Cervical Cancer Risk Classification: Analisi e Previsione

Bagnalasta Matteo[1+], Biondi Stefano[2*], Pierri Luca[3*]

Abstract

Per molto tempo il tumore della cervice ha rappresentato la più frequente forma di cancro per le donne. Nei Paesi in via di sviluppo questo tumore è ancora la seconda causa di morte per cancro, mentre nel mondo occidentale il numero dei casi e dei decessi continuano a diminuire grazie all'introduzione di test mirati e puntuali: è il caso del Pap-test e della Colposcopia, entrambi esami di diagnosi precoce molto efficaci.
Nonostante sia una delle tipologie di cancro più prevenibile, uccide ancora circa $4\,000$ donne negli Stati Uniti, $300\,000$ nel mondo.
Nella seguente relazione sarà portata a termine un'ampia analisi sui principali rischi che contribuiscono allo sviluppo del cancro alla cervice; rischi emersi a seguito di numerosi e costosi test. Proprio a fronte di ciò, l'obiettivo finale della stessa sarà trovare il predittore che utilizzi il minore numero possibile di rischi (features) che, contemporaneamente, mantenga massima la corretta previsione del cancro (Biopsy $= 1$).

[1] **Autore corrispondente**: m.bagnalasta@campus.unimib.it
[2] **Autore corrispondente**: s.biondi7@campus.unimib.it
[3] **Autore corrispondente**: l.pierri1@campus.unimib.it
[+] *Fisica, Dipartimento di Fisica "G. Occhialini", Univeristà di Milano Bicocca, Italia*
[*] *Data Science, Dipartimento di Computer Science, Univeristà Milano Bicocca, Italia*

Contents

Introduzione

Con circa 570 000 nuovi casi nel 2018, il tumore della cervice è il quarto tipo di cancro più diffuso nella popolazione femminile [1].
In Italia ogni anno vengono diagnosticati circa 2 300 nuovi casi, prevalentemente allo stadio iniziale. Tuttavia, una donna su 10 000 riceve una diagnosi già quando il tumore è in forma avanzata[2].

Quando il tumore viene diagnosticato in fase avanzata il tasso di mortalità è elevato. L'alto indice di mortalità può essere ridotto attraverso un approccio che include prevenzione, diagnosi precoce, programmi di screening e trattamenti efficaci.
La diagnosi in fase molto precoce o precancerosa è possibile ricorrendo con regolarità ad uno screening con il Pap-test o con l'HPV-test.
Ci sono altri fattori o esami medici che permettono di prevedere l'esistenza del cancro alla cervice in maniera più mirata?
Utilizzando un dataset composto dalle rilevazioni di 858 individui, tramite tecniche di Machine Learning abbiamo cercato di predire l'attributo binario "Biopsy".
Nella valutazione dei predittore abbiamo posto particolare attenzione alla recall, valore che indica la bassa presenza di falsi negativi, e alla precision, valore che indica l'alto numero di predizioni corrette. Per confrontare i predittori abbiamo quindi utilizzato come misura di sintesi la media armonica di precision e recall, che vogliamo massimizzata.

L'analisi è stata eseguita usando Knime, un software open-source, che permette di sviluppare e visualizzare il workflow del progetto in Machine Learning. Inoltre abbiamo utilizzato uno script Python per la pulizia dei dati iniziali prendendo spunto dal lavoro pubblicato su Kaggle da Atakan Söztekin [4]

1. Dataset e Statistiche Descrittive

1.1 Panoramica

Conclusi i convenevoli, è tempo di fornire una descrizione di massima dei dati utilizzati. Il Dataset è disponibile pubblicamente sulla piattaforma Kaggle ed è composto da 858 righe e 36 colonne [3]. Considerando che descrivere singolarmente un numero così elevato di colonne sarebbe deleterio per il lettore, seguirà una breve descrizioni delle variabili principali:

