Breast Cancer Diagnosis Wisconsis

Anastasia Marzi, Alessandro Riboni, Davide Sangalli, Federico Signoretta, Diana Tenca

Negli Stati Uniti, il cancro al seno è quello più comunemente diagnosticato nelle donne, dopo il cancro alla pelle. Questa tipologia di cancro può essere riscontrata sia negli uomini che nelle donne, anche se è molto più comune nelle donne.
Grazie al sostanziale supporto ricevuto per la lotta contro il cancro e ai fondi per la ricerca, negli ultimi anni ci sono stati molti progressi nei campi della diagnosi e dei diversi trattamenti per questa determinata tipologia di cancro. Il tasso di sopravvivenza sta aumentando e il numero di morti associato con questa malattia sta decrescendo stabilmente, questo grazie a fattori come la possibilità di diagnosticare il cancro tempestivamente, un nuovo approccio personalizzato al trattamento e una migliore comprensione della malattia.
Il nostro obiettivo in questo progetto è quello di riuscire a capire se un tumore è benigno oppure maligno, utilizzando solamente i risultati di un'agobiopsia, senza dover fare ulteriori test. Vorremmo individuare alcuni algoritmi di classificazione e clusterizzazione che funzionino in maniera ottimale in modo tale da poterli consigliare agli ospedali.

Indice

1 Introduzione

La diagnosi del tumore al seno è tradizionalmente effettuata attraverso una biopsia totale, una procedura chirurgica invasiva. La tecnica di aspirazione tramite un ago sottile (FNAs) determina un modo meno invasivo per esaminare una piccola quantità di tessuto della massa tumorale; comunque, la diagnosi con questa procedura ha ottenuto risultati di diverso tipo. Esaminando attentamente sia le caratteristiche delle singole cellule che le importanti features contestuali, quali, per esempio, la grandezza del grumo (massa) di cellule, gli scienziati di alcune istituzioni specializzate sono stati in grado di diagnosticare con successo il tipo di tumore (benigno/maligno) utilizzando le tecniche di FNAs.
La procedura di diagnosi comincia con l'aspirazione di una piccola quantità di fluido prelevato dalla massa tumorale presente nel seno. Il materiale aspirato viene successivamente inserito all'interno di un vetrino e studiato. L'immagine viene analizzata attraverso un microscopio Olympus alla cui sommità è collegata una color video camera JVC TK-1070U.
Inizialmente, è necessario fornire un contorno approssimato della massa tumorale. A tal fine, è stato sviluppato un'interfaccia grafica che permette all'utente di dare in input un'approssimazione del bordo di un numero sufficiente di nuclei così da ottenere un campione rappresentativo. Applicando tale interfaccia si ottiene un'immagine del tutto analoga a quella della Figura 1. Partendo da questa approssimazione, il bordo reale è trovato utilizzando un modello di "contorno attivo" noto in letteratura come *"snake"*. Si tratta di una spline [1] deformabile che cerca di minimizzare la funzione energia lungo una porzione di linea chiusa. La funzione energia è così definita:

$$E = \int_s \Big(\alpha E_{cont}(s) + \beta E_{curv}(s) + \gamma E_{image}(s) \Big) ds$$

[1] In analisi matematica, una spline è una funzione costituita da polinomi che ha come scopo l'interpolazione di una serie di punti, detti nodi, appartenenti ad un intervallo, in modo tale che la funzione sia continua, almeno fino ad un dato ordine di derivata, in tutti i punti dell'intervallo.

dove s è la lunghezza dell'arco preso in considerazione, α, β e γ sono i pesi rispettivamente di E_{cont}, E_{curv} e E_{image} e sono costanti scelte empiricamente.
I tre termini E_{cont}, E_{curv} e E_{image} rappresentano diverse proprietà dell'energia. E_{cont} e E_{curv} sono due misure geometriche entrambe costruite per penalizzare le discontinuità della curva. Nel primo caso, viene studiata la distanza tra due punti dello snake mentre, nel secondo, quella tra i punti del bordo e il centro della massa. La E_{image} misura il livello di discontinuità della scala di grigi lungo lo snake.

