MODELLO DI MACHINE LEARNING PER LA PREDIZIONE DEL CANCRO AL SENO

FEDERICA D'ANTICO, 829572
MARCO LATELLA, 829498
MICHELE LEPORATI, 834976

Contenuti della relazione

Descrizione del dominio di riferimento e obiettivi dell'elaborato

A partire dai dati contenuti nel dataset "Breast Cancer Prediction Dataset" l'obiettivo dell'elaborato è quello di riconoscere la malignità di un nodulo sospetto, attraverso lo sviluppo di un modello basato su apprendimento supervisionato. La predizione avviene sulla base delle caratteristiche del nodulo.

A tal fine verranno implementati due differenti modelli e attraverso il loro confronto sarà poi possibile considerare quello che presenta le caratteristiche migliori.

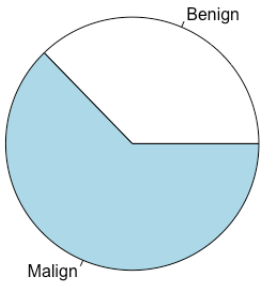
Scelte di design per la creazione del data set, eventuali ipotesi o assunzioni

Il dataset "Breast Cancer Prediction Dataset" è reperibile sul sito Kaggle al seguente link: https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset . È composto da 569 istanze caratterizzate da informazioni riscontrabili a seguito dell'individuazione di un nodulo con esame radiologico.

Variabile	Descrizione	Tipo
mean_radius	Raggio medio del nodulo sospetto	float
mean_texture	Texture media del nodulo sospetto	float
mean_perimeter	Perimetro medio del nodulo sospetto	float
mean_area	Area media del nodulo sospetto	float
mean_smoothness	Levigatezza media del nodulo sospetto	float
diagnosis	Esito della malignità del nodulo sospetto	int (1, 0)

Analisi preliminare

Da una prima semplice analisi si è riscontrato che il dataset risulta essere leggermente sbilanciato in quanto vi è una presenza maggiore di diagnosi con esito "Malign". Come è possibile vedere dal grafico vi è una prevalenza di diagnosi "Malign" (di circa il 63%) rispetto a "Benign" (circa 37%).



A primo impatto è possibile intuire che i modelli di machine learning che verranno addestrati su questo dataset potrebbero essere condizionati da questo sbilanciamento; saranno, dunque, più facilmente portati a classificare un'istanza come appartenente alla classe "Malign". Prendendo in considerazione il dominio applicativo, il comportamento appena descritto non è necessariamente negativo: un falso-positivo è un errore accettabile in confronto a un falso-negativo.

Introduzione nuove variabili

L'analisi del dominio e il confronto con altri dataset riguardanti lo stesso campo applicativo hanno portato all'introduzione di due nuove variabili. In questo ambito infatti è possibile trovare anche delle misurazioni sul volume e sulla compattezza del nodulo sospetto. A seguito di un ulteriore approfondimento si è constatato che queste misurazioni sono calcolabili attraverso valori presenti nel dataset.

Di seguito sono indicate le formule con le quali si ottengono i valori.

Variabile	Descrizione	Tipo	Formula
volume	Volume del nodulo sospetto	float	$4/3*\Pi *r^3$
compactness	Compattezza del nodulo sospetto	float	$P^2/A-1$

