# Cervical Cancer Risk Analysis

Ilaria Marini[1], Davide Meloni[1], Alberto Raimondi[1], Camilla Scuffi[1]

**Abstract**
Il cancro alla cervice è una forma tumorale che si sviluppa abbastanza lentamente e nelle sue prime fasi è in genere asintomatica, quindi di difficile identificazione se non mediante un esame citologico della cervice, che consente di riconoscere precocemente le eventuali cellule anomale e di poterle trattare efficacemente. La sintomaticità di questa patologia, individuata da sanguinamento vaginale e dolore, avviene solo in fase avanzata di carcinoma pertanto i trattamenti, a questo stadio, possono risultare inutili. Tuttavia molte persone evitano di sottoporsi a screening a causa dell'elevato costo dei test diagnostici.
Quindi utilizzando degli algoritmi di classificazione supervisionata si vuole stabilire se l'insorgenza del cancro alla cervice possa essere predetta a partire da informazioni su alcune caratteristiche demografiche, abitudini personali e dati medici di una determinata persona e ci si propone di valutare l'efficacia dei test diagnostici nell'individuazione della patologia.
Nel report vengono mostrati i risultati di tutte le possibili combinazioni che si ottengono confrontando otto diversi algoritmi su cinque differenti configurazioni del dataset. Dopo tale comparazione e l'analisi delle principali misure di affidabilità della predizione, viene approfondito l'algoritmo migliore: Adaptive Boosting (AdaBoost), ottenuto con il filtro ReliefF e la metodologia SMOTE.

[1] *Dipartimento di Informatica Sistemistica e Comunicazione, Università degli studi Milano-Bicocca, Milan, Italy*

## Indice



## Introduzione

Il cancro alla cervice è una patologia dell'apparato genito-riproduttivo femminile che colpisce la parte inferiore dell'utero, detta "collo" o "cervice".

Fino a metà del '900 il cancro alla cervice era tra le prime cause di morte per cancro tra le donne, successivamente, grazie allo sviluppo del Pap test (esame citologico in grado di individuare alterazioni delle cellule della cervice) i decessi sono significativamente diminuiti. Ad oggi negli Stati Uniti questo tumore è la terza neoplasia ginecologica più frequente e l'ottava neoplasia più diffusa tra le donne. L'American Cancer Society stima che nel 2018 negli Stati Uniti si verificheranno 13240 nuovi casi di cancro alla cervice e si conteranno 4170 decessi.

Il cancro della cervice è un carcinoma generalmente causato dall'infezione da papillomavirus umano (Bosch and de Sanjose, 2007); raramente, si tratta di un adenocarcinoma. Dalla sua natura infettiva ne deriva la rilevanza della somministrazione del vaccino anti HPV in ottica di prevenzione. Tale misura preventiva è già in adozione nelle nazioni Europee, Australia, Stati Uniti e negli altri paesi sviluppati, mentre sta vedendo una diffusione più lenta nei paesi in via di sviluppo.

L'HPV si trasmette per via sessuale, pertanto i principali fattori di rischio associati al cancro alla cervice sono primo rapporto in giovane età ed un elevato numero di partners sessuali. La diffusione di questo tipo di cancro è infatti maggiore nei paesi del terzo mondo, in particolare Africa e Sudamerica, dove l'utilizzo del preservativo come difesa dalle malattie sessualmente trasmissibili è relativamente minore.

La displasia cervicale è solitamente asintomatica, la paziente che non si sia sottoposta ad esami mirati inizia ad allarmarsi solo a seguito di un sanguinamento vaginale post-coitale, che avviene però solo ad uno stadio avanzato della malattia.

Il nostro studio si propone di ottenere un modello di classificazione che consenta di stabilire se una paziente sia portata o meno a sviluppare il cancro della cervice. A tal proposito si è utilizzato un dataset contenente dati raccolti presso un ospedale venezuelano e utilizzati per uno studio della patologia (Fernandes et al., 2007).

## 1. Data Understanding

### 1.1 Descrizione delle variabili

Il dataset è stato sviluppato dall'Ospedale Universitario di Caceres e diffuso da Kaggle, esso include dati su 858 pazienti. In particolare, è costituito da:

- informazioni personali sulle pazienti: Età, Numero di partner sessuali, Età del primo rapporto sessuale e Numero di gravidanze;
- abitudini personali: Fumare, Utilizzare contraccettivi (ormonali e/o IUD);
- informazioni mediche: Malattie sessualmente trasmissibili, Diagnosi;
- test diagnostici: Test di Hinselmann, Test di Schiller, Citologia e Biopsia.

Per meglio comprendere il significato delle variabili, si fornisce una spiegazione generale dei diversi test diagnostici utilizzati nel dataset.

Il test di Hinselmann (o colposcopia) è un esame che consente la visione ingrandita della cervice uterina, permettendo quindi una rilevazione di eventuali anomalie: lesioni, alterazioni o neoplasie.

Il test di Schiller è un esame che ha lo scopo di individuare aree della cervice uterina con delle anomalie attraverso l'applicazione di una soluzione di iodio.

L'esame citologico consiste nel prelievo di una piccola quantità di cellule del collo dell'utero con un tampone cervicale e nell'analisi di tali tessuti al microscopio.

La biopsia è un esame istologico che ha lo scopo di confermare i sospetti di malignità del carcinoma a cui possono condurre i test citologici, di Hinsellman e di Schiller. Si tratta di un esame piuttosto invasivo che comporta l'asportazione di uno o più frammenti di tessuto attraverso un piccolo intervento chirurgico.