- *Age: numeric-discrete*
  Età della paziente
- *Number of sexual partners: numeric-discrete*
  Numero di partner sessuali che la paziente ha avuto.
- *First sexual intercourse: numeric-discrete*
  Età nella quale la paziente ha praticato per la prima volta attività sessuale
- *Num of pregnancies: numeric-discrete*
  Numero di gravidanze della paziente
- *Smokes: double-binary*
  Indica se la paziente è fumatrice/non fumatrice. Accoppiata con la variabile *Smokes (years)* che da quanti anni la paziente fuma e la variabile *Smokes (pack/years)* che indica la media dei pacchetti consumati ogni anno.
- *Hormonal Contraceptives: double-binary*
  Indica se la paziente fa uso di contraccettivi ormonali o meno. Come la variabile descritta subito sopra, anche qui è presente la variabile companion *Hormonal Contraceptives (years)*.
- *IUD: double-binary*
  Indica la presenza o meno di un piccolo dispositivo per la prevenzione della gravidanza nell'utero della paziente. Accoppiata con la variabilile *IUD: years*.
- *STDs: double-binary*
  Indica se la paziente è affetta o meno da malattie sessualmente trasmissibili.
- *Dx: integer-binary*
  Il test DX per l'oncotipo analizza l'attività di un gruppo di geni che possono influenzare il comportamento del cancro e la risposta al trattamento.
- *Hinselmann: integer- binary*
  Risultato dell'Hinselmann test (colposcopia).
- *Schiller: integer-binary*
  Risultato dello Schiller test.
- *Citology: integer-binary*
  Risultato dell'esame di citologia delle urine, utilizzato come indagine pre-cancro.
- *Biopsy: integer-binary*
  Variabile target che indica il risultato della Biopsia.

Sono poi presenti differenti colonne che fanno riferimento alle variabili *STDs* e *DX*. Ognuna di esse ha infatti come label di colonna una delle due variabili seguito dalla specifica malattia.

2. Statistiche Preliminari e Preprocessing

2.1 Preparazione dei Dati

Procedendo con un'analisi descrittiva preliminare dei dati così come il CSV reperito su Kaggle fornisce, è possibile notare la presenza di due problemi non banali: la discreta quantità di valori mancanti e una forte *Class Imbalance* che affligge la variabile target.
L'approccio seguito è stato guidato dal tipo e dalla numerosità di dati mancanti della variabile.
In accordo con il lavoro di Söztekin abbiamo identificato le colonne *Age, Number of sexual partners, first sexual intercourse, Num of pregnancies* come quelle con un numero di dati mancanti inferiore a 100. Per non modificare la distribuzione delle variabili abbiamo sostituito i valori nulli con la mediana della colonna.