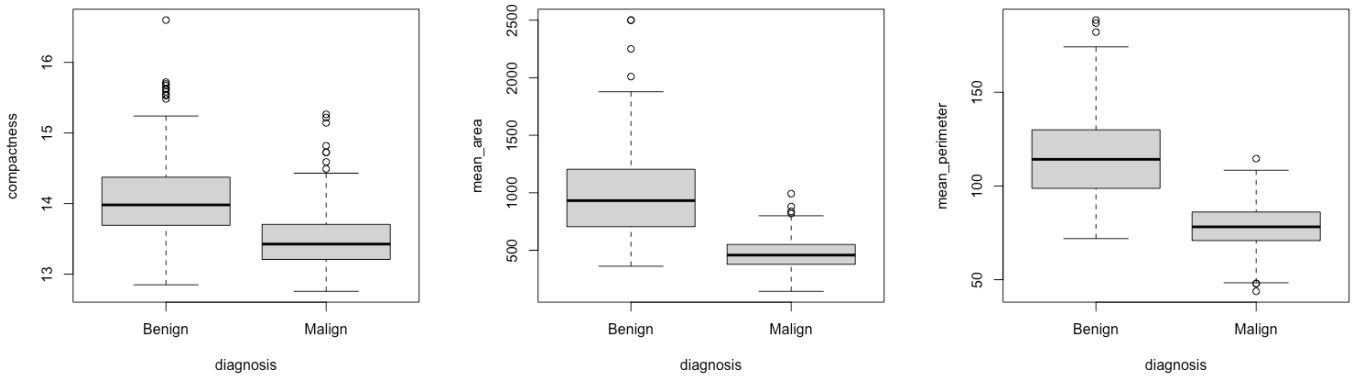
Nelle successive fasi del lavoro si è provveduto a valutare l'impatto che queste aggiunte hanno avuto nello spazio di rappresentazione.

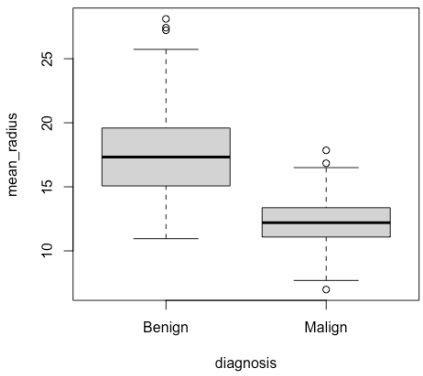
Analisi esplorativa

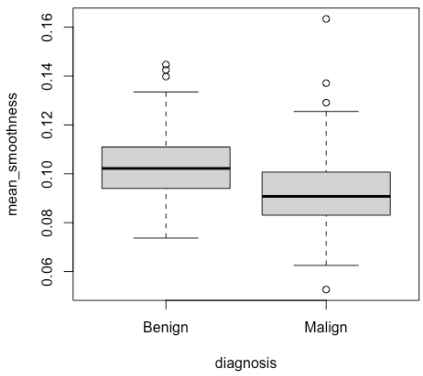
La prima fase dell'analisi esplorativa è stata svolta prendendo in considerazione e analizzando la correlazione tra i vari attributi e l'etichetta target e la correlazione tra le diverse features, rispettivamente tramite analisi univariata e multivariata.

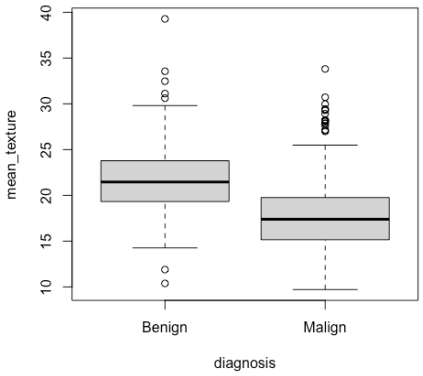
Analisi univariata

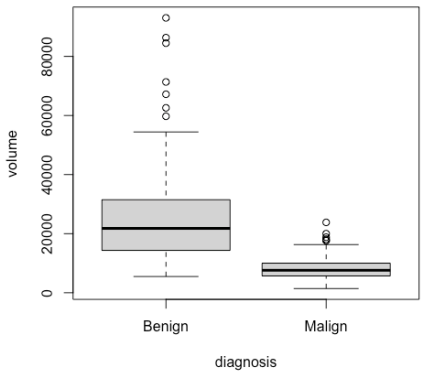
Di seguito sono stati riportati i boxplot ottenuti dall'analisi univariata. Da notare come le features "mean_radius", "mean_perimeter", "mean_area" e "volume" potrebbero essere significative nel discriminare le istanze e tra le varie covariate quella che spicca maggiormente è proprio quella introdotta precedentemente ("Volume").





