Breast cancer’s classification

Davide Luperi, 826249

Francesca Motta, 830107

Filippo Angelico, 849105

Abstract

Il lavoro si è svolto in merito all’analisi di un dataset che raccoglie informazioni su pazienti con cancro al seno.
Per ognuna delle 569 pazienti il cancro è stato diagnosticato come maligno o benigno.
L’obiettivo dell’analisi è di classificare il cancro in base alle informazioni raccolte relative a 30 caratteristiche di nuclei cellulari ottenuti da immagini digitalizzate di un agoaspirato (FNA) di una massa mammaria.

1 Introduzione

Lo scopo dell’analisi è di classificare il tipo di cancro. Nello specifico, non ci si è limitati a massimizzare l’accuracy ma, valutando il contesto ci si è focalizzati anche sulla specificity.
In questo report si trova la spiegazione delle scelte fatte durante le varie fasi dell’analisi. Dopo aver introdotto il dataset, sono riportate le scelte fatte in fase di preprocessing (in particolare per la selezione delle variabili). Successivamente, sono stati introdotti gli algortimi, tra i quali è stato scelto quello più adatto alla risoluzione del problema di classificazione. Questo lavoro si è concentrato sugli algoritmi di Support Vector Machine (lineare e non), Random Forest e K-Nearest Neighbors.
Infine, sono esposti i risultati trovati e degli spunti per lavori successivi.

2 Materiali

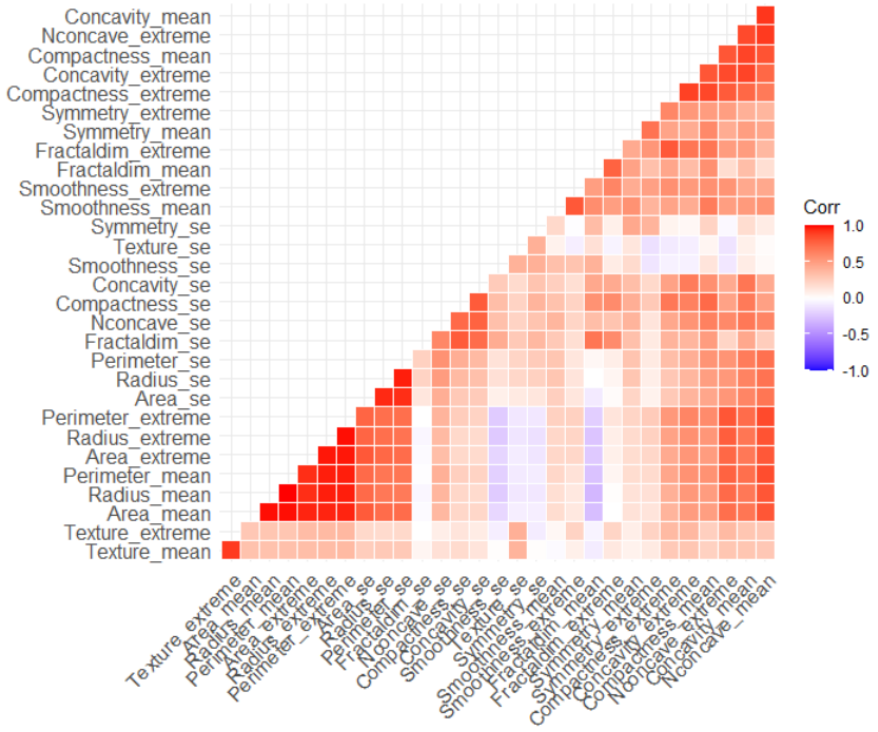
Il dataset su cui è stato condotto il lavoro presenta 32 variabili che raccolgono informazioni provenienti da uno studio su un campione di 569 osservazioni.