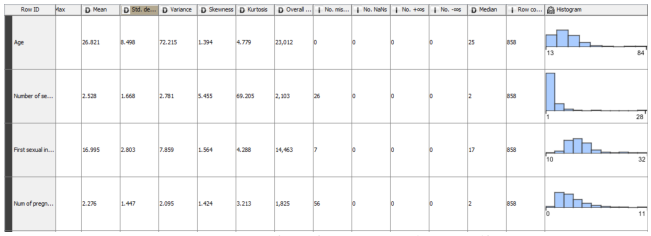


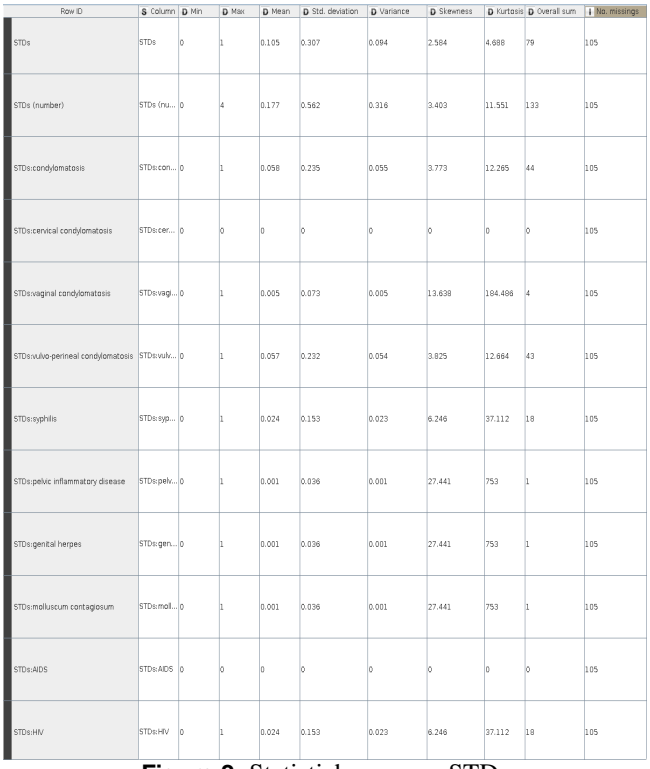
Figure 1. Sostituzione con la mediana

Proseguendo nell'analisi delle variabili con dati mancanti, concentriamoci su quelle di tipo dicotomico. Non essendo più possibile utilizzare la mediana, Söztekin utilizza la matrice di correlazione per fare confronti tra la variabile dicotomica a cui sostituire il dato mancante e quelle numeriche complete. Le correlazioni però non sono sufficienti per decidere che valore inserire e i confronti da lui fatti, a volte, non sono del tutto chiari, quindi abbiamo deciso di proseguire diversamente.
Le variabili maggiormente afflitte dai missing values sono quelle che riguardano le malattie sessualmente trasmissibili (*STDs*). In particolare le colonne *STDs:Time since first diagnosis* e *STDS:Time since last diagnosis* hanno 787 valori nulli, più del 90%, di conseguenza sono state eliminate. Tutte le restanti *STDs*, tranne il *STDs (number)* che non assume valori nulli, sono caratterizzate dagli stessi 105 valori mancanti. Considerando il forte sbilanciamento del numero di valori negativi rispetto a quelli positivi decidiamo di rimuovere le righe con i valori nulli in questione. Valorizzare senza una chiara strategia il 12% dei valori inserirebbe un bias nei dati e quindi in tutta la successiva analisi. Inoltre, rimuovendo i record in questione, la presenza dei valori nulli nelle restanti colonne è ridotta al minimo e più facilmente gestibile.
Le restanti colonne sono 3 gruppi di attributi:

- Smokes, Smokes (years) e Smokes (pack/years)

- Hormonal Contraceptives e Hormonal Contraceptive (years)
- IUD e IUD (years)

Il primo attributo è dicotomico e rappresenta se il paziente è coinvolto in ciò che esso rappresenta mentre le sucessive sono valori rapportati ad esso. Per questo motivo se nella colonna dicotomica il valore mancante è stato sostituito con 0 allora nelle restanti colonne del gruppo verrà inserito 0, viceversa, se nella colonna dicotomica il valore mancante è stato sostituito con 1 allora nelle restanti colonne del gruppo verrà inserita la mediana della colonna.
Gli attributi *Smokes* e *IUD* hanno una predominanza di 0, più dell'86%, quindi si è deciso di sostituire i rispettivi 10 e 16 valori mancanti con 0.
L'attributo *Hormonal Contraceptives* invece ha il 64% di valori positivi, quindi, in accordo con regola di sostituzione del valore più frequente, si sono sostituiti i 13 valori mancanti con 1.



Row ID	Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum	No. missings
STDs	STDs	0	1	0.105	0.307	0.094	2.584	4.688	79	105
STDs (number)	STDs (nu...	0	4	0.177	0.562	0.316	3.403	11.551	133	105
STDs:condylomatosis	STDs:con...	0	1	0.058	0.235	0.055	3.773	12.265	44	105
STDs:cervical condylomatosis	STDs:cer...	0	0	0	0	0	0	0	0	105
STDs:vaginal condylomatosis	STDs:vagi...	0	1	0.005	0.073	0.005	13.638	184.486	4	105
STDs:vulvo-perineal condylomatosis	STDs:vulv...	0	1	0.057	0.232	0.054	3.825	12.664	43	105
STDs:syphilis	STDs:syp...	0	1	0.024	0.153	0.023	6.246	37.112	18	105
STDs:pelvic inflammatory disease	STDs:pelv...	0	1	0.001	0.036	0.001	27.441	753	1	105
STDs:genital herpes	STDs:gen...	0	1	0.001	0.036	0.001	27.441	753	1	105
STDs:molluscum contagiosum	STDs:moll...	0	1	0.001	0.036	0.001	27.441	753	1	105
STDs:AIDS	STDs:AIDS	0	0	0	0	0	0	0	0	105
STDs:HIV	STDs:HIV	0	1	0.024	0.153	0.023	6.246	37.112	18	105

Figure 2. Statistiche gruppo STDs

Per lo sbilanciamento delle frequenze della colonna *Biopsy* poco o nulla può essere portato a termine nello step di preprocessing. Bisognerà tenere conto della problematica durante la valutazione degli algoritmi predittivi. Facendo particolare attenzione alla tecnica della *matrice di costo*, del *Cumulative Gain* e della *ROC Curve*.

2.2 Normalizer e Shuffle

Non avendo informazioni riguardo la distribuzione dei valori delle features si è deciso di utilizzare la normalizzazione come tecnica di Features Scaling. Abbiamo quindi normalizzato le features del dataset in valori compresi tra 0 e 1. In questo modo, in fase di confronto tra variabili diverse, queste ultime rimarranno proporzionali le une con le altre evitando che differenti unità di grandezza generino differenti pesi di importanza. Quindi per ogni colonna del dataset abbiamo applicato la sostituzione:

$$y_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}, \qquad \forall\, i \in \text{feature}$$

Per ogni feature nel dataset.

