale fra 0 e 1. Nel caso in esame, il Parameter Optimization Loop ha ottenuto un valore ottimo di K pari a 3. La tecnica SMOTE permette di risolvere il problema dello sbilanciamento delle classi tramite oversampling senza però cadere nella trappola dell'overfitting che una semplice estrazione con reinserimento potrebbe portare. È bene considerare che questa estrazione casuale potrebbe causare problemi nel caso di variabili categoriali ed è stata creata una versione leggermente modificata dell'algoritmo chiamata SMOTE-NC per risolvere il problema. [5]
Al contrario, il ricorso alla tecnica dell'*undersampling* è stato valutato inadeguato. Infatti, la presenza di scarse osservazioni aventi modalità della variabile dipendente

positiva (soltanto 55) avrebbe poi reso inefficiente e poco significativo l'utilizzo di qualsiasi modello di classificazione.

3. DATA MODELING

Sono stati applicati modelli di classificazione appartenenti a quattro diverse categorie:

- **Modelli probabilistici:** modelli che, sotto ipotesi di indipendenza degli attributi, permettono di ottenere risultati accurati avvalendosi di concetti propri della logica bayesiana, quali le probabilità *a priori* e *a posteriori*. Fra di essi, è stato implementato il modello *Naïve Bayes*, che assegna al data point la classe $\hat{y}$ di output con la più grande probabilità *a posteriori*, $\hat{y} = \ argmax_{k=\{1,\dots,K\}} p(C_k)\ \Pi_i\ p(x_i \mid C_k)$.

- **Modelli euristici:** questi modelli, spesso, non possono fornire risultati ottimali ma sono in grado di ottenere approssimazioni ragionevoli senza richiedere sforzi computazionali eccessivi o ipotesi restrittive sui dati di input. Fra questi modelli, durante l'analisi sono stati utilizzati: il *Decision Tree*, nell'implementazione *J48* di *Weka*, il cui obiettivo è quello di separare le unità in base alla variabile risposta avvalendosi di alberi binari che suddividono in maniera binaria l'albero in base a dei valori di cutoff delle altre variabili e la sua naturale estensione, il *Random Forest*, che genera un certo numero di alberi decisionali e fornisce in output la *moda* delle previsioni dei singoli alberi. Inoltre, in quest'ultimo caso, per scegliere correttamente l'iperparametro relativo al numero di alberi da combinare, è stato utilizzato il *Parameter Optimization Loop* di KNIME, implementando un algoritmo di *grid search* per trovare il valore ottimo. [4]

- **Separation classifiers:** modelli che tentano di mappare i dati di input in uno spazio di dimensione superiore che separi meglio le variabili in base alla variabile risposta, avvalendosi delle cosiddette *kernel functions*. Nel lavoro, è stata esaminata una *Support Vector Machine*, con due diversi kernel: uno polinomiale, e uno *RBF* (*Radial Basis Function*). Il kernel polinomiale utilizzato è definito dalla funzione $K(x,y) = (x^T y)^2$, quello *RBF* dalla funzione $K(x,y) = exp\ (-\frac{||x-y||^2}{2\sigma^2})$. Quest'ultimo è stato alla fine giudicato come il migliore.

- **Modelli di regressione:** modelli che sotto ipotesi piuttosto stringenti sulle variabili in input, risultano molto flessibili e facilmente comprensibili in quanto è possibile misurare l'effetto delle diverse variabili nella classificazione a partire dai coefficienti assegnati dal modello. Sono stati utilizzati due differenti tipi di regressione logistica: uno implementato nel nodo standard di KNIME, e uno con l'implementazione di Weka, che si avvale inoltre di una logica di Boosting, combinando fra loro diversi classificatori deboli e pesando i risultati con opportuni pesi.
  La regressione logistica standard cerca dei coefficienti β tali che le probabilità stimate $p(x) = \frac{1}{1-\exp(-\beta x)}$ siano il meno lontane possibile dalla vera classe. Questa approssimazione è ottenuta tramite algoritmi numerici, e differenti algoritmi possono portare a risultati leggermente diversi.
  L'implementazione standard di KNIME ha ottenuto risultati migliori.