### 1.2 Esplorazione del dataset

I grafici che seguono illustrano la percentuale di malate di cancro in alcune categorie di pazienti. Si nota che le fumatrici ammalate sono relativamente di più rispetto alle non fumatrici. Il cancro inoltre è relativamente più frequente per chi utilizza la spirale contraccettiva. In linea con quanto la ricerca medica afferma il cancro è più frequente per chi ha avuto un maggior numero di gravidanze. Per quanto riguarda invece il fattore di rischio età del primo rapporto, i risultati del dataset sembrano parzialmente smentire quanto sostenuto dalla medicina.

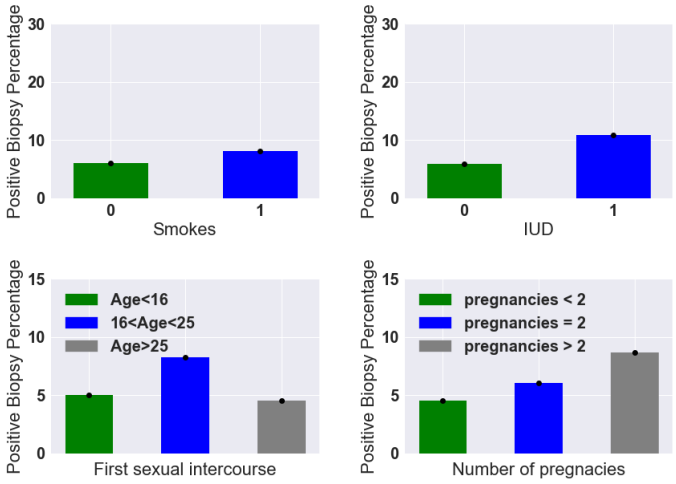



**Figura 1.** Grafici di analisi descrittiva mostranti l'incidenza del cancro in diverse categorie di pazienti.

### 1.3 Scopo dell'elaborato

L'obiettivo di questo lavoro è quello di determinare un modello di classificazione per poter stabilire se una paziente sia affetta dal cancro alla cervice.

Per quanto appreso sui test diagnostici, la scelta della variabile target è ricaduta su "Biopsy", in quanto è l'esame che con maggior grado di affidabilità può escludere o confermare una diagnosi di tumore alla cervice. Inoltre la scelta è motivata dal fatto che, a differenza dei test di Hinselmann e Schiller e dell'esame citologico, la biopsia risulta essere un intervento particolarmente invasivo. Pertanto si è ritenuto interessante porsi la domanda sulla possibilità di individuare un modello di classificazione che stabilisca se una persona sia affetta da cancro alla cervice, avendo a disposizione solo informazioni personali, storia medica e risultati dei test diagnostici meno invasivi di una paziente.

## 2. Data Preprocessing

Prima di poter utilizzare efficacemente le tecniche di Machine Learning per estrarre conoscenza dal dataset, è necessario affrontare alcune problematiche che si sono presentate nel corso dell'esplorazione dei dati e che non possono essere sorvolate. Nella fattispecie, verranno affrontate le seguenti procedure:

1. trattamento dei valori mancanti;
2. bilanciamento del dataset;
3. selezione degli attributi rilevanti.

È poi necessario precisare che durante l'esplorazione iniziale del dataset si è osservato che a due pazienti era associata un'età inferiore a quella del primo rapporto sessuale, pertanto sono stati eliminati i record corrispondenti. Inoltre, si sono

eliminate le colonne ”STDs:AIDS” e ”STDs: cervical condylomatosis” poiché esse presentavano il solo valore nullo per tutti i record contenuti nel dataset.

## 2.1 Missing Value Handling

L’esplorazione del dataset ha portato immediatamente alla luce la presenza di numerosi valori mancanti, come riportato in Tab.1.

**Missing Values**

| Attribute | Attribute Type | No. Missing Values | Procedura adottata |
|---|---|---|---|
| No. sexual partners | ratio | 26 | Mediana |
| First sexual intercourse | ratio | 7 | Mediana |
| No. pregnancies | ratio | 56 | Mediana |
| Smokes | binary | 13 | Moda: 0 |
| Smokes (years) | ratio | 13 | Valore costante: 0 |
| Smokes (packs/year) | ratio | 13 | Valore costante: 0 |
| Hormonal Contraceptives | binary | 108 | Moda: 1 |
| Hormonal Contraceptives (years) | ratio | 108 | Mediana |
| IUD | binary | 108 | Moda: 0 |
| IUD (years) | ratio | 105 | Valore costante: 0 |
| STDs | binary | 105 | Valore costante: 0 |
| STDs (number) | ratio | 105 | Valore costante: 0 |
| STDs:condylomatosis | binary | 105 | Valore costante: 0 |
| STDs:cervical condylomatosis | binary | 105 | Valore costante: 0 |
| STDs:vaginal condylomatosis | binary | 105 | Valore costante: 0 |
| STDs:vulvo-perineal condylomatosis | binary | 105 | Valore costante: 0 |
| STDs:syphilis | binary | 105 | Valore costante: 0 |
| STDs:pelvic inflammatory disease | binary | 105 | Valore costante: 0 |
| STDs:genital herpes | binary | 105 | Valore costante: 0 |
| STDs:molluscum contagiosum | binary | 105 | Valore costante: 0 |
| STDs:HIV | binary | 105 | Valore costante: 0 |
| STDs:HPV | binary | 105 | Valore costante: 0 |
| STDs: Time since first diagnosis | ratio | 787 | Esclusione attributo |
| STDs: Time since last diagnosis | ratio | 787 | Esclusione attributo |

**Tabella 1.** La tabella sopra riportata mostra gli attributi del dataset che contengono valori mancanti, il loro tipo, il numero di missing values e la procedura adottata per il trattamento di quest’ultimi.