Infine abbiamo mescolato il dataset evitando il mantenimento di particolari pattern che possono essersi creati nella fase di estrazione. Così da non rischiare di identificare delle correlazioni inesistenti.

2.3 Statistiche Descrittive post Preprocessing

La gestione dei valori nulli e la normalizzazione delle variabili numeriche hanno determinato un cambiamento del dataset che però non ha influito sulle distribuzioni delle variabili.

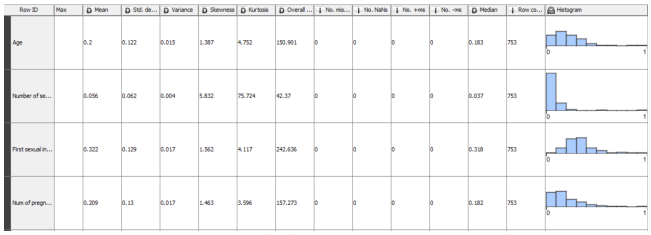


Figure 3. Statistiche post Preprocessing

Avendo eliminato tutti i valori mancanti è possibile effettuare alcune considerazioni sulla matrice di correlazione, utilizzando l'indice di correlazione di *Pearson*.
Come ci si aspettava, alcune delle variabili appartenenti alla macro categoria STDs hanno un alto indice di correlazione.

Row ID	STDs:con...	STDs:vulv...
STDs:condylo...	1	0.988
STDs:vulvo-p...	0.988	1

Figure 4. Correlazione variabili STDs

Da notare la discreta correlazione tra le variabili relative ai risultati dei test Hinselmann, *Schiller*, *Citology* e la variabile target *Biopsy*.
Nel seguito della relazione questa correlazione verrà ripresa per trarre conclusioni e decisioni sull'analisi.

Row ID	Hinselmann	Schiller	Citology	Biopsy
Hinselmann	1	0.653	0.197	0.556
Schiller	0.653	1	0.376	0.735
Citology	0.197	0.376	1	0.346
Biopsy	0.556	0.735	0.346	1

Figure 5. Variabili Correlate con Biopsy

3. Analisi dei Dati e Machine Learning

3.1 Algoritmi Valutati

Il primo passo consiste nel decidere quali algoritmi di Machine Learning utilizzare tenendo in considerazione la tipologia delle nostre features (input) e della variabile che vogliamo predire (Biopsy).
Tutte le variabili sono numeriche o dicotomiche, con una netta maggioranza delle seconde sulle prime. Sono stati dunque utilizzati:

- *Decision Tree J48*
- *Random Forest*
- *Naive Bayes*
- *Multi-Layer Perceptron*
- *Logistic Regression*
- *Support Vector Machine*

Per ogni algoritmo verrà performata sia una previsione univariata che una multivariata. La valutazione verterà sulle misure di *Accuracy*, *Precision Recall* ed *F-Measure*. Come è stato già osservato in precedenza, considerato l'elevata *class imbalance* della variabile target, l'*Accuracy* non risulta essere una buona misura di valutazione, in quanto la *majority class* (0, nel caso di *Biopsy*) nel training set domina il processo di learning per gli algoritmi scelti.
L'algoritmo con le performances migliori verrà esaminato nel dettaglio, a livello di modello. Emergerà come i coefficienti dominanti della previsione siano le colonne relative ai tre test medici già osservate in precedenza. Dopo aver rimosso tutte le colonne eccetto le tre in questione, verranno eseguiti nuovamente tutti gli Algoritmi per osservare se *sia possibile ottenere le stesse (o superiori) performances*.

3.2 Misure di Validazione

Come Misure di valutazione dei predittori si sono utilizzate l'*Accurcy*, la *F-Measure*, la *Recall* e la *Precision*. Per calcolare questi indici abbiamo utilizzato $2/3$ dei dati come *Training Set* e il resto come *Test Set* con una 3-*fold Cross Validation* con *Stratified Sampling* per evitare problematiche di over e under-fitting. Sono stati addestrati i predittori con l'evaluator univariato *InfoGainAttributeEval* e con l'evaluator multivariato *CfsSubsetEval*. La feature selection multivariata con evaluator *WrapperSubsetEval* non ha mai predetto valori positivi per *Biopsy* quindi è stata scartata.

