BREAST CANCER WISCONSIN

Bettini Ivo Junior (matricola 806878)
Cocca Umberto (matricola 807191)
Traversa Silvia (matricola 816435)

INTRODUZIONE

Per il nostro progetto abbiamo scelto il dataset Breast Cancer Wisconsin (Diagnostic).
Il dataset contiene 32 attributi, i quali descrivono le caratteristiche dei nuclei cellulari presenti in immagini digitalizzate di un preparato biologico agoaspirato fine (FNA) di una massa di una ghiandola mammaria.

Per ogni nucleo cellulare sono state effettuate 10 misurazioni:

1. radius (media delle distanze dal centro ai punti sul perimetro)
2. texture (deviazione standard dei valori della scala dei grigi)
3. perimeter
4. area
5. smoothness (variazione locale delle lunghezze del raggio)
6. compactness (perimetro^2 / area - 1.0)
7. concavity (gravità delle porzioni concave del contorno)
8. concave points (numero di porzioni concave del contorno)
9. symmetry
10. fractal dimension ("approssimazione coastline" - 1)

Per ognuna di queste vengono riportati i valori della media, dell'errore standard e il valore peggiore/più grande (media dei tre valori maggiori).

L'obiettivo del nostro elaborato è poter classificare la tipologia di una massa tumorale individuata nella mammella in benigna (B) o maligna (M).

CREAZIONE DEL DATASET

Dopo l'importazione del dataset si è verificata l'eventuale presenza di valori nulli e si è riscontrata la loro assenza, motivo per il quale non è stato necessario ricorrere ad alcun metodo di rimozione o approssimazione di valori mancanti.
È stato necessario, però, eliminare la prima colonna (e quindi passare da 31 a 30 attributi) poiché conteneva l'id del paziente sottoposto all'esame.

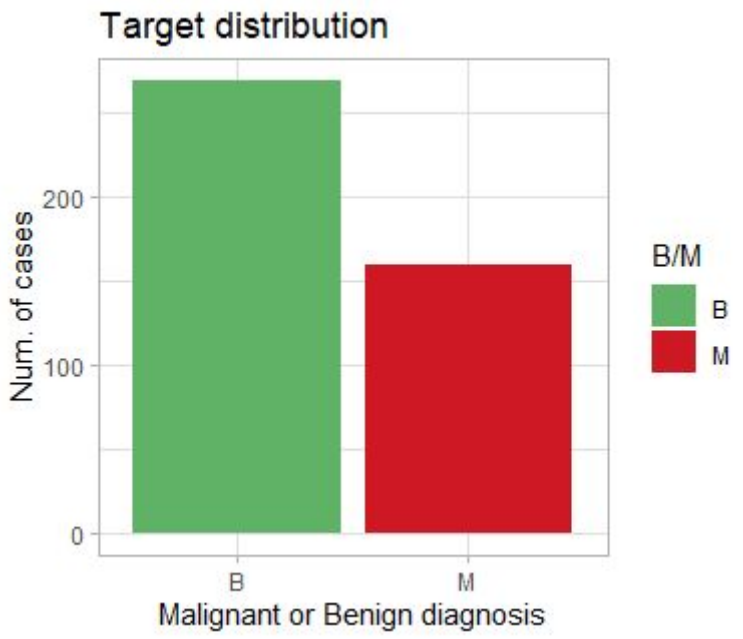
```
      radius_mean      texture_mean    perimeter_mean
        "numeric"         "numeric"         "numeric"
        area_mean   smoothness_mean  compactness_mean
        "numeric"         "numeric"         "numeric"
  concavity_mean concave.points_mean     symmetry_mean
        "numeric"         "numeric"         "numeric"
fractal_dimension_mean    radius_se        texture_se
        "numeric"         "numeric"         "numeric"
     perimeter_se         area_se      smoothness_se
        "numeric"         "numeric"         "numeric"
  compactness_se      concavity_se concave.points_se
        "numeric"         "numeric"         "numeric"
     symmetry_se fractal_dimension_se      radius_worst
        "numeric"         "numeric"         "numeric"
   texture_worst   perimeter_worst        area_worst
        "numeric"         "numeric"         "numeric"
 smoothness_worst  compactness_worst   concavity_worst
        "numeric"         "numeric"         "numeric"
concave.points_worst   symmetry_worst fractal_dimension_worst
        "numeric"         "numeric"         "numeric"
```

Osservando la tipologia dei dati presenti non è stato ritenuto necessario applicare alcuna modifica riguardo questo aspetto.

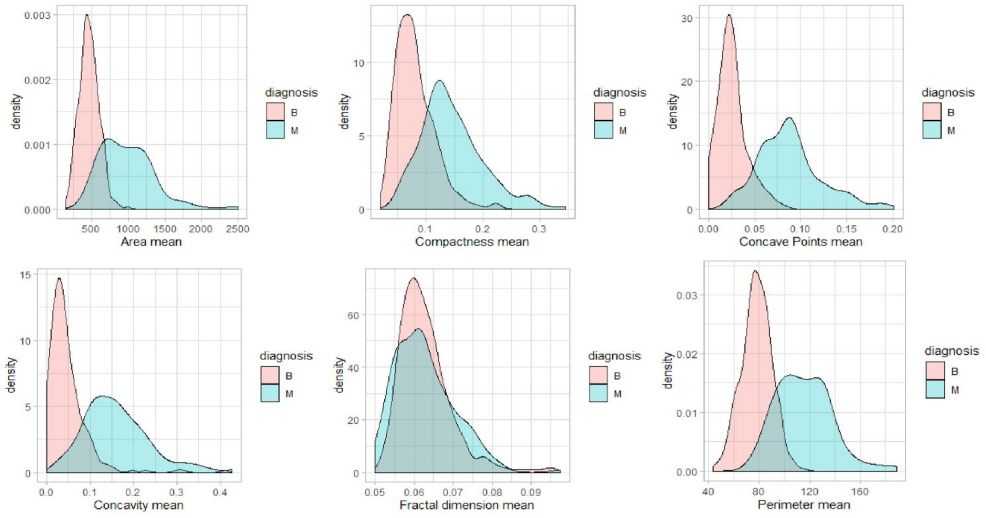
Dopo questi controlli possiamo dunque affermare che il dataset contiene 569 righe e 30 attributi, con nessun valore nullo.