Il trattamento dei valori mancanti è stato affrontato in modo specifico per ogni attributo.

Per gli attributi quantitativi discreti Numero di partner sessuali, Età relativa al primo rapporto sessuale e Numero di gravidanze si è deciso di sostituire i dati mancanti con il valore mediano per ogni variabile: la sostituzione con la media avrebbe risentito della presenza di outliers individuabili attraverso l’osservazione dei boxplots delle singole distribuzioni.

Per quanto riguarda invece le variabili binarie si è optato per il rimpiazzo tramite la moda, che maggiormente si addice a variabili categoriche. Poiché il valore della moda dell’attributo ”Hormonal Contraceptives” è risultato essere 1 si è reso necessario sostituire i valori mancanti con la mediana anche per l’attributo continuo ”Hormonal Contraceptives (years)”. I missing values sugli attributi ”Smokes (years)” e ”STDs (number)” sono invece stati trattati inserendo il valore fisso 0 poiché gli attributi binari associati ”Smokes”e ”STDs” sono risultati avere moda 0.

Gli attributi ”STDs: Time since first/last diagnosis” sono presenti nel dataset solo per i record per i quali ”STDs: No. Diagnosis” è diverso da zero; di conseguenza vi sono valori mancanti per la maggior parte delle righe. Si è quindi optato per l’eliminazione dell’attributo in questione in quanto la sostituzione di tutti i missing values con il valore fisso 0 sarebbe stata privo di significato ed avrebbe condotto ad un forte bias, oltre che ad una variabilità veramente ridotta.

## 2.2 Data Balancing

Come spesso accade per i dati in campo medico il dataset di partenza presenta un forte sbilanciamento delle frequenze tra le due classi: la modalità associata ad esito negativo della biopsia è infatti molto più numerosa, nella fattispecie sono solo 55 su 858 i pazienti con esito positivo all’esame della biopsia (circa il 6.4% dei record nel dataset). Ciò costituisce un problema per l’applicazione di algoritmi di machine learning perché essi assumono che le classi siano bilanciate ed in caso contrario producono risultati distorti in quanto predicono un eccessivo numero di casi abbinati alla classe più frequente. É necessario quindi bilanciare il numero di record nel dataset di train in funzione della variabile target ”Biopsy”. A tal scopo si è utilizzata la procedura SMOTE (Synthetic Minority Over-sampling Technique), nodo nativo del software KNIME©. Come esposto da Bowyer et al. (2011), SMOTE è una tecnica di *over-sampling* della classe in minoranza del dataset. Invece di replicare dati già presenti, questo metodo crea dei nuovi record ”sintetici”. L’algoritmo analizza diversi campioni della classe sbilanciata e computa un massimo di 5 nearest neighbors per introdurre nuovi record posizionati nel segmento che collega fra di loro due nearest neighbors. Ovviamente la procedura SMOTE è applicabile solo nel training set, dove la sua funzione è particolarmente utile per i motivi appena esposti; al contrario, applicare la SMOTE all’intero dataset porterebbe a spiacevoli episodi di over-fitting, andando a creare strutture fittizie non presenti nel dataset iniziale e che quindi non sarebbero più rappresentative delle dinamiche reali che si vogliono analizzare.

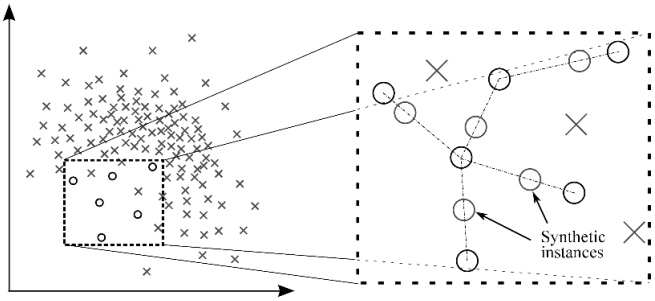



**Figura 2.** In figura è rappresentata graficamente la logica su cui si basa la procedura SMOTE: si selezionano uno a uno i record relativi alla classe minoritaria presenti nel dataset di train, si trovano quelli più vicini nello spazio degli attributi e si introducono dei dati artefatti lungo le rette che collegano due record vicini.

## 2.3 Feature Selection

La selezione degli attributi è una fase cruciale del flusso di Machine Learning, infatti si è potuto verificare direttamente come una scelta oculata degli attributi rilevanti ai fini della classificazione influisca in modo significativo nelle predizioni della variabile target.

É possibile avere una prima idea di quali attributi siano importanti nell'inferenza della variabile target "Biopsy" computando la matrice di correlazione fra gli attributi, mostrata graficamente in Fig.3.

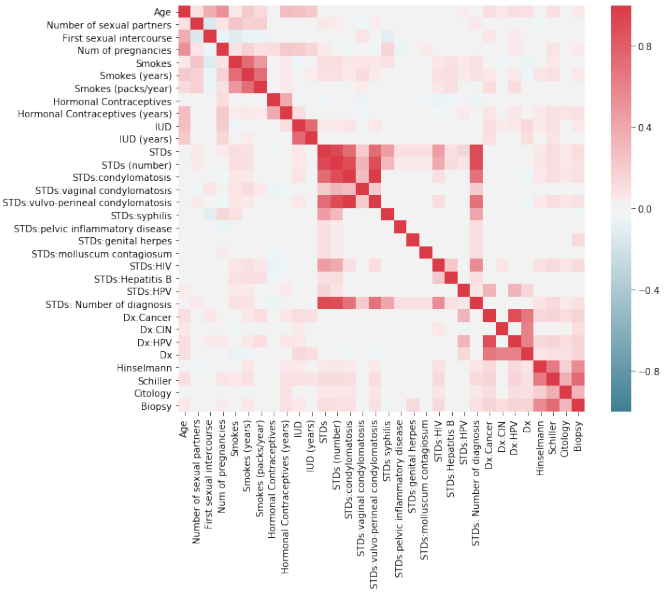



**Figura 3.** La matrice di correlazione mostra una stretta dipendenza fra i quattro esami di diagnosi del cancro cervicale.