3.3 Dataset Completo: Performance dei Modelli

Essendo un problema con molta class imbalance l'Accuracy perde di valore nella corretta valutazione del predittore. La *F-Measure*, media armonica di Recall e Precision, invece, insieme alla *Recall* presa singolarmente, diventano fondamentali per una corretta ed oggettiva valutazione dei predittori.

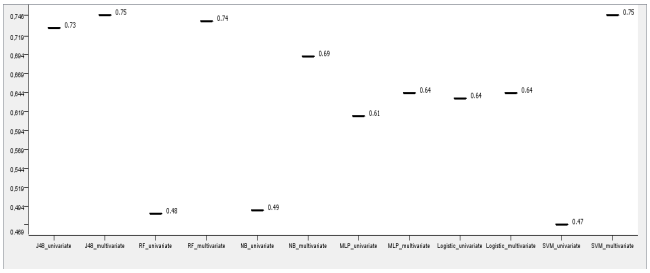


Figure 6. F-Measure

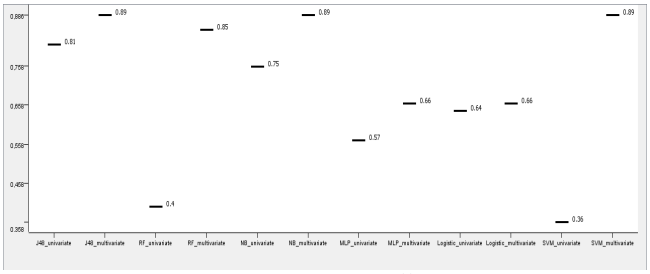


Figure 7. Recall

Possiamo subito notare che i risultati massimi li otteniamo con il *J48* e con il *Support Vector Machine* entrambi con la multivariate futures selection con evaluator *CfsSubsetEval*. E' importante mantenere un'alta *Recall*, cioè bassa percentuale di *Falsi Negativi*, perchè rappresenta il miglioramento rispetto ad un classificamento casuale. In questo specifico ambito sanitario significa minimizzare il numero di pazienti che hanno il cancro ma a cui non viene diagnosticato. Questo fattore è importantissimo non solo per un beneficio economico, ma ovviamente anche per un beneficio per la qualità della vita dei pazienti. Aumenta infatti la probabilità di guarigione e diminuisce l'impatto invasivo che una cura ha sul paziente in una fase iniziale.

Per quanto riguarda le altre 2 misure possiamo osservare che i due predittori scelti massimizzano anche l'*Accuracy*

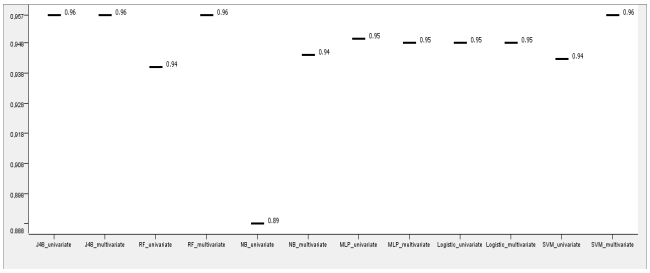


Figure 8. Accuracy

mentre la *Precision* non è massima ma rimane nella media dei predittori analizzati.

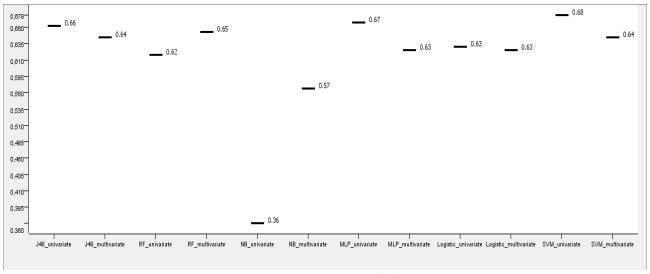


Figure 9. Precision

In definitiva procediamo nell'analisi scegliendo il *J48* e il *Support Vector Machine* rinunciando a 4 centesimi di *Precision* in cambio della massima *Recall* e *F-Measure*.

3.4 Migliori 2 Features: Performance dei Modelli

Analizzando le features prese in considerazione dai modelli dopo la selection, possiamo notare che sono state scelte le stesse variabili:

- STDs genital herpes
- Dx Cancer
- Schiller
- Citology

Inoltre l'albero decisionale generato dal *J48* utilizza come unica variabile la *Schiller*. Quindi il *J48* equivale a utilizzare la variabile *Schiller* come predittore.

