Report Machine Learning

Cervical Cancer Risk Factors for Biopsy

Pietro Carmine Valenti (807548), Fabio Pasquale Lombardi (860550), Paolo Lindia (860507)

Università degli Studi di Milano-Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione (DISCO)
CdLM in Data Science

Indice

1 Introduzione **3**

2 *Dataset description* **3**

3 *Data preprocessing* **4**

3.1 Analisi esplorativa dei dati 4

3.2 Selezione delle variabili 5

3.3 Trattamento dei valori mancanti 5

3.4 Normalizzazione 5

4 *Class imbalance problem* **5**

4.1 *Oversampling* 6

4.2 *Cost-Sensitive learning* 6

5 *Data modeling* **7**

6 *Evaluation* **7**

6.1 Regressione logistica e SMO 9

6.2 Rilevanza delle variabili 10

7 Conclusioni e possibili sviluppi futuri **11**

Abstract

Negli ultimi anni la *data analysis*, in particolare il *machine learning*, sono diventati strumenti essenziali per lo sviluppo dei diversi settori dell'attività umana; uno di quelli che più ne beneficia è senza dubbio il campo medico e sanitario. La possibilità di prevedere o classificare patologie rende l'analisi dei dati un sostegno essenziale agli studi in campo medico: la classificazione per la presenza di malattie e l'individuazione dei fattori maggiormente influenti sono solo alcuni esempi di ciò che lo studio dei dati può offrire a tale settore. Il presente report riassume un procedimento di classificazione per la presenza o meno di cancro alla cervice uterina, una grave malattia che nuoce alla salute di molte donne; sono stati sviluppati 6 classificatori differenti al fine di ottimizzare al massimo la capacità di prevedere la presenza di tale patologia.

1 Introduzione

Una neoplasia o tumore indica, in patologia, «*una massa di tessuto che cresce in eccesso ed in modo scoordinato rispetto ai tessuti normali, e che persiste in questo stato dopo la cessazione degli stimoli che hanno indotto il processo*» - *Rupert Allan Willis* [4]. Una neoplasia può essere classificata secondo diverse caratteristiche:

- Il tipo istologico delle cellule proliferanti;
- L'aggressività e il decorso clinico previsto: tumori benigni (non cancerosi) e tumori maligni (cancerosi o cancro);
- La staduazione tumorale per quanto riguarda i tumori maligni.

Il carcinoma alla cervice uterina rientra nella classe dei tumori maligni. Nei decenni scorsi tale patologia era conosciuta come un killer silenzioso tra le donne perché non presenta alcun sintomo fino a quando non è in una fase avanzata. Ciò rendeva la malattia difficile da individuare precocemente e complicava il trattamento. Lo sviluppo di nuove tecniche quali lo *screening* e il *Pap-test*, il tasso di mortalità per cancro alla cervice uterina è diminuito significativamente del 74% tra il 1955 e il 1992. Sebbene tale patologia oggi sia più prevedibile, ogni anno, negli Stati Uniti, continuano ad essere diagnosticati 11.000 casi, dei quali circa il 35% decede; mentre nel mondo l'ammontare dei decessi sfiora i 300.000 ogni anno. Diverse sono le cause di questa patologia, tra le quali: la genetica, le malattie sessualmente trasmissibili, il fumo, l'immunodepressione e l'uso di contraccettivi. Uno studio di un modello predittivo (di classificazione) completamente basato su queste cause potrebbe rivelarsi utile per fornire un sostegno al medico, ai test e all'individuazione delle cause più rilevanti.

Nella presente analisi si è cercato di dare una risposta a tali temi, sviluppando un processo di classificazione (composto da 6 metodi) atto a prevedere il possibile risultato della biopsia: i dati di partenza sono afferenti a 858 pazienti tailandesi, di età compresa tra i 13 e gli 84 anni, delle quali si sono rilevate 32 variabili potenzialmente influenzati. Tra le possibili variabili *target* (risultato del test citologico, di Schiller, di Hinselmann e della biopsia) è stato deciso di impiegare la biopsia poiché il responso dà la (quasi) certezza della presenza di cancro. Si è inoltre deciso di non utilizzare i 3 test come variabili esplicative poiché avrebbero influenzato pesantemente il modello, togliendo rilevanza alle altre. Poichè il principale problema nella struttura dei dati risulta la *class imbalance*, ovvero i pazienti positivi sono solamente il 6.4% del totale, sono stati applicati due diversi metodi per la risoluzione di tale criticità: l'*oversampling* e *Cost-Sensitive Learning*.