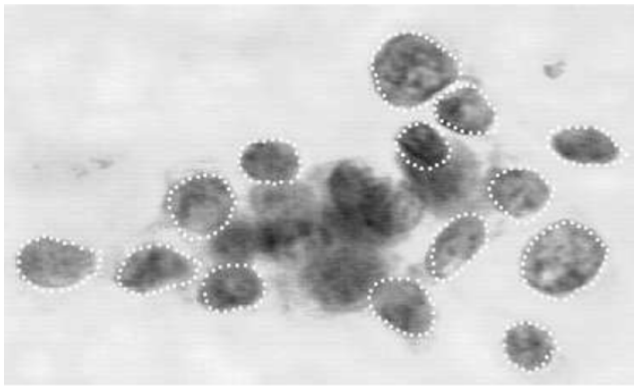

Figura 1: Approssimazione iniziale dei bordi delle cellule tumorali.

1.1 Presentazione Dataset

Il datataset che abbiamo scelto di analizzare è il "*Breast Cancer Diagnosis Wisconsis*". Questo è composto da 569 records, ognuno dei quali rappresenta le analisi di agobiopsia effettuate su un paziente diverso, il quale viene identificato con l'id inserito nella prima colonna. Compensiamo il fatto che non ci sia una singola misura di forma che sembri catturare l'idea di "irregolarità" sfruttando diverse features per rappresentare la forma del contorno dello snake.
In particolare, il sistema di diagnostica descritto precedentemente estrae 10 diverse caratteristiche dai confini del nucleo della cellula generati dagli snake. Tutte queste caratteristiche sono numericamente modellate in modo tale che il valore più grande indichi tipicamente una probabilità più alta che il tumore sia maligno.
Le caratteristiche sono:

- Il *raggio*: il raggio di un singolo nucleo è misurato facendo la media della lunghezza dei segmenti radiali che collegano il centroide dello snake a tutti i singoli punti snake posti sul perimetro;

- La *texture*: è la deviazione standard calcolata sul valore della scala di grigi visualizzati nella radiografia;

- Il *perimetro* $2p$: la distanza totale fra i punti snake costituisce il perimetro nucleare;

- L'*area* A: l'area nucleare è misurata semplicemente contando il numero di pixels nella parte interiore dello snake e aggiungendone metà nel perimetro;

- La *compattezza*: Il perimetro e l'area vengono combinati per dare una misura di compattezza del nucleo della cellula usando la formula $\frac{(2p)^2}{A}$. Questo numero senza dimensione è minimizzato da un disco circolare e cresce con l'irregolarità dei bordi. Comunque, questa misura di forma, oltre ad aumentare con l'alzarsi del grado di malignità del tumore, aumenta anche per nuclei di cellule allungati, il quale non indica necessariamente una crescita della possibilità di un tumore maligno. La compattezza è anche distorta se stiamo studiando cellule piccole perché diminuisce l'accuratezza imposta dalla digitalizzazione del campione.

- La *regolarità del bordo*: la regolarità del contorno nucleare è quantificata misurando la differenza fra la lunghezza della linea radiale e la media della lunghezza delle linee che lo circondano. In poche parole, si misurano le variazioni locali nella lunghezza del raggio. Questa è simile alla computazione dell'energia di curvatura, E_{curv}, negli snakes.

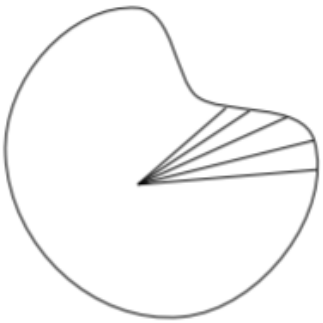

Figura 2: Regolarità del bordo

- La *concavità*: in un ulteriore tentativo di collezionare informazioni sulla forma misuriamo il numero e la 'severità' delle concavità o delle indentazioni nel nucleo della cellula. Disegniamo delle corde fra punti non adiacenti dello snake e misuriamo la distanza alla quale il bordo reale del nucleo giace nella parte interna di ogni corda. Questo parametro è enormemente influenzato dalla lunghezza di queste corde, infatti più piccole sono le corde meglio è in grado di catturare piccole concavità. Si

è deciso di enfatizzare le piccole indentazioni, poichè grandi irregolarità della forma sono delineate da altre feature.