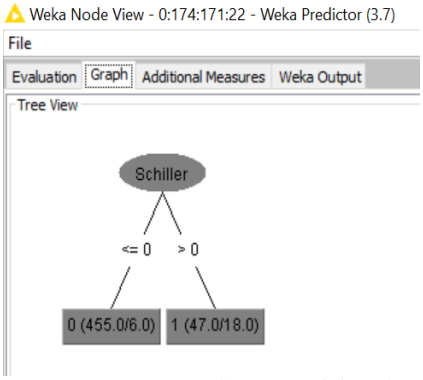


Figure 10. J48 - Albero Decisionale

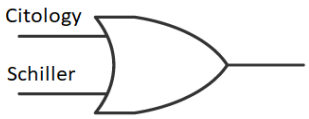
In accordo quindi con la matrice di correlazione abbiamo deciso di fare una feature selection manuale mantenendo unicamente le variabili *Schiller* e *Citology*.
A conferma del fatto che sono le uniche 2 variabili rilevanti al fine di predire la *Biopsy*, emerge come vengano massimizzate tutte e 4 le misure di analisi per tutti i predittori tranne il *Naive Bayes* che perde di *Precision* e quindi diminuisce anche la *F-Measure*.

Row ID	Accuracy	F-meas...	Recall	Precision
J48_univariate	0.958	0.746	0.887	0.644
J48_multivariate	0.958	0.746	0.887	0.644
RF_univariate	0.958	0.746	0.887	0.644
RF_multivariate	0.958	0.746	0.887	0.644
NB_univariate	0.947	0.701	0.887	0.58
NB_multivariate	0.947	0.701	0.887	0.58
MLP_univariate	0.958	0.746	0.887	0.644
MLP_multivari...	0.958	0.746	0.887	0.644
Logistic_univa...	0.958	0.746	0.887	0.644
Logistic_multi...	0.958	0.746	0.887	0.644
SVM_univariate	0.958	0.746	0.887	0.644
SVM_multivari...	0.958	0.746	0.887	0.644

Figure 11. Selettori dopo Futures Selection Manuale

3.5 OR: Performance del Modello

La feature selection manuale performata al punto 3.4 ed i relativi risultati in termini di performance, hanno fatto si che venisse confermato quello che era già emerso in fase di analisi di correlazione lineare tra le variabili: le variabili *Schiller* e *Citology* sono quasi esclusivamente le uniche responsabili della previsione. Può essere senz'altro interessante utilizzare le due variabili in una funzione OR manuale e confrontarne il potere predittivo con gli algoritmi di Machine Learning già utilizzati in precedenza.



Mantenendo la stessa strategia di 3-Fold cross validation, i risultati che otteniamo rispecchiano le aspettative: il valore predittivo della variabile *Biopsy* è contenuto in queste due variabili.

J48_Accuracy	0.958
J48_Recall	0.887
J48_F-Measure	0.746
J48_Precision	0.644
OR_Accuracy	0.936
OR_Recall	0.906
OR_Precision	0.535
OR_F-Measure	0.669

Figure 12. Misure J48 vs OR

Rispetto l'algoritmo *J48* la Recall aumenta dell' 1.9% mentre la *Precision* peggiora di quasi l'11%.
La differenza in media della *Precision* non è statisticamente rilevante mentre il numero di *Falsi Positivi* aumentano peggiorando il nostro potere predittivo complessivo.
Probabilmente la soluzione migliore non è affidarsi alla semplice regola empirica ma continuare a sfruttare gli algoritmi di Machine Learning.

3.6 Costo, Gain e ROC

Per poter scegliere tra i predittori visti in precedenza abbiamo supposto di avere una matrice di costo tipica per i problemi con una forte *Class Imbalance*.

		PREDICTED CLASS	
		-1	+1
ACTUAL CLASS	-1	0	1
	+1	100	-1

Figure 13. Matrice di Costo

Avendo questi valore di costo

Row ID	Value
SVM_Cost	579
Schiller_Cost	579
OR_Cost	495

Figure 14. Valore di Costo dei Selettori

saremmo tentati di scegliere l'algoritmo OR. Notiamo però che la differenza tra i valori è minore di 100, che equivale a dire che l'impiego dell'algoritmo OR comporta un *False Negative* in meno. In effetti il problema sta nell'identificare quanto impatto (costo) hanno i *False Positive*. Per capirlo basta calcolare per quale valore di costo associato ai *False Positive* il costo totale dell'algoritmo OR diventa maggiore del costo totale dell'algoritmo *J48*. Utilizzando le *confusion* e la *cost matrix* con incognito il valore di costo dei *FP* abbiamo

$$452 + 43x > 553 + 26x$$

$$x > 5,941$$

Quindi con un costo dei *FP* maggiore o uguale al 6% del costo dei *FN* abbiamo un'inversione di scelta di algoritmo.