2 *Dataset description*

Si è scelto di utilizzare la versione 6 del dataset "Cervical Cancer Risk Classification", reso disponibile sulla piattaforma Kaggle [5]. 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 De-

vice).
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. La biopsia viene eseguita al fine di confermare (Biopsy=1) o escludere (Biopsy=0) un sospetto di malattia basato su uno dei tre precedenti test (*Hinselmann, Schiller, Citology*)

3 *Data preprocessing*

La prima fase dello studio consiste nell'insieme delle analisi esplorative e nella risoluzione delle eventuali problematiche legate alla struttura dei dati.

3.1 Analisi esplorativa dei dati

Durante questa fase si è prestata attenzione alla forma dei dati originali e alle principali statistiche descrittive, concentrandosi sia sulla variabile *target*, sia sulle variabili indipendenti; dall'analisi delle statistiche descrittive e delle *scatter-matrix* non sono emerse particolari inconsistenze o anomalie nei dati oltre alla presenza di valori mancanti e di classi sbilanciate per la variabile *target* (93.6% di negativi contro il 6.4% di positivi). Le analisi effettuate sono:

- Studio delle principali statistiche descrittive per le variabili numeriche.
- Analisi grafica della dispersione e della relazione tra le variabili (*scatter-matrix*).
- Analisi grafica della distribuzione delle variabili numeriche attraverso Box plot, e delle va-

riabili binarie (compresa la variabile target) mediante grafico a barre.

- Analisi delle correlazioni lineari tra le variabili numeriche (correlogramma).

3.2 Selezione delle variabili

La *feature selection* è una fase fondamentale per lo sviluppo di un buon modello di classificazione, e risulta ancora più cruciale quando si hanno a disposizione un numero consistente di variabili. Come procedimento si è optato per una rimozione ragionata di alcune *features* poiché il loro contenuto informativo non ne giustifica la presenza nel modello. Tali variabili risultano:

- Le due variabili numeriche *Time since first diagnosis* e *Time since last diagnosis* poiché sono caratterizzate da una presenza di valori mancanti che eccede di gran lunga la soglia scelta del 10% (787 *missing* su 858).
- Le variabili *STDs: cervical condylomatosis*, *STDs: vaginal condylomatosis*, *STDs: AIDS*, *STDs: Hepatites B*, *STDs: HPV*, *STDs: molluscum contagiosus*, *STDs: genital herpes*, *STDs: pelvic infiammatory disease* poichè caratterizzate da la quasi totale assenza di osservazioni positive: in particolare è stata utilizzata come soglia 0,05% di presenza di valori positivi sul totale.
- *STDs, STDs: number, STDs: number of diagnosis* poichè riassumono le altre STDs.
- IUD poichè variabile binaria la cui informazione è contenuta anche in IUD (Years).
- *Smokes* poichè variabile binaria la cui informazione è contenuta anche in *Smokes(Years), Smokes(Packs/Years).*
- Dx poichè variabile binaria la cui informazione è contenuta anche in *Dx:Cancer, Dx:HPV, Dx:HIV.*
- *Hormonal contraceptives* poichè variabile binaria la cui informazione è contenuta anche in *Hormonal contraceptives(Years).*
- Le tre variabili target alternative: Hinselmann, Schiller e *citology.*

Eliminate tali variabili il Dataset finale è composto di 14 variabili esplicative.

3.3 Trattamento dei valori mancanti

Questa fase è stata caratterizzata in un primo momento dalla discretizzazione della variabile *Age*; la quale è stata suddivisa in 6 classi di uguale numerosità attraverso il metodo *Equal frequency unsupervised discretization.* La discretizzazione è stata di supporto all'imputazione dei valori mancanti, poiché si è scelto di sostituire questi ultimi con la mediana condizionata alle classi d'età appartenenti.

3.4 Normalizzazione

Il processo di normalizzazione è stato applicato alle variabili quantitative, al fine di eliminare il cosiddetto effetto di scala tra le differenti variabili, il quale potrebbe influenzare la classificazione della variabile *target.* Per fare ciò è stato applicata la *Z-Score Normalization*:

$$z_i = \frac{x_i - \mu}{\sigma}$$

con x_i i-esimo record della variabile, μ media e σ deviazione standard.

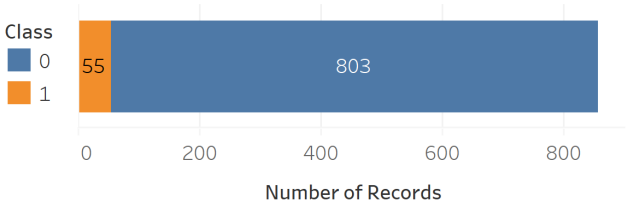
la quale restituisce per ogni variabile numerica, una distribuzione avente media 0 e varianza 1.

