Confronto tra modelli di classificazione per la previsione del cancro alla cervice

Marco Distrutti[1], Nada Ksha[1], Artjoms Olzojevs[1], Luca Maggi[1]

Sommario

L'analisi è stata svolta su un dataset fornito dall'*Hospital Universitario de Caracas in Venezuela*, riguardante i fattori di rischio del cancro alla cervice, studiando dati anagrafici, storia clinica dei pazienti e del loro stile di vita.
Poiché il dataset è risultato essere sbilanciato, le misure adottate nell'analisi sono state adeguate a questa condizione. Applicando tecniche di classificazione di diversa natura per potere individuare le pazienti affette da cancro alla cervice, sono stati ricavati otto modelli di machine learning. La validazione dei modelli è stata svolta con un processo di *Cross Validation*, prima a 5 partizioni e in seguito a 10. Successivamente, è stato eseguito un confronto tra i classificatori e, ai migliori tra essi, è stato applicato un *Wrapper* di *Feature Selection*. Due modelli in particolare, *Random Forest* e *Support Vector Machine,* hanno infine primeggiato. La selezione degli attributi ha mostrato una forte relazione tra una diagnosi positiva di cancro alla cervice e l'esposizione al Papilloma Virus Umano (HPV).

Keywords
Machine Learning — Medicine — Cervical Cancer — STDs

[1]*Dipartimento di Informatica, Università degli Studi di Milano-Bicocca, CdLM: Data Science*

Indice

1. Introduzione

Il cancro alla cervice è considerato uno dei tumori più diffusi in tutto il mondo. È una tipologia di tumore asintomatica, soprattutto nelle sue prime fasi. È perciò necessario prevenire la malattia grazie agli esami medici. In questo report si analizza un dataset di 36 variabili e 857 record, riguardanti dati anagrafici, storia clinica dei pazienti e del loro stile di vita.
Il report è strutturato nel seguente modo:
dopo aver introdotto il dataset, sono state descritte le variabili ed è stata svolta un'analisi preliminare sui dati, in seguito sono stati riportati i modelli utilizzati e le misure di performance calcolate. In ultimo, sono state riportate le analisi ed i risultati ottenuti.

2. Descrizione del dataset

Descrizione delle variabili:

1. Age: età
2. Number of sexual partners: numero dei partner sessuali
3. First sexual intercourse: età primo rapporto sessuale
4. Num of pregnancies: numero (quantità) di gravidanze
5. Smokes: variabile binaria (Y/N) che indica se si tratta di una fumatrice
6. Smokes (years): numeri di anni da fumatrice
7. Smokes (packs/year): quantità dei pacchetti di sigarette consumate/all'anno
8. Hormonal Contraceptives: variabile binaria che indica l'utilizzo o meno di contraccettivi ormonali
9. Hormonal Contraceptives (years): durata degli anni dell'utilizzo di contraccettivi ormonali
10. IUD: variabile binaria che indica l'utilizzo della spirale intrauterina
11. IUD (years): durata degli anni di utilizzo della spirale intrauterina
12. STDs: variabile binaria che indica la presenza o meno di una malattia sessualmente trasmissibile
13. STDs (number): somma delle malattie sessualmente trasmissibili
14. STDs - condylomatosis: variabile binaria che indica la presenza di almeno una delle due forme (cervicale o vaginale) di tale malattia (sessualmente trasmissibile)

15. STDs - cervical condylomatosis: variabile binaria che indica la presenza di tale malattia (sessualmente trasmissibile)
16. STDs - vaginal condylomatosis: variabile binaria che indica la presenza di tale malattia (sessualmente trasmissibile)
17. STDs - vulvo-perineal condylomatosis: variabile binaria che indica la presenza di tale malattia (sessualmente trasmissibile)
18. STDs - syphilis: variabile binaria che indica la presenza di tale malattia (sessualmente trasmissibile)
19. STDs - pelvic inflammatory disease: variabile binaria che indica la presenza di tale malattia (sessualmente trasmissibile)
20. STDs - genital herpes: variabile binaria che indica la presenza di tale malattia (sessualmente trasmissibile)
21. STDs - molluscum contagiosum: variabile binaria che indica la presenza di tale malattia (sessualmente trasmissibile)
22. STDs - IDS: variabile binaria che indica la presenza di tale malattia (sessualmente trasmissibile)
23. STDs - HIV: variabile binaria che indica la presenza di tale malattia (sessualmente trasmissibile)
24. STDs - Hepatitis B: variabile binaria che indica la presenza di tale malattia (sessualmente trasmissibile)
25. STDs - HPV: variabile binaria che indica la presenza di tale malattia (sessualmente trasmissibile)
26. STDs - Number of diagnosis: somma totale delle malattie sessualmente trasmissibile
27. STDs - Time since first diagnosis: numero di anni dalla prima diagnosi di almeno una delle malattie sessualmente trasmissibili
28. STDs - Time since last diagnosis: numero di anni dall'ultima diagnosi di una delle malattie sessualmente trasmissibili
29. Dx - Cancer: variabile binaria che indica la presenza di un cancro
30. Dx -CIN: variabile binaria che indica la prensenza o meno di *Cervical intraepithelial neoplasia*
31. Dx - HPV: variabile binaria che indica la prensenza o meno dell'*HPV*
32. Dx - variabile binaria che indica la prensenza o meno di almeno una tra la variabile Dx- HPV e Dx-CIN
33. Hinselmann: variabile binaria che indica l'esecuzione del test diagnostico. Serve per esaminare una zona illuminata della vista della cervice, vagina e vulva.
34. Schiller: variabile binaria che indica l'esecuzione o meno di un esame medico che consiste nell'avere una soluzione con iodio che viene applicata alla cervice per diagnosticare il cancro alla cervice
35. Cytology: target variable: variabile binaria che indica l'esecuzione o meno del cosìddetto PaP test, il quale serve per indicare una possibile presenza anormale di cellule nella cervice che possono portare a sviluppare un cancro.
36. Biopsy: è una variabile binaria che indica l'esecuzione o meno di una procedura chirurgica nella quale si asporta una parte del tessuto della cervice per esaminare l'anomalità riscontrata durante l'esame della *cytology*.

2. Preprocessing

Prima di procedere con l'implementazione delle tecniche di Machine Learning, è stata eseguita una fase di *Preprocessing*.

Analisi degli outlier

Per le variabili discrete è stata effettuata un'analisi degli outlier. Il risultato dell'analisi ha mostrato la presenza di pochi outlier e per tale motivo i record sono stati studiati singolarmente procedendo con l'eliminazione di quelli errati sulla base delle conoscenze di dominio.

Analisi dei Missing Values

È stato ispezionato il dataset con l'obiettivo di identificare i *Missing Values*. Si è notato che 124 record presentavano su più attributi numerosi *Missing Values*: dopo un'ulteriore analisi è stato deciso di eliminarli in quanto ritenuti non significativi. Per i record che invece contenevano pochi *Missing Values* è stata effettuata la sostituzione con la mediana per gli attributi numerici e con la moda per gli attributi binari perché trattasi del metodo più conveniente su un dataset sbilanciato.

Valutazione dello sbilanciamento del dataset

Lo studio delle frequenze assunte dalla variabile di classe ha evidenziato la natura fortemente sbilanciata del dataset, come mostrato nella Figura 1.

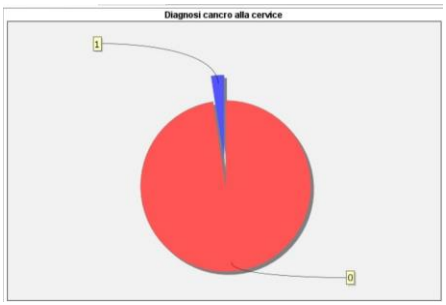

Figura 1: Sbilanciamento Variabile di Classe

Analisi correlazione e selezione delle variabili

Alcune variabili evidentemente non correlate con l'insorgenza di un cancro sono state eliminate. Queste riguardano gli attributi inerenti ad esami svolti per accertare la presenza di un cancro. Inoltre due variabili, "STDs - Time since first diagnosis" e "STDs - Time since last diagnosis" sono state rimosse poiché presentano la quasi totalità dei campi mancanti.

Un'analisi di correlazione è stata condotta sugli attributi restanti, essa ha condotto alla matrice di correlazione in Figura 2. Dopo aver impostato una soglia dello 0.8 per la correlazione tra le variabili si è proceduto ad eliminare quelle risultate ridondanti cercando così di ottenere un sottoinsieme di attributi non correlati fra loro, ma correlati alla variabile di classe.

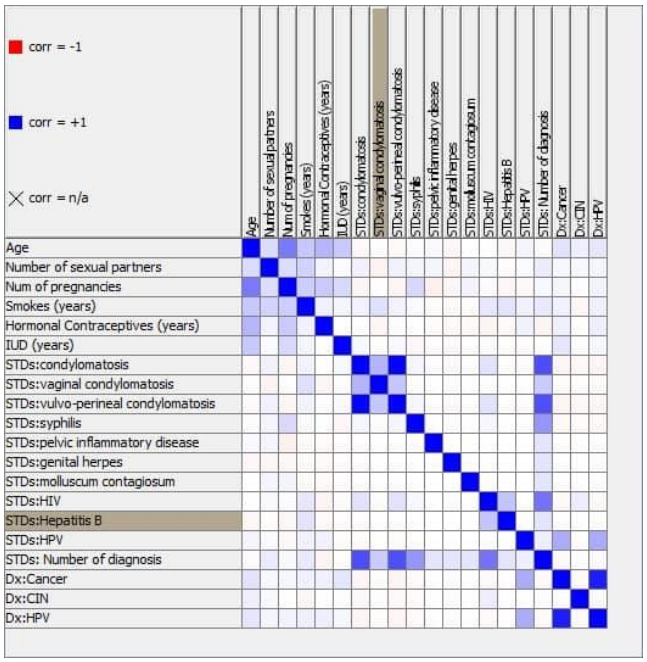

Figura 2: Matrice di Correlazione

3. Modelli

I modelli impiegati sono stati selezionati per favorire il confronto tra le differenti tecniche di classificazione statistiche al fine di individuare la più adatta. In particolare, si è posta attenzione ai seguenti approcci.[1]

- **Euristici:** tra questi la scelta è ricaduta su un modello semplice, il *Decision Tree* (J48), e la sua iterazione più complessa, il *Random Forest* (RF). Per quanto riguarda il primo è stato potato ad una soglia di confidenza dello 0.25 per ridurne complessità ed overfitting ai dati. Il numero di alberi creati dall'algoritmo *Random Forest* è stato limitato a 10.

- **Basati sulla regressione:** dovendo prevedere un attributo di classe binario il modello impiegato per questa tecnica di classificazione è quello *Logistico* (LOG).

- **Basati sulla separazione**: la partizione dello spazio degli attributi è stata svolta mediante due modelli, un *Support Vector Machine* (SMO) e un *Artificial Neural Network*. Il primo è stato creato seguendo il nodo "*SMO(3.7)*" del pacchetto *Weka* di *Knime* e utilizzando il kernel *puk*. Il secondo è un *Multilayer Perceptron* (MLP) con undici *hidden layers* basato sulla tecnica di classificazione delle istanze fondata sul *back-propagation error*.

- **Probabilistici:** come rappresentanti di questa tecnica di classificazione sono stati selezionati il modello *Naive Bayes* e due suoi sviluppi, il *Tree Augmented Naive Bayes* (NBTree), un modello che genera un albero decisionale con classificatori *Naive Bayes* come foglie, e la *Bayesian Network* (BNet).

Cross Validation

La validazione dei modelli è stata svolta mediante *cross validation*, prima di tipo *5-fold* poi *10-fold*. Il dataset, quindi, è stato suddiviso nel numero di partizioni indicate, ognuna contenente un numero quasi costante di record. Per ogni iterazione una delle partizioni è stata utilizzata come *test set* mentre le altre sono servite da *training set*. Il risultato finale è ottenuto come media dei risultati intermedi, uno per ogni *test set*.

Questo approccio è preferibile per il maggiore controllo che offre rispetto ad altri come l*'holdout* o la sua versione iterata. Esso infatti garantisce che ogni record del dataset sia incluso nel *training set* lo stesso numero di volte e che sia altresì contenuto nel *test set* esattamente una volta. Non essendo disponibile un'euristica assoluta per la scelta delle partizioni da effettuare, rappresentate dalla variabile "K", si è deciso di replicare la tecnica due volte utilizzando i valori definiti empiricamente dalla letteratura come migliori. Inoltre, poiché il dataset a disposizione si configura come fortemente sbilanciato, ogni sua partizione è stata selezionata mediante una tecnica di campionamento stratificato proporzionale, in modo da garantire che le proporzioni delle classi tra il dataset di *training* e quello di *test* fossero mantenute.

Feature Selection

Infine, ai migliori tra i modelli individuati è stato applicato un processo di filtraggio e selezione delle variabili. Questo è stato svolto seguendo un approccio *Wrapper*, ovvero utilizzando il classificatore stesso per trovare il sottoinsieme ottimale degli attributi disponibili. Nello specifico si è optato per una *Forward feature selection* che, partendo dal modello nullo, lo popolasse di attributi cercando di massimizzare il coefficiente *Kappa di Cohen*, scelto appositamente per via della natura sbilanciata dei dati a disposizione. I risultati così ottenuti si propongono di migliorare l'interpretabilità dei modelli sottostanti, oltre che sviluppare classificatori caratterizzati da un costo computazionale minore e, potenzialmente, da un minor overfitting ai dati.

4. Misure di performance

La scelta delle metriche da adottare per la valutazione dei modelli è stata fortemente influenzata dalla natura sbilanciata del dataset impiegato. Questa caratteristica

[1] A seguito di ogni modello è riportato fra parentesi l'acronimo con cui lo si può ritrovare nei grafici.

non ha permesso di usufruire delle misure classiche di *Accuracy* ed *Error*.
Al loro posto sono state selezionate delle misure adatte al tipo di dataset in uso[2]:

- *Precision:* definita come la frazione dei record effettivamente positivi nel gruppo dei record che il classificatore ha definito come tali.

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

- *Recall:* misura della frazione di record positivi predetti correttamente dal classificatore.

$$r = \frac{TP}{TP + FN}$$

- *F1 measure:* definita matematicamente come la media armonica delle due misure precedenti, ha lo scopo di riassumerle sinteticamente.

$$F1 = \frac{2rp}{r + p}$$

- *Kappa di Cohen:* coefficiente statistico utilizzato per comparare l'*Accuracy* ottenuta da un modello rispetto a quella attesa nel caso di una classificazione casuale.

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Sotto il profilo grafico, infine, lo sbilanciamento nei dati ha portato a preferire l'impiego di una curva PR (*Precision-Recall*) al posto di una più comune curva ROC (*Receiver Operating Characteristic*); perciò anche l'AUC (*Area Under the Curve*) a cui si farà in seguito riferimento è da interpretarsi come relativa ad essa.

5. Analisi e risultati

K-fold Cross Validation, K = 5
I risultati ottenuti in seguito all'impiego di questa procedura sono riassunti nella Tabella 1.

Classificatore	Precision	Recall	F1measure	AUC
Decision Tree	1.00	0.389	0.560	0.478
Naive Bayes	0.395	0.833	0.536	0.295
Support Vector Machine	0.875	0.778	0.824	0.829
Logistico	0.824	0.778	0.800	0.627
Naive Bayes Tree	0.889	0.889	0.889	0.717
Random Forest	0.846	0.611	0.710	0.713
Multilayer Perceptron	0.889	0.889	0.889	0.709
Bayesian Network	0.889	0.889	0.889	0.717

Tabella 1: Risultati K-fold Cross Validation K = 5

[2] Le formule sono basate sulla matrice di confusione.

- TP = True positive records
- FP = False positive records
- FN = False negative records
- Pr(a) = *Accuracy* osservata
- Pr(e) = *Accuracy* attesa nel caso di un classificatore random

In generale la maggior parte dei modelli ha restituito valori similmente alti sia per la *Precision* che per la *Recall*. Eccezione a questa tendenza sono stati i due modelli più elementari del gruppo, il *Naive Bayes* e il *Decision Tree*. Il primo ha raggiunto una *Recall* alta, ma, contemporaneamente, una bassa *Precision*; il secondo, invece, ha restituito un'immagine speculare. Data la semplicità dei modelli in esame è ragionevole che possano aver accentuato una misura a discapito dell'altra. Il primo ha classificato un numero esagerato di osservazioni come positive, aumentando così di molto i FP e, al contempo, riducendo i FN; il secondo ha seguito una logica opposta. Inoltre, l'assunzione di indipendenza tra i predittori, implicita nel modello *Naive Bayes*, non è rispettata, come evidenziato dalla matrice di correlazione in Figura 2. Infine, è opportuno evidenziare come la natura sbilanciata dei dati a disposizione sia una delle possibili cause dei risultati ottenuti da questi due modelli, che si sono comunque rivelati più sensibili a questo problema nonostante gli sforzi compiuti per attenuarlo.
Per favorire un confronto più agevole tra i classificatori sono state scelte due misure di sintesi. In Figura 3 è mostrato un paragone tra i valori della F1 measure.

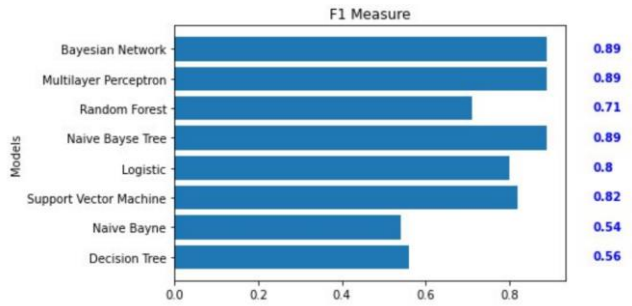

Figura 3: Confronto F1 measure, Cross Validation K = 5

Da questi è possibile notare come i modelli che sembrano aver performato meglio sono stati il *Tree Augmented Naive Bayes*, il *Multilayer Perceptron* e la *Bayesian Network*, seguiti dal *Support Vector Machine* e dal modello *Logistico*, tutti comunque sopra ad una soglia dello 0.8. La sintesi offerta dalla *F1 measure* è completata dalla Figura 4 in cui è proposto il grafico della *Precision Recall curve* per ognuno dei classificatori. Oltre l'andamento delle curve è significativa l'area da esse sottesa (AUC) che funge da utile misura per un confronto.

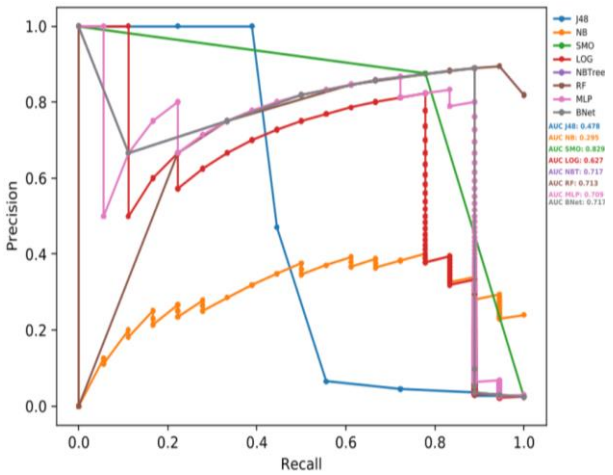

Figura 4: Precision Recall Curve, Cross Validation K = 5

Questo elemento aggiuntivo conferma l'ottimo risultato dei primi tre classificatori individuati precedentemente, ma evidenzia anche come il classificatore *logistico* fosse sopravvalutato dalla *F1 measure*, mentre il *Support Vector Machine* ne venisse sottostimato. Ponendo una soglia empirica ad un valore di 0.7, non solo l'analisi dell'AUC permette di eliminare il modello *logistico* dalla selezione dei migliori, ma porta anche all'attenzione il *Random Forest.* Malgrado il suo valore alto, però, soffre di una relativa bassa *Recall*, evidenziata sia dalla *F1 measure* che dall'andamento iniziale della *PR curve*. Generalmente questo modello non è consigliato per dataset sbilanciati, ma il campionamento stratificato proporzionale applicato in fase di partizionamento del dataset ha corretto significativamente questo problema.
Interessante notare come i due modelli basati sulla regola di decisione di *Bayes*, il *Tree Augmented Naive Bayes* e la *Bayesian Network* abbiano ottenuto risultati perfettamente similari sia nella tabella che nel grafico. Questo deriva dalla loro natura comune, essendo il primo derivabile da una semplificazione del secondo.

K-fold Cross Validation, K = 10
Raddoppiando i partizionamenti del dataset si sono ottenuti i risultati riassunti nella Tabella 2.

Classificatore	Precision	Recall	F1measure	AUC
Decision Tree	1.00	0.278	0.435	0.331
Naive Bayes	**0.375**	**0.833**	**0.517**	**0.282**
Support Vector Machine	0.875	0.778	0.824	0.829
Logistico	**0.833**	**0.833**	**0.833**	**0.684**
Naive Bayes Tree	0.889	0.889	0.889	0.635
Random Forest	**0.867**	**0.722**	**0.788**	**0.925**
Multilayer Perceptron	0.889	0.889	0.889	0.729
Bayesian Network	**0.889**	**0.889**	**0.889**	**0.635**

Tabella 2: Risultati K-fold Cross Validation, K = 10

Da questi si può evincere come aumentare il numero delle partizioni, riducendo quindi il numero di record al loro interno, non abbia portato significativi cambiamenti. Tutte le tendenze prima analizzate vengono sostanzialmente confermate, ciò è evidente anche dal confronto delle *F1 measure* nella Figura 5.

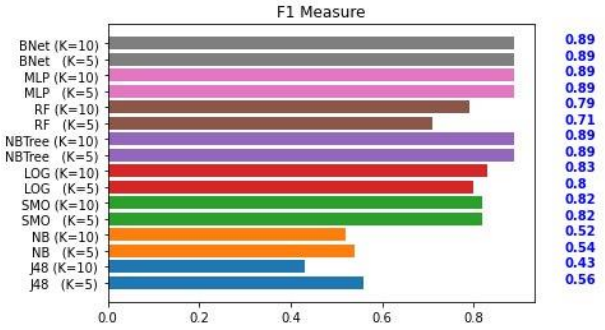

Figura 5: Confronto F1 measure tra le due Cross Validation

Gli unici cambiamenti significativi riguardano due modelli: il *Decision Tree*, che si attesta comunque a valori trascurabili, e il *Random Forest.*
Quest'ultimo subisce infatti una rivalutazione; la *Recall* relativamente bassa mostrata precedentemente è aumentata, migliorando a sua volta l'*F1 measure*. Ancora più convincente a riguardo è la *PR curve*, mostrata in figura 6.

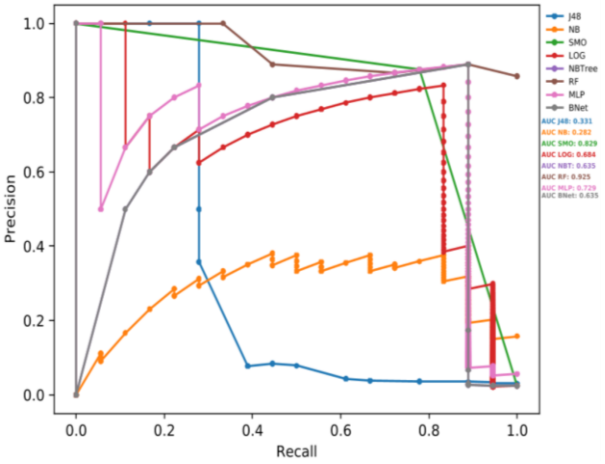

Figura 6: Precision Recall Curve, Cross Validation K = 10

Da questa si può notare come il valore per l'AUC del modello *Random Forest* sia divenuto il più elevato in assoluto. Ciò riflette l'aumento di rappresentatività dell'attributo della classe minoritaria ottenuto incrementando il numero di *fold*, campionato mediante metodo stratificato proporzionale.
Al contrario il modello *Bayesian Network*, pur mantenendo un alto valore nella *F1 measure*, ha subito un crollo nella AUC. Ciò è comprensibile osservando la *PR curve* che lo caratterizza. Essa è composta da valori bassi sia sulla *Precision* che sulla *Recall* nelle sue prime fasi. La somiglianza con l'altro modello basato anch'esso sulla regola di *Bayes*, il *Tree Augmented Naive Bayes*, è ancora mantenuta. Disponendo di un dataset sufficientemente limitato, l'uso di una tecnica di *Cross Validation* a dieci partizionamenti è preferibile in quanto permette un generale miglioramento dei modelli più rilevanti, senza comunque presentare problemi computazionali. Come considerazione finale, si può affermare che i modelli da ritenersi migliori, sia per bilanciamento tra *Precision* e *Recall* che per valore della AUC, sono il *Random Forest*, il *Support Vector Machine* ed il *Multilayer Perceptron*.

Feature Selection

Il processo precedentemente descritto di *Forward Feature Selection* è stato applicato ai tre modelli di cui sopra, con lo scopo di migliorarne performance e comprensibilità.

In seguito a ciò si è registrato un drastico calo nelle variabili necessarie per il loro addestramento. Sia il *Support Vector Machine* che il *Multilayer Perceptron* hanno conservato un solo attributo, quello inerente alla presenza o meno di una diagnosi per Papilloma Virus. Il modello *Random Forest* ha invece conservato quattro attributi:

- Numero di gravidanze
- Anni di utilizzo dei contraccettivi a base ormonale
- Anni di utilizzo della spirale intrauterina
- Diagnosi del Papilloma Virus

I risultati prodotti basandosi sul nuovo sottoinsieme di variabili sono riportati nella Tabella 3.

Classificatore	Precision	Recall	F1measure	AUC
Support Vector Machine	0.889	0.889	0.889	0.890
Random Forest	**0.895**	**0.944**	**0.919**	**0.929**
Multilayer Perceptron	0.889	0.889	0.889	0.640

Tabella 3: Risultati Feature Selection

Rispetto al caso precedente, tutti i modelli hanno migliorato le proprie performance in termini di *Precision* e *Recall*. L'AUC è aumentata per tutti tranne che per il *Multilayer Perceptron*, attestandosi a valori molto alti per il modello *Random Forest* in particolare. Questo cambiamento si evince dal grafico della Figura 7.

Dall'analisi di questi fattori risulta evidente come i modelli *Random Forest* e *Support Vector Machine* siano quelli che consentono di ottenere risultati migliori e più comprensibili per la previsione del cancro, soprattutto utilizzando un numero ridotto di variabili.

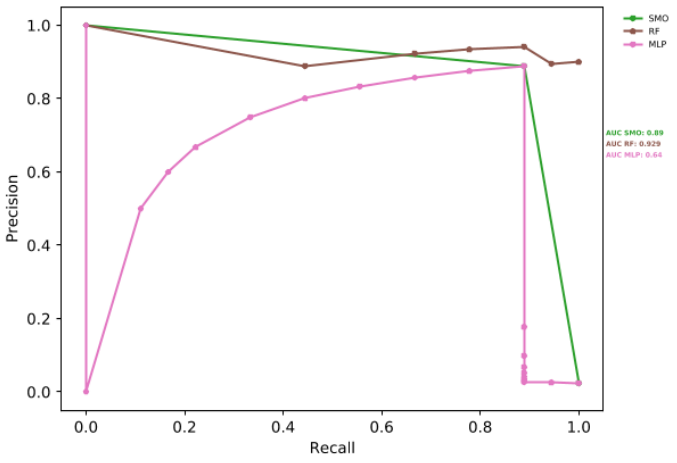

Figura 7: Precision Recall Curve Feature Selection

Conclusioni

Nel report sono stati sviluppati otto modelli di classificazione, validati con due differenti approcci di *Cross Validation*. Le metriche usate per la loro valutazione sono state selezionate per tenere conto dello sbilanciamento del dataset. Da questo si è giunti alla conclusione che tre modelli in particolare fossero i migliori: *Random Forest*, *Support Vector Machine* e il *Multilayer Perceptron*. A questi è stato applicato un *Wrapper* di *Feature Selection* per poterne aumentare l'interpretabilità. Alla fine di questo processo, una volta individuato un sottoinsieme di attributi ottimale per ogni classificatore, i modelli che si sono dimostrati migliori sono stati il *Random Forest* e il *Support Vector Machine*. Attraverso questi è possibile fornire una stima accurata della presenza o meno di cancro alla cervice nelle pazienti. Tra le variabili che permettono questa previsione, la principale risulta essere l'esposizione al Papilloma Virus. I limiti riscontrati nello svolgimento di questa analisi si possono riassumere nell'esigua natura del dataset a disposizione e nel suo sbilanciamento. Questi due fattori hanno comunque influito sulla performance dei modelli nonostante si sia cercato di minimizzarne gli effetti. La tecnica di campionamento stratificato proporzionale ha corretto la naturale tendenza a sottostimare l'attributo di classe del modello *Decision Tree* e di tutti i suoi derivati.

Riferimenti bibliografici

- https://pubmed.ncbi.nlm.nih.gov/15950092
- https://tobaccoatlas.org/country/venezuela
- https://www.kaggle.com/loveall/cervical-cancer-risk classification
- An Introduction to Statistical Learning, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. Springer New York Heidelberg Dordrecht London, 2017
- Cohen's kappa coefficient as a performance measure for feature selection, Susana M. Vieira, Student Member, IEEE, Uzay Kaymak, Member, IEEE and Joao M.C. Sousa, Member, IEEE
- The Relationship Between Precision-Recall and ROC Curves, Jesse Davis, Mark Goadrich. Department of Computer Sciences and Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, 1210 West Dayton Street, Madison, WI, 53706 USA
- Tackling the Poor Assumptions of Naive Bayes Text Classifiers, Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger, Artificial Intelligence Laboratory; Massachusetts Institute of Technology; Cambridge, MA 021