Qui sotto sono elencate le variabili presenti nel dataset di partenza:

- id: ID number del paziente
- diagnosis: tipologia di cancro (M = maligno, B = benigno)

Le prossime varibili riguardano le caratteristiche rilevate del nucleo della cellula tumorale:

- Radius: distanza dal centro ad un punto del perimetro del nucleo
- Texture: deviazione standard della scala di grigio
- Perimeter: perimetro del nucleo
- Area: area del nucleo
- Smoothness: variazione locale della lunghezza del raggio del nucleo
- Compactness: perimeter^2 / area - 1.0
- Concavity: intensità delle parti concave della superficie del nucleo
- Nconcave: numero delle parti concave del contorno del nucleo
- Symmetry
- Fractaldim



Per ognuna di queste grandezze sono riportate tre variabili:

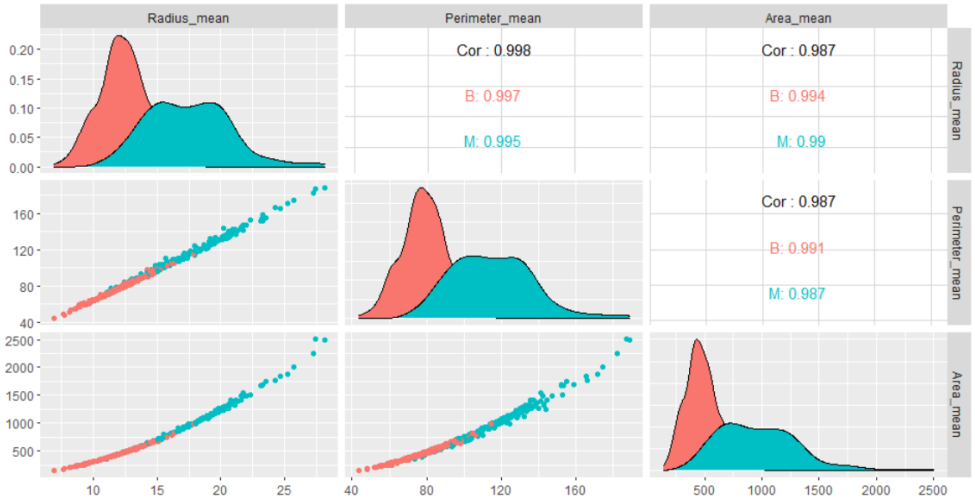
- Mean: la grandezza mediamente rilevata nel paziente in esame
- Se: la deviazione standard dei valori trovati nel paziente in esame
- Extreme: il valore massimo o peggiore trovato nella paziente

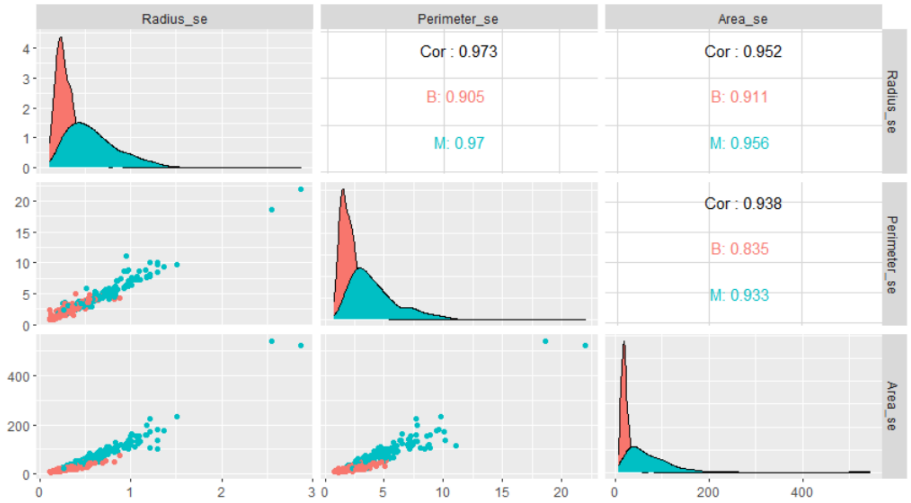
2 Fase di pre-processing

L'analisi della correlazione ha evidenziato la presenza di gruppi di variabili altamente correlate tra loro, come si nota dal grafico soprariportato.