4 *Class imbalance problem*

Una tra le principali criticità che ha caratterizzato il progetto risulta quella delle classi sbilanciate, problematica con il quale si ha spesso a che fare quando si sviluppano analisi riguardanti malattie rare; nello specifico i dati presentano un forte sbilanciamento della variabile *target* verso la classe "assenza di cancro" (classe *0*), la quale si manifesta nel 93.6% delle osservazioni. In assenza di provvedimenti adeguati volti a ribilanciare, almeno in parte, il dataset vi è il rischio che il classificatore non riesca a prevedere correttamente la classe minoritaria (classe *1*), la cui corretta classificazione, in caso di presenza di malattie, è cruciale.

Per ovviare a tale problematica si è optato per due metodi di risoluzione differenti:

- Sovracampionamento (*Oversampling*)
- Apprendimento basato sulla matrice dei costi (*Cost-Sensitive learning*)



Number of records for each class

4.1 *Oversampling*

Il ricampionamento consiste in tecniche basate sull'utilizzo di sottoinsiemi di dati i quali possono essere selezionati casualmente o secondo una procedura sistematica, allo scopo di approssimare alcune caratteristiche della distribuzione. In particolare, nel caso di classi sbilanciate, tale approccio viene applicato per il ribilanciamento delle classi; per fare ciò esistono due approcci differenti: l'*undersampling*, che consiste nell'eliminazione casuale di alcune osservazioni della classe maggioritaria, e l'*oversampling*, il quale invece aggiunge osservazioni alla classe minoritaria. Per il progetto si è optato per il secondo metodo in quanto l'esigua numerosità di osservazioni non avrebbe consentito di applicare un sottocampionamento.

Per applicare l'*oversampling* si è deciso per l'utilizzo dell'algoritmo *Synthetic Minority Oversampling Technique* (SMOTE), il quale si basa su un metodo più elaborato rispetto alla semplice aggiunta di osservazioni tramite campionamento casuale con reinserimento. Lo SMOTE genera osservazioni "sintetiche" a partire dalla classe di minoranza e li aggiunge al set di dati esistenti. I record artificiali della classe di minoranza vengono generati basandosi sulla similarità nello spazio dei predittori. Per ciascun record x_i appartenente alla classe di minoranza vengono creati k osservazioni delle quali vengono selezionate solo le più simili (Chawala [et al.], 2002 [1]). Il metodo per definire le nuove osservazioni si basa sull'algoritmo *K-Nearest Neighbour* (KNN), il quale seleziona le k osservazioni (della classe minoritaria) più vicine ad x_i e ne viene selezionato casualmente uno (x_j). A questo punto viene generata la nuova osservazione tramite la relazione:

$$x_s = x_i + (x_j - x_i)\delta_s$$

dove x_s rappresenta la nuova osservazione generata e δ_s un peso casuale nell'intervallo [0,1].

Per ovviare al rischio di ottenere stime troppo ottimistiche (*overly-optimistic*) si è deciso di applicare lo SMOTE solamente al *training-set*; tale decisione deriva dalla forte probabilità di *overfitting* derivante dall'utilizzo indiscriminato sul dataset completo, e in particolare alla "corruzione" del *test-set* (Santos M. S. [et al.], 2018 [3]).

4.2 *Cost-Sensitive learning*

Il secondo metodo utilizzato per il ribilanciamento delle classi della variabile *target* si basa sull'utilizzo di una matrice dei costi: il tipo di classificazione che ne deriva viene denominato *Cost-Sensitive Learning*. L'assunzione che ne giustifica l'impiego è che, per lo specifico caso della diagnosi di malattie, sia molto più grave ottenere un falso negativo (FN) che un falso positivo (FP); a seguito di questo ragionamento è possibile indirizzare l'algoritmo di classificazione in modo da ridurre il rischio di un'osservazione classificata erroneamente negativa.

Come anticipato il metodo si serve di una matrice dei costi 2x2 la quale ha sulla diagonale *true positive/negative* pesi pari a 0 (non esiste un costo per classificazioni corrette), mentre sulla diagonale *false positive/negative* due valori C_{FP} e C_{FN} equivalenti ai costi sostenuti in caso di errata classificazione. Per le caratteristiche intrinseche del fenomeno studiato, è stato ritenuto necessario dare un costo superiore ai FN rispetto ai FP. La *Cost-Matrix* assume quindi la seguente forma:

		Prediction outcome p	**Prediction outcome** n
actual value	p′	TP Cost:0	FN Cost:10
	n′	FP Cost:1	TN Cost:0

La scelta del costo 10 per i FN ha una duplice motivazione:

- La prima è, come sopracitato, l'assegnazione di un peso molto maggiore per i FN.
- La seconda è di carattere tecnico: in seguito ad una serie di iterazioni, nel quale viene cambiato il valore di C_{FN}, il valore 10 risulta essere ottimale in termini di miglioramento della performance dei classificatori misurata tramite *Sensitivity*, senza un eccessiva perdita di *Accuracy*.