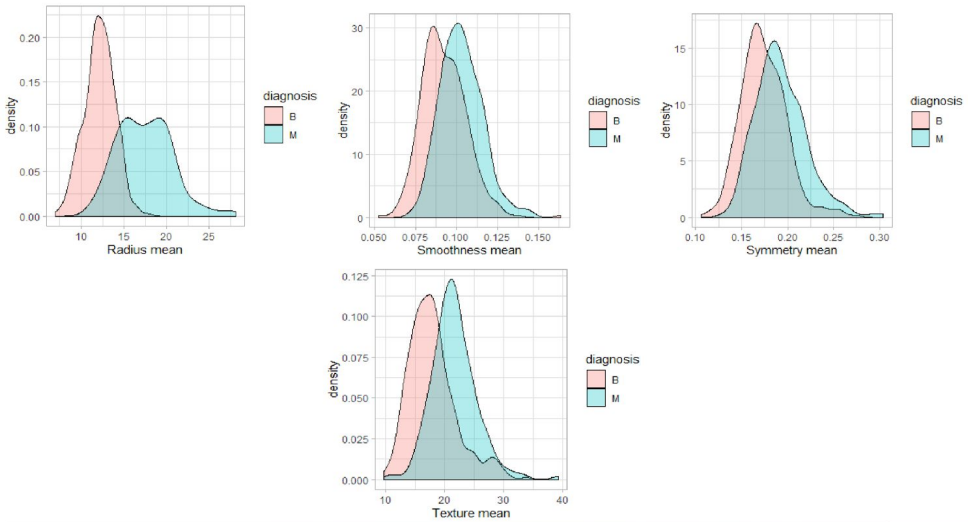
ANALISI ESPLORATIVA

Come primo passo della nostra analisi esplorativa dei dati abbiamo deciso di visualizzare la distribuzione del nostro target, evidenziare dunque la differenza di casi di diagnosi di tumori benigni e maligni (62.7% contro 37.2%). Quello che abbiamo ottenuto è il seguente grafico:

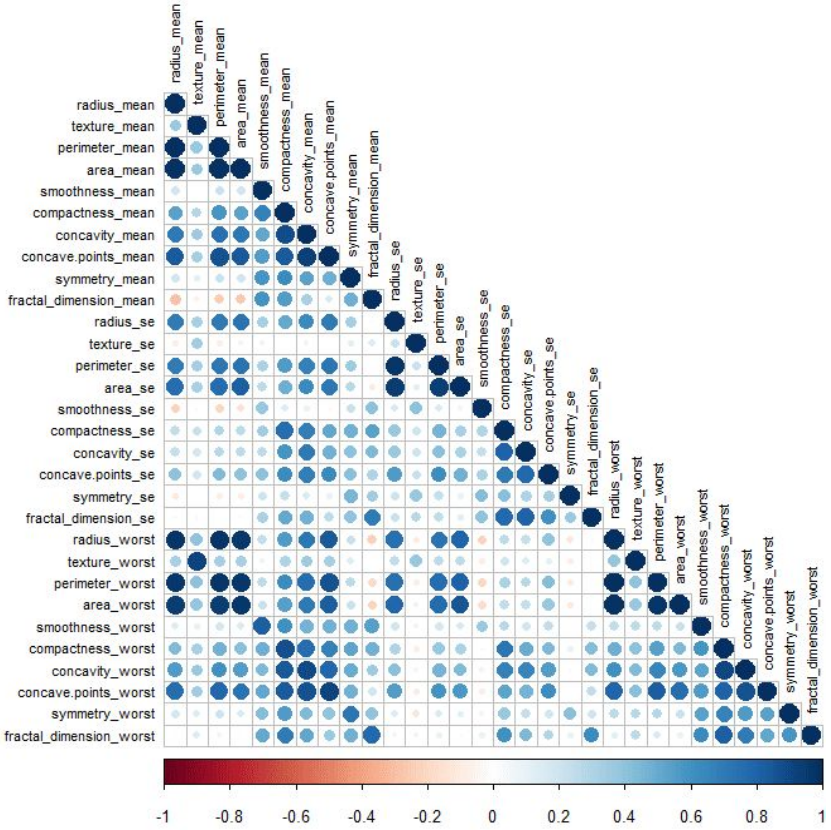


Abbiamo voluto visualizzare la distribuzione di un sottogruppo di variabili (le misurazioni medie) rispetto al target tramite dei density plot.