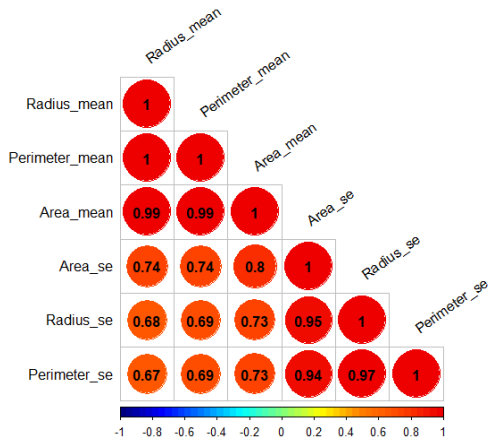
In particolare, si può concentrare l'attenzione su alcuni gruppi di variabili. Queste analisi esplorative risultano molto utili in vista della feature selection da attuare per la costruzione dell'opportuno classificatore.

Per esempio, le tre variabili qui riportate misurano rispettivamente il raggio, il perimetro e l'area del nucleo delle cellule tumurali: come ci si aspetta, sono altamente correlate. Inoltre, si intuisce dal grafico la loro elevata capacità discriminatoria rispetto alla classe della variabile risposta.

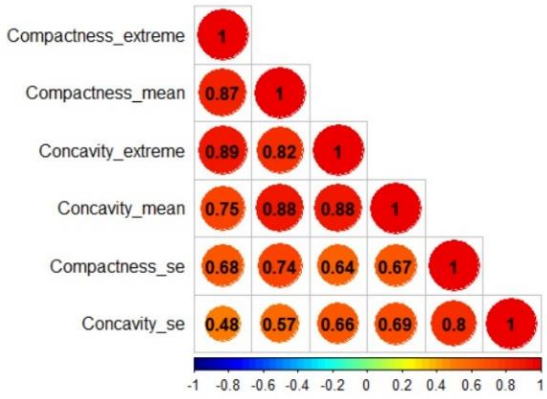




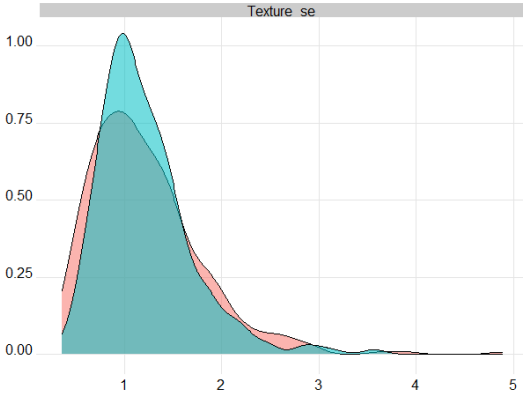
Dal grafico a lato si vede che alcune variabili potrebbero dare un contributo maggiore alla classificazione rispetto ad altre. Infatti, in queste variabili le classi di risposta sono abbastanza sovrapposte a differenza delle variabili illustrate nel grafico precedente.



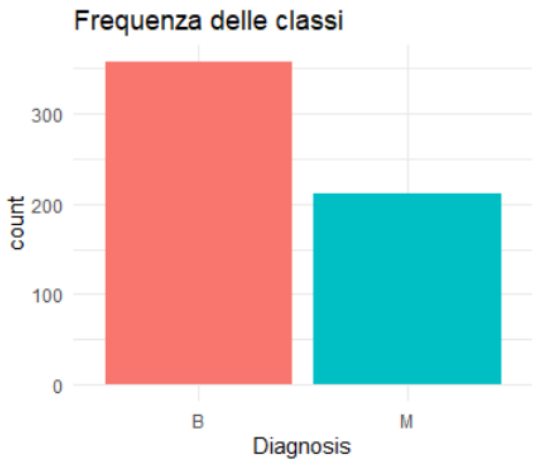
Per una rapida visione della correlazione tra le variabili di questi due gruppi si può far riferimento al grafico a sinistra. Altri sottogruppi di variabili presentano alta correlazione: a titolo di esempio è riportato nel grafico a destra un ulteriore sottogruppo di variabili molto correlate.