Va tuttavia evidenziato che C_{FN} ha un significato puramente teorico in quanto il peso dell'eventualità di un FN non ha un costo materiale e/o quantificabile; è stato ritenuto maggiormente grave prevedere il risultato negativo di una biopsia quando il paziente presenta la malattia.

5 *Data modeling*

Il corpo del progetto consiste in un sistema di vari algoritmi di classificazione il cui impiego è volto all'individuazione del miglior classificatore per il fenomeno in analisi.

Sono stati utilizzati classificatori appartenenti a quattro diversi gruppi:

- Modelli euristici: sono modelli che forniscono risultati approssimati, ma spesso risultano meno onerosi computazionalmente e in termini di assunzioni. In particolare sono stati impiegati 2 classificatori: *Random Forest* e *Decision Tree* (J48-Weka)
- Modelli separatori: sono modelli volti all'indivuazione di funzioni matematiche che separino lo spazio delle osservazioni in maniera ottimale, ovvero al fine di minimizzare il numero di previsioni errate. In particolare sono stati applicati 2 classificatori: il *Sequential minimal optimization* (SMO), ovvero una particolare tipologia di *Support Vector Machine*, e il *Multi-Layer Perceptron* (MLP), una rete neurale applicata con 10 neuroni.
- Modelli di regressione: sono classificatori basati sull'utilizzo di regressioni statistiche. In particolare è stata applicata la *Logistic Regression*, in cui la variabile dipendente (dicotomica) viene ricondotta ad una distribuzione binomiale e l'algoritmo di previsione si basa sull'impiego di probabilità condizionate.
- Modelli probabilistici: sono modelli che si servono di un sistema probabilistico basato sul teorema di Bayes:

  $$P[A_i|B] = \frac{P[B|A_i] * P[A_i]}{\sum_{j=1}^{N} P[B|A_j] * P[A_j]}$$

  con $P[A_i]$ probabilità a priori, $P[B|A_i]$ verosimiglianze, $P[A_i|B]$ probabilità a posteriori. In particolare è stato impiegato il classificatore *Naive Bayes*.

Tutti e 6 i modelli descritti sono stati implementati sia con il supporto del metodo di *oversampling* (Sez. 4.1), sia sotto forma di *Cost-Sensitive Learning* (Sez. 4.2).

6 *Evaluation*

Dopo il processo di classificazione appena descritto si è passati alla fase di valutazione della *performance* dei modelli. A tale scopo sono stati impiegati indicatori e grafici basati sulla *Confusion Matrix*; definiti TN=*True Negative*, TP=*True Positive*, FN=*False Negative*, FP=*False Positive* sono stati utilizzati:

- *Accuracy*: è un indicatore che misura l'accuratezza di previsione dei modelli. Viene calcolata come la percentuale di osservazioni classificate correttamente sul totale delle previsioni:

  $$Acc = \frac{TN + TP}{TN + TP + FN + FP}$$

- *Sensitivity/Recall*: rappresentano la capacità del classificatore di prevedere i valori positivi (*True Positive Rate*, TPR):

  $$Sensitivity = \frac{TP}{TP + FN} = Recall$$

- *Specificity*: rappresenta la capacità del classificatore nel prevedere correttamente i valori negativi (*True Negative Rate*, TNR):

$$Specificity = \frac{TN}{TN + FP}$$

- *Precision*: rappresenta la quota di valori positivi previsti correttamente sul totale delle previsioni positive (indicatore dell'affidabilità nella previsione dei positivi):

$$Precision = \frac{TP}{TP + FP}$$

- *F-Measure*: è un indicatore composto il quale fornisce una valutazione generale della qualità del classificatore. Si calcola come media armonica tra *Recall* (R) e *Precision* (P):

$$F = \frac{2 * R * P}{R + P}$$

- F_2-*Measure*: poiché si è optato per privilegiare la *Recall* è stata utilizzata anche una trasformata F_β (con $\beta = 2$) che ricalcola la *F-Measure* come segue:

$$F = \frac{(\beta^2 + 1) * R * P}{R + \beta^2 * P}$$

- ROC-*Curve* e AUC: la curva ROC (*Receiver Operating Characteristic*) è un grafico che evidenzia la capacità del classificatore di fare previsioni meglio o peggio di una classificazione casuale; sull'asse delle ordinate troviamo il *True Positive Rate* (TPR, che equivale alla *sensitivity*) e su quello delle ascisse il *False Positive Rate* (FPR, che equivale a 1-*specificity*). L'AUC (*Area Under Curve*) rappresenta la porzione di spazio sottostante alla ROC; è un indicatore sintetico riferito alla curva e rappresenta la probabilità che un modello classifichi un'istanza positiva scelta casualmente meglio di un'istanza negativa scelta anch'essa in maniera casuale.

Nonostante la valutazione della qualità dei classificatori sia avvenuta mediante numerosi indicatori vengono ritenuti maggiormente informativi *Sensitivity/Recall*, *F-Measure* e $F_2-Measure$; le motivazioni, come anticipato nelle precedenti sezioni, dipendono dalla natura del fenomeno, e dall'importanza che si da al minimizzare il numero di previsioni erroneamente classificate come negative (FN).

Di seguito vengono mostrati i risultati in due tabelle, una per ciascun metodo di *Class imbalance problem solving* (Sez. 4), le quali mostrano tutte le statistiche appena descritte, per ciascuno dei 6 modelli di classificazione.

Evaluation (Oversampling)

Classifier	Recall	Precision	Sensitivity	Specifity	F-measure	F2-measure	Accuracy	AUC
Random Forest	0.055	0.09	0.055	0.962	0.069	0.0602	0.904	0.521
SMO	0.333	0.15	0.333	0.871	0.207	0.268	0.837	0.602
Logistic Regression	0.333	0.128	0.333	0.845	0.185	0.252	0.812	0.565
Multilayer Perceptron	0.222	0.105	0.222	0.871	0.143	0.182	0.83	0.62
Decision Tree (J48)	0.111	0.118	0.111	0.943	0.114	0.112	0.89	0.561
Naive Bayes	0.278	0.106	0.278	0.841	0.154	0.210	0.805	0.594

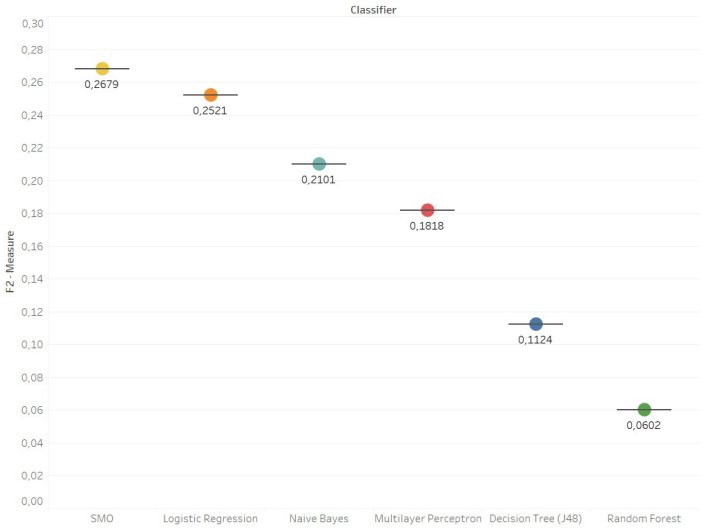
Evaluation (Cost-Sensitive Learning)

Classifier	Recall	Precision	Sensitivity	Specifity	F-measure	F2-measure	Accuracy	AUC
RandomForest	0.056	0.091	0.056	0.962	0.069	0.06	0.904	0.556
SMO	0.278	0.147	0.278	0.890	0.192	0.236	0.851	0.584
Logistic Regression	0.278	0.139	0.278	0.882	0.185	0.231	0.844	0.526
Multilayer Perceptron	0.111	0.0769	0.111	0.909	0.0909	0.102	0.858	0.547
Decision Tree (J48)	0.111	0.074	0.111	0.905	0.089	0.101	0.855	0.508
Naive Bayes	0.167	0.071	0.167	0.852	0.1	0.131	0.809	0.551

