Diagnosi di malattie cardiache tramite tecniche di Machine Learning

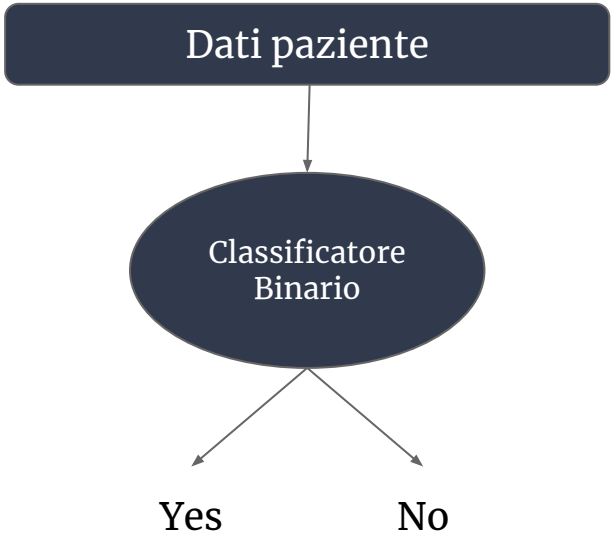
A cura di:

Luca Cogo

Comprendere il problema

Le malattie cardiache sono la prima causa di morte in tutto il mondo: secondo le stime ogni anno sono più di 17 milioni le morti legate a cardiopatie, circa il 31% di tutti i decessi.

Lo scopo di questo progetto è quello di utilizzare dei modelli di machine learning per realizzare un classificatore binario che sia in grado, sulla base dei dati clinici dei pazienti, di diagnosticare malattie cardiache potenzialmente fatali.



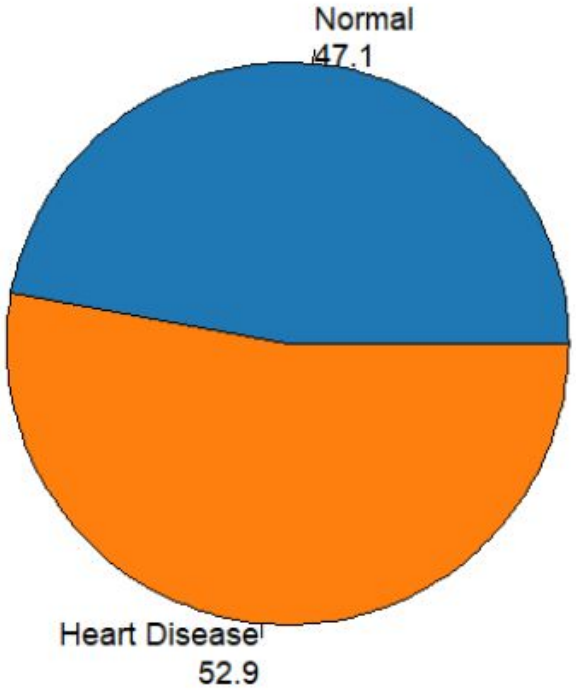
Il dataset

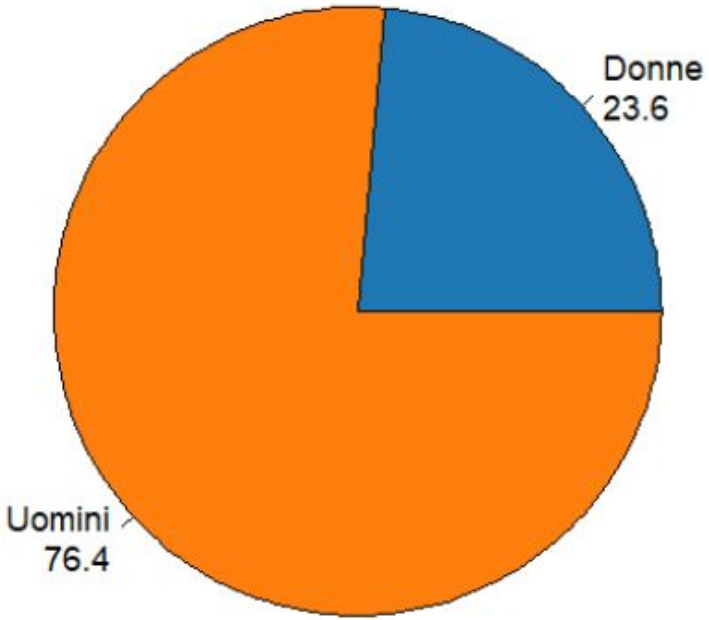
Il dataset utilizzato è stato reperito dal sito Kaggle ed è stato realizzato unendo i dati di diversi dataset indipendenti tra loro. Si è ottenuto quindi un dataset da 1190 istanze e 11 features:

- Età
- Sesso (0 = Femmina, 1 = Maschio)
- Chest Pain Type (1 = dolore anginoso tipico, 2 = dolore anginoso atipico, 3 = dolore non anginoso, 4 = nessun dolore)
- Pressione sanguigna a riposo (mm/Hg)
- Colesterolo (mg/dL)
- Glicemia (1 se > 120 mg/dL, 0 altrimenti)
- Elettrocardiogramma a riposo(0 = normale, 1 = anomalia ST-T, 2 = ipertrofia ventricolare)
- Battiti cardiaci massimi
- Angina pectoris dopo esercizi (0 = assente, 1 = presente)
- Sottoslivellamento ST (mm)
- Inclinazione tangente ST durante esercizi (0 = normale, 1 = ascendente, 2 = orizzontale, 3 = discendente)

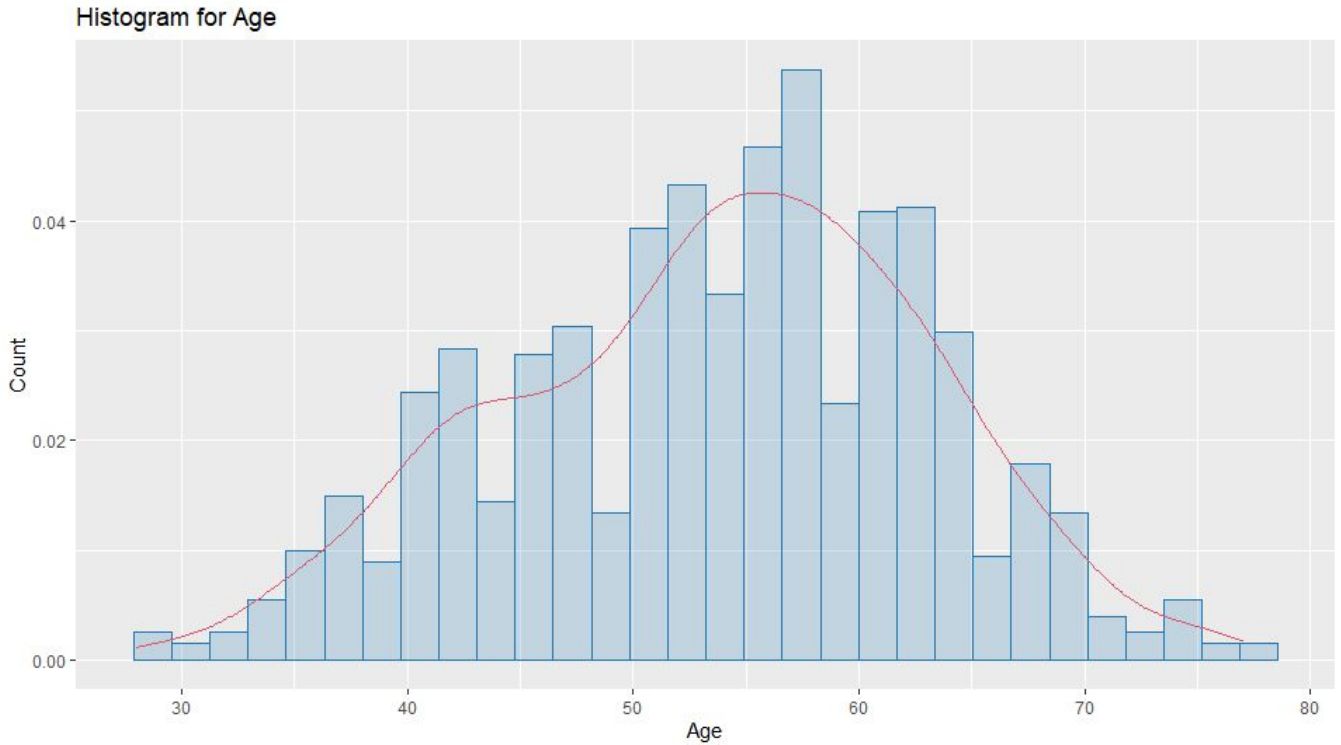
In aggiunta, è presente anche l'etichetta target (1= individuo a rischio, 0 altrimenti)