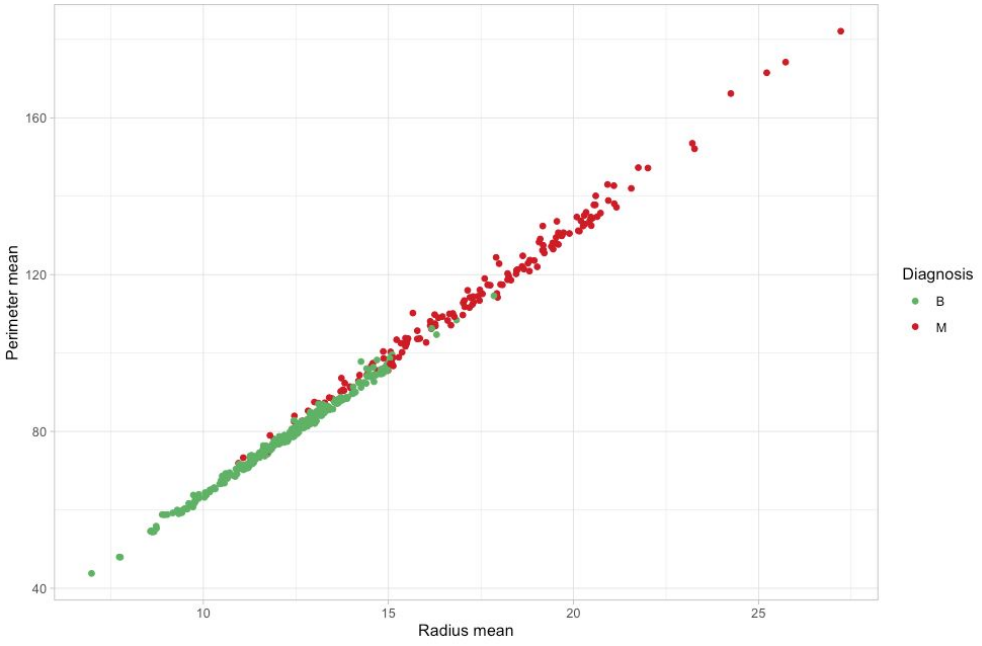


Successivamente abbiamo deciso di analizzare la correlazione fra gli attributi del dataset, in modo da individuare possibili coppie di attributi fortemente correlati.

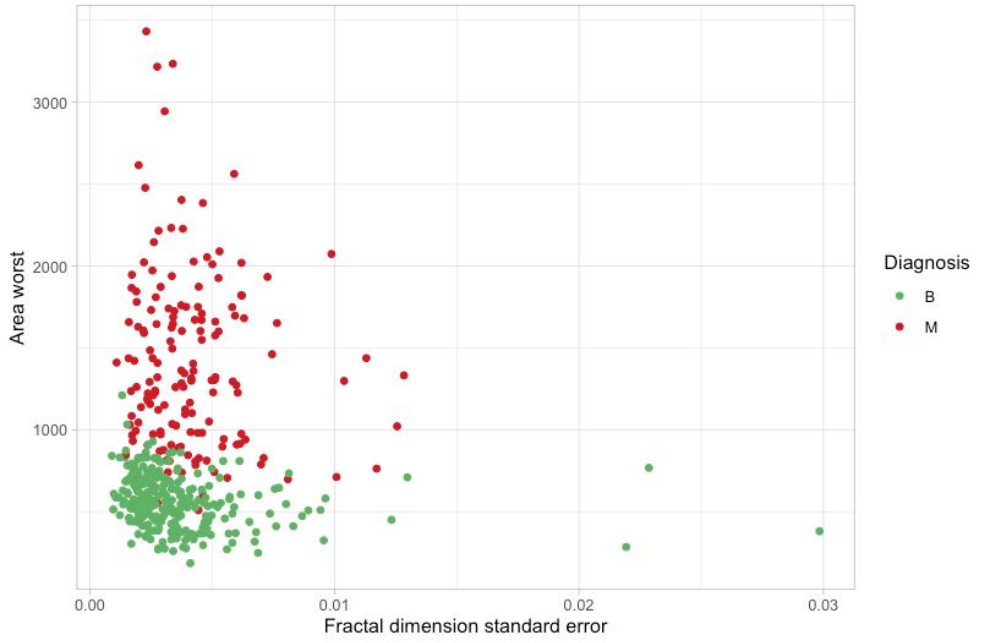


Alla luce del risultato ottenuto abbiamo deciso di visualizzare graficamente le relazioni esistenti tra alcune variabili. Per fare ciò, abbiamo voluto evidenziare un caso di alta correlazione e un caso di non correlazione fra due variabili.

Nel caso di correlazione alta abbiamo scelto le variabili *perimeter_mean* e *radius_mean*, ottenendo il seguente risultato:



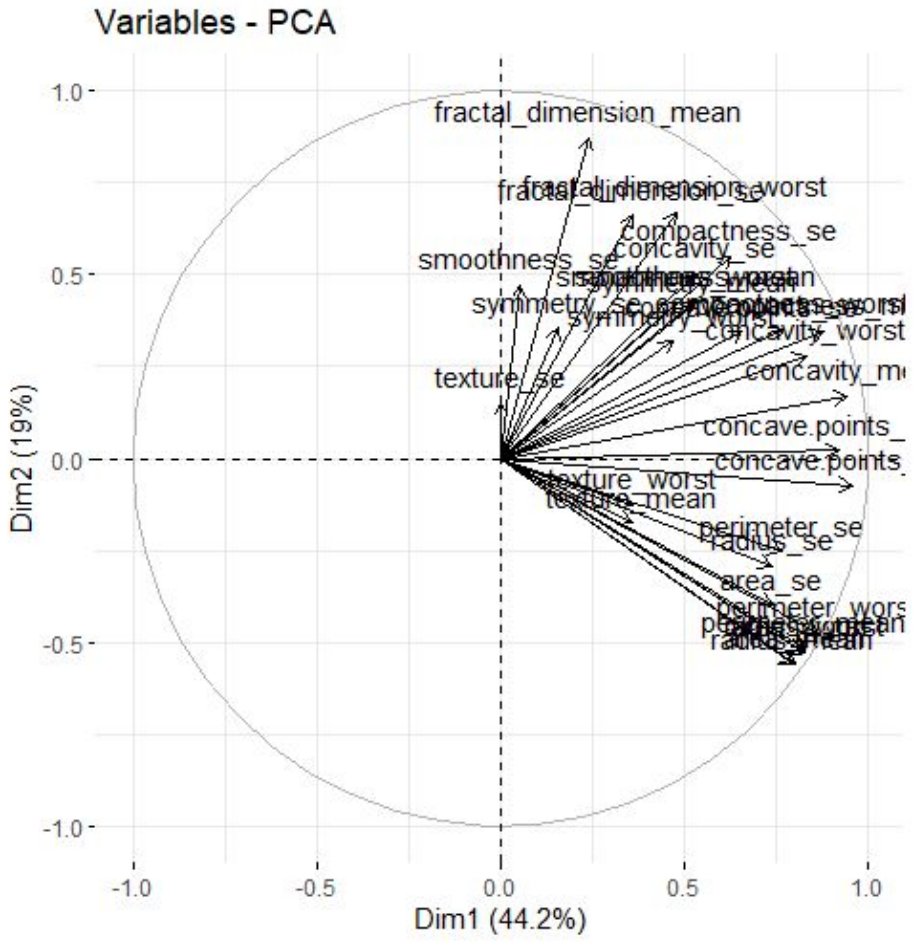
Per quanto riguarda, invece, la bassa correlazione abbiamo analizzato le variabili *area_worst* e *fractal_dimension_standard_error*. Abbiamo ottenuto ciò:



A colpo d'occhio si può notare la correlazione dalla distribuzione dei punti nel grafico: dove è alta si segue la forma di una retta, dove è bassa i punti sono dispersi.

Dopo aver notato la forte correlazione esistente fra alcune delle variabili del nostro dataset, abbiamo deciso di implementare la PCA (Principal Component Analysis) per verificare la possibilità di poter ridurre il numero di feature necessarie per poter addestrare i nostri modelli con un'accuratezza soddisfacente.

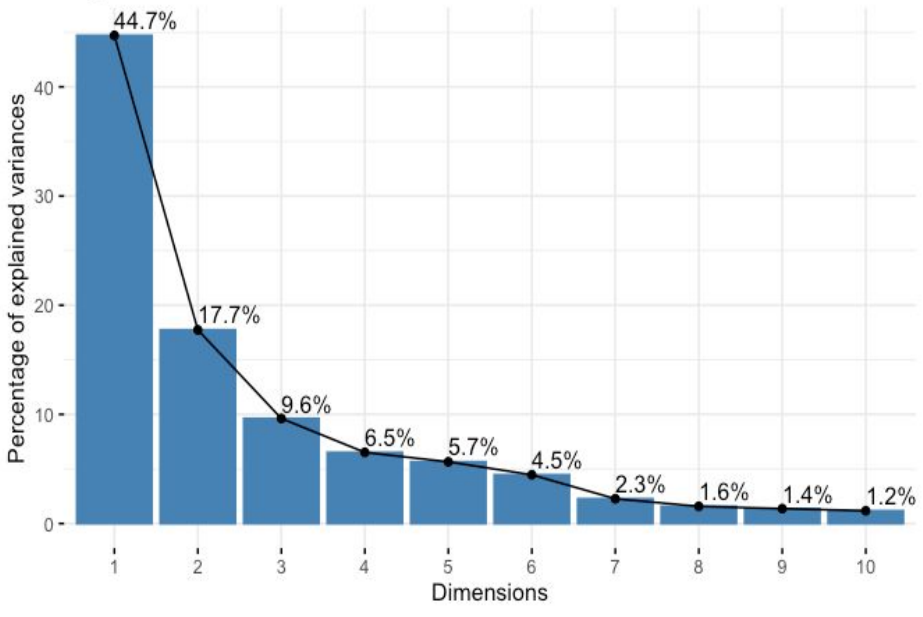
Questa analisi consiste in una trasformazione lineare che mappa il dataset in un nuovo sistema di coordinate dove, dalla prima all'ultima, troviamo le variabili con varianza più significativa in ordine decrescente. Nel grafico seguente sono rappresentate le variabili estratte dalla PCA e, in particolare, quelle che si trovano nello stesso quadrante hanno una forte correlazione.



È necessario effettuare uno studio degli autovalori per poter individuare le dimensioni con varianza maggiore, ossia quelle che portano più informazione. Gli autovalori, infatti, sono delle misurazioni che calcolano la quantità di varianza contenuta nelle componenti principali (PC). Otteniamo il seguente risultato:

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	1.341230e+01	4.470767e+01	44.70767
Dim.2	5.322313e+00	1.774104e+01	62.44871
Dim.3	2.887710e+00	9.625699e+00	72.07441
Dim.4	1.955722e+00	6.519074e+00	78.59348
Dim.5	1.695961e+00	5.653205e+00	84.24669
Dim.6	1.339653e+00	4.465511e+00	88.71220
Dim.7	6.808593e-01	2.269531e+00	90.98173
Dim.8	4.723430e-01	1.574477e+00	92.55620
Dim.9	4.096183e-01	1.365394e+00	93.92160
Dim.10	3.503131e-01	1.167710e+00	95.08931
Dim.11	3.212507e-01	1.070836e+00	96.16014
Dim.12	2.741948e-01	9.139828e-01	97.07413
Dim.13	2.243637e-01	7.478791e-01	97.82201
Dim.14	1.626528e-01	5.421760e-01	98.36418
Dim.15	9.417819e-02	3.139273e-01	98.67811
Dim.16	8.605493e-02	2.868498e-01	98.96496
Dim.17	6.054598e-02	2.018199e-01	99.16678
Dim.18	5.219103e-02	1.739701e-01	99.34075
Dim.19	3.729911e-02	1.243304e-01	99.46508
Dim.20	3.311969e-02	1.103990e-01	99.57548
Dim.21	3.020199e-02	1.006733e-01	99.67615
Dim.22	2.463084e-02	8.210281e-02	99.75825
Dim.23	2.148986e-02	7.163287e-02	99.82989
Dim.24	1.706740e-02	5.689134e-02	99.88678
Dim.25	1.581307e-02	5.271023e-02	99.93949
Dim.26	9.008969e-03	3.002990e-02	99.96952
Dim.27	6.737998e-03	2.245999e-02	99.99198
Dim.28	1.640102e-03	5.467007e-03	99.99745
Dim.29	6.466754e-04	2.155585e-03	99.99960
Dim.30	1.194461e-04	3.981537e-04	100.00000

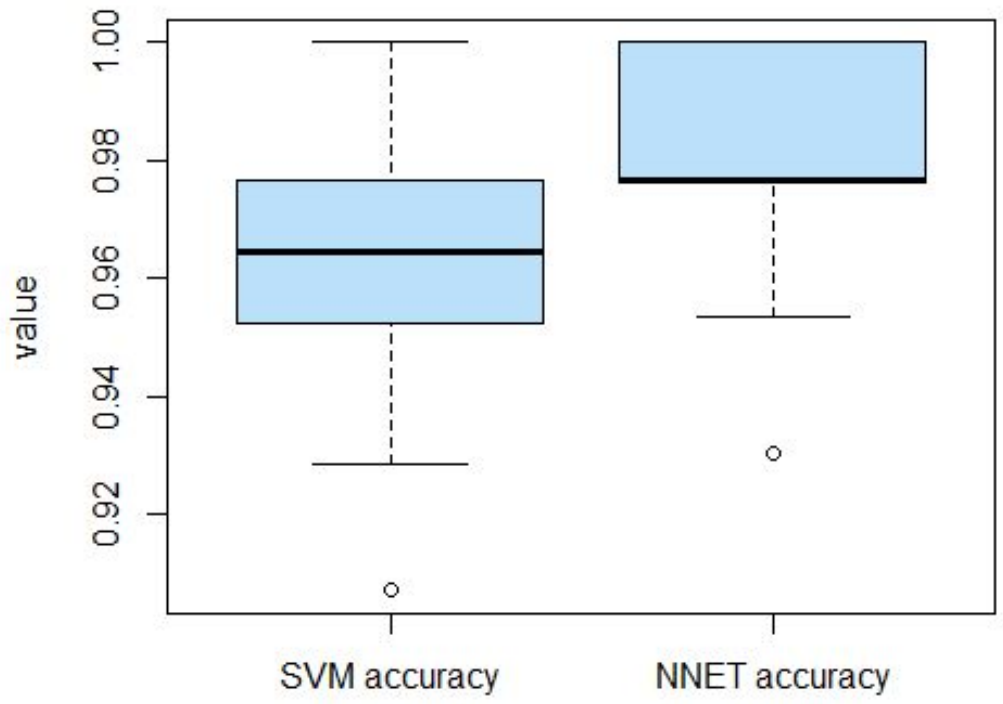
Si può notare che nelle prime 10 dimensioni si ha una varianza cumulativa del 95%, ergo quelle dimensioni bastano per garantire, appunto, il 95% delle informazioni che possono fornire i nostri attributi. Graficamente possiamo visualizzare le percentuali dei componenti della PCA nel seguente modo:



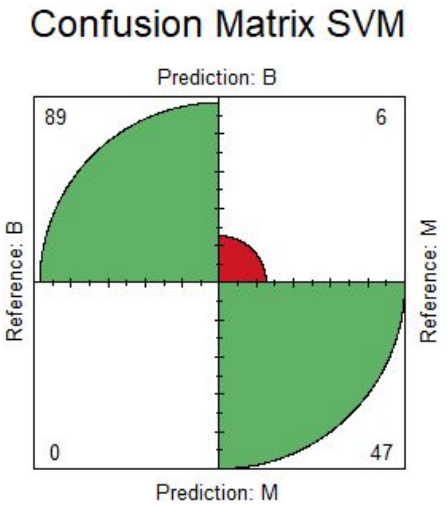
Dopo aver applicato la PCA, attraverso di essa abbiamo effettuato delle predizioni sul trainset e testset. Facendo questo passaggio abbiamo generato nuovi valori, nello specifico un nuovo trainset e testset, che abbiamo utilizzato per la creazione dei nostri modelli.

Effettuando poi uno studio dell'accuratezza dei due modelli di machine learning scelti (che verranno illustrati più in dettaglio successivamente) abbiamo ottenuto un'accuratezza di circa 96%. Questo risultato è molto interessante poiché dimostra che anche generando nuovi dati con la PCA e utilizzando dunque solo un terzo degli attributi, riusciamo a mantenere un'accuratezza elevata.