L'obiettivo della Feature Selection è quello di trovare variabili poco correlate fra loro ma considerevolmente dipendenti dalla variabile target. Usare il minor numero di attributi possibili per raggiungere un risultato convincente, ovvero riducendo la dimensionalità dello spazio degli attributi, garantisce un costo computazionale minore per il processo di classificazione e contribuisce ad aumentare l'interpretabilità dei risultati. Si è pertanto deciso di confrontare diverse modalità di feature selection in modo tale da valutare quale fosse la migliore nella previsione della variabile target Biopsy. Per fare ciò si sono raffrontate le aree sottese dalle ROC curves e le metriche di accuratezza, precisione e Recall di ogni classificatore per ogni diversa modalità. I contesti in cui sono stati valutati i classificatori sono i seguenti:

- tutte le variabili;
- feature seleciton con filtro multivariato "ReliefF";
- tutte le variabili con dati incrementati sinteticamente tramite metodologia SMOTE;
- feature seleciton con filtro multivariato "ReliefF" con dati incrementati sinteticamente tramite metodologia SMOTE;
- backward feature selection con wrapper utilizzando il classificatore utilizzato successivamente per il training.

Per poter confrontare le prestazioni dei differenti algoritmi di learning si è deciso di adottare il metodo di "filter" multivariato ReliefF attribute selection, in primis per la sua capacità di lavorare sia con binari che continui. Inoltre i risultati dell'algoritmo ReliefF sono indipendenti dal classificatore utilizzato e sono prodotti tenendo congiuntamente in considerazione la rilevanza delle variabili per prevedere la classe target (considerate anche in sinergia) e l' eventuale ridondanza di quest'ultime, cioè la sovrapposizione del loro contributo esplicativo con quello di altre variabili.

| Rank | Attribute |
|---|---|
| 0.3825 | Schiller test |
| 0.1297 | Hinselmann test |
| 0.1137 | Citology test |
| 0.0729 | First sexual intercourse |
| 0.0662 | Hormonal Contraceptives |
| 0.0555 | Number of pregnancies |
| 0.0509 | Age |
| 0.0493 | IUD |
| 0.0405 | Hormonal Contraceptives years |
| 0.0387 | Smokes |
| 0.0294 | STDs |
| 0.0247 | previous Cancer |
| 0.0247 | previous HPV |
| 0.0227 | Number of sexual partners |
| ... | Altri |

**Tabella 2.** Attributi più importanti secondo il filtro ReliefF

Si è optato per evitare l'utilizzo della metodologia della Principal Component Analysis perché essa avrebbe ridotto di molto l'interpretabilità del modello finale che è stato ritenuto importante per l'eventuale utilizzo del nostro modello in un contesto reale.

## 3. Data Classification

### 3.1 Learning Algortithm Selection

Grazie alla compattezza del dataset si sono potuti provare agevolmente più algoritmi di classificazione in modo computazionalmente economico. Se il numero di record fosse stato più elevato, la selezione del modello sarebbe dovuta essere preceduta da una fase di sampling per poter valutare velocemente le differenze di performance tra i possibili classificatori.

Si è deciso di valutare diverse categorie di algoritmi sia euristici che probabilistici, in particolare la decisione si è poi concentrata sui seguenti classificatori:

- Primal Estimated sub-Gradient Solver for SVM (Spegasos);
- Adaptive Boosting (AdaBoost);
- Random forest a 10 alberi (RandomForest);
- Radial Basis Function Network (RBFN);
- Decision tree with Naive Byaes classifier at the leaves (NBtree);
- Logistic regression (Logisitc);
- Naive Bayes (NaiveBayes);
- Decision tree (J48).

Si è poi proceduto a valutare ogni algoritmo di classificazione in ogni configurazione di feature selection e ribilanciamento precedentemente descritte e ne sono stati valutati i risultati.

## 3.2 Cross Validation

Una volta selezionati gli algoritmi, essi vengono valutati attraverso la Cross Validation, la quale permette di allenare e testare il modello predittivo attraverso diversi sottoinsiemi del dataset iniziale, garantendo che ciascun record sia incluso nel training set lo stesso numero di volte e nel test set esattamente una sola volta (Tan et al., 2005). Nel dettaglio, il processo di Cross Validation consiste nella costruzione di una partizione dell'insieme di dati iniziali in K sottinsiemi disgiunti di simile numerosità e, ad ogni iterazione, l' i-esimo sottoinsieme del dataset iniziale (i=1,...,K) costituisce il validation set mentre l'unione dei rimanenti sottoinsiemi rappresentano il training set.

Il vantaggio di sottoporre l'algoritmo a K fasi di training è dovuto al fatto che le misure di performance del modello di classificazione, che si ottengono come media aritmetica delle singole misure calcolate ad ogni passo, rappresentano una valutazione più attendibile del modello predittivo. Infatti si evitano sia i problemi di overfitting sia quelli di campionamento affetto da bias che si potrebbe verificare quando si esegue una sola partizione iniziale del dataset in training set e validation set.

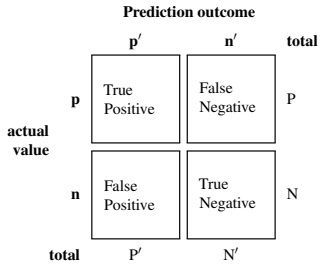
In questa analisi, si è scelto di iterare il processo di validazione dieci volte e, poiché le classi della variabile target sono fortemente sbilanciate, è stato adottato uno schema di campionamento stratificato in modo tale da garantire che ciascun sottoinsieme creato mantenga le proporzioni iniziali delle due classi di valori per "Biopsy".