Analisi dei dati: Target e Sex

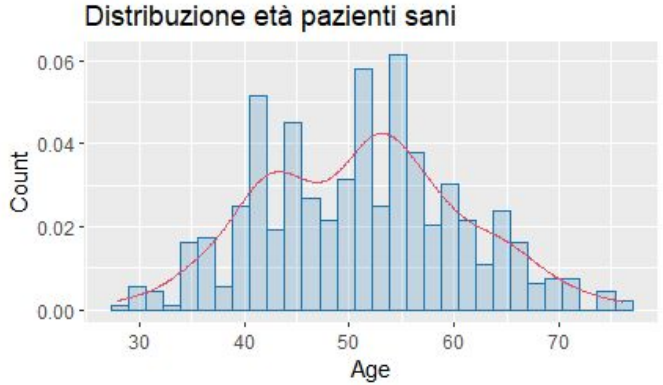


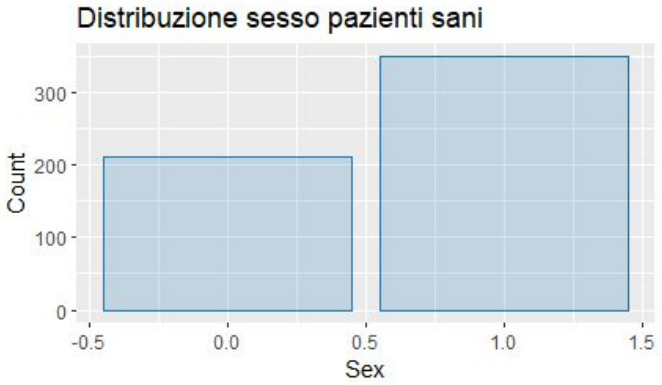


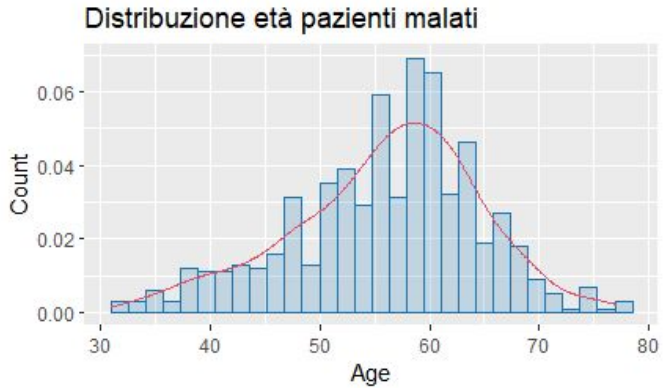