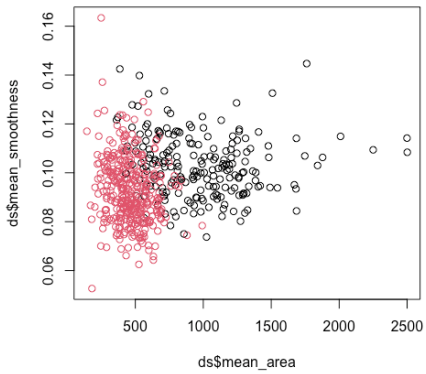


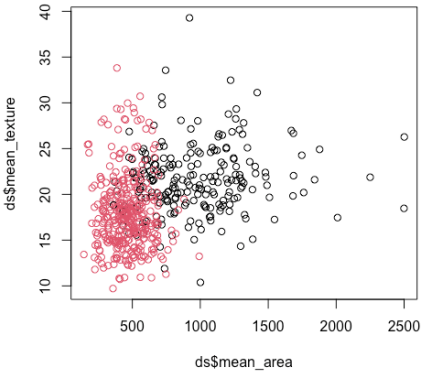


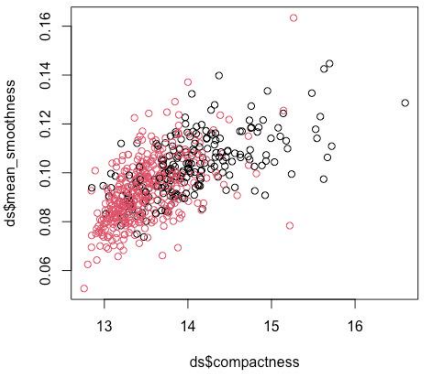


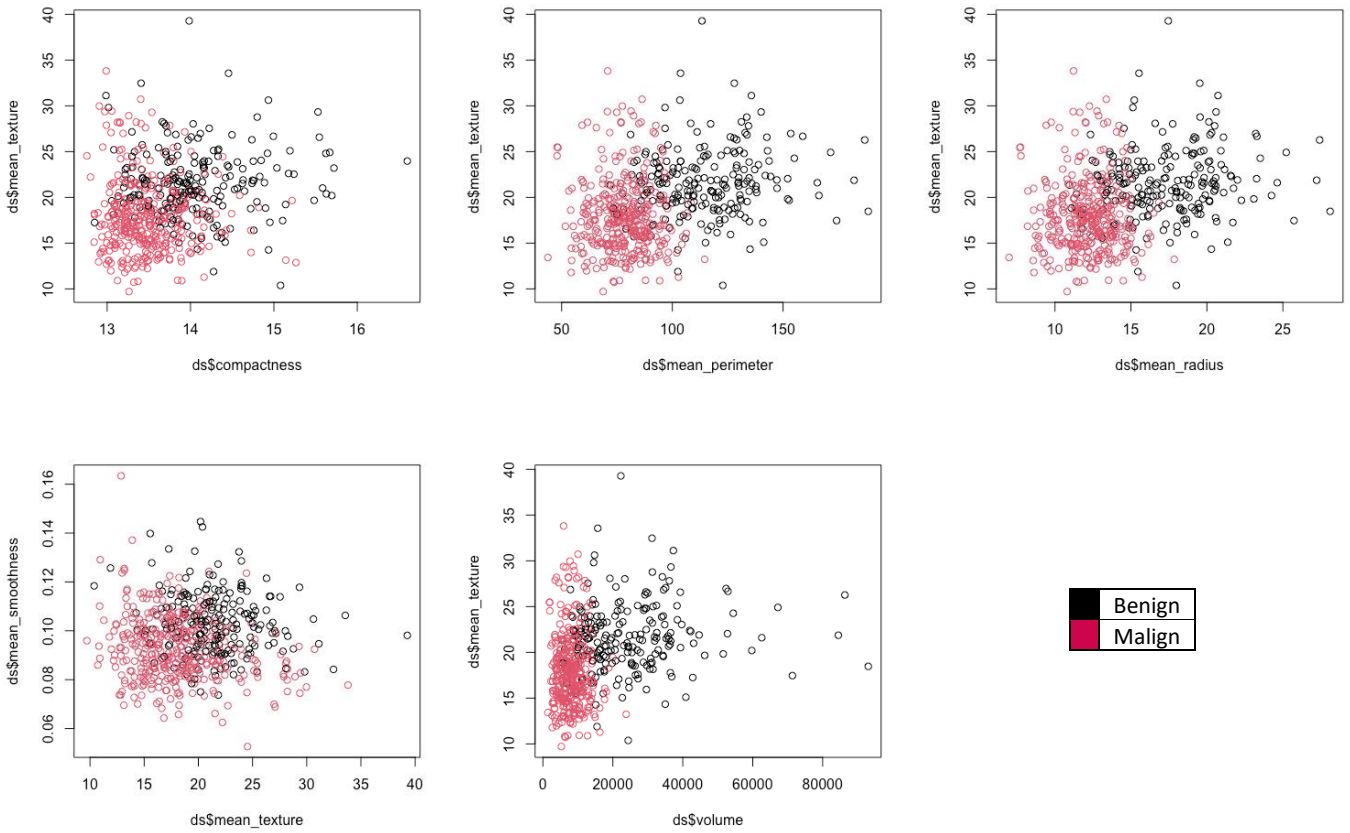
Analisi multivariata

Di seguito vengono riportati i grafici risultanti dall'analisi multivariata. Da questi si può notare che anche effettuando un confronto di correlazione tra coppie variabili, si è comunque in grado di evidenziare elementi discriminanti capaci di determinare la natura benevola o malevola di un nodulo. In particolare, questo concetto porta alla luce le quattro variabili citate durante l'analisi univariata.



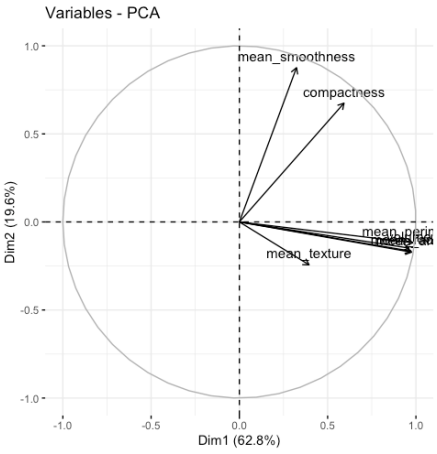


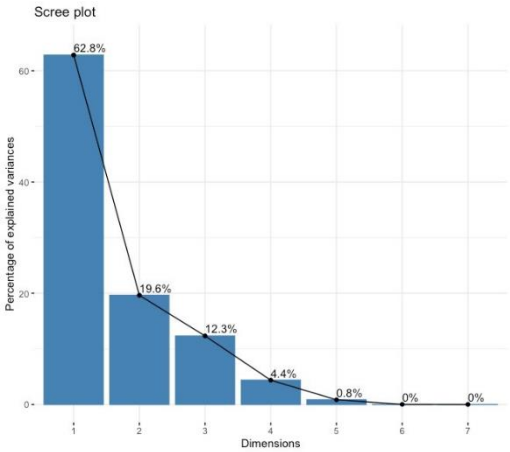




PCA

La seconda fase dell'analisi esplorativa è stata svolta sugli attributi del dataset tramite la Principal Component Analysis. Attraverso questa analisi è stato possibile osservare come le prime tre dimensioni ricavate spieghino circa il 94% della varianza. Riportando graficamente gli autovalori delle covariate è stato notato come gli attributi "mean_perimeter", "mean_area", "mean_radius" e "volume" siano positivamente correlati tra loro. La correlazione individuata e la varianza spiegata hanno portato alla rimozione delle prime tre variabili sopra citate. L'obiettivo perseguito è stato quello di lasciare un attributo ("volume") che racchiudesse gli altri e che quindi permettesse di ridurre lo spazio delle features senza perdere informazione significativa.





Da quest'analisi è emerso inoltre come l'introduzione della features "compactness" abbia influito positivamente sulla rappresentazione dell'informazione.