## 3.3 Classification Evaluation

Dopo aver ottenuto le diverse predizioni della variabile target per i diversi algoritmi è necessario confrontare i risultati ottenuti. Il modo migliore per valutare diverse predizioni di una variabile è basarsi sul numero di record correttamente classificati dal modello e su quelli che non sono risultati compatibili con le previsioni.

### 3.3.1 La valutazione di un problema binario

Per quanto riguarda la classificazione di una variabile di output binaria, essa può assumere due soli valori; vengono solitamente chiamati classe positiva i record relativi al valore in minoranza o perlomeno quelli su cui l'analisi vuole concentrarsi in dettaglio, invece i record associati al valore restante vanno a formare la classe negativa. Dunque il modello predittivo può cadere in due tipi di errori: esso può classificare positivi dei valori che invece sono negativi, o viceversa capita classifichi come negativi elementi della classe positiva; i primi vengono chiamati "False Positive" (FP), mentre i secondi sono detti "False Negative"(FN). Allo stesso modo i record positivi classificati correttamente andranno a far parte dei "True Positive"(TP), mentre quelli negativi vengono detti "True Negative"(TN). Queste quattro partizioni dei record vanno a formare la cosiddetta matrice di confusione (Confusion Matrix) data da




dove nella diagonale maggiore vengono posizionati i record correttamente classificati dal modello, mentre sulla diagonale opposta sono riportati i record erroneamente predetti.

### 3.3.2 Comparare le performance di più modelli

Con le quattro quantità appena introdotte è possibile calcolare l'accuratezza di una predizione, questa rappresenta un importante indice per valutare la bontà di un algoritmo, infatti essa è definta come

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \in [0,1], \tag{1}$$

ovvero il rapporto tra i valori correttamente predetti e tutti quelli contenuti nel dataset a meno dei record scartati nel trattamento dei valori mancanti. L'accuratezza di un modello può variare tra 0 e 1 e più è prossima all'unità maggiore è la capacità predittiva del modello. Quando però una classe è in minoranza, come già accennato, il valore dell'accuratezza raggiunta non basta a confrontare la prestazione di più modelli. Massimizzare l'accuratezza non è sufficiente poiché il dataset è profondamente sbilanciato rispetto alla variabile target "Biopsy", di conseguenza risulta fondamentale introdurre l'indice di Recall, dato da

$$Recall = \frac{TP}{TP+FN} \quad \in [0,1], \tag{2}$$

e la precisione, definita come

$$Precision = \frac{TP}{TP+FP} \quad \in [0,1], \tag{3}$$

dove TP indica il numero di valori positivi correttamente classificati dal modello di classificazione, FN rappresenta il numero di falsi negativi e FP il numero di falsi positivi; per il modo in cui sono definiti, recall e precision possono assumere valori compresi fra zero e uno. Per un indice che prende in considerazione entrambe le due grandezze viene introdotta la F-measure, una media armonica di recall e precision, data da

$$Fmeasure = 2\frac{p \cdot r}{p+r} \quad \in [0,1], \tag{4}$$

dove $p$ e $r$ rappresentano rispettivamente Precision e Recall.

Ci si è posti l'obiettivo di minimizzare il numero di falsi negativi ottenuti giustificando questa scelta con il fatto che seppure la biopsia sia un esame impegnativo per la paziente esso costituirebbe il male minore rispetto ad un'eventuale

mancata diagnosi. Alla luce di ciò nel seguito verrà attribuita più importanza al valore del Recall piuttosto che della Precision.

Un altro metodo per la valutazione dei classificatori per problemi di predizione a 2 classi è il calcolo dell'area sottesa alle cosiddette curve di ROC (Receiver Operating Characteristic).

Come mostrato in Fig. 4, questi schemi grafici mostrano la relazione fra il tasso di falsi positivi (False Positive Rate) e quello di veri positivi (True Positive Rate) per diversi modelli di classificazione. Una curva di ROC ha due casi limite: la retta diagonale con area 0.5 che rappresenta il modello randomico e una linea spezzata che dall'origine sale fino al punto (0,1) per poi proseguire orizzontalmente fino al punto (1,1), quest'ultima rappresenta il caso di classificatore perfetto.

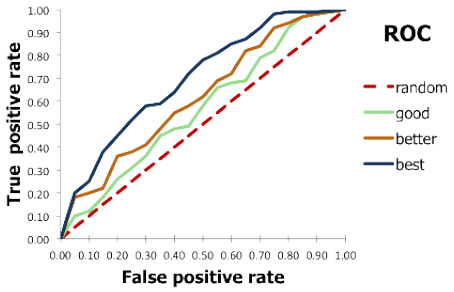



**Figura 4.** Il grafico mostra diverse curve di ROC messe a confronto tra loro e con il modello di classificatore casuale.

## 4. Flusso di KNIME

Dopo aver caricato il dataset è stata affrontata una fase di preprocessing in cui è stato trattato il problema dei missing values e in cui sono state normalizzate le variabili quantitative con valori compresi nell'intervallo [0,1].

Successivamente, si mostra un esempio in cui è stato affrontato il problema della feature selection: utilizzando il metodo wrapper, si mostra il valore della recall a seconda del numero di attributi scelti dal modello di classificazione.