Analisi dei dati: Età



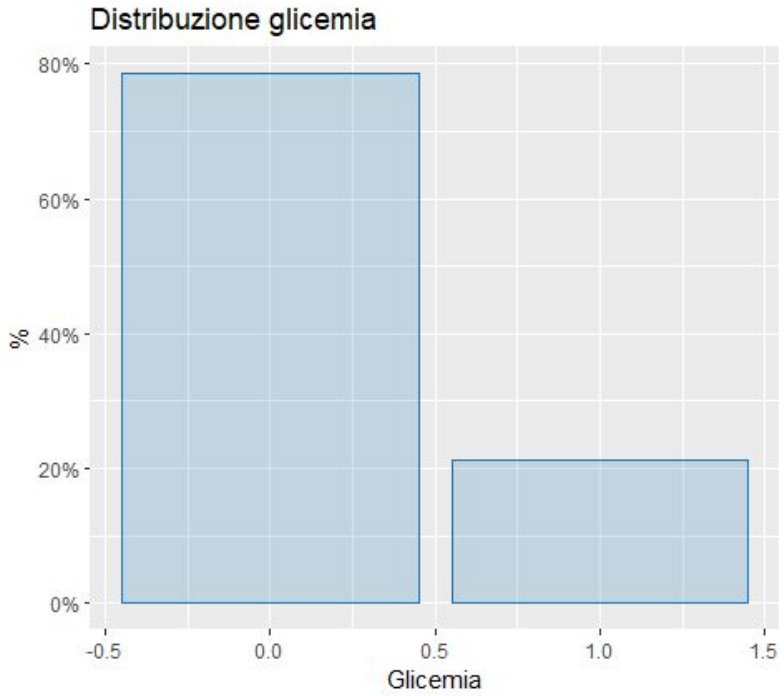
Analisi dei dati: Età e sesso

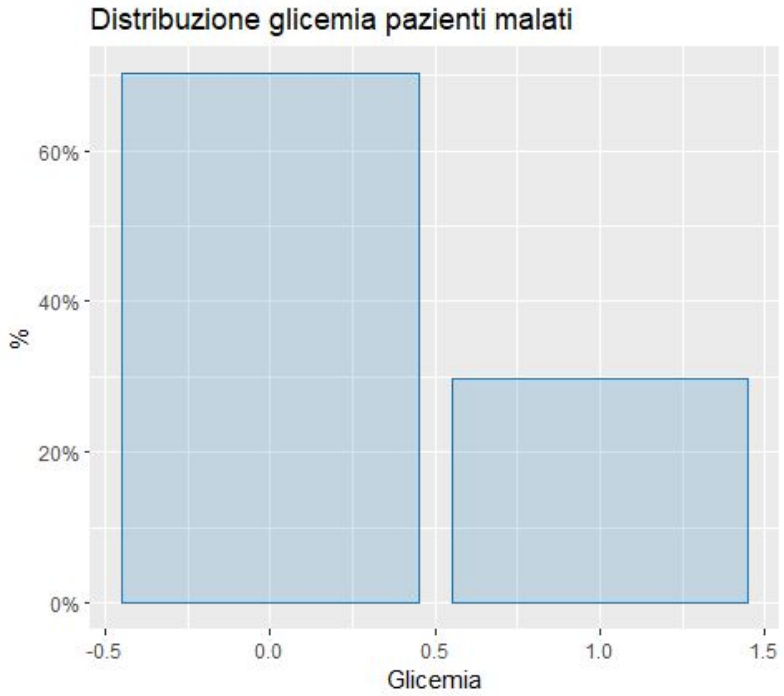