Figura 3: Concavità del bordo

- I *punti concavi*: questa feature è simile alla concavità, ma misura solamente il numero delle concavità sul bordo, piuttosto che la magnitudine.
- La *simmetria*: per misurare la simmetria, si prende in considerazione l'asse maggiore o la corda più lunga passante per il centro e si misura la differenza di lunghezza fra le linee perpendicolari all'asse che partono da essa e si estendono fino al bordo della cellula in entrambe le direzioni. Si presta speciale attenzione per i casi dove l'asse maggiore taglia il bordo della cellula a causa della concavità.

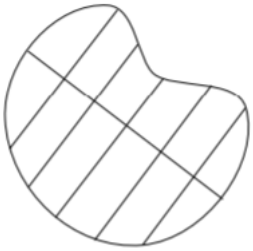

Figura 4: Simmetria

- La *dimensione frattale*: la dimensione frattale di una cellula è approssimata utilizzando "*l'approssimazione della linea di costa*" descritta da Mandelbrot. Il perimetro del nucleo è misurato usando 'righelli' sempre più grandi. Come aumenta la grandezza del righello, decresce la precisione della misurazione e anche il perimetro osservato. Plottando questi valori su una scala logaritmica e misurando la pendenza verso il basso otteniamo (il negativo di) un'approssimazione alla dimensione frattale. Come con tutte le misure di forma, un valore maggiore corrisponde ad un contorno meno regolare e quindi ad una maggiore probabilità di malignità del tumore.

Figura 5: Dimensione frattale

Tutte le misure di forma sono verificate utilizzando il fantasma di una cellula idealizzata. Si nota come queste misure aumentino quando il bordo diventa meno regolare e siano largamente non correlate alla dimensione del contorno.
Per ogni massa tumorale consideriamo queste 10 caratteristiche principali e per ognuna di esse calcoliamo tre valori differenti: la media, la deviazione standard e il "valore peggiore", cioè la media dei tre valori più grandi. Quest'ultimo valore è il più utile intuitivamente per questo problema, poiché solo poche cellule maligne potrebbero essere raccolte all'interno di un dato campione.
A questi primi 31 attributi si aggiunge infine la colonna della diagnosi, in cui è specificato se si tratta di un tumore maligno (M) o di un tumore benigno (B). Nel dataset che abbiamo a disposizione il numero di tumori maligni è 212 e rappresenta il 37% del dataset totale, contro i 357 tumori benigni (63%). Ci troviamo di fronte ad un problema di classe sbilanciato solo in minima parte, ma abbiamo deciso di considerarlo e analizzarlo come tale in quanto, ponendoci nell'ottica di un medico, abbiamo supposto fosse più importante predire correttamente la classe maligna del tumore.

2 Preprocessing

In questa fase iniziale di analisi, la prima cosa che bisogna controllare è la presenza o meno di valori mancanti che potrebbero influire sui nostri risultati, distorcendoli e facendoci arrivare a conclusioni completamente diverse se non addirittura errate.
Nel nostro caso siamo fortunati perché i dati raccolti sono completi e quindi possiamo usarli senza dover apportare alcuna modifica. In seguito, abbiamo rivolto la nostra attenzione ai vari attributi del dataset rendendoci conto che la colonna contenente l'id del paziente risulta ininfluente e quindi abbiamo deciso di eliminarla.
Infine, all'interno del nostro workflow abbiamo inserito un nodo Shuffle per evitare che la disposizione dei dati influenzasse i nostri risultati.

3 Classification

Inizialmente abbiamo optato per una classificazione senza nessuna feature selection, utilizzando, quindi, tutti e 30 gli attributi, e con una semplice partizione in training set e test set dei dati (rispettivamente 67% e 33%). Come abbiamo già esposto, lo scopo della classificazione è quello di riuscire a predire la tipologia del tumore, maligno o benigno, con una particolare attenzione alla corretta predizione della classe maligna, la nostra classe rara.
Nella letteratura del Machine Learning troviamo moltissimi classificatori e per la nostra analisi abbiamo scelto di utilizzarne 8 in particolare. Questi classificatori possono essere raggruppati in 4 macrogruppi e divisi in base alle loro caratteristiche principali. Si dividono in:

- Modelli Euristici: questo gruppo di modelli viene utilizzato quando problemi di complessità computazionale rendono impossibile la ricerca di una soluzione esatta. Gli algoritmi euristici, in base a conoscenze del problema e dall'esperienza, possono essere in grado di trovare soluzioni approssimate il cui valore è abbastanza vicino a quello ottimo. All'interno di questo gruppo abbiamo deciso di applicare l'albero decisionale *J48* e il *Random Forest*. In particolare, un albero decisionale è un albero binario in cui ogni nodo divide i pattern sulla base di un criterio su una singola feature. É uno strumento molto utilizzato in applicazioni di pattern categorici o misti. Questo algoritmo è considerato euristico per la scelta che bisogna fare nel momento in cui bisogna decidere l'algoritmo con cui splittare l'albero ad ogni livello.
  Il Random Forest è, invece, multiclassificatore, cioè un approccio dove diversi classificatori sono utilizzati per eseguire la classificazione dei pattern; le decisioni dei singoli classificatori sono fuse ad un certo livello del processo di classificazione. Questo appartiene alla famiglia dei metodi di *Bagging (Bootstrap Aggregating)* e in esso i singoli classificatori sono appunto alberi decisionali.

- Modelli basati sulla regressione: questi modelli usano una probabilità condizionale parametrica. La regressione assegna un valore continuo ad un pattern. É utile per la predizione di valori continui. Risolvere un problema di regressione corrisponde ad apprendere una funzione approssimante delle coppie input output date. In particolare alcuni utilizzano la regressione logistica e infatti il modello da noi utilizzato è il *Bayesian Logistic Regression*;

- Modelli Probabilistici: questi sfruttano la formula di Bayes e calcolano la probabilità a posteriori per classificare i record. Se tutte le distribuzioni in gioco sono note, l'approccio Bayesiano costituisce la migliore regola di classificazione possibile.
  Formula di Bayes:

$$P(Y|\mathbf{X}) = \frac{P(\mathbf{X}|Y)P(Y)}{P(\mathbf{X})}$$

  Dato un pattern da classificare abbiamo bisogno quindi della probabilità a priori $P(Y)$ e della densità di probabilità condizionali $P(\mathbf{X}|Y)$. In questo caso la regola di Bayes massimizza la probabilità a posteriori, cioè massimizza la densità di probabilità condizionale tenendo comunque conto della probabilità a priori delle classi. La regola si dimostra ottima in quanto minimizza l'errore di classificazione.
  Il modello che abbiamo deciso di utilizzare è il *Naive Bayes*. Comunque anche il modello precedentemente citato nella regressione si basa sulla formula di Bayes.

- Modelli basati sulla separazione: di questo gruppo fanno parte tutti quei modelli che per svolgere la classificazione partizionano lo spazio degli attributi, come per esempio le Support Vector Machine. L'SVM è uno degli strumenti più utilizzati per la classificazione di pattern. Invece di stimare la densità di probabilità delle classi, come avviene nei modelli probabilistici, si è pensato di risolvere direttamente il problema di determinare le superfici decisionali tra le classi(*classification boundaries*). La Support Vector Machine nasce come classificatore binario (2 classi), estendibile a più classi. Nel nostro caso abbiamo una SVM non linare (cioè la superficie di separazione non è più un iperpiano come nel caso lineare ma una superficie complessa) senza ipotesi sulla separabilità dei pattern. In questi casi bisogna risolvere un problema di programmazione matematica sfruttando una Funzione Kernel che è la funzione di similarità calcolata nello spazio di origine degli attributi. Se i pattern del training set sono classificati con ampio margine si può sperare che anche pattern del test set vicini al confine tra le classi siano gestiti correttamente. All'interno della classe delle SVM abbiamo deciso di utilizzare tre diversi tipi di Kernel: il "*Poly Kernel*", il "*Puk Kernel*" e l' *SPegasos*.
  Esiste anche un'altra tipologia di classificatori all'interno di questa classe: le Reti di Neuroni Artificiali. In particolare, abbiamo utilizzato il *Multi Layer Perceptron* (MLP), nel quale ogni neurone artificiale è rappresentato come un nodo di un grafo ordinato che comunica in maniera unidirezionale, dall'attributo di input X all'attributo di Classe Y.