Nella sezione "Classification process" sono stati confrontati 8 diversi algoritmi di classificazione (SPegasos, AdaBoost, RandomForest, RBFN, NBtree, Logistic, NaiveBayes, J48) con diverse configurazioni:

- vengono utilizzate tutte le variabili senza applicare nessun metodo di feature selection;
- si utilizza un filtro multivariato "ReliefF" per la scelta degli attributi più significativi;
- si utilizza il metodo SMOTE per il ribilanciamento delle classi;
- si utilizzano sia il filtro multivariato "ReliefF" sia il metodo SMOTE;
- si utilizza il metodo wrapper per la feature selection.

Nella sezione "classification evaluation" vengono mostrati i risultati delle misure di Accuracy, Recall, Precision e F-Measure e delle ROC curve per i diversi algoritmi. In particolare, nella figura 5 vengono riportati i grafici di tali risultati per 4 algoritmi (SPegasos, AdaBoost, NaiveBayeS e J48).

Infine, nella sezione superiore del workflow si mostrano i risultati della Cross Validation utilizzando l'algoritmo migliore in termini di Recall, AUC ed interpretabilità: AdaBoost, utilizzato considerando sia il metodo SMOTE sia il filtro multivariato "ReliefF". Inoltre, si vuole mostrare se, con l'algoritmo migliore, sia possibile prevedere in modo significativo il valore di "Biopsy" senza utilizzare tutti gli altri test diagnostici. Pertanto, tale algoritmo è stato eseguito considerando 4 diversi insiemi di attributi in input:

- tutte le variabili (quindi, includendo tutti i test diagnostici: Hinselman, Schiller e Citologia);
- viene escluso l'attributo "Schiller" (quindi, si utilizzano solo i test diagnostici: Hinsellman, Citology);
- vengono esclusi gli attributi "citology" e "Schiller" (quindi, si utilizza solo il test diagnostico: Hinselman);
- vengono esclusi gli attributi "Hinselman" e "Schiller" (quindi, si utilizza solo il test diagnostico: Citologia)

e nella figura 7 vengono riportati i valori di Accuracy, Precision, Recall e F-Measure.

## 5. Risultati e Discussione

Si riportano di seguito i risultati dei modelli più rilevanti per poi procedere con la loro discussione. Si rimanda al flusso KNIME per i risultadi degli altri modelli.

| | Accuracy | Recall | Precision | F-Measure | AUC |
|---|---|---|---|---|---|
| **S_PEGASOS** | 0.954 | 0.745 | 0.621 | 0.678 | 0.86 |
| **SP_FILTER** | 0.954 | 0.745 | 0.621 | 0.678 | 0.86 |
| **SP_SMOTE** | 0.959 | 0.855 | 0.635 | 0.729 | 0.91 |
| **SP_SMOTE+FILTER** | 0.806 | 0.873 | 0.232 | 0.366 | 0.84 |
| **SP_WRAPPER** | 0.961 | 0.873 | 0.649 | 0.744 | 0.92 |

**Tabella 3.** Spegasos classifier results

| | Accuracy | Recall | Precision | F-Measure | AUC |
|---|---|---|---|---|---|
| **ADABOOST** | 0.954 | 0.709 | 0.629 | 0.667 | 0.88 |
| **ADAB_FILTER** | 0.954 | 0.709 | 0.629 | 0.667 | 0.88 |
| **ADAB_SMOTE** | 0.958 | 0.764 | 0.646 | 0.7 | 0.90 |
| **ADAB_SMOTE+FILTER** | 0.960 | 0.873 | 0.64 | 0.738 | 0.88 |
| **ADAB_WRAPPER** | 0.961 | 0.873 | 0.649 | 0.744 | 0.87 |

**Tabella 4.** Adaboost classifier results

| | Accuracy | Recall | Precision | F-Measure | AUC |
|---|---|---|---|---|---|
| **NAIVE_BAYES** | 0.884 | 0.727 | 0.323 | 0.447 | 0.87 |
| **NB_FILTER** | 0.884 | 0.727 | 0.323 | 0.447 | 0.87 |
| **NB_SMOTE** | 0.864 | 0.745 | 0.287 | 0.414 | 0.85 |
| **NB_SMOTE+FILTER** | 0.877 | 0.855 | 0.326 | 0.472 | 0.91 |
| **NB_WRAPPER** | 0.957 | 0.873 | 0.615 | 0.722 | 0.90 |

**Tabella 5.** Naive Bayes classifier results

| | Accuracy | Recall | Precision | F-Measure | AUC |
|---|---|---|---|---|---|
| **J48** | 0.946 | 0.527 | 0.592 | 0.558 | 0.69 |
| **J48_FILTER** | 0.949 | 0.564 | 0.608 | 0.585 | 0.73 |
| **J48_SMOTE** | 0.943 | 0.473 | 0.565 | 0.515 | 0.63 |
| **J48_SMOTE+FILTER** | 0.944 | 0.727 | 0.548 | 0.625 | 0.87 |
| **J48_WRAPPER** | 0.961 | 0.873 | 0.649 | 0.744 | 0.87 |

**Tabella 6.** J48 classifier results

### 5.1 Comparazione dei diversi algoritmi di learning e configurazioni

Si osserva immediatamente che ogni modello ha una Accuracy molto alta, questo è dovuto allo sbilanciamento del dataset

spiegato in precedenza che permette a modelli di ipotesi nulla di ottenere risultati a prima vista molto positivi. L'Accuracy è stata quindi poco considerata come indice di paragone per i modelli ed al suo posto, come già accennato, si sono utilizzati Recall, F-Measure e l'area sottesa dalla ROC curve (AUC).