Descrizione e motivazione dei modelli di machine learning scelti

Rete neurale

La rete neurale è un modello predittivo di classificazione e fa parte dei modelli supervisionati. La rete è composta da un numero variabile di elementi di elaborazione altamente interconnessi, detti neuroni, che lavorano in parallelo per risolvere un problema specifico.

Si è scelto di utilizzare questo modello in quanto, essendo notoriamente uno dei modelli di machine learning più accurati e data la sensibilità dell'argomento, ci è sembrato corretto inserire questo modello tra i sistemi presi in considerazione.

Albero di decisione

L'albero decisionale è un modello predittivo di classificazione e fa parte dei modelli supervisionati. Un albero di decisione è un grafo rappresentate le decisioni e le loro possibili conseguenze, prende in ingresso un oggetto o una situazione, descritta mediante un insieme di attributi, e restituisce una decisione. Esso è composto da nodi, rami e foglie e rappresenta qualsiasi formula booleana.

Tenendo in considerazione il dominio applicativo e la natura binaria del problema, ci è sembrato consono optare per questo modello in quanto risulta essere un buon sistema di previsione per la classificazione binaria delle istanze.

Esperimenti e confronto tra modelli

Prima di eseguire la 10-fold cross validation per la rete neurale, sono stati fatti diversi test al fine di trovare i parametri ottimali con cui settare la rete. Sono state testate dieci diverse reti con un livello nascosto variando il numero di neuroni nascosti da 1 a 10 e il parametro di decadimento dei pesi da 0.1 a 0.9. Tramite questi test è emerso che i parametri ottimali risultano essere 2 neuroni nascosti e decadimento dei pesi pari a 0.1. La rete neurale è stata allenata tramite la funzione *train()* della libreria *caret* impostando come metodo *nnet*.

L'albero decisionale è stato allenato con la stessa funzione usata per la rete neurale, ma utilizzando il metodo *rpart*.

10-fold cross validation e matrice complessiva

Di seguito sono riportate le matrici di confusione finali ricavate dalla 10-fold cross validation.

Rete neurale		
pred\ref	benigno	maligno
benigno	61	29
maligno	2	74

Albero decisionale		
pred\ref	benigno	maligno
benigno	44	10
maligno	19	93

Nello specifico è possibile visionare le matrici di confusione delle singole iterazioni eseguite sulle 10 fold ottenute durante la fase di validation della rete neurale.

Rete neurale – fold 1		
pred\ref	benigno	maligno
benigno	14	3
maligno	1	23

Rete neurale – fold 2		
pred\ref	benigno	maligno
benigno	10	0
maligno	5	25

Rete neurale – fold 3		
pred\ref	benigno	maligno
benigno	13	1
maligno	2	24

Rete neurale – fold 4		
pred\ref	benigno	maligno
benigno	12	0
maligno	3	26

Rete neurale – fold 5		
pred\ref	benigno	maligno
benigno	13	0
maligno	2	26

Rete neurale – fold 6		
pred\ref	benigno	maligno
benigno	11	2
maligno	4	23

Rete neurale – fold 7		
pred\ref	benigno	maligno
benigno	15	1
maligno	0	25

Rete neurale – fold 8		
pred\ref	benigno	maligno
benigno	14	1
maligno	0	24

Rete neurale – fold 9		
pred\ref	benigno	maligno
benigno	11	1
maligno	4	24

Rete neurale – fold 10		
pred\ref	benigno	maligno
benigno	14	1
maligno	1	24

Vengono riportate, inoltre, le diverse misure di performance per ogni fold.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Accuracy	0.9024	0.8750	0.9250	0.9268	0.9512	0.8500	0.9756	0.9744	0.8750	0.9500
Precision	0.9583	0.8333	0.9231	0.8966	0.9286	0.8519	1.0000	1.0000	0.8571	0.9600
Recall	0.8846	1.0000	0.9600	1.0000	1.0000	0.9200	0.9615	0.9600	0.9600	0.9600
F-measure	0.9200	0.9091	0.9412	0.9455	0.9630	0.8846	0.9804	0.9796	0.9057	0.9600

Allo stesso modo sono riportate le matrici di confusione delle singole iterazioni eseguite sulle 10 fold ottenute durante la fase di validation dell'albero decisionale.