Dopo aver applicato tutti i classificatori scelti al nostro dataset, abbiamo dovuto confrontarli per vedere quale ottenesse i risultati migliori.
Per ognuno di essi abbiamo calcolato i valori di:

- *Precision*:
$$\frac{TP}{TP + FP}$$
cioè la frazione che indica quanti tra i records che il classificatore classifica come positivi lo siano realmente;

- *Recall*:
$$\frac{TP}{TP + FN}$$
cioè la frazione che indica quanti tra i record effettivamente positivi siano stati riconosciuti come tali;

- F_1-measure, che è una combinazione delle prima due misure:
$$F_1 = \frac{2rp}{r + p}$$

Comunque, avendo deciso di studiare il problema come un class imbalance, la misura che, a nostro avviso, meglio esprime la performance dei classificatori é la F_1-measure, quindi è quella a cui abbiamo scelto di dare maggior peso.
Per prima cosa, abbiamo utilizzato il modello Hold-Out per stabilire se ci fossero o meno dei classificatori che overfittavano, ossia che sono eccessivamente ottimisti sul training set. Successivamente, per stabilire quale metodologia utilizzare tra l'Iterated Hold-Out e la K-Folds Cross Validation, abbiamo condotto alcune prove giungendo alla seguente conclusione: a causa della presenza di outliers, il metodo dell'Iterated Hold-Out risulta poco affidabile poichè, ad ogni lancio del workflow, i risultati ottenuti cambiano sensibilmente. Come conseguenza di queste considerazioni, abbiamo deciso di non utilizzarlo nel corso della nostra analisi. Invece, poichè il nostro data set non è particolarmente esteso, siamo disposti ad accettare l'aumento del costo computazionale del modello a favore della diminuzione della distorsione della stima della performance dello stesso e per questo motivo abbiamo deciso di applicare anche il K-Folds Cross Validation.

Di seguito, vengono esposte come prima cosa le considerazioni derivanti dall'applicazione dell'Hold-Out al nostro dataset. Dopo alcuni test, abbiamo notato come il modello *Bayesian Logistic Regression* non funzioni bene in quanto non si riesce a calcolare la F_1-measure perchè la precision è indefinita (questo succede quando il numero di True Positive e False Positive è uguale a 0, quindi il classificatore non classifica nulla come istanza positiva). Per questo motivo abbiamo deciso di non includerlo nè nelle successive tabelle nè nei boxplot, in modo da non distorcere la composizione del grafico.

Classificatori	**Train Set**	**Test Set**
NB	0.91	0.89
Poly Kernel	0.97	0.96
SPegasos	0.98	0.98
MLP	0.99	0.56
J48	0.99	0.91
Puk Kernel	0.99	0.95
RF	1.0	0.9

Osservando la tabella che riporta i valori della F_1-measure dei diversi classificatori, si evince che il modello *MLP* overfitta: ha una F_1-measure pari a 0.99 sul train set e a 0.56 sul test set. Anche gli altri modelli, ad esclusione dell'*SPegasos* overfittano leggermente sul training set, ma con una differenza non particolarmente significativa. Il modello migliore risulta essere, basandosi sulla F_1-measure, l'*SPegasos*, unico modello che ha un'ottima performance sia sul train che sul test set.

In seguito, come anticipato nel paragrafo precedente, abbiamo applicato anche il metodo della *K-Folds Cross Validation*, utilizzando uno stratified sampling con 10 fogli. Osservando il boxplot si nota che, anche utilizzando questa tecnica, l'*SPegasos* risulta uno dei modelli migliori assieme al *Poly Kernel* con una F_1-measure pari a 0.97. Tuttavia, fatta eccezione per il modello *MLP* che, con una F_1-measure pari a 0.55, riconferma la sua inadeguatezza per la risoluzione del nostro problema, anche gli altri modelli hanno una buona performance con valori di F_1-measure che variano tra 0.91 e 0.96.

Classificatori	**F_1-measure**	**Recall**	**Precision**
MLP	0.55	0.99	0.38
Naive Bayes	0.91	0.89	0.93
J48	0.93	0.91	0.94
RF	0.95	0.94	0.97
Puk Kernel	0.96	0.95	0.97
Poly Kernel	0.97	0.94	1.0
SPegasos	0.97	0.96	0.98

Come ultimo passo, abbiamo osservato la ROC curve ottenuta dopo aver applicato la K-Folds Cross Validation. Senza particolari sorprese, si vede immediatamente che il modello *Bayesian Logistic Regression* sia del tutto inadatto al nostro problema. Ha addirittura una performance peggiore del modello *Zero-Rule*, cioè il campionamento casuale dei dati.
Interessante invece, il comportamento del modello *MultiLayer Perceptrion* che sembrerebbe avere

un'ottima performance, al contrario di quello che si è dedotto dai risultati dei boxplot. Il motivo di questo comportamento inaspettato è da ricercarsi nella natura della Roc Curve stessa: sull'asse delle ordinate viene rappresentata la recall del modello che, nel caso del *MLP*, è molto alta e questo giustifica il risultato del grafico. Il modello classifica tutti i record come positivi ossia, tutte le masse come maligne, ottenendo ottimi valori di recall, ma pessimi di precision.
Per quanto concerne i modelli migliori, la Roc Curve evidenzia comportamenti molto simili suggerendo che l'analisi sulle F_1-measure e le conclusioni da essa tratte siano corrette.