- **Reti Neurali:** modelli tanto potenti quanto *black-box*. Permettono di approssimare qualsiasi funzione tramite l'interconnessione di *neuroni artificiali*, creando un *percettrone.*
  Tuttavia, hanno il difetto di essere difficilmente interpretabili. Nel corso del lavoro, è stato utilizzato un *Multilayer Perceptron* standard.

4. MODELS EVALUATION

4.1 Cross Validation

Per ottenere una validazione dei classificatori utilizzati, è stato utilizzato un approccio basato sulla *Cross Validation*, il cosiddetto K-*fold cross validation*. Questa tecnica statistica suddivide il dataset in *K* partizioni di eguale numerosità e assicura che tutti i record vengano utilizzati almeno una volta sia nel *training test* che nel *test* set. Nel dividere il dataset in K partizioni di eguale ampiezza, si è scelto di utilizzare ancora una volta il campionamento stratificato al fine di mantenere la proporzione delle classi della variabile dipendente nei K subset. Ad ogni passo, K-1 partizioni entrano a far parte del training set, su cui l'algoritmo di classificazione viene allenato, mentre la rimanente k-esima partizione è utilizzata come test set, blocco di osservazioni di cui il modello predice il valore dell'attributo di classe.
Nel caso in esame per decidere il valore di K, che solitamente risulta essere 3, 5 o 10, è stato utilizzato il nodo *parameter optimization loop* per capire quale fosse il valore ottimale di iterazioni da effettuare.
Tuttavia, il processo è risultato computazionalmente troppo dispendioso quindi si è scelto di imputare il numero di iterazioni della *Cross Validation* a K=5.

In conclusione, adottare la K-*fold cross validation* è stato utile per evitare che le misure di accuratezza utilizzate poi per valutare i modelli di classificazione fossero influenzate dal processo di stima dei modelli stessi.

4.2 Measure comparison: Precision, Recall, F_1 Measure

Essendo in presenza di un dataset avente classi sbilanciate, l'*accuracy*, ovvero la capacità di effettuare previsioni corrette su nuovi records, non è significativa per misurare la "bontà" del modello poiché, in questo caso, si basa sulla cosiddetta *ZeroR rule,* secondo il quale il classificatore considera solamente i records del dataset appartenenti alla classe più frequente.

Le misure più adatte per analizzare la capacità predittiva del modello, quindi, risultano essere la *precision* e la *recall.* In particolare, la *precision* descrive la frazione di record effettivamente positivi tra tutti quelli predetti come tali dal modello e, quindi, più è elevata, minore è il la quantità di osservazioni erroneamente classificate come positive.

$$\boldsymbol{precision} = \frac{TP}{TP + FP}$$

La *recall*, invece, rappresenta la porzione di record positivi correttamente classificati dal modello rispetto al totale di osservazioni realmente positive.

Premesso ciò, un valore di *recall* alto indica una bassa percentuale di record positivi classificati in modo errato.

$$\boldsymbol{recall} = \frac{TP}{TP + FN}$$

In molti casi, le due misure possono essere contraddicenti poiché, aumentando i veri positivi (TP) della classe più rara, si determina un miglioramento della *recall* ed un conseguente peggioramento della *precision* perché possono aumentare i falsi positivi (FP).

A tal proposito, per evitare incongruenze, diventa opportuno considerare le due metriche simultaneamente.

In particolare, si può utilizzare la F_1 *Measure* che riassume *precision* e *recall,* e si calcola tramite la media armonica delle due:

$$\boldsymbol{F_1} = \frac{2 * r * p}{r + p}$$

dove *r* indica la *recall* e p indica la *precision.* Un alto valore di F_1 implica un valore ragionevolmente alto per entrambe le metriche.

Per individuare il classificatore migliore, quindi, è stato effettuato un confronto tra le tre quantità sopra descritte, prendendo in esame i risultati relativi ad ogni modello stimato.