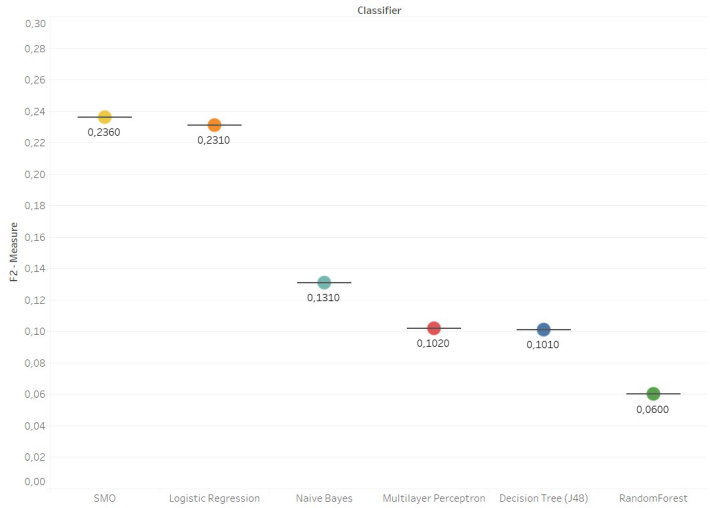
Il confronto tra le due tabelle fornisce una duplice informazione circa la *performance* delle tecniche di classificazione:

- La prima riguarda i classificatori più performanti, ovvero SMO e Regressione Logistica; seguendo il ragionamento esposto precedentemente si può considerare la F_2-*Measure* come indicatore più significativo al fine della valutazione della *performance* dei modelli.
- La seconda riguarda il miglior metodo per la risoluzione dello sbilanciamento delle classi nella variabile dipendente *Biopsy*; si può notare come generalmente, e in particolare per SMO e logistica, gli indicatori mostrino come il metodo di *Oversampling* (SMOTE) dia risultati più allettanti rispetto al *Cost-Sensitive Learning*.

Di seguito vengono mostrati due grafici nel quale si mettonoo in risalto le differenze per l'indicatore F_2-*Measure*, il primo per l' *Oversampling*, il secondo per il *Cost-Sensitive learning*



F_2-Measure Box Plot (Oversampling)



F_2-Measure Box Plot (Cost-Sensitive Learning)

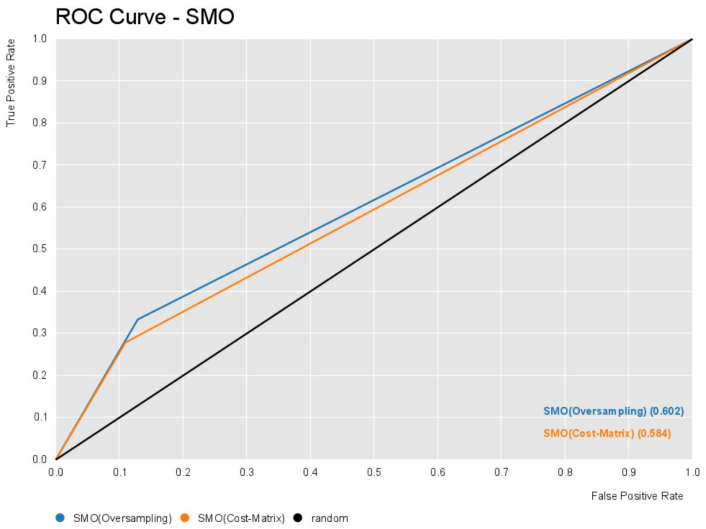
Ancora una volta viene messa in luce la superiorità degli algoritmi SMO e logistica, e in particolare del metodo che impiega la tecnica di SMOTE.

6.1 Regressione logistica e SMO

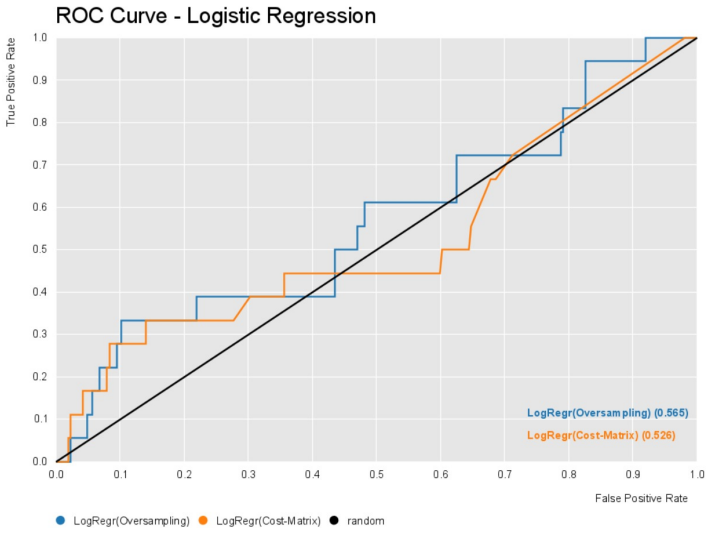
Può risultare interessante presentare un confronto tra i due classificatori che hanno dato i risultati migliori in termini di *performance*: la Regressione Logistica e la SMO. Vengono sotto presentati la tabella riguardante gli indicatori sulla qualità della classificazione e i grafici che confrontano le curve ROC. Si presta particolare attenzione al confronto tra gli stessi due classificatori applicati sia con lo SMOTE, sia con la *Cost-Matrix*.