Inoltre, dalla distribuzione marginale di alcune variabili sono evidenti distribuzioni delle classi di risposta molto sovrapposte come si vede nel grafico sottostante: queste in fase di selezione non sono quindi state scelte tra le variabili utili per la costruzione del classificatore.



Analizzando così tutte le varibili a disposizione, il lavoro si dirama in due direzioni. In prima analisi, si è tentato la ricerca di un classificatore sulla base di otto sole variabili. In seconda battuta, nel tentativo di migliorare l'accuracy della classificazione, la selezione è ricaduta su tredici variabili nella speranza di non generare classificatori troppo complessi ma in grado di discriminare meglio tra i casi benigni e maligni di tumore.



Inoltre, l'analisi esplorativa ha evidenziato lo sbilanciamento delle classi della variabile risposta. Per risolvere questo problema, il training set risulta un campione retrospettivo del dataset a disposizione: cioè è stato generato in modo tale che le classi siano presenti in ugual misura.

3 Algoritmi di classificazione

I classificatori presi in analisi si basano su diversi algoritmi:

- Support Vector Machine (SVM)
- Random Forest (RF)
- K-Nearest Neighbors (KNN)

Il dataset è stato diviso in training set e test set in modo che il training fosse bilanciato rispetto alla variabile risposta. Nello specifico, si è campionato l'80% delle unità statistiche appartenenti alla classe minoritaria e si è completato il training set campionando lo stesso numero di osservazioni dalla classe maggioritaria.

La procedura di stima Cross-Validation è stata di tipo 10-fold ed è stata ripetuta per 30 volte per evitare che l'assegnazione delle unità statistiche ai diversi sottoinsiemi creati influenzasse la stima del classificatore.

3.1 Support Vector Machine

Il primo classificatore preso in analisi è quello basato sulla SVM. Si è considerato il caso lineare ed anche le due trasformazioni kernel più comuni: polynomial e radial.

Per ogni classificatore analizzato, si è considerata una griglia di valori per gli iperparametri e si sono scelti quelli che massimizzavano l'accuracy.

In prima istanza, si è considerato il dataset composto da otto variabili esplicative. Si è iniziato stimando il Linear SVM. Nonostante l'accuracy sul validation set fosse soddisfacente (95,25%), si è cercato un miglioramento del risultato provando a stimare

anche con classificatori più complessi che considerassero una trasformazione kernel. Questi ultimi hanno performato in modo analogo e con un leggero miglioramento rispetto al classificatore lineare.

Successivamente, si è spostata l'attenzione sul dataset composto da dodici covariate. C'è stato un miglioramento in termini di accuracy per ognuno dei tre classificatori ed analogamente a prima i due classificatori implementati con i kernel hanno una simile performance.

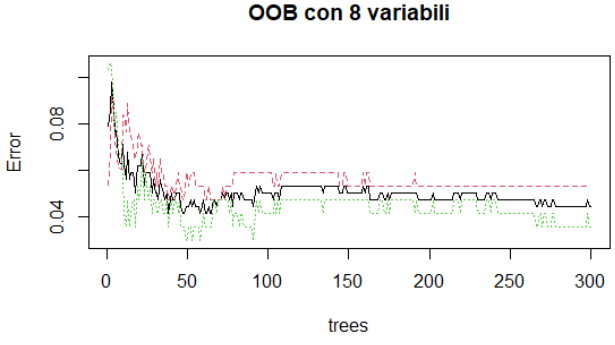
Nella tabella sottostante sono riportati i risultati ottenuti per tutti i classificatori esaminati:

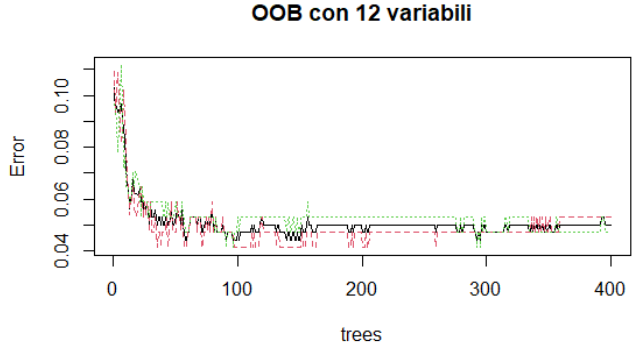
	8 ESPLICATIVE			12 ESPLICATIVE		
	Iperparametri	Accuracy	Specificity	Iperparametri	Accuracy	Specificity
L-svm	C=0.1	0.9525	0.9706	C=1	0.9559	0.9588
P-svm	C=1 Grade=2 Scale=0.1	0.9560	0.9824	C=1 Grade=2 Scale=0.1	0.9631	0.9882
R-svm	C=1 Sigma=0.1	0.9564	0.9588	C=10 Sigma=0.01	0.9625	0.9706

3.2 Random Forest

Gli iperparametri sono stati così scelti:

- Come numero di variabili campionate casualmente come candidate per ogni split di ogni albero è stata scelta la radice quadrata del numero delle esplicative, utilizzato di default dalla funzione randomForest nell'omonimo pacchetto di R;
- Per il numero totale di alberi considerati, si è iniziato valutando 1000 e si sono fatte più prove cercando di minimizzare questo numero senza però incrementare significativamente l'errore OOB. Le scelte fatte non hanno riscontrato un peggioramento in termini di accuracy.

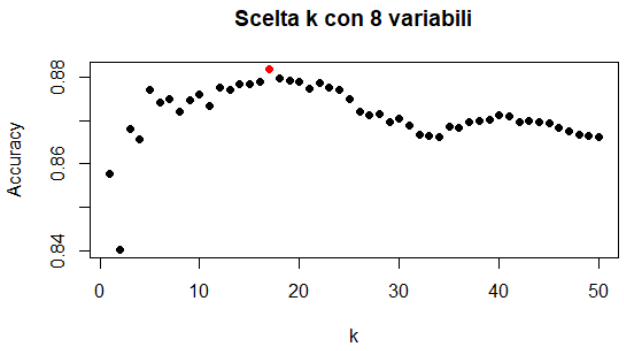


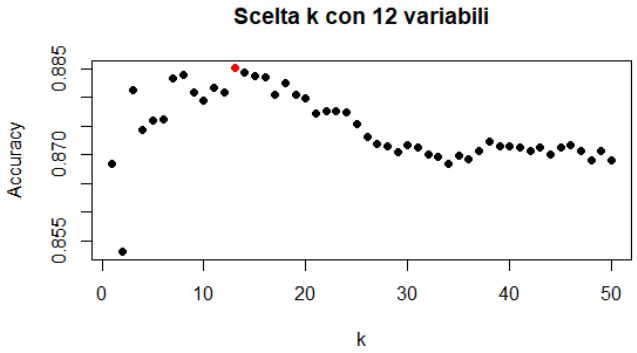


Il RF valutato sul dataset con otto esplicative ha portato a un'accuracy del 95,29% e una specificity pari a 93,77%. Con il RF implementato con dodici esplicative si ottinene un'accuracy del 95,88% e una specificity pari a 94,83%.

3.2 K-Nearest Neighbors

Anche in questo caso si sono valutati i due casi che differiscono per il numero di esplicative utilizzate. L'iperparametro è stato scelto su una griglia di valori usando come criterio la massimizzazione dell'accuracy (si vedano i grafici sottoriportati).





Il k ottimo per il classificatore KNN costruito con il dataset contenente otto esplicative è risultato pari a 17 e si è ottenuta un'accuracy del 88,17%. Invece, implementando il classificatore con il dataset con dodici variabili si è ottunuto il k ottimale pari a 13 ed un'accuracy uguale a 88,51%.