Row ID	Value
SVM_Cost	709
Schiller_Cost	709
OR_Cost	710

Figure 15. Valore di Costo dei Selettori con FP=6

Nel nostro caso un *FP* equivale a diagnosticare il cancro ad una persona che non lo ha con conseguente costo per esami futuri e costo nella qualità della vita del paziente. L'analisi quindi con una matrice di costo che tenga in considerazione anche di questa possibilità ci sembra opportuna.

Per essere il più oggettivi possibile quindi prendiamo in considerazione il *Gain Chart* e la *ROC Curve*.

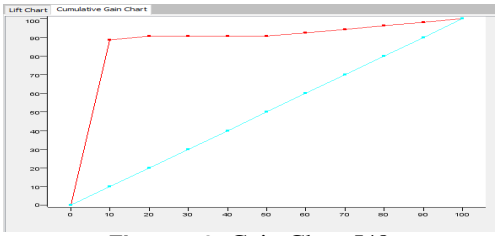


Figure 16. Gain Chart J48

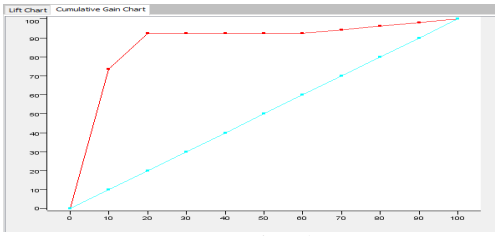


Figure 17. Gain Chart OR

Non riportiamo quello del *SVM* perchè praticamente identico a quello del *J48 (Schiller)*.
I *Gain Chart* in figura sono molto simili. Possiamo dire che per basse percentuali del dataset, fino al 20%, conviene utilizzare il *J48 (Schiller)*: anche se minima torviamo una maggiore percentuale di malati. Dopodichè la scelta è ininfluente.

Analizzando infine la *ROC Curve* per gli algoritmi scelti notiamo, come da teoria, la dipendenza dal *Gain Chart*. Ricordiamo che una curva con area sottesa maggiore equivale ad avere un predittore migliore in quanto avrà una percentuale di *True Positive* maggiore a parità di *False Positive*
In accordo con quanto detto, il *J48* e il *SVM* hanno un'area sottesa la *ROC Curve* maggiore, anche se non di molto, rispetto l'algoritmo *OR*.

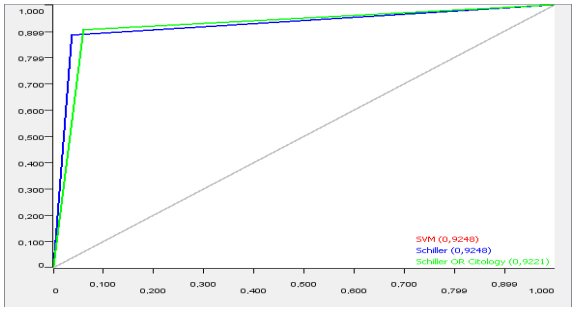


Figure 18. ROC Curves

4. Conclusion and Suggestion

In conclusione tra il *SVM* e il *J48* non ci sono differenze significative tali da indurci a sceglierne uno piuttosto che l'altro. Il *J48* però utilizza un'unica variabile: il risultato del test *Schiller*. In una logica di economia, in cui meno test eseguo e meglio è per il paziente, che subisce un trattamento meno invasivo, e per la sanità pubblica, che risparmia, scelgliamo il *J48*.

L'algoritmo *OR* guadagna *Recall* aumentando il numero di *False Positive*. In pratica aumenta significativamente il numero di pazienti sani predetti come malati per predirne solamente uno in più che è realmente malato. La scelta, come detto precedentemente, ricade sul costo di trattamento di un *False Positive* rispetto ad un *False Negative*. Se il rapporto è maggiore o uguale del 6% allora conviene utilizzare il *J48*. In definitiva, senza conoscere la matrice di costo, ci sentiamo di scegliere il *J48* che è più equilibrato dell'algoritmo *OR*, il quale ha una *Recall* più alta ma non in modo significativo mentre la *Precision* è significativamente più bassa.