SMO and Logistic Regression Comparison

Classifier	Recall	Precision	Sensitivity	Specifity	F-measure	F2-measure	Accuracy	AUC
Seq. Minimal Optim. (Oversampling)	0.333	0.15	0.333	0.871	0.207	0.268	0.837	0.602
Logistic Regression (Oversampling)	0.333	0.128	0.333	0.845	0.185	0.252	0.8125	0.565
Seq. Minimal Optim. (Cost-Matrix)	0.278	0.147	0.278	0.890	0.192	0.236	0.851	0.584
Logistic Regression (Cost-Matrix)	0.278	0.139	0.278	0.882	0.185	0.231	0.844	0.527



ROC-Curve: Sequential minimal optimization



ROC-Curve: Logistic Regression

Nuovamente si può notare la superiorità in termini di indicatori e ROC-*Curve* dei modelli che utilizzano l*Oversampling* ed in particolare dell'algoritmo *Sequential minimal optimization*.

6.2 Rilevanza delle variabili

Lo scopo del progetto non si limita solamente all'individuazione del miglior algoritmo per prevedere il risultato di una biopsia; risulta infatti chiaro che anche il miglior modello di classificazione non può sostituire il parere di un esperto del settore, e che non si può affidare ad una tecnica di apprendimento automatizzato decisioni che riguardano la vita di un paziente. Proprio per tale ragione il riscontro di una *performance* non particolarmente allettante non pregiudica l'importanza dei risultati ottenuti.

La tematica che probabilmente può dare un contributo maggiore allo sviluppo e al progresso del settore medico riguarda l'individuazione dei principali fattori influenzanti, ovvero determinanti per l'aumento della probabilità di contrarre la malattia. A tale fine si è deciso di estrarre l'influenza che ciascuna variabile ha sulla probabilità di ottenere un risultato positivo alla biopsia (e quindi di contrarre il cancro alla cervice), utilizzando i due classificatori che forniscono informazioni afferenti a tale scopo: il *RandomForest* e la Regressione Logistica.

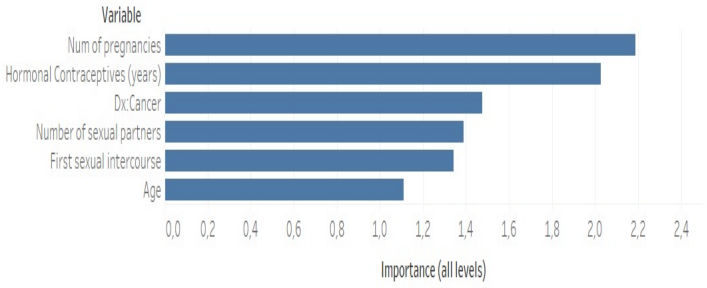
Per quanto riguarda il primo l'informazione desiderata risulta ricavabile dalla divisione dei nodi degli alberi (*splits*); in particolare l'algoritmo di *RandomForest* implementato dal software Knime fornisce informazioni circa gli *splits* e i *candidates* per i primi 3 livelli. L'importanza di ciascuna variabile all'interno del modello viene calcolata a ciascun livello con la formula:

$$Importance_i = \frac{splits_i}{candidates_i}$$

dove con i si intende il livello i-esimo (i=0,1,2). Dopo avere calcolato l'*importance* per ogni variabile su ciascuno dei tre livelli, è stato creato un indice aggregato dato dalla somma di tali valori:

$$Importance_{tot} = \sum_{i=0}^{2} Importance_i$$

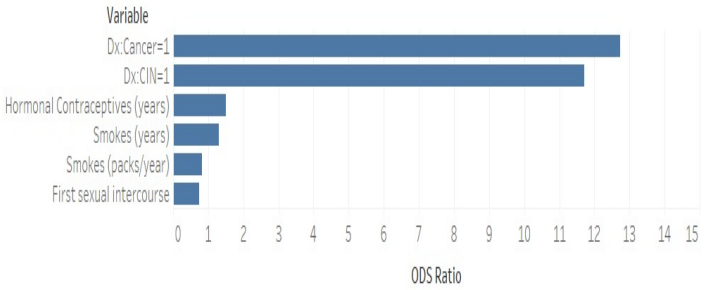
A questo punto sono state estratte le 6 variabili con il valore di $Importance_{tot}$ più elevato, le quali sono rappresentate nel grafico sottostante



Feature Relevance: RandomForest

La criticità di tale tecnica è che non si hanno né informazioni circa il moltiplicatore che determina un aumento (o una diminuzione) della probabilità di avere un risultato della biopsia positivo, né circa il segno della relazione tra le variabili esplicative e la variabile *target*.