Si nota innanzitutto che all'aumentare della F-measure non aumenta necessariamente anche l'AUC. Questo è dovuto al fatto che l'AUC va a considerare il rapporto tra i True Positive e i False Positive ignorando completamente i False Negative. Essendo l'obiettivo del nostro lavoro la riduzione al minimo delle diagnosi falsamente negative si è scelto il modello migliore basandoci principalmente su F-measure e recall e solo in secondo luogo sulla AUC. In Fig. 5 sono mostrati i risultati in termini di Accuracy, Recall e Precision della classificazione con i quattro algoritmi più prestanti, mentre in Fig. 4 sono mostrate le rispettive curve di ROC. La metodologia ReliefF utilizzata singolarmente non porta a consistenti benefici in nessuna delle configurazioni per nessuno dei modelli, così come il vantaggio derivante dall'utilizzo unicamente della metodologia SMOTE è poco rilevante, tuttavia se si combinano le due metodologie i modelli migliorano sensibilmente ad esclusione del classificatore SPegasos che peggiora drasticamente. L'ultima configurazione basata sul metodo wrapper porta consistenti miglioramenti ad alcuni classificatori e nessuno ad altri, nessun modello tuttavia presenta peggioramenti.

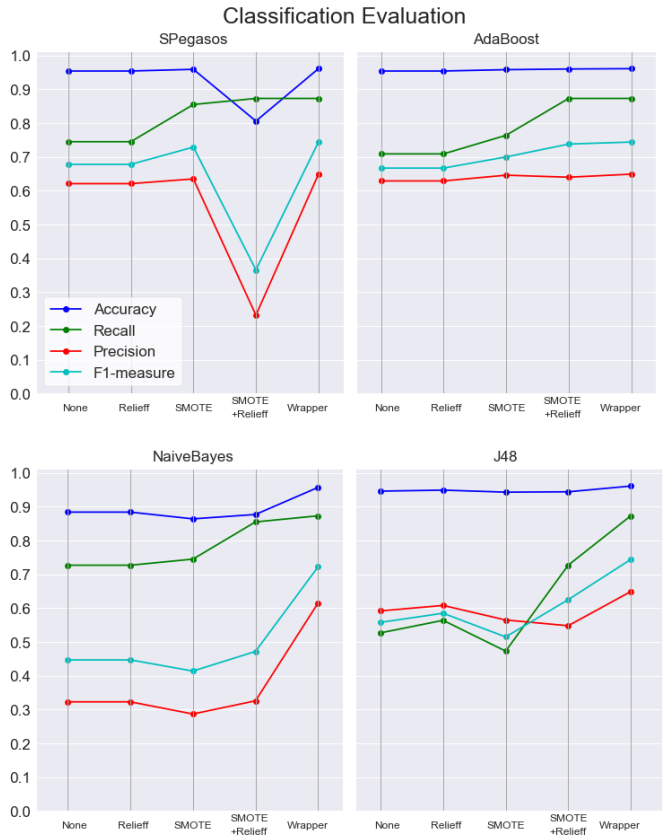



**Figura 5.** Le performance dei classificatori al variare della configurazione utilizzata.

Nonostante modelli basati su wrapper siano spesso molto performanti essi basano la loro decisione su un numero limitatissimo di attributi, spesso solo il più significativo (Schiller), ciò ha portato a scegliere l'utilizzo di una metodologia di filter per il nostro algoritmo finale in modo da consentire una miglior interpretabilità del modello e un'applicabilità di quest'ultimo anche nel caso in cui non si disponga del risultato del test di Schiller.

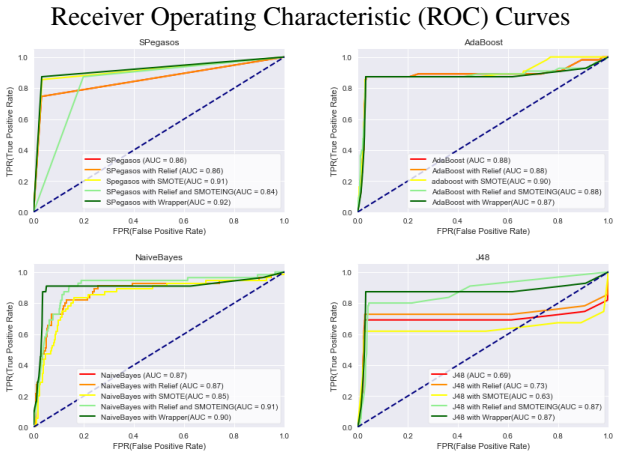



**Figura 6.** Le curve di ROC relative agli algoritmi di learning più soddisfacenti.

## 5.2 La configurazione migliore

La configurazione migliore per prevedere la variabile target biopsy e mantenere un alto livello di interpretabilità è risultata l'utlizzo del classificatore AdaBoost con 10 iterazioni unito al suo training su di un dataset con record sintetici ottenuti con metodologia SMOTE. Successivamente si è testata la resistenza di questa configurazione in assenza delle variabili citology, Heinselmann e Schiller in modo progressivo valutando la corrispondente variazione nelle metriche, come riportato nella figura 7 e nella tabella 7. Si osserva chiaramente l'importanza della variabile Schiller nell'efficacia del modello.