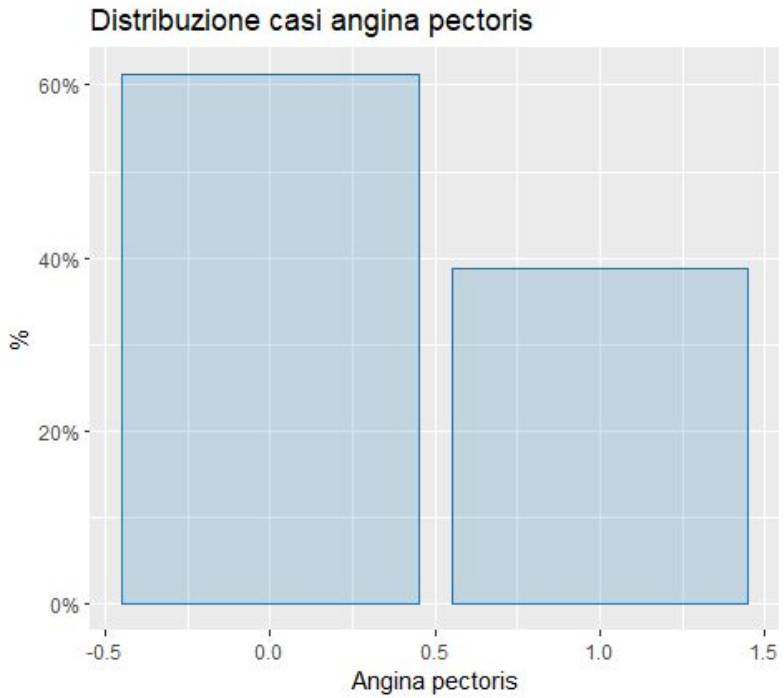


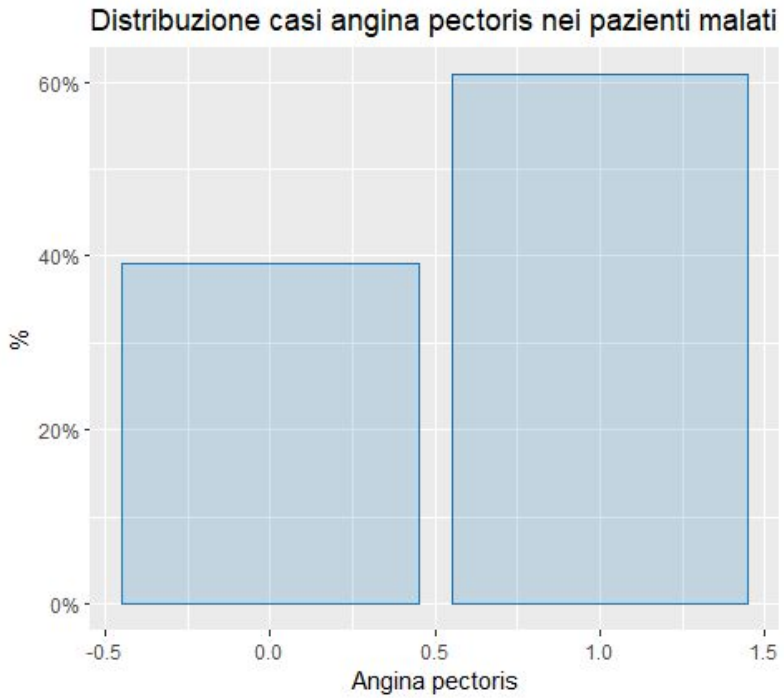
Analisi dei dati: Variabili categoriche



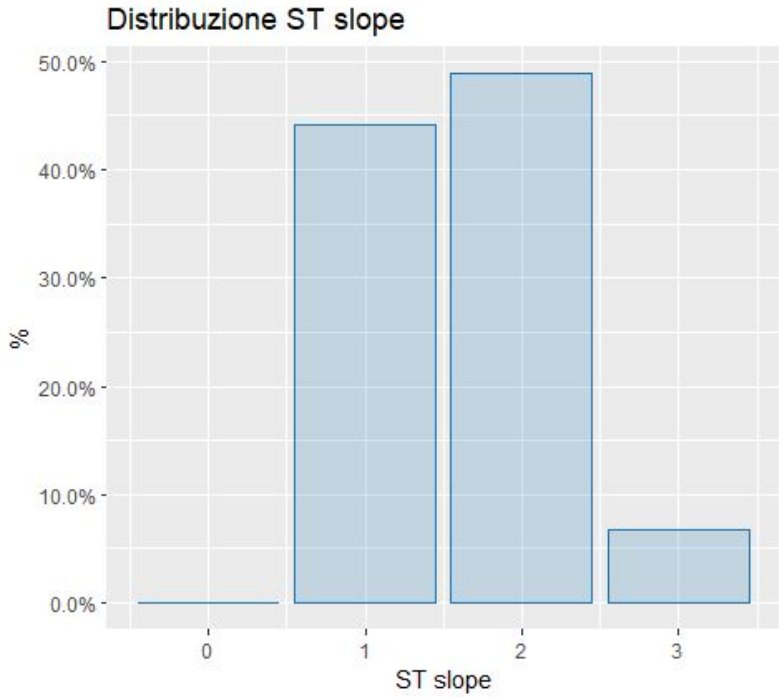