Per migliorare il modello predittivo che ha preso in considerazione un unico test medico si consiglia di ampliare il numero del dataset facendo attenzione a non lasciare informazioni mancanti che diminuiscono la grandezza dello stesso. Inoltre è consigliato focalizzarsi maggiormente sulle cause e sui rischi maggiormente descritti in letteratura specializzata riguardo questo tipo di tumore. Sicuramente variabili come l'*età*, *numero di partner sessuali*, *insufficienze immunitarie*, *parenti con tumore della cervice*,... sono interessanti

da approfondire alla ricerca di correlazioni che aiutino nella prevenzione e diagnostica precoce[2].

References

[1] https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/

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

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

[4] https://www.kaggle.com/atakansoztekin/cancer-data-analysis-and-model-implementation

Appendice

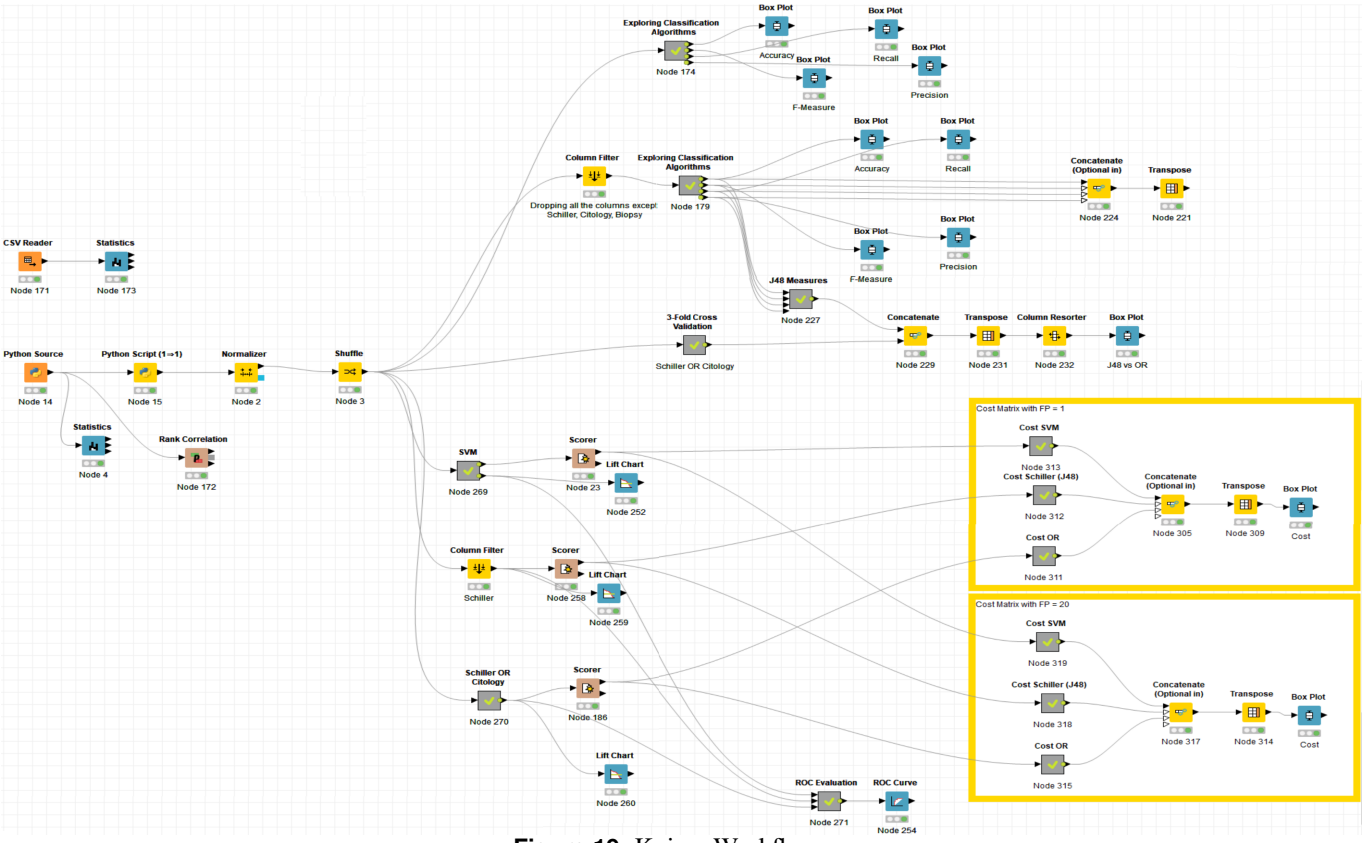


Figure 19. Knime Workflow