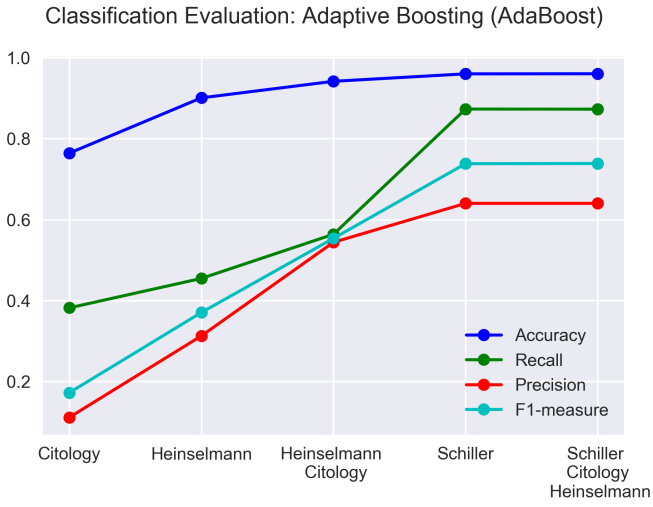



**Figura 7.** Variazione delle performance del classificatore AdaBoost al variare degli attributi utilizzati.

| | Accuracy | Recall | Precision | F-measure |
|---|---|---|---|---|
| **Citology** | 0.764 | 0.382 | 0.111 | 0.172 |
| **Heinselmann** | 0.901 | 0.455 | 0.312 | 0.370 |
| **Heinselmann+Citology** | 0.942 | 0.564 | 0.544 | 0.554 |
| **Schiller** | 0.960 | 0.873 | 0.640 | 0.738 |
| **Heinselmann +Citology + Schiller** | 0.960 | 0.873 | 0.640 | 0.738 |

**Tabella 7.** Miglioramento con la progressiva aggiunta di attributi riferiti ai test diagnostici.

Interessante notare che nonostante questa metodologia scelta non vada a limitare le variabili utilizzate come il metodo wrapper il risultato va in comunque ad essere influenzato totalmente dalla variabile Schiller che sovrasta tutte le altre.

Si riporta infine la confusion matrix finale ottenuta con il modello scelto e tutte le variabili. Il numero di falsi negativi è stato ridotto al minimo di sette casi su ottecentocinquantasei.

| | | Valore predetto | | |
|---|---|---|---|---|
| | | Negativo | Positivo | Totale |
| Valore reale | Negativo | 774 | 27 | 801 |
| | Positivo | 7 | 48 | 55 |
| | Totale | 781 | 75 | 856 |

## 6. Conclusioni

L'obiettivo, prefissato all'inizio della ricerca, di riuscire a prevedere efficacemente il valore di "Biopsy" ha raggiunto risultati soddisfacenti infatti il modello fornito da AdaBoost, utilizzando gli attributi relativi ai test diagnostici meno invasivi della biopsia, oltre a fornire un valore elevato di Accuracy (96%), raggiunge un grado di Recall dell'87%.

In accordo con quanto ci si aspettava tra le variabili disponibili nel dataset, quelle più importanti in funzione della previsione del cancro sono risultate essere gli esiti dei test diagnostici: "Schiller", "Hinselmann" e "Citology".

Considerato che la letteratura medica evidenzia come principali fattori di rischio per il cancro alla cervice il numero di partners sessuali l'età del primo rapporto sessuale ci si sarebbe attesi inoltre che quest'ultimi avessero avuto un contributo esplicativo relativamente più elevato nella predizione dell'esito della biopsia rispetto ad altre variabili. Abbiamo riscontrato che nel caso in esame il numero di partner sessuali è poco utile nel determinare l'esito della biopsia: ha correlazione lineare nulla e non è presente tra le prime dieci variabili in termini di importanza selezionate dal filtro ReliefF. La variabile relativa all'età del primo rapporto invece, pur presentando un coefficiente di correlazione lineare veramente basso (0.008), risulta la più importante esplicativa dopo i test diagnostici secondo il ranking del filtro ReliefF, si può quindi ipotizzare un legame non lineare con la variabile target e/o un effetto sinergico con altre variabili.

Tuttavia non è stato possibile ottenere un modello che prevedesse efficacemente la malattia in assenza delle variabili indicanti i test diagnostici.

Il risultato ottenuto con l'utilizzo del metodo wrapper fa cogliere l'importanza del test di Schiller nel predire l'esito di una biopsia per il cancro alla cervice e fa giungere alla conclusione che una semplice prevenzione basata su caratteristiche demografiche dei pazienti potrebbe non essere sufficiente a trovare tutti i casi di cancro alla cervice in tempo utile. Si consiglia quindi una prevenzione svolta attraverso test di Schiller in modo da avere dei risultati efficaci senza passare direttamente alla biopsia ed evitando test citologici e di Hinselmann che comporterebbero spese e inconvenienti ulteriori senza migliorare le previsioni.

### Prospettive future

Il dataset preso in analisi contava meno di un migliaio di record: disporre di un numero maggiore di dati potrebbe rendere più efficaci le tecniche di Machine Learning che sono state utilizzate. Sarebbe possibile avanzare modelli più articolati, dove l'importanza delle diverse variabili demografiche e dei vari test diagnostici potrebbe essere messa ancora più in luce con metodologie di clustering o analisi associativa. Inoltre sarebbe interessante avere maggiori informazioni circa le pazienti, per esempio, risulterebbe forse utile conoscere la zona dove vivono o il loro reddito medio in modo da studiare come la qualità di vita di una donna incida sulle sue possibilità di contrarre una patologia tumorale alla cervice.

## Bibliografia

F. X. Bosch and S. de Sanjose. The epidemiology of human papillomavirus infection and cervical cancer. *Dis. Markers*, 23, 2007. URL https://www.ncbi.nlm.nih.gov/pubmed/17627057.

Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority oversampling technique. *CoRR*, abs/1106.1813, 2011. URL http://arxiv.org/abs/1106.1813.

Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Transfer learning with partial observability applied to cervical cancer screening. 2007. URL https://pdfs.semanticscholar.org/1c02/438ba4dfa775399ba414508e9cd335b69012.pdf.

Kaggle. Cervical cancer risk classification. URL https://www.kaggle.com/loveall/cervical-cancer-risk-classification.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005. ISBN 0321321367.