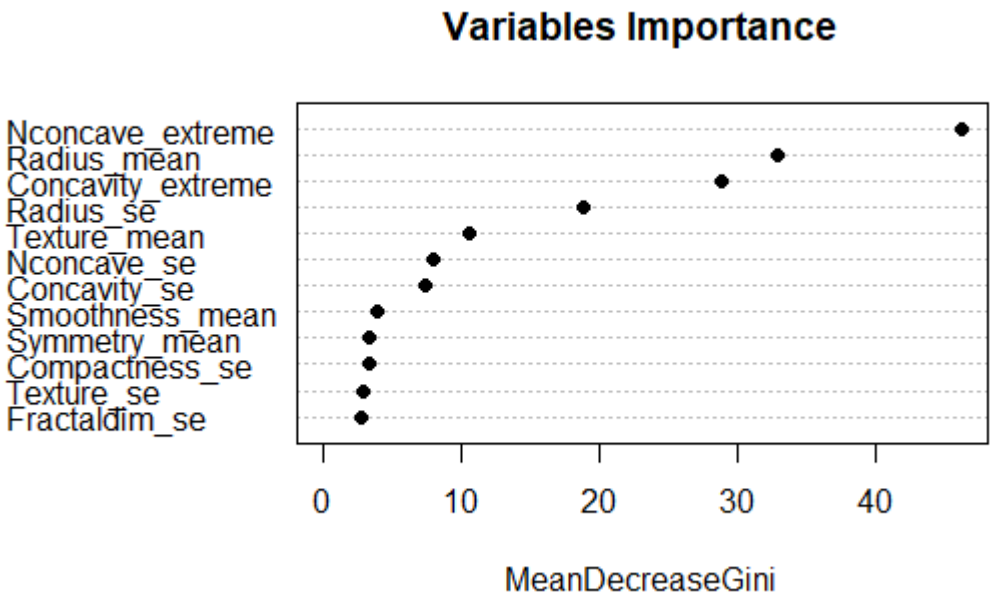
4 Risultati

Considerando le dimensioni ridotte del dataset e nel rispetto degli obiettvi prefissati, tra questi classificatori si è preferito validare sul test set il SVM Polynomial mapping valutato sul dataset con 12 variabili esplicative sebbene sia computazionalmente più oneroso rispetto ad altri presi in considerazione. Il classificatore ha infatti ottenuto i migliori risultati sia in termini di accuracy (96,31%) sia di specifity (98,82%).

Ponendo l'attenzione su classificatori che permettano l'interpretazione delle classi tramite le variabili esplicative considerate, si è scelto di validare sul test set anche il classificatore basato sul Random Forest. Questo è coerente con la scelta in fase di variable selection di affidarsi ad analisi esplorative relative alle distribuzioni marginali ed alle correlazioni.

L'accuracy ottenuta sul test set dal SVM Polynomial mapping è pari a 98,25%. Invece, la specificity, di importante rilevanza per il problema di corretta specificazione di un tumore maligno, è pari a 97,62% (si è osservato un solo falso negativo sui 42 veri negativi presenti nel test set).

Il Random Forest implementato con 12 esplicative ha riportato sul test set un'accuracy uguale a 96,51% e una specificity pari a 92,86%. Come ci si aspettava, i risultati sono peggiori rispetto al classificatore precedentemente analizzato. Il vantaggio di questo classificatore è di fornire un'indicazione riguardo all'importanza delle singole esplicative nel discriminare tra le classi della varibile dipendente.



5 Discussione

Il miglior classificatore trovato, in termini di specificity, riesce a individuare correttamente un cancro al seno maligno con il 97,62% di probabilità.
Considerando le dimensioni ridotte del dataset, la scelta del classificatore è ricaduta su uno di tipo non lineare. Se si fosse interessati a garantire un minor sforzo computazionale, la scelta potrebbe ricadere sul Linear SVM che riporta comunque risultati soddisfacenti.
In questo lavoro, la variable selection è stata determinata con la sola analisi esplorativa: analizzando le distribuzioni marginali e le correlazioni. Un approccio alternativo che potrebbe essere spunto per lavori successivi è quello di eseguire una feature selection sulla base della pca o comunque considerando combinazioni delle variabili di partenza.