L'accuratezza media ottenuta con la 10 - fold cross validation (su tre ripetizioni) della SVM è 96,1%, mentre con la rete neurale NNET è 97,9%, e si può visualizzare il confronto nel seguente grafico:

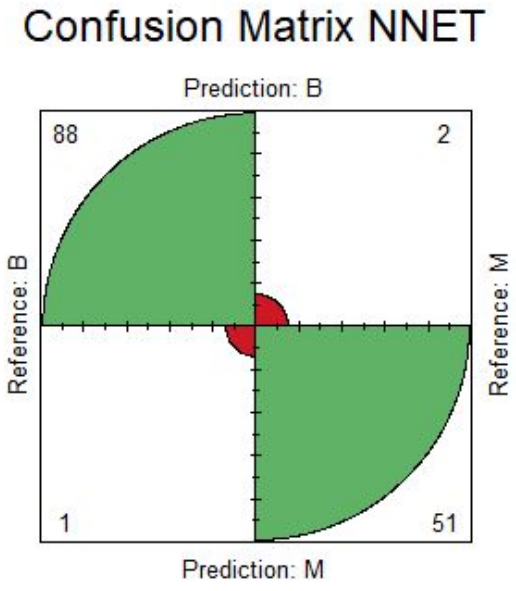


La matrice di confusione per la SVM ottenuta dopo la predizione sul testset è la seguente:



Con un livello di accuratezza del 95% si ottengono zero falsi positivi e sei falsi negativi sottolineando ulteriormente l'accuratezza del modello.

Con la rete neurale NNET abbiamo ottenuto un'accuratezza del 98%, un risultato migliore. La matrice di confusione riporta i seguenti risultati:



Anche in questo caso il risultato è ottimo, avendo ottenuto un solo falso positivo e solo due falsi negativi.

MODELLI DI MACHINE LEARNING

I modelli di machine learning che abbiamo deciso di utilizzare sono due modelli studiati a lezione, la rete neurale e la support vector machine.

La scelta è ricaduta su questi in quanto utilizzano due approcci che possiamo definire contrapposti per l'apprendimento: infatti, mentre la rete neurale simula attraverso un modello matematico il comportamento di un sistema fisico e consente di approssimare in uno specifico contesto la corrispondenza esistente tra un ingresso e un'uscita, la svm cerca di costruire un iperpiano di separazione che massimizzi la distanza tra elementi di classi diverse.

ESPERIMENTI

Abbiamo eseguito diversi esperimenti per valutare la performance dei modelli della support vector machine e la rete neurale. Eseguiamo una 10 - fold cross validation (su tre ripetizioni), la matrice di confusione sui dati di testing e poi sono state calcolate ulteriori stime delle seguenti performance: Precision, Recall, F - Measure, ROC e AUC.

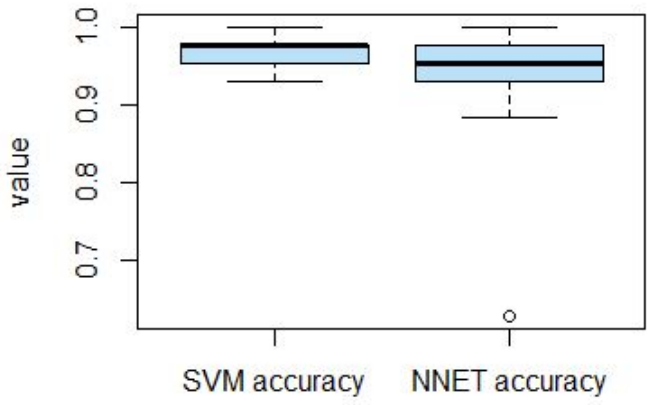
Spiegheremo di seguito ognuno di questi esperimenti.

10 - fold cross validation

Nella fase 10 - fold cross validation il training set viene suddiviso in 10 parti in modo randomico e per 10 volte vengono usate 9 parti come training set e 1 parte come testing sempre diversi a ogni iterazione. Nella funzione del controllo dell'addestramento è stato inoltre settato il parametro ripetizioni uguale a 3, in questo modo la 10 - fold CV viene ripetuta per tre volte, tutto al fine di ottenere un livello di accuratezza più preciso.

Addestramento dei modelli

Addestriamo entrambi i modelli usando tutti gli attributi del trainset. Nonostante i differenti approcci sintetizzati precedentemente, i risultati che otteniamo sono molto buoni e simili. Infatti in entrambi i casi otteniamo, per mezzo della 10 - fold cross validation (su tre ripetizioni), un'accuratezza media di poco superiore al 94%.

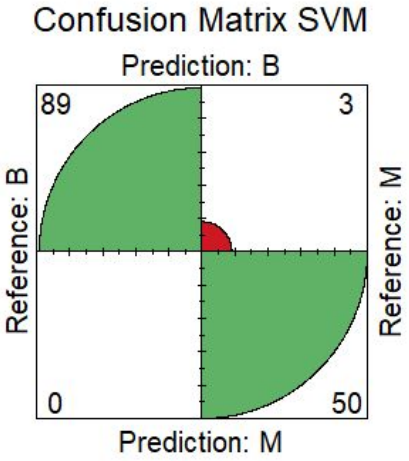


Nello specifico, con il modello della support vector machine si passa da un minimo di 92% a un massimo di 100% e per la rete neurale da un minimo dell'88% a un massimo del 100% con la rete neurale. Le medie, invece, sono circa 97% e 95% per la support vector machine e la rete neurale rispettivamente. Notiamo quindi che la svm ha un comportamente leggermente migliore rispetto alla rete neurale.