Albero decisionale – fold 1		
pred\ref	benigno	maligno
benigno	32	6
maligno	13	72

Albero decisionale – fold 2		
pred\ref	benigno	maligno
benigno	26	0
maligno	19	78

Albero decisionale – fold 3		
pred\ref	benigno	maligno
benigno	26	2
maligno	19	73

Albero decisionale – fold 4		
pred\ref	benigno	maligno
benigno	33	3
maligno	13	72

Albero decisionale – fold 5		
pred\ref	benigno	maligno
benigno	36	6
maligno	9	72

Albero decisionale – fold 6		
pred\ref	benigno	maligno
benigno	24	1
maligno	21	74

Albero decisionale – fold 7		
pred\ref	benigno	maligno
benigno	35	3
maligno	10	72

Albero decisionale – fold 8		
pred\ref	benigno	maligno
benigno	26	2
maligno	16	76

Albero decisionale – fold 9		
pred\ref	benigno	maligno
benigno	35	5
maligno	10	70

Albero decisionale – fold 10		
pred\ref	benigno	maligno
benigno	30	4
maligno	15	71

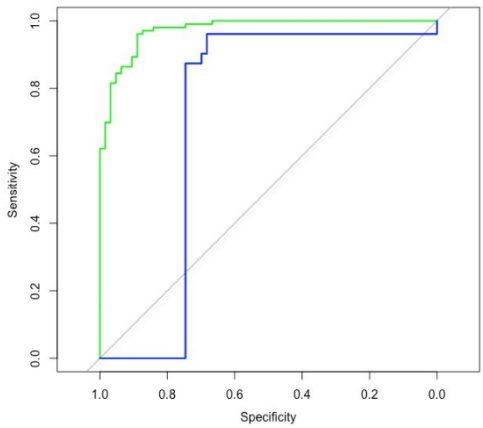
Sono stati riportati inoltre le diverse misure di performance per ogni fold.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Accuracy	0.8455	0.8455	0.8250	0.8667	0.8780	0.8167	0.8917	0.8500	0.8750	0.8417
Precision	0.8471	0.8041	0.7935	0.8471	0.8889	0.7789	0.8780	0.8610	0.8750	0.8256
Recall	0.9231	1.0000	0.9733	0.9600	0.9231	0.9867	0.9600	0.9744	0.9333	0.9467
F-measure	0.8834	0.8914	0.8743	0.9000	0.9057	0.8706	0.9172	0.8941	0.9032	0.8820

Analisi dei risultati ottenuti

Accuracy, precision, recall, f-measure, ROC e AUC

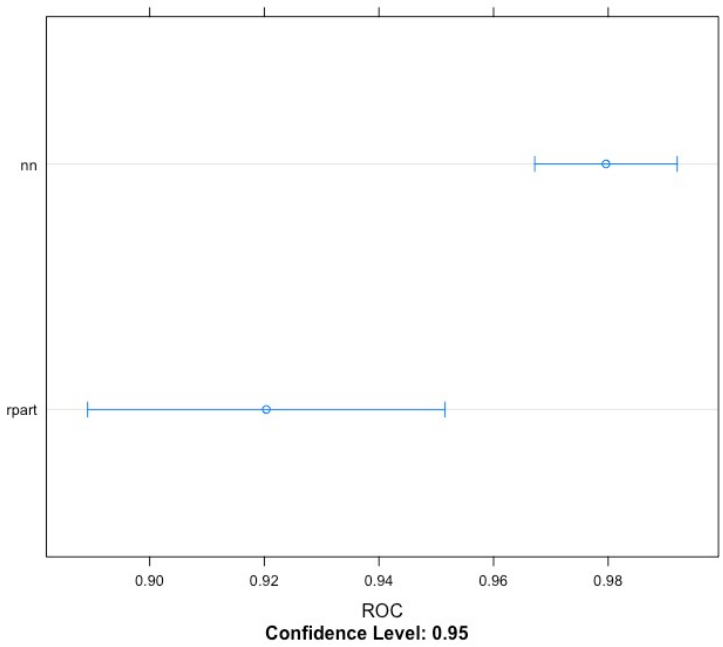
- Precision è il rapporto tra valori vero-positivi e la somma tra valori vero-positivi e falso-positivi.
- Recall è il rapporto tra valori vero-positivi e la somma tra valori vero-positivi e falso-negativi.
- F-measure misura l'accuratezza di un test. Tiene in considerazione precisone e recupero di un test.
- ROC illustra le prestazioni di un sistema di classificazione binario e traccia il tasso positivo reale contro il tasso di falsi positivi per diversi punti di taglio
- AUC viene utilizzata per misurare la performance di un modello di classificazione e ci calcola a partire dalla curva ROC in quanto equivale all'area al di sotto della curva.



