Relazione Progetto Statistica Numerica

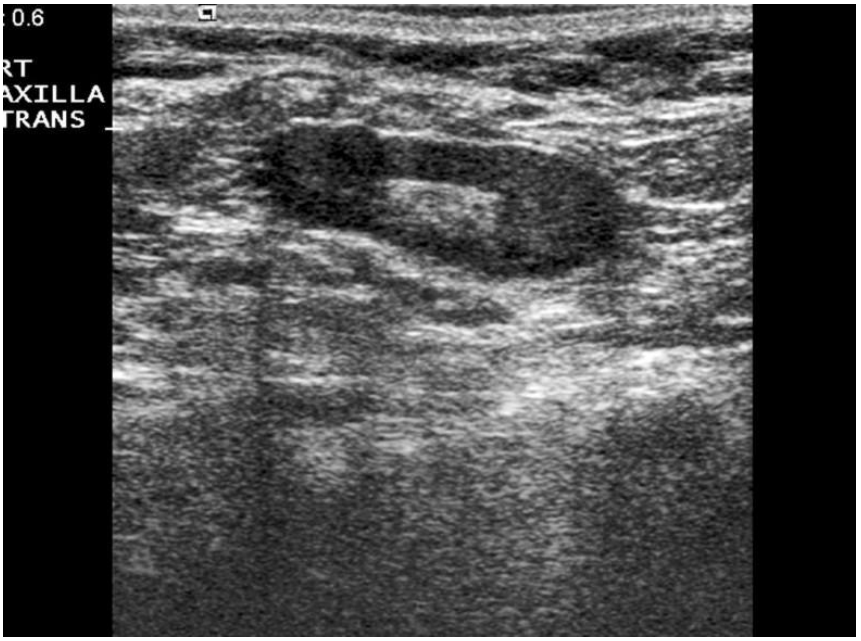
Il dataset preso in esame è il Breast Cancer Wisconsin (Diagnostic):
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
Il dataset presenta una colonna (Diagnosis) in cui viene diagnosticato come benigno/maligno (indicate con M/B) il tumore preso in esame.
Le colonne successive presentano i parametri riguardanti la forma e la dimensione della massa tumorale in esame, raggruppati rispettivamente considerando: media (mean), errore standard (se) e il più grande valore riscontrato nei campioni del valore medio (worst).

Il dataset non presenta dati mancanti (NaN).

Per visualizzare i dati, abbiamo optato per ricreare tanti grafici di dispersione confrontando ogni colonna con un'altra, mentre abbiamo utilizzato dei grafici a barre per avere un'idea della distribuzione della colonna in esame.

Descrizione del Dataset

Le Features del Dataset sono calcolate a partire da una digitalizzazione dell' immagine di una massa tumorale ottenuta con una FNA (fine needle aspiration). Le Features descrivono le caratteristiche dei nuclei delle cellule presenti nell'immagine.



sono presenti 32 attributi, ovvero:

un numero di identificazione (ID),
diagnosis (dove viene indicata la diagnosi del tumore preso in oggetto,
ovvero se benigno (B) oppure maligno (M)),
30 altri tipi di dato valutati divisi in:
radius → distanza media tra perimetro e centro
texture →(visualizzata in scala di grigi)

perimeter
area
smoothness
compactness
concavity → indica il livello di concavità delle zone concave
concave points → indica il numero dei punti di concavità rilevati
symmetry
fractal dimension

ognuno di questi dati sono presenti nel dataset come: errore standard, risultato peggiore e media. Quindi a partire da 10 caratteristiche si è arrivati ai 30 attributi del dataset.

I valori di: radius, perimeter, area sono stati considerati come dati ottenibili da esami invasivi, quindi si è stabilito di non considerarli in sede di feature selection.

Selezione, caricamento, Pre processing
Viene controllata la presenza di valori NaN e non ve ne sono presenti. Dopodichè viene creato un nuovo dataset dove per comodità' di analisi abbiamo assegnato i valori espressi in lettere della diagnosi (B, M) con valori numerici: '0' per esprimere diagnosi benigna, ed '1' per esprimere diagnosi maligna.

SPLITTING
Con lo splitting dividiamo il dataset in training set, validation set, test set. Abbiamo 409 elementi per il train set, 80 per il test set ed 80 per il validation set.
La divisione in set dei dati è stata eseguita nel seguente modo:

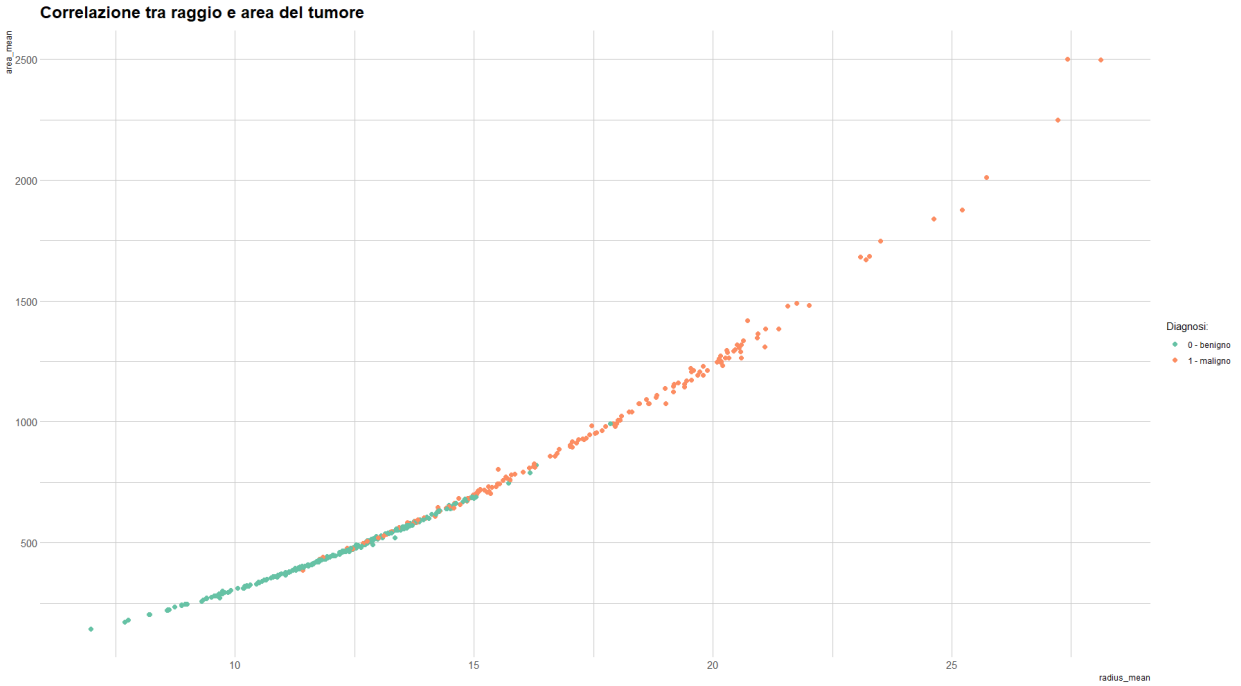
DESCRIZIONE	PERCENTUALE	NUMERO DATI
totale dati	100%	569
train set	~72%	409
test set	~14%	80
validation set	~14%	80

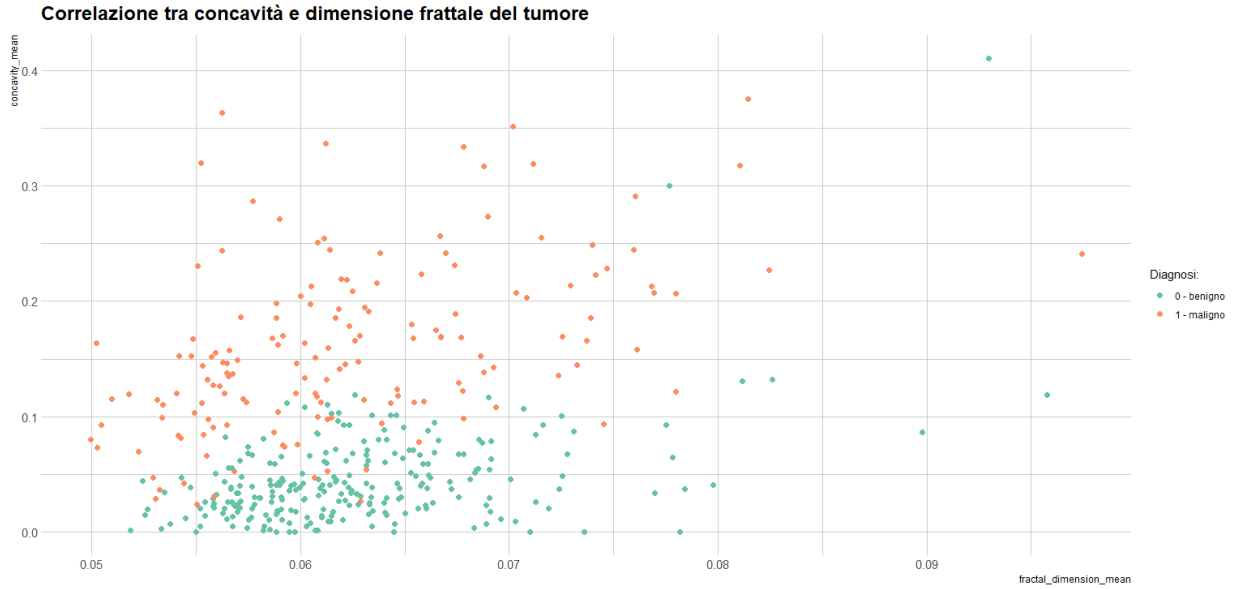
Training set: usato per addestrare il modello
Test set: utilizzato per valutare il modello sulle predizioni future (contiene una 'simulazione' di possibili dati futuri, ossia dati che si ricaveranno in seguito ma che ancora non si hanno a disposizione)
Validation set: usato per scegliere il modello ottimale poiché il test set rappresenta una collezione di dati che tecnicamente non si hanno a disposizione nel momento della model selection.

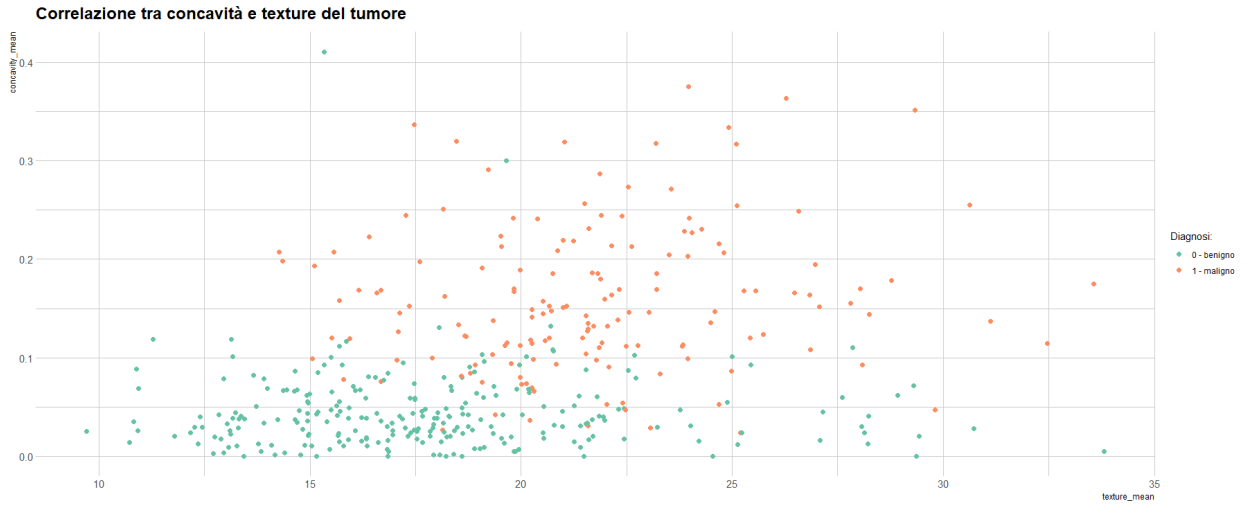
VISUALIZZAZIONE

Sono stati creati una serie di grafici di dispersione con i risultati della diagnosi per comprendere graficamente la correlazione dei dati, ed una serie di grafici a barre per osservare la distribuzione dei dati
Vengono sotto riportati alcuni esempi dei grafici ottenuti (i restanti vengono consegnati in una cartella dedicata).

Grafici di dispersione:

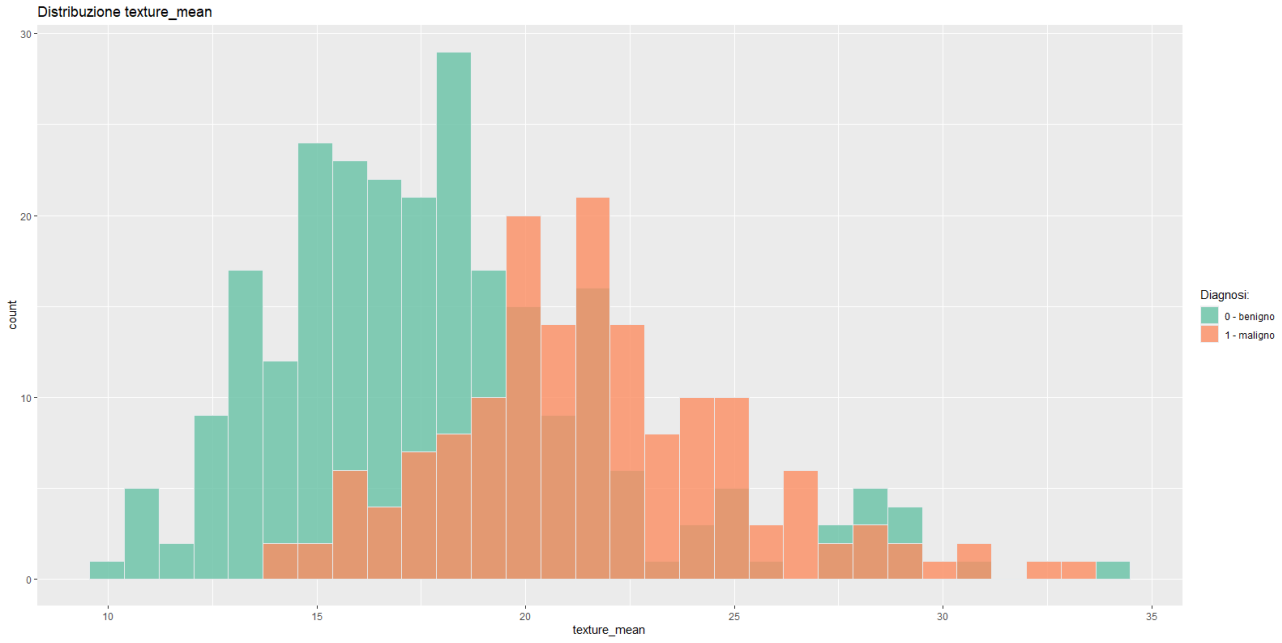


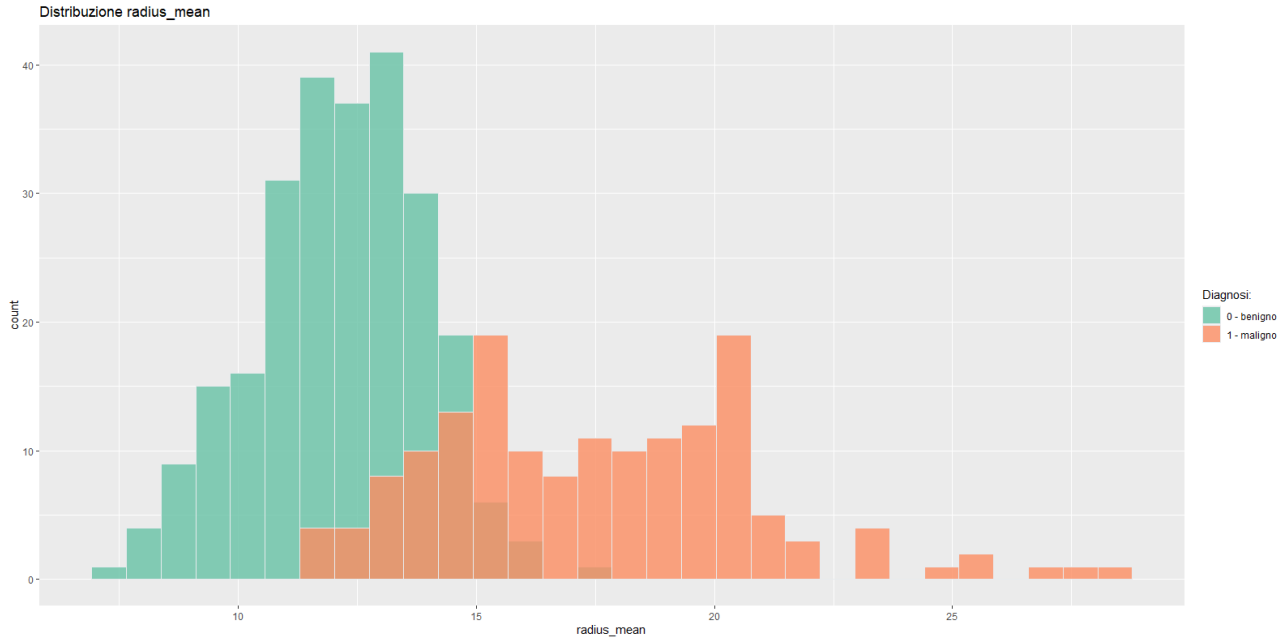


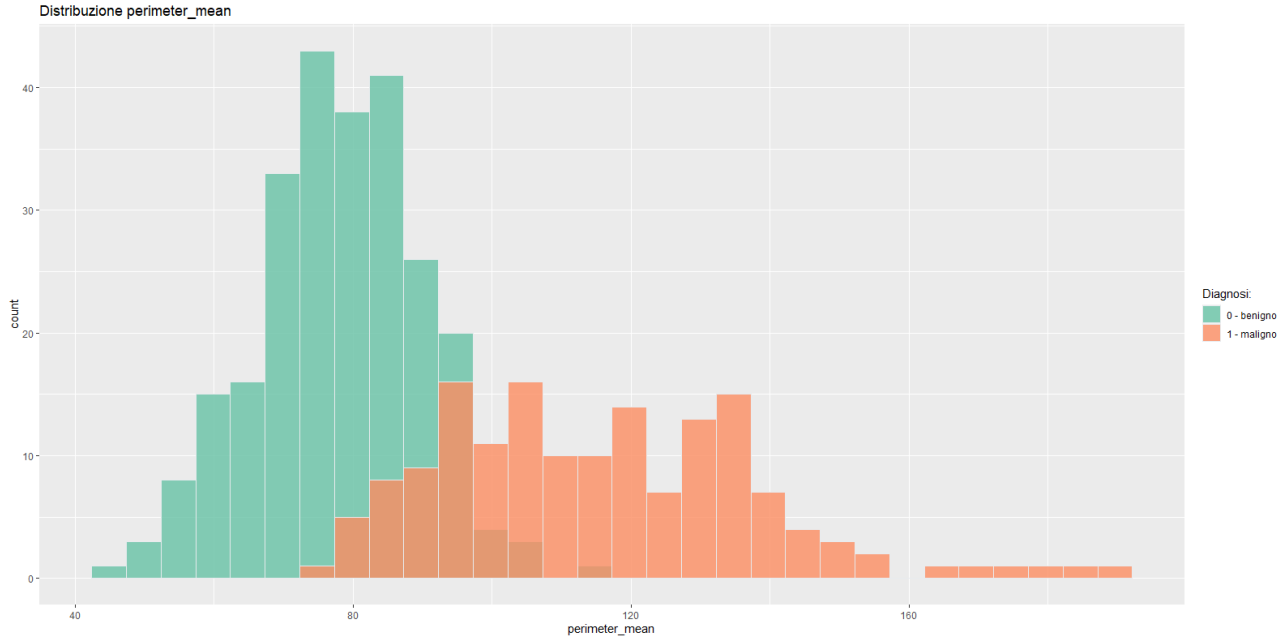


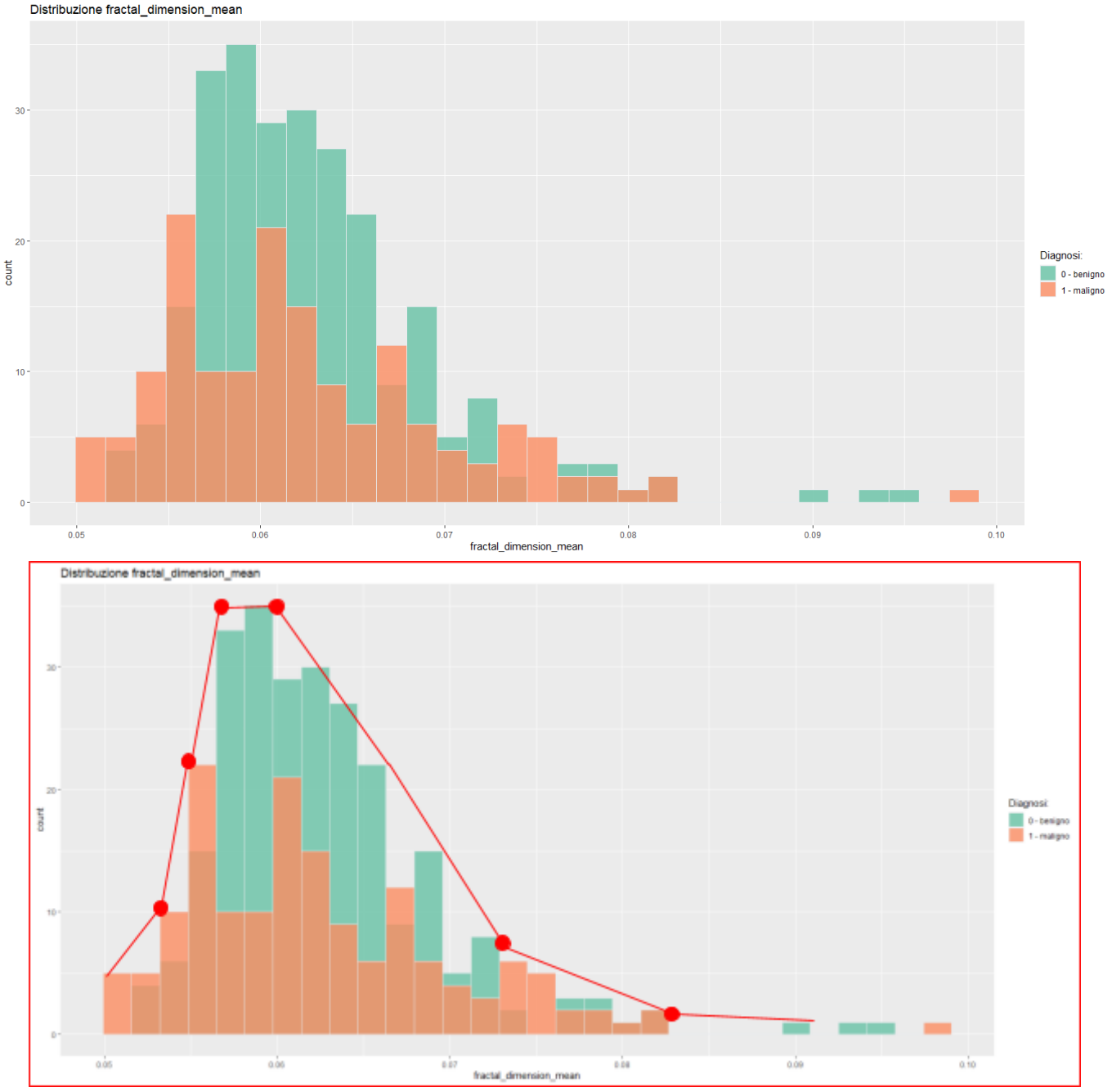
Si può osservare da questi grafici di dispersione una netta separazione delle diverse diagnosi del tumore (benigno e maligno).
Ad esempio, nella correlazione tra raggio e area (primo grafico) si può notare come al raggiungimento di una certa dimensione del tumore (raggio e area) le diagnosi risultanti sono sempre di tumore maligno.

Grafici a barre:



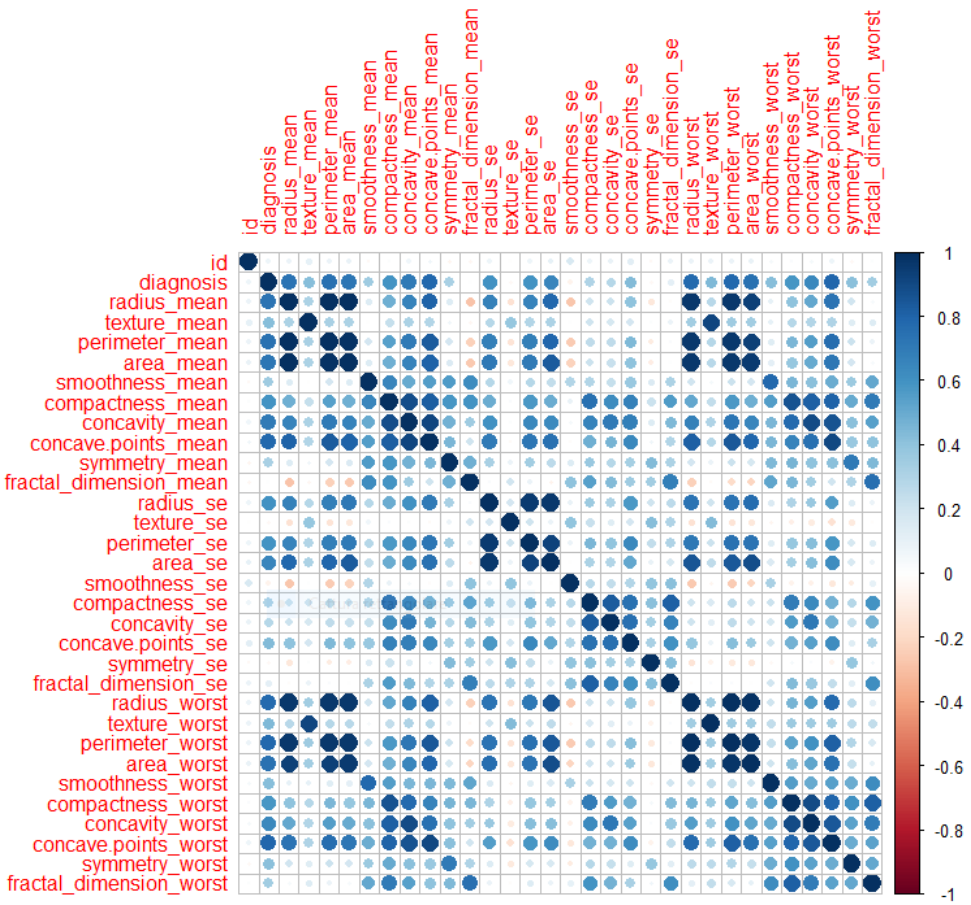

Distribuzione radius_mean
count
radius_mean
Diagnosi:
0 - benigno
1 - maligno


Distribuzione perimeter_mean
count
perimeter_mean
Diagnosi:
0 - benigno
1 - maligno



La maggior parte delle distribuzioni ottenute appaiono asimmetriche a destra (distribuzione chi-quadro) con tendenze leptocurtiche.

Matrice di correlazione:



La matrice di correlazione ci mostra che non vi sono correlazioni negative particolarmente forti.

La netta separazione delle diagnosi positive e negative, osservata precedentemente nei grafici di dispersione, ci fa pensare a quanto siano correlate alcune colonne tra di loro al fine della diagnosi; infatti si possono notare nella matrice di correlazione colonne fortemente correlate (positivamente) proprio dove i dati in esame rappresentano parametri riguardanti la dimensione, come ad esempio raggio, perimetro e area.

Una considerazione è che i valori medi ("_mean") sono correlati fortemente con i rispettivi valori "_worst" (ossia i peggiori/più alti).
Questo fa intuire che la media utilizzata per i valori medi ("_mean") sia particolarmente soggetta a variazione da parte dei valori estremi, ad esempio la media aritmetica.

Osservando la Diagnosis, si può notare che è particolarmente correlata rispetto alle dimensioni (raggio, perimetro e area). Quindi probabilmente a una massa tumorale maggiore, corrisponde una diagnosi maligna.

ADDESTRAMENTO DEL MODELLO

Il modello SVM è stato addestrato utilizzando il Training set tramite un kernel polinomiale passandogli costo '1', grado '3' considerando tutti i parametri presenti nel dataset originale.

```
model.SVM <- svm(diagnosis ~ ., num.df, kernel="polynomial", cost=1, degree=3)
```

VALUTARE LA PERFORMANCE DEL MODELLO

Con la funzione di Misclassification Rate (MR.polynomial) viene definita la percentuale di predizioni errate fatte dalla funzione predittore.

Tramite l' Accuracy (Acc.polynomial) è definita l'esattezza delle predizioni fatte dalla funzione predittore.

Con il model.SVM viene predetto il test set e, tale predizione, viene valutata tramite le funzioni descritte sopra.

valori ottenuti:

```
> MR.polynomial
[1] 0.09535452

> Acc.polynomial
[1] 0.9046455
```

È possibile provare a modificare gli iperparametri del modello, al fine di ottenere performance migliori

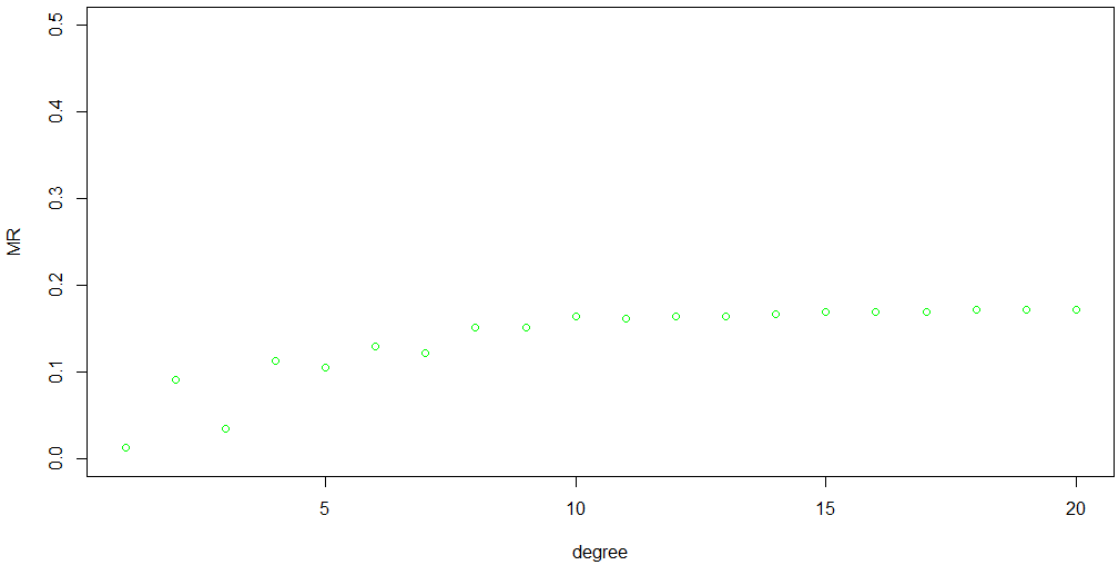
HYPERPARAMETER TUNING
provato ad addestrare con altri tipi di kernel ed altri parametri di costo, grado, e gamma

Aumentando il costo da 2 a 5 (degree 3) si osserva che la performance migliora sensibilmente:

```
> MR.polynomial
[1] 0.03422983
> Acc.polynomial
[1] 0.9657702
```

al costo 10 il miglioramento non è più netto:

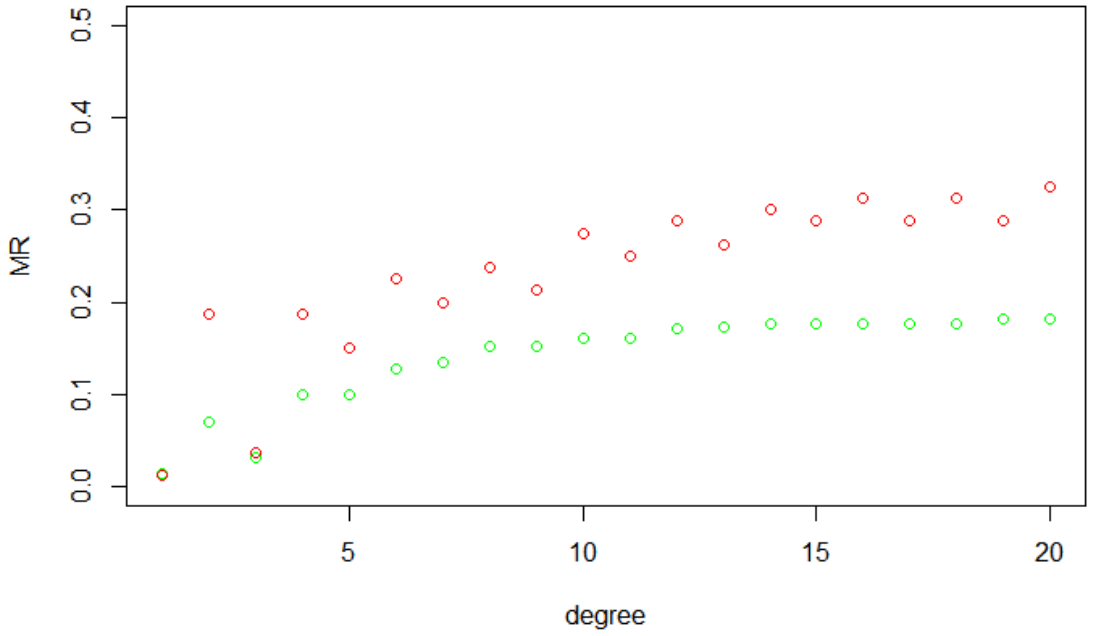
```
> MR.polynomial
[1] 0.03178484
> Acc.polynomial
[1] 0.9682152
```



Implementato un ciclo for che permette di variare il grado della funzione, mantenendo il costo costante di 10, si è riuscito a determinare l'andamento del MR al variare del grado (da 1 a 20).

Si nota che il MR si abbassa ad un grado del polinomio minore (raggiungendo quasi lo 0 con grado 1), il che fa pensare che il kernel polinomiale non sia il più adatto per ottenere le migliori performance per un dataset di tale semplicità.

Come step successivo si è verificato il comportamento del modello polinomiale al variare del grado anche con predizioni fatte sul test set (ossia il dataset che dovrebbe simulare la raccolta di dati "futuri")



Il grafico mostra il comportamento del modello sul train set (verde) e sul test set (rosso).

Cambiando kernel, passando da polinomiale a radiale, otteniamo performance perfette:

```
#---MODELLO CON KERNEL RADIALE---
model.SVM <- svm(diagnosis ~., num.df, kernel = "radial", cost = 10, gamma = 0.3)
#summary(model.SVM)

y.pred <- predict(model.SVM, num.df)
#y.pred

MR.radial <- MR(y.pred, num.df$diagnosis)
Acc.radial <- Acc(y.pred, num.df$diagnosis)
MR.radial
Acc.radial
```

```
> MR.radial
[1] 0
> Acc.radial
[1] 1
```

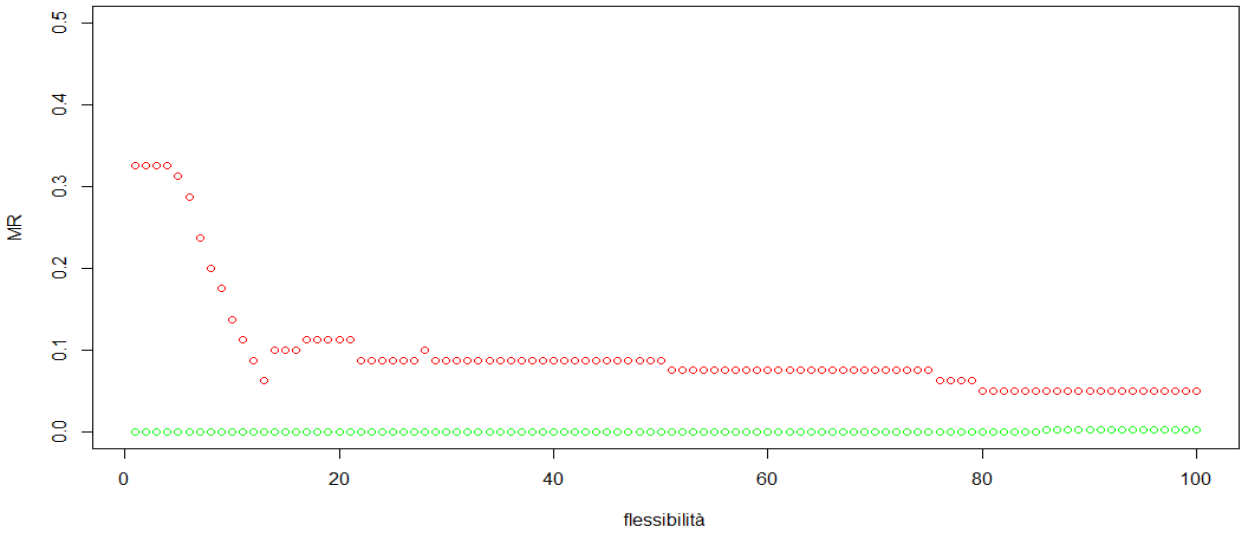
Con un modello più flessibile si ottengono delle performance perfette (100% accuratezza)

VALUTAZIONE DELLA PERFORMANCE

Si valuta il modello al variare del gamma in modo tale da capire quale sia il parametro gamma che permette di ottenere il modello ottimale.

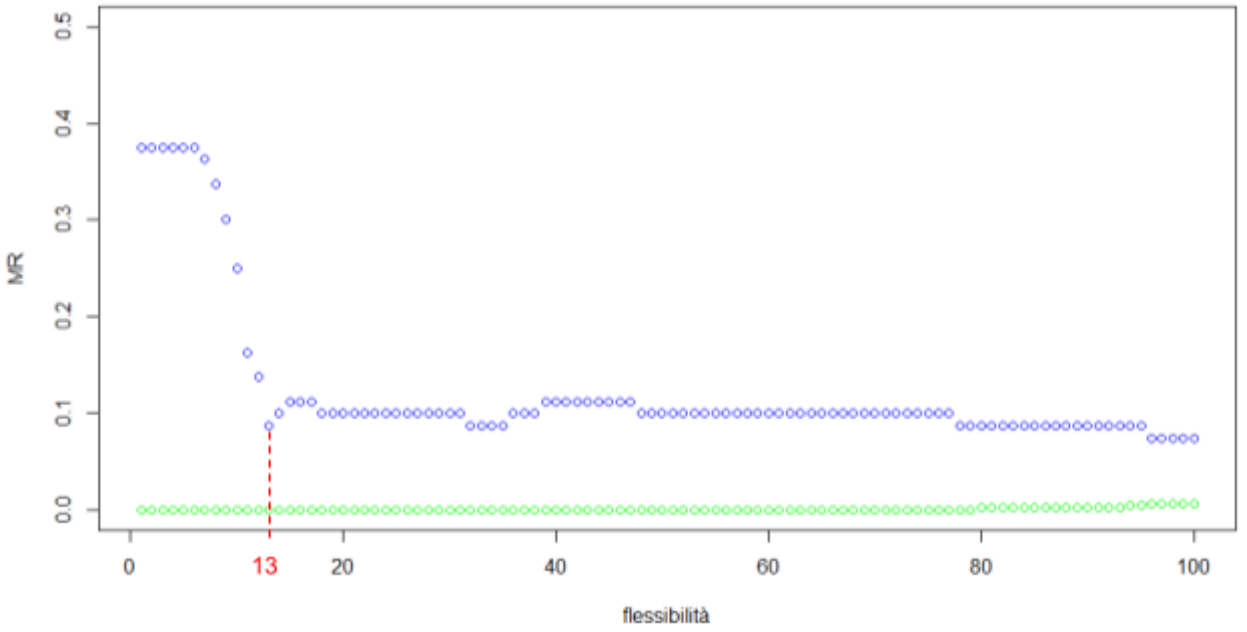
Si esegue dunque un ciclo for come segue:

```
# Come cambia la performance del modello con kernel radiale al variare di gamma
MR.radial.total <- 1:100
MR.radial.test <- 1:100

for (d in 1:100) {

  model.SVM <- svm(diagnosis ~., num.df, kernel = "radial", cost = 10, gamma = 5/d)
  y.pred <- predict(model.SVM, num.df)
  MR.radial <- MR(y.pred, num.df$diagnosis)
  MR.radial.total[d] <- MR.radial

  model.SVM <- svm(diagnosis ~., num.df, kernel = "radial", cost = 10, gamma = 5/d)
  y.pred <- predict(model.SVM, df.test)
  MR.radial <- MR(y.pred, df.test$diagnosis)
  MR.radial.test[d] <- MR.radial
}

# Visualizzazione grafica della variazione della performance
plot(MR.radial.total, type='p', xlab="flessibilità", ylab="MR", ylim=c(0, 0.5), col='green')
points(MR.radial.test, type='p', col='red')
```



il grafico mostra il comportamento del modello sul test set (in rosso)

L'orientamento della flessibilità mostrata nel grafico indica che al crescere della x la flessibilità diminuisce e viceversa aumenta.
Questo grafico denota un comportamento di overfit (ossia quando il modello è troppo flessibile rispetto al dataset), mentre non sembra verificarsi un comportamento di underfit (ossia quando il modello non è abbastanza flessibile rispetto alla complessità del dataset).



il grafico mostra il comportamento del modello sul validation set (in blu)

In questo grafico si osserva un comportamento di overfit con gamma 5/13 o maggiore (verso sinistra) e si osserva anche che con gamma uguale a 5/13 si ottiene il MR (Misclassification Rate) più basso e quindi il modello "migliore".

Abbiamo scelto quindi come modello ottimale:

```
#---MODEL SELECTION---
model.SVM <- svm(diagnosis ~., num.df, kernel = "radial", cost = 10, gamma = 5/13)
y.pred <- predict(model.SVM, df.test)

MR.radial <- MR(y.pred, df.test$diagnosis)
Acc.radial <- Acc(y.pred, df.test$diagnosis)
MR.radial
Acc.radial
```

ottenendo il seguente MR e la seguente Accuracy:
> MR.radial
[1] 0.0625
> Acc.radial
[1] 0.9375

INTERPRETAZIONE PROBABILISTICA

> MR(y.total, df.test$diagnosis)
[1] 0.0875
> Acc(y.total, df.test$diagnosis)
[1] 0.9125

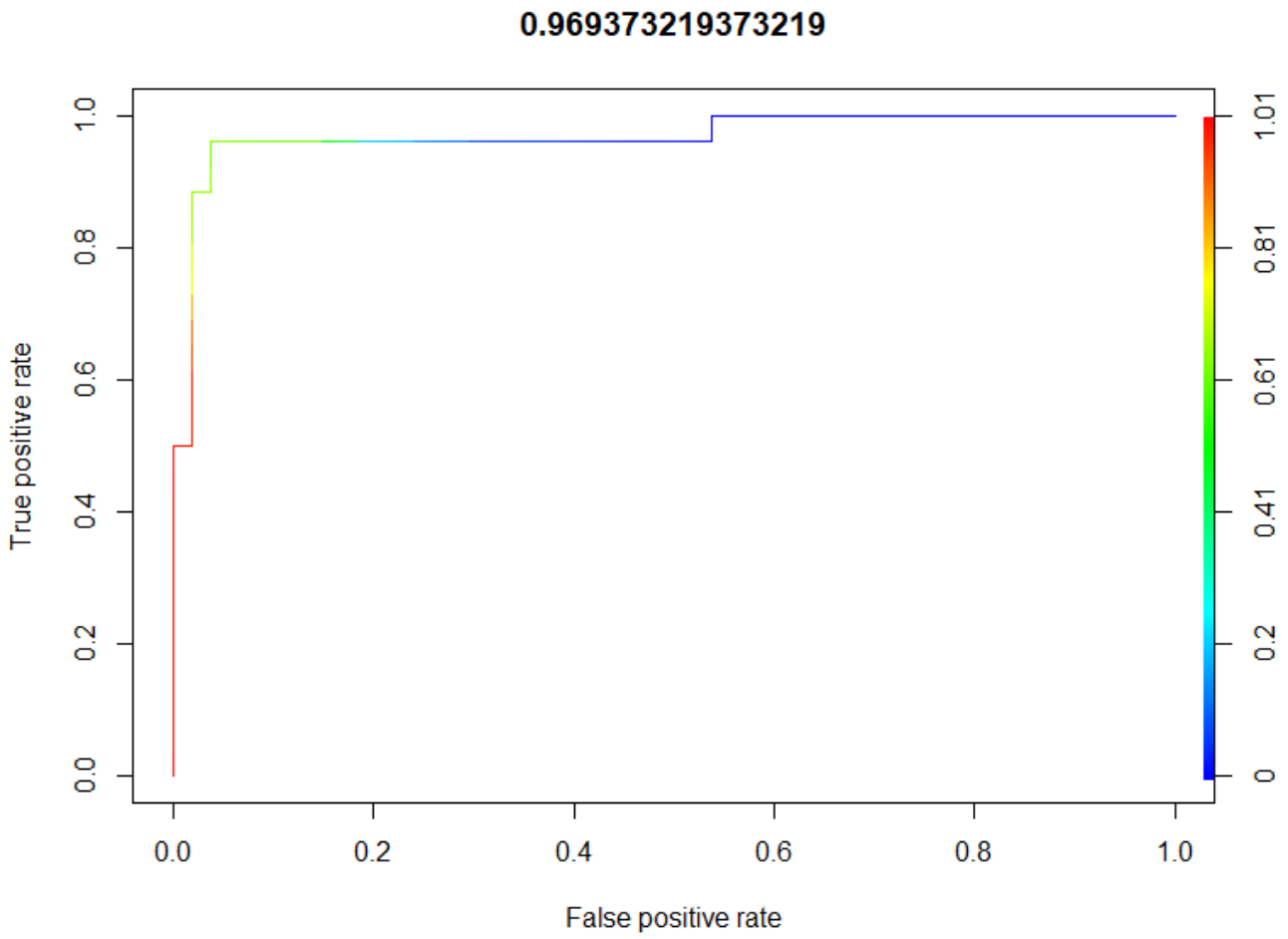
y.pred	**0**	**1**
0	**53**	**6**
1	**1**	**20**

00 : true positive (53)
01: false positive (6)
10: false negative (1)
11 : true negative (20)

La somma dei valori sulla diagonale principale (quelli predetti in maniera corretta) è di gran lunga maggiore rispetto alla somma di quelli della diagonale opposta, confermando l'ottima qualità del modello.

True Positive Rate (Sensitività): **~ 0,9814**
False Negative Rate: **~ 0,0186**
True Negative Rate (Specificità): **~ 0,7692**
False Positive Rate: **~ 0,2308**
Precisione (PPV): **~0,8983**

Curva di ROC



Dal grafico della curva di Roc si nota che l'area sotto la curva approssima al ~97% il rettangolo (più l'area sotto la curva approssima l'intero rettangolo, più il modello scelto sarà ottimale).

AUC (Area Under The ROC Curve): **0,969373219373219**

STUDIO STATISTICO SUI RISULTATI

L'esecuzione del modello una sola volta non è sufficiente per dare una valutazione corretta del modello, essendo aleatori i dati utilizzati.
Quindi si implementa un ciclo for in cui vengono ripetute per 100 volte le fasi precedenti addestramento e testing salvando i valori del Misclassification Rate all'interno di un array costituente l'SRS(100).
Si utilizzerà il SRS ottenuto per ricavare i dati statistici caratterizzanti.

STATISTICA DESCRITTIVA ED INFERENZIALE

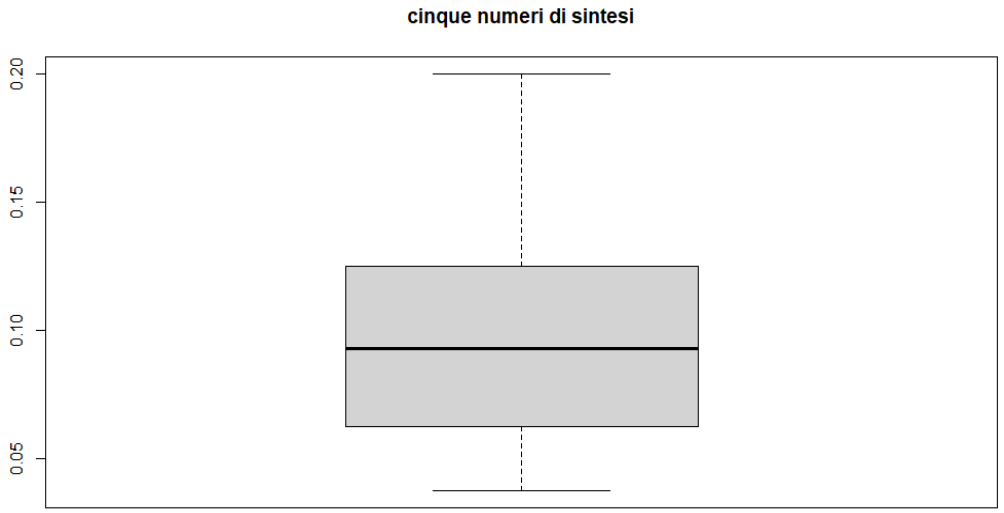
Tabella dei calcoli statistici eseguiti:

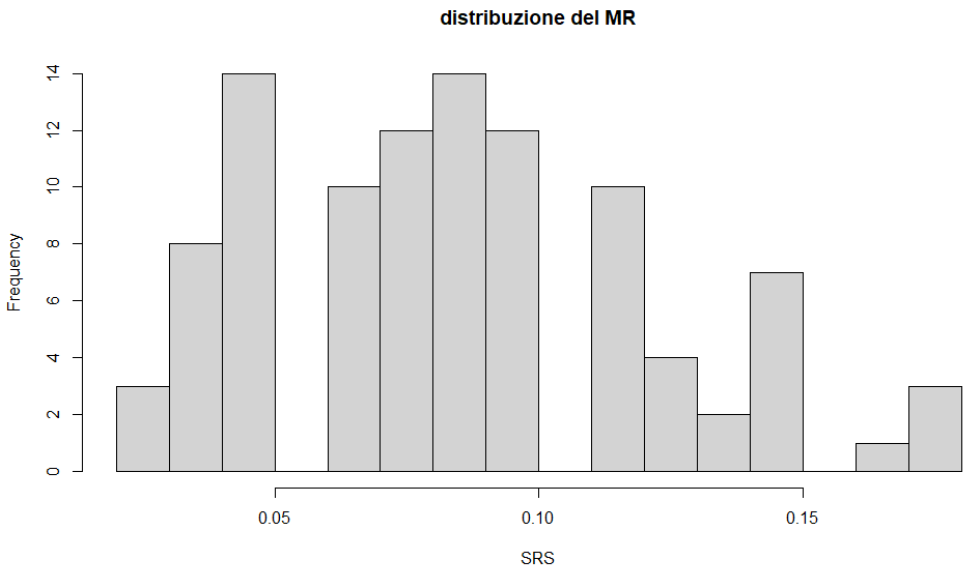
	risultato
calcolo della media semplice	0.086625
mediana	0.0875
media troncata	0.08421875
varianza del SRS	0.001354972
deviazione standard del SRS	0.03680994
range interquartile (IQR)	0.053125

Statistica ordinata: ordinamento dei dati in modo crescente, per ottenere informazioni sulla distribuzione dei dati e capire dove sono concentrati. Calcoliamo quindi i Quantili

0%	25%	50%	75%	100%
0.0375	0.0625	0.0875	0.1250	0.2000

Min.	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max.**
0.02500	0.05937	0.08750	0.08662	0.11250	0.17500





la distribuzione ottenuta presenta asimmetria anche se la mediana ha un valore molto simile alla media

Utilizzando la statistica inferenziale selezioniamo 100 campioni di dati. La media e la deviazione standard sono i valori ricavati nella precedente parte di statistica descrittiva. Calcolando l' intervallo di confidenza con il valore del livello di confidenza a = 0.05% otteniamo:

Intervallo di confidenza:

estremo superiore	0.08872735
estremo inferiore	0.08403104

Creiamo un simple random sample di 100 valori, calcoliamo la media 1000 volte si ottiene che il 95.2% delle volte il valore rientra nell'intervallo.

FEATURE SELECTION

Si esegue la feature selection, per osservare l'accuratezza del modello considerando solamente alcuni tipi di dato. Questo procedimento è utile ad esempio per ponderare la rilevanza di dati ottenibili tramite procedure invasive rispetto a quelli ottenibili da procedure non invasive. Nella tabella sottostante sono descritti i tipi di dati utilizzati e i relativi risultati ottenuti in termini di Accuracy e Misclassification Rate del modello.

Dati utilizzati	Misclassification rate	Accuracy
con tutte le colonne in input	0.0625	0.9375
prendendo solo le colonne "_mean" (media)	0.125	0.875
prendendo solo le colonne "_se" (standard error)	0.15	0.85

prendendo solo le colonne “_worst” (dato più alto/peggiore)	0.025	0.975
prendendo solo colonne di dati ottenibili da procedure meno invasive	0.075	0.925
prendendo solo colonne di dati ottenibili da procedure più invasive (di quantità)	0.0125	0.9875
prendendo un solo dato: area (sia “_mean”, sia “_se” che “_worst”)	0.025	0.975

Il miglior risultato ottenuto è mediante i dati ottenibili tramite procedure invasive (es. biopsie) riguardanti dunque le dimensioni accurate della massa tumorale.
Tali valori tendono a determinare correttamente la diagnosi, come presunto all’inizio tramite la correlation matrix.
I valori dello standard error, di contro, portano a performance peggiori essendo esso stesso una stima dello scarto quadratico medio.
Nonostante tutto si sono ottenuti ottimi risultati in ogni caso analizzato.