Le misure sovracitate sono riassunte nella tabella in Figura 4.2.1.

Classifiers	Recall	Precision	F-Measure
Logistic Regression	0.1	0.278	0.147
Simple Logistic	0.053	0.222	0.086
Random Forest	0.058	0.556	0.086
Decision Tree	0.075	0.389	0.126
Naive-Bayes	0.128	0.333	0.185
Support Vector Machine	0.163	0.389	0.151
Multilayer Perceptron	0.114	0.222	0.151

Figura 4.2.1: Recall, Precision e F-Measure dei modelli di classificazione.

Come si può notare anche dal *Boxplot* in Figura 4.2.2, osservando i valori della *precision,* il risultato migliore è ottenuto dal classificatore *Random Forest,* con un valore pari a 0.56, seguito dalla *Support Vector Machine, con* una precision pari a 0.39.

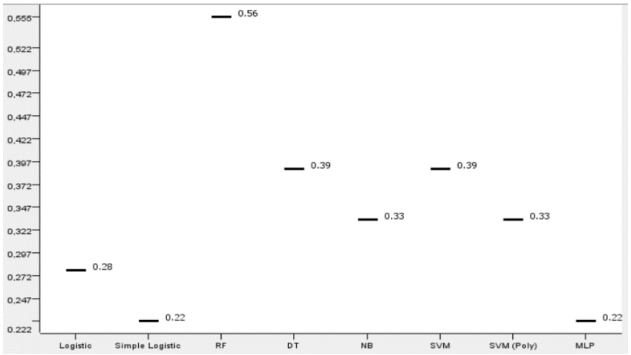

Figura 4.2.2: Boxplot con i valori della precision dei modelli di classificazione.

Successivamente, sono stati analizzati i valori della F_1 *Measure* che, come detto in precedenza, tiene conto sia della *precision* che della *recall* calcolate sul modello.

Osservando la F_1 *Measure,* si nota che il classificatore migliore è il *Support Vector Machine*, con una quantità pari a 0.23, coerente con i risultati emersi precedentemente.

Per quanto riguarda il *Random Forest*, la F_1 *Measure* è molto bassa, sintomo di una pessima misura di *recall*.

Secondo la F_1 *Measure,* perciò, i classificatori relativamente più precisi risultano essere:

- Support Vector Machine (0.23)
- Naive-Bayes (0.185)

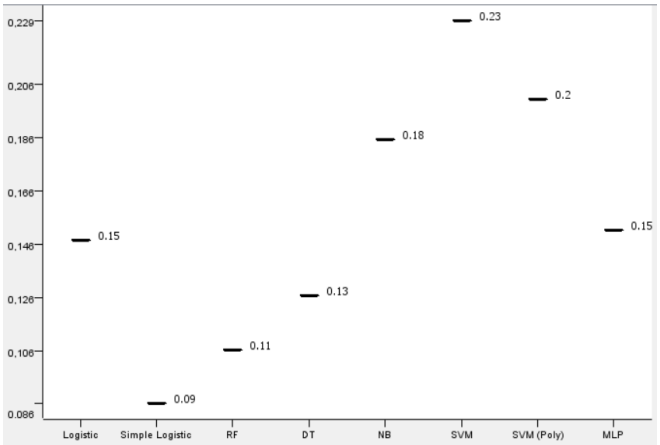

Figura 4.2.3: Boxplot con i valori della F-measure dei modelli di classificazione.

Come descritto nel paragrafo successivo, per effettuare un'analisi grafica più approfondita, sono state confrontate le *ROC curve* dei tre modelli migliori selezionati secondo le misure di "bontà".