Matrice di confusione

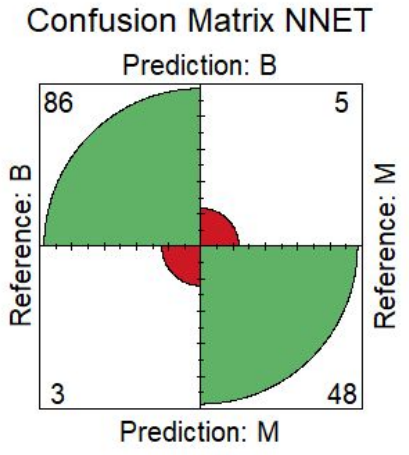
Con la matrice di confusione vogliamo calcolare l'accuratezza statistica della predizione dei dati che esegue la nostra macchina. Questo calcolo sarà effettuato sui modelli di machine learning che abbiamo scelto (SVM e rete neurale).

La support vector machine ha ottenuto dei piccoli punti percentuali maggiori di accuratezza rispetto alla rete neurale, vediamo di seguito la sua matrice di confusione ottenuta dalla predizione sui dati di test:



Dall'accuratezza alta non potevamo che aspettarci dei risultati ottimi nella predizione e infatti notiamo che degli 89 casi riguardanti tumori benigni, tutti gli 89 sono stati predetti correttamente. In parallelo vediamo dei 53 casi maligni ci sono 50 casi classificati correttamente e 3 casi predetti come benigni seppur maligni.

La matrice di confusione della rete neurale è la seguente:



Anche qui i risultati sono molto buoni, infatti, vengono predetti correttamente 86 casi su 89 come benigni e 3 casi come maligni seppur benigni. Al contrario per quanto riguarda il predire i casi maligni la rete neurale ha predetto correttamente 48 casi su 53, ma 5 casi sono stati classificati come benigni seppur rappresentano i maligni.
In entrambe le classificazioni si ottengono dei risultati migliori con la support vector machine.

Curva ROC e valore AUC

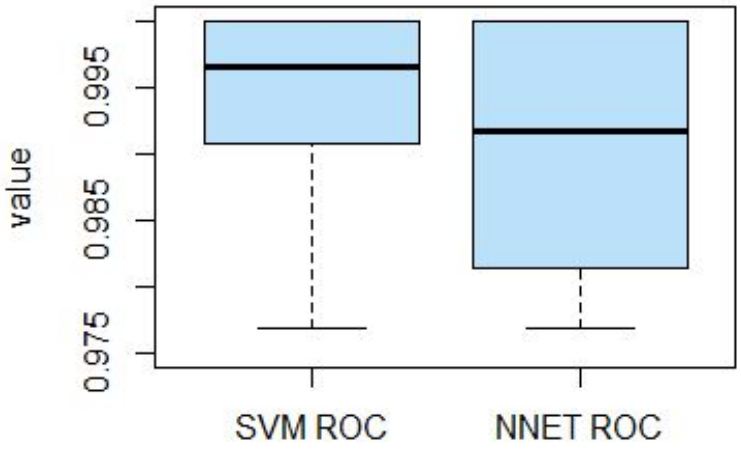
La curva ROC è un grafico rappresentato in un piano dove lungo i due assi si rappresentano la sensibilità e 1-specificità, rispettivamente rappresentati da True Positive Rate (TPR, frazione di veri positivi) e False Positive Rate (FPR, frazione di falsi positivi). In altre parole, si studiano i rapporti fra allarmi veri (hit rate) e falsi allarmi.

Attraverso l'analisi delle curve ROC si valuta la capacità del classificatore di distinguere con sufficiente chiarezza, calcolando l'area sottesa alla curva ROC (Area Under Curve, AUC). Il valore di AUC, compreso tra 0 e 1, equivale infatti alla probabilità che il risultato del classificatore applicato ad un individuo estratto a caso dal gruppo dei malati sia superiore a quello ottenuto applicandolo ad un individuo estratto a caso dal gruppo dei sani.

Le curve ROC passano per i punti (0,0) e (1,1), e hanno inoltre due condizioni le quali rappresentano due curve limite:

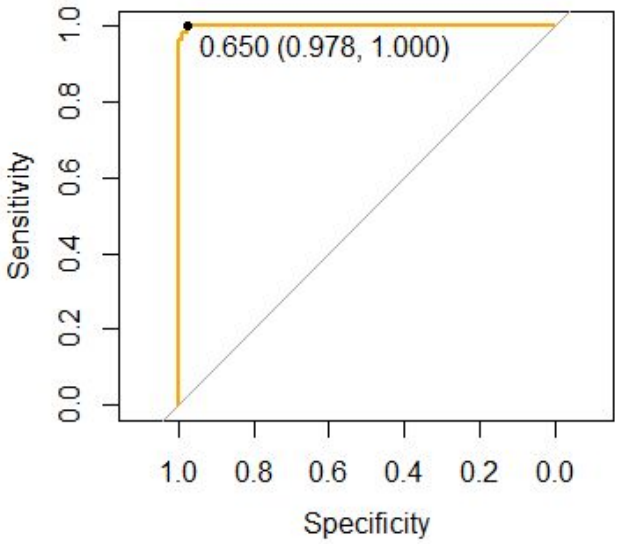
Una taglia il grafico a 45°, passando per l'origine. Questa retta rappresenta il caso del classificatore casuale, la cui area sottesa AUC è pari a 0,5.
La seconda curva è rappresentata dal segmento che dall'origine sale al punto (0,1) e da quello che congiunge il punto (0,1) a (1,1), avendo un'area sottesa di valore pari a 1 rappresenta il classificatore perfetto.