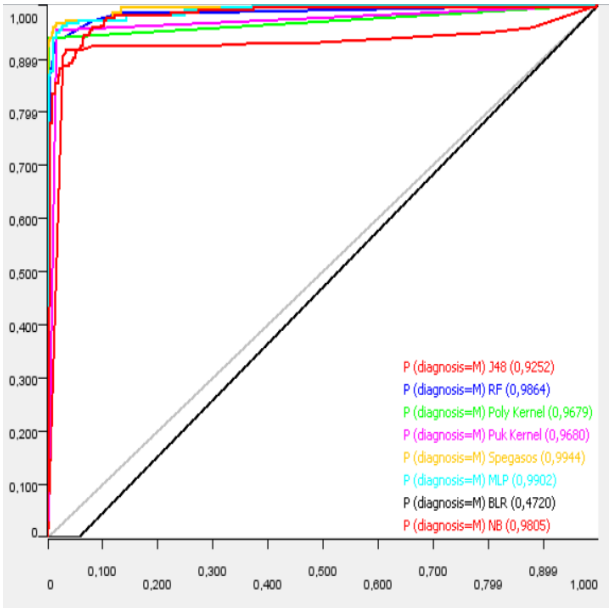

Figura 6: K-Folds Cross Validation-Roc Curve.

Visto l'alto numero di attributi presenti nel dataset, abbiamo deciso di provare ad applicare la Feature Selection così da eliminare gli attributi ridondanti e irrilevanti, diminuendo il costo computazionale e migliorando l'interpretabilità del classificatore.

3.1 Feature Selection

In seguito a numerose prove, nelle quali i modelli wrapper e filter univariati non ci fornivano risultati ottimali, abbiamo deciso di utilizzare dei modelli filter multivariati. Dopo aver applicato i modelli *J48*, *Random Forest* e l'albero decisionale *FT*, avendo ottenuto in output gli stessi attributi, abbiamo preso la decisione di proseguire il nostro studio utilizzando esclusivamente l'albero decisionale *FT*.
Alla fine, abbiamo ricavato un subset ottimale composto da 7 delle 32 features di partenza: *concavepoints_mean*, *area_se*,*texture_worst*, *perimeter_worst*, *area_worst*, *concavity_worst* e *concavepoints_worst*.

Una volta identificate le feature ottimali, abbiamo riapplicato il ragionamento condotto precedentemente applicando il metodo dell'Hold-Out per lo studio dell'overfitting dei modelli e quello della K-Folds Cross Validation per valutarne le performance.
Abbiamo scelto di escludere, anche in questo caso, il modello *Bayesian Logistic Regression* per l'impossibilità di calcolare l'F_1-measure che potrebbe distorcere i nostri risultati. Osservando i valori ottenuti dall'applicazione dell'Hold-Out è emerso che il modello *MLP* overfitta anzi, la sua prestazione sul test set peggiora rispetto al caso precedente. Tutti i modelli sono soggetti ad un leggero overfitting ad eccezione dell'*SPegasos* e del *Puk Kernel* che mantengono inalterato il valore della F_1-measure sia sul train che sul test set.

Classificatori	**Train Set**	**Test Set**
NB	0.94	0.9
Poly Kernel	0.96	0.93
Puk Kernel	0.96	0.96
SPegasos	0.96	0.96
MLP	0.97	0.27
J48	0.98	0.94
RF	1.0	0.93

Anche in questo caso abbiamo proseguito il nostro studio applicando la K-Folds Cross Validation tenendo in considerazioni solo il set ottimale di feature. I risultati di Precision, Recall e F_1-measure ottenuti sono riportati nella seguente tabella:

Classificatori	**F_1-measure**	**Recall**	**Precision**
MLP	0.18	0.1	1.0
Naive Bayes	0.93	0.92	0.94
J48	0.93	0.91	0.95
RF	0.94	0.92	0.97
Poly Kernel	0.95	0.93	0.98
Puk Kernel	0.95	0.95	0.98
SPegasos	0.96	0.95	0.98

Salta subito all'occhio che, anche in questo caso, il modello *MLP* funziona male dal punto di vista della F_1-measure.
Per tutti gli altri modelli, invece, il valore della F_1-measure varia solo leggermente, in positivo o in negativo, rispetto a quello calcolato con il metodo Hold-Out.
Al primo posto troviamo comunque l'*SPegasos*, subito seguito dalle altre due Support Vector Machine (*Puk Kernel* e *Poly Kernel*).
Per terminare l'analisi ed effettuare il confronto fra i diversi classificatori, anche in questo caso, abbiamo calcolato la Roc Curve.

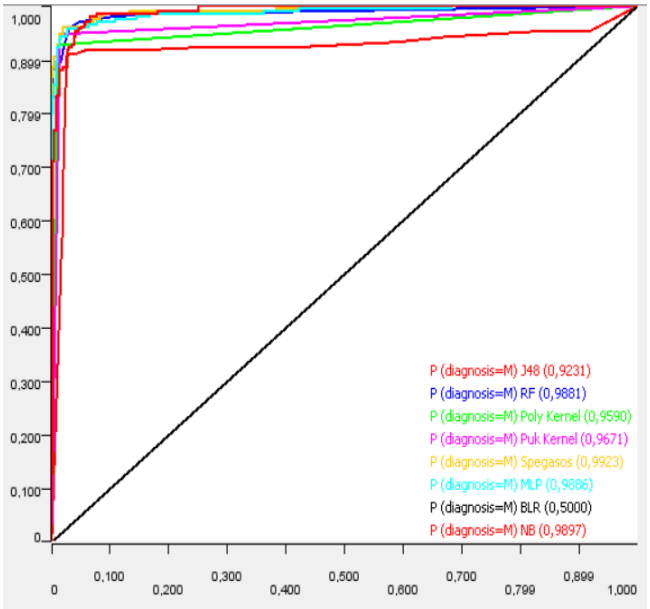

Figura 7: Roc Curve dopo l'applicazione della Feature Selection.

Osservando la Figura 7 si nota, senza particolari sorprese, che il *BLR* non funziona bene, infatti si trova più o meno a cavallo della Zero-Rule.
Anche in questo caso si vede come ci sia una differenza, per quanto riguarda l'*MLP*, fra i risultati della F_1-measure e quello che risalta nel grafico. Questo viene spiegato con le stesse giustificazioni precedenti.
Per il resto il grafico rispecchia quello che avevamo dedotto dalla tabella calcolata con la K-Folds Cross Validation e cioè che gli altri diversi classificatori funzionano tutti relativamente bene e si giocano il posto di miglior classificatore a seconda del livello di False Positive Rate (cioè la frazione di record erroneamente classificati come positivi fra tutti i record effettivamente negativi) che siamo disposti accettare. Anche una lieve variazione di questo valore potrebbe far cambiare il risultato e quindi non c'è modo di definire il classificatore migliore in maniera assoluta.

3.2 Conclusioni

In conclusione, i modelli *SVM*, il *Poly Kernel*, il *Puk Kernel* e l'*SPegasos* sono quelli con la miglior performance sia applicando la feature selection che non facendolo mentre il modello *MLP*, assieme al *Bayesian Logistic Regression*, risulta, in ogni caso, del tutto inadatto al problema.
Sebbene la feature selection non porti un miglioramento delle performance dei classificatori, essa permette, diminuendo il numero di attributi impiegati dal classificatore, di migliorare l'interpretabilità del modello e, in previsione di un'analisi futura, di diminuire il costo del collezionamento dei dati. Inoltre, l'applicazione della feature selection attenua leggermente l'overfitting dei modelli.