4.3 ROC Curve

Una tecnica grafica molto utilizzata per confrontare i modelli è la curva di *ROC*, *receiver operating characteristic.* Essa viene rappresentata considerando tutti i possibili valori del test e, per ognuno di questi, viene calcolata la proporzione di veri positivi e quella di falsi positivi. Congiungendo i punti che mettono in rapporto le due proporzioni, si ottiene la curva e, in particolare, l'area sottostante ad essa indica una misura di accuratezza del modello, l'*AUC* (A*rea Under the Curve*). Nel caso di un classificatore perfetto, l'area sottostante la curva di *ROC* è pari a 1 mentre, nel caso di un modello avente basse capacità predittive, è inferiore a 0.5.

È stata, quindi, rappresentata la curva di ROC per tre diversi modelli:

- *Random Forest* poiché avente una *precision* molto alta;
- *Support Vector Machine*, poiché avente una F_1 *Measure* superiore agli altri;
- *Logistic Regression* perché possiede una F_1 *Measure* accettabile ed è utile a fini interpretativi del modello.

Osservando il grafico in *Figura 4.3.1*, si nota che la *Random Forest* ha un *AUC* molto basso (0.4751), concorde ad una F_1 *Measure* estremamente bassa e, quindi, viene scartata. Ancora una volta, il classificatore migliore, invece, risulta essere il *Support Vector Machine*, avente un *AUC* pari a 0.63. Inoltre, viene tenuto in considerazione anche il modello logistico poiché, oltre ad avere un *AUC* relativamente alto rispetto agli altri, permette anche una facile interpretazione dei parametri del modello al fine di individuare le principali cause tumorali, come descritto nella parte seguente.

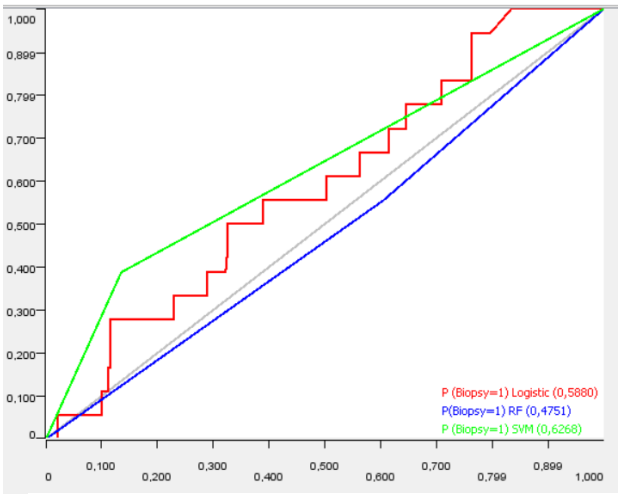

Figura 4.3.1: Curva ROC.

5. CONCLUSIONI

L'obiettivo di questa analisi è quello di mettere in luce l'importanza delle tecniche di machine learning sullo studio di fenomeni di natura medica. Nello specifico, si sono affrontate le problematiche relative alla classificazione di fenomeni sbilanciati, poiché maggiormente comuni nello studio delle malattie rare.

L' obiettivo del lavoro non mirava semplicemente alla soluzione di un problema di classificazione, ma a fornire le cause che maggiormente influenzano il rischio di avere il tumore. Per raggiungere lo scopo prefissato, è stato utilizzato il modello logistico in quanto permette un'ottima interpretazione dei parametri.

Dall'immagine in Figura 5.1, si possono notare i fattori che maggiormente influenzano il rischio di cancro. Focalizzandoci solo sulle principali 5, sottolineiamo il contributo di:

- **CIN**: a parità di tutte le restanti esplicative, il rischio di contrarre il tumore per soggetti a cui è stata diagnosticata la neoplasia o CIN, è 4 volte più alto rispetto all'assenza di essa.