Graficamente, ecco di seguito rappresentato il confronto fra l'addestramento del modello con la metrica ROC nel caso di SVM e di NNET:

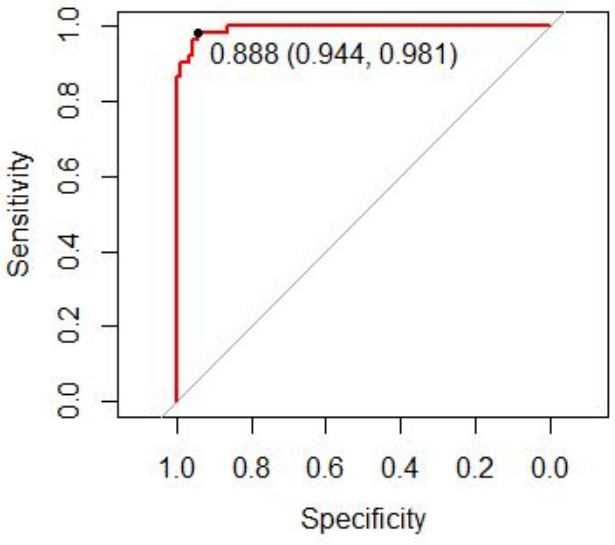


Sia nel caso del modello SVM che nel modello della rete neurale otteniamo degli ottimi risultati, infatti le due curve si distaccano di molto dalla retta che indica un classificatore casuale, raggiungendo quasi il caso perfetto.

Curva ROC della SVM con area under the curve (AUC): 0.9994



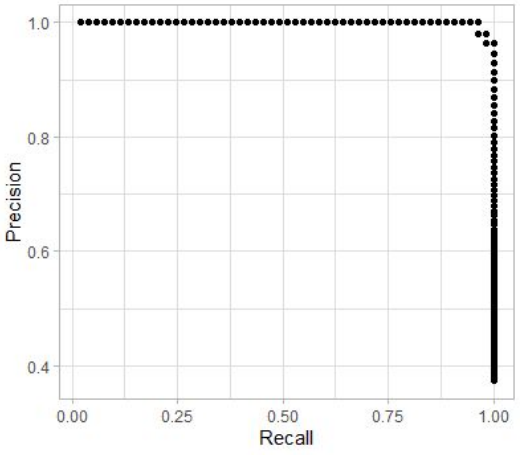
Curva ROC della rete neurale con area under the curve (AUC): 0.9936

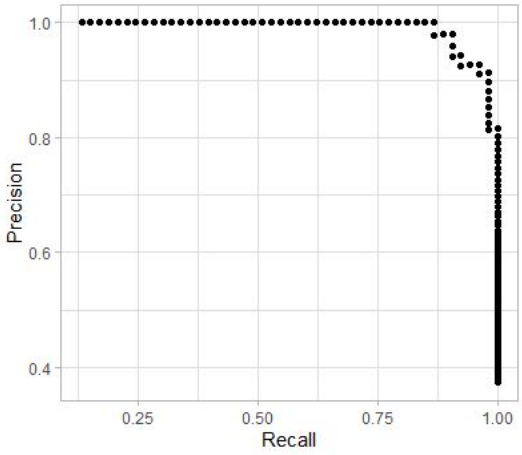


Precision, Recall e F - Measure

La Precision (chiamata anche valore predittivo positivo) è la frazione di istanze rilevanti tra le istanze recuperate, mentre il Recall (noto anche come sensibilità) è la frazione della quantità totale di istanze rilevanti effettivamente recuperate. Sia la precision che il recall si basano quindi su una comprensione e una misura di pertinenza.

Di seguito sono riportati i grafici che rappresentano la relazione fra precision e recall per la SVM (sinistra) e NNET (destra):





La F - Measure (chiamata anche F1 score) è la misura di accuratezza di un determinato test. La misura tiene in considerazione la precisione e il recupero del test, dove la precisione è il rapporto fra numero di veri positivi e il numero di tutti i risultati positivi, mentre il recupero è il rapporto fra il numero di veri positivi e il numero di tutti i test che sarebbero dovuti risultare positivi. La F - Measure viene calcolata tramite le media armonica di precision e recall.

I valori della F - Measure per la SVM e NNET sono rispettivamente 98% e 95%.

ANALISI FINALI E CONCLUSIONI

Ricapitolando tutto quello che abbiamo descritto finora, possiamo affermare di aver ottenuto dei risultati soddisfacenti. Siamo stati in grado di mantenere il 95% dell'informazione contenuta nel dataset riducendo gli attributi da 31 a 10 con l'uso della PCA.

La support vector machine si è comportata meglio rispetto alla rete neurale. Non possiamo dire che la svm sia sempre migliore poiché, se i due modelli venissero addestrati nuovamente oppure se i modelli fossero addestrati su un nuovo sottoinsieme del dataset, questo potrebbe portare a livelli di accuratezza diversi, dove la rete neurale potrebbe rivelarsi migliore.

Possiamo certamente affermare, però, che in entrambi i casi i risultati hanno raggiunto dei livelli di accuratezza molto alti, ma non così alti da superare nettamente quelli della PCA. Nel caso della valutazione dell'accuratezza attraverso la 10 - fold cross validation, infatti, si ottengono risultati migliori con la PCA, 97%, contro il 95% su tutto l'insieme degli attributi.

Infine concludiamo che, visto la tematica delicata nei casi di modellizzazione di tumori, è molto importante tenere sotto controllo la misura della Recall, la quale ci permette di identificare la percentuale di positivi effettivi identificati correttamente. Nel nostro caso con i modelli addestrati su tutti gli attributi con la rete neurale abbiamo una recall del 97%, ciò significa che il modello identifica correttamente il 97% dei tumori benigni. Diversamente il modello della support vector machine ha una recall del 100%, ovvero identifica correttamente il 100% dei tumori benigni.