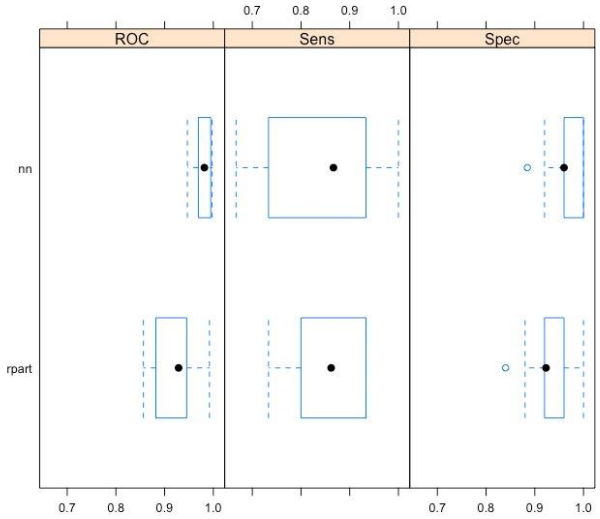
Rete Neurale	
Accuracy	0.8133
Precision	0.9737
Recall	0.7184
F-measure	0.8268
AUC	0.9736

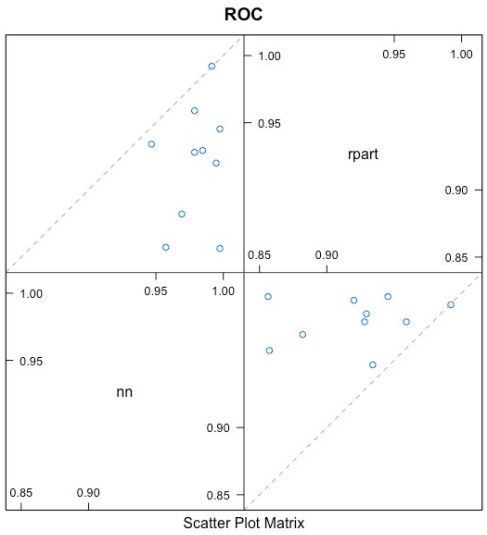
Albero Decisionale	
Accuracy	0.8253
Precision	0.8304
Recall	0.9029
F-measure	0.8651
AUC	0.8373

ROC Rete Neurale
ROC Albero Decisionale



Confrontando i margini di confidenza dei due modelli da noi utilizzati è possibile notare come la rete neurale abbia un margine molto più stretto rispetto a quello dell'albero di decisione. Inoltre, i margini dei due sistemi sono completamente distinti (non sovrapposti)





Tempi di esecuzione

	Everything	FinalModel	Prediction
NN	1.491	0.036	NA
RPART	1.541	0.006	NA

Conclusioni

Analizzando i risultati ottenuti è possibile evidenziare come la rete neurale sia in grado di ottenere dei risultati migliori e più stabili rispetto all'albero di decisione. Infatti, a fronte di un valore di accuracy simile vi è una maggior precisione da parte della rete nel predire la classe "Malign". Questo, tenendo conto del dominio applicativo è un ottimo risultato in quanto individuare con precisione un nodulo maligno ci è sembrato più significativo che predire la natura benevola dello stesso.