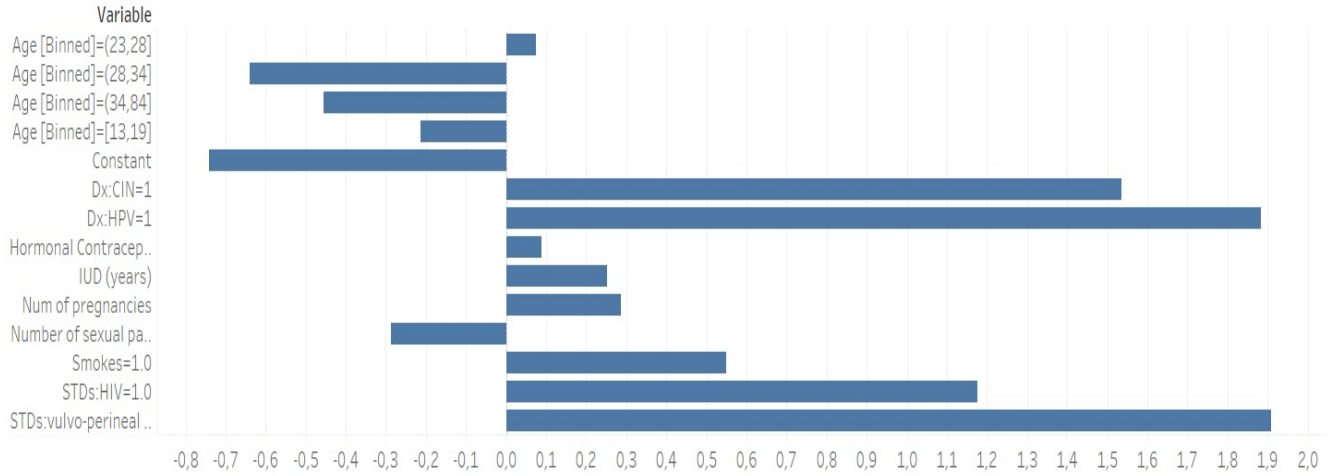

Figura 5.1: Stima dei coefficienti del modello logistico, Evento: Biopsy=1

- **HPV**: noto anche come *Papilloma Virus,* sembrerebbe la malattia che più di ogni altra aumenta la propensione al tumore. Infatti, a parità di tutte le variabili la presenza del virus, implica un aumento dell'attitudine al cancro più di 6 volte rispetto a donne che non sono state affette da HPV.
- **Smokes**: alle donne fumatrici è associato un rischio più alto di cambiamenti precancerosi della cervice rispetto alle non fumatrici. Infatti, l'attitudine alla manifestazione del cancro risulta essere 1.7 volte più alta.
- **HIV:** le donne che manifestano un maggior rischio di infezione, così come *un'insufficienza immunitaria* che può essere legata a diverse cause (per esempio un'infezione da HIV), hanno una propensione al cancro 3 volte più alta.
- **VULVO-PERINEAL:** la presenza di malattie sessualmente trasmissibili, tra cui la *condilomatosi genitale*, mostra una propensione al tumore maggiore rispetto alle donne che non hanno riscontrato l'infezione.

Grazie al modello logistico sono state individuate le variabili che possono influenzare in maniera significativa la possibilità di contrarre il tumore. Campagne di sensibilizzazione posso sfruttare le informazioni ottenute, cercando di aumentare la diffusione della vaccinazione da HPV, poiché responsabile della maggior parte dei tumori della cervice.

Quest'indagine ha considerato come variabile risposta la biopsia del tumore, ciò nonostante sviluppi futuri del problema potrebbero essere effettuati considerando altri tipi di test per il riconoscimento del cancro quali: Hinselmann, Schiller, o di tipo citologico. Un ulteriore sviluppo successivo del problema potrebbe essere focalizzare l'attenzione solo su donne che hanno un'alta attività sessuale o riuscire a collezionare dati sullo stato di salute famigliare, al fine di identificare una possibile causa ereditaria della malattia.

RIFERIMENTI

[1] Kaggle: https://www.kaggle.com/loveall/cervical-cancer-risk-classification

[2] WHO World Health Organization: https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/

[3] AIRC: https://www.airc.it/cancro/informazioni-tumori/guida-ai-tumori/tumore-alla-cervice-uterina

[4] https://stackabuse.com/cross-validation-and-grid-search-for-model-selection-in-python/

[5]https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a-html/chawla2002.html