Il secondo metodo, basato su regressione logistica, trova fondamento sull'analisi dell'*Odds-Ratio*, indicatore ottenuto come esponenziale dei coefficienti della regressione stessa; per una maggiore consistenza dei risultati si è deciso di selezionare solo i coefficienti ritenuti significativi dal *p-value* (<0.2) basato sullo *z-score* (indicatore di significatività dei coefficienti). I risultati sono espressi nel grafico sottostante:



Feature Relevance: Logistic Regression

L'*Odds-Ratio* da informazione sia circa il moltiplicatore, sia circa il segno della relazione: un valore compreso tra 0 e 1 indica una relazione negativa, ovvero una variabile il cui aumento porta ad una riduzione della probabilità di risultare positivi ad una biopsia, mentre un rapporto maggiore di uno riceve un'interpretazione opposta. L'*Odds-ratio* indica inoltre il moltiplicatore che rappresenta l'aumento di probabilità di risultato positivo alla biopsia.

Per la maggiore informatività sui coefficienti di rilevanza delle variabili, la logistica può essere considerata il metodo migliore per la valutazione dei fattori maggiormente influenzanti; in particolare si evidenzia un fortissimo perso delle variabili *Dx:Cancer* (=1) e *Dx:CIN* (=1), riferite rispettivamente alla presenza di ulteriori forme di cancro e alla presenza di neoplasia intraepiteliale cervicale. Dai risultati si evince inoltre una tendenziale diminuzione della probabilità di contrarre cancro alla cervice all'aumentare dell'età del primo rapporto sessuale. L'unico coefficiente che sembra scontrarsi con la logica e con l'evidenza scientifica riguarda il numero di pacchetti di sigarette all'anno, il cui *Odds-Ratio* mostra una relazione negativa.

7 Conclusioni e possibili sviluppi futuri

Nonostante la *performance* predittiva dei classificatori non abbia evidenziato risultati di livello elevato, lo studio presenta comunque una buona rilevanza per quanto concerne i fattori influenti. L'eccessivo sbilanciamento delle classi non ha infatti consentito di ottenere un sistema di classificazione capace di massimizzare il potere predittivo dei modelli impiegati.

L'individuazione di un'insieme di fattori, che influenzano la probabilità di contrarre il cancro alla cervice uterina, risulta invece componente estremamente utile a sostegno del settore medico sanitario; lo studio presentato infatti mira a fornire un supporto agli esperti di dominio, i quali possono beneficiare di un'analisi dei dati che consenta loro di sensibilizzare riguardo il tema trattato. I pazienti che si riconoscano come appartenenti ad una categoria a rischio, in base a tali fattori, potrebbero quindi essere indirizzati verso un monitoraggio più frequente (ad esempio tramite analisi periodiche come il PAP-test), e ad una maggiore attenzione verso comportamenti potenzialmente rischiosi.

Una maggiore disponibilità di dati, utili al miglioramento dell'analisi, potrebbe sicuramente migliorare la capacità del modello, in quanto si sarebbe in condizione di applicare tecniche meno distorsive per la soluzione delle problematiche legate alle classi sbilanciate (*undersampling*). Un'ulteriore possibile sviluppo potrebbe consistere in una raccolta di dati maggiormente ramificata (in termini geografici, di età...), e che consenta allargare il campo visivo rispetto a tale tematiche.

Riferimenti bibliografici

[1] Chawala N. W., Bowyer K., *SMOTE: Synthetic Minority Over-sampling Technique*, art. in *Journal of Artificial Intelligence Research*, DOI: 10.1613/jair.953, 2002.

[2] Klersy C., Scudeller L., *Interpretare i modelli prognostici multivariati: il modello logistico* in GIAC, Vol. 5, Numero 4, 2002.

[3] Santos M. S., Soares J. P., Abreu P. H., Araujo H., Santos J., *Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches*, art. in *IEEE Computational Intelligence Magazine*, DOI: 10.1109/MCI.2018.2866730, 2018.

[4] In *Robbins Basic Pathology*, 8ª edizione, Saunders/Elsevier 2007, cap. 6.

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