4 Clustering

Per svolgere l'analisi di classificazione abbiamo dovuto tenere in considerazione la colonna *Diagnosis*, in cui è presente la variabile di classe. Ci siamo domandati, quindi, se potevamo trovare un modo per predire la stessa variabile nella maniera ottimale senza doverla usare come input in un algoritmo, utilizzando quindi una tecnica non supervisionata. Abbiamo allora deciso di sfruttare un altro importante strumento del Machine Learning: il *Clustering*.
Abbiamo provato ad utilizzare diversi algoritmi di clustering con l'obiettivo di creare 2 gruppi esclusivi, per poi controllare se corrispondessero o meno alla distinzione iniziale benigno/maligno.
Il quesito che ci siamo posti riguarda la possibilità di identificare l'attributo di classe *Diagnosis* considerando solo alcuni degli attributi presenti nel dataset. Gli attributi presi in considerazione sono stati selezionati attraverso un confronto con un esperto di dominio che abbiamo consultato per rendere più completa la nostra analisi (Dott. Nicolò Silvestro Sutera - Medicina Generale). In particolare, sono stati selezionati gli attributi che vengono maggiormente tenuti in considerazione durante lo studio e l'analisi di una mammografia: *texture_mean*, *texture_se*, *smoothness_mean*, *smoothness_se*, *concave_points_mean*, *concave_points_se*, *symmetry_mean*, *symmetry_se*, *radius_se*, *perimeter_mean*, *perimeter_se* e *texture_worst*.
Tali attributi sono stati normalizzati nell'intervallo $[0, 1]$ per rendere più omogeneo il processo di *clustering*.
Dunque, procedendo su questa strada, abbiamo utilizzato due algoritmi di clustering in grado di individuare due gruppi esclusivi e partizionali, ovvero:

- *K-means method*
- *K-medoids method* (PAM, *Partitioning Around Medoids)*

dove K è il numero di cluster che si vogliono ottenere. Entrambi questi metodi sono tecniche euristiche che si basano sul concetto di prototipo: la richiesta che facciamo agli elementi è che siano più simili al prototipo del clustering a cui sono associati rispetto ai prototipi degli altri. Il K-means si basa sul concetto di centroide, cioè ogni cluster è rappresentato dal valore medio degli oggetti nel cluster, mentre il K-medoids ha al centro il concetto di medoide, che è uno degli oggetti localizzati vicino al centro del cluster.

In generale, esistono diversi metodi per verificare la qualità dei cluster che vengono prodotti da un algoritmo. Possono essere utilizzate misure interne, esterne e relative.
Nel nostro particolare caso, sono state adottate delle misure esterne, poiché conoscevamo a priori il numero di cluster di cui avevamo bisogno (*k=2*, gruppo M e gruppo B) e grazie alla conoscenza pregressa della colonna '*Diagnosis*' siamo anche a conoscenza del gruppo al quale le singole osservazioni dovrebbero appartenere.
Fissato $K = 2$ sono state calcolate, per entrambi gli algoritmi, le misure esterne di *Jaccard*, *Rand* ed, infine, *Fowlkes and Mallows*. Osservando tali misure possiamo notare una prevalenza, seppur mediocre, del *K-means method* rispetto a tutti e tre gli indici.

	Rand	Jaccard	Fowlkes and Mallows
K-means	0.85	0.77	0.87
K-medoids	0.84	0.75	0.86

Da questa osservazione possiamo affermare che il *K-means method* potrebbe essere un buon supporto medico per la determinazione della tipologia di tumore (maligno o benigno).

4.1 Conclusioni

In generale, possiamo sostenere che la metodologia del *clustering*, considerando gli attributi sopracitati, è in grado di identificare correttamente buona parte delle osservazioni presenti nel dataset e, dunque, è in grado di fornire un valido strumento per l'analisi di questo fenomeno, soprattutto utilizzando il metodo individuato.
In conclusione, considerando il fatto che per l'applicazione degli algoritmi di clustering non è stata utilizzata la variabile di classe, possiamo affermare di aver ottenuto degli ottimi risultati, che però non possono essere paragonati a quelli ottenuti dai migliori algoritmi di classificazione. Quindi, avendo a disposizione un adeguato numero di dati, divisibile in training e test set, su cui è possibile sviluppando un modello basandosi sul valore dell'attributo di classe è sicuramente preferibile utilizzare tecniche di classificazione rispetto a quelle di clustering.

Riferimenti bibliografici

[1] https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

[2] *Nuclear Feature Extraction For Breast Tumor Diagnosis* - W.Nick Street, William H.Wolberg, O.L.Mangasarian

[3] Appunti del Corso di Machine Learning del professor Davide Maltoni - Università di Bologna

[4] Appunti del Corso di Machine Learning del professor Fabio Stella - Università degli Studi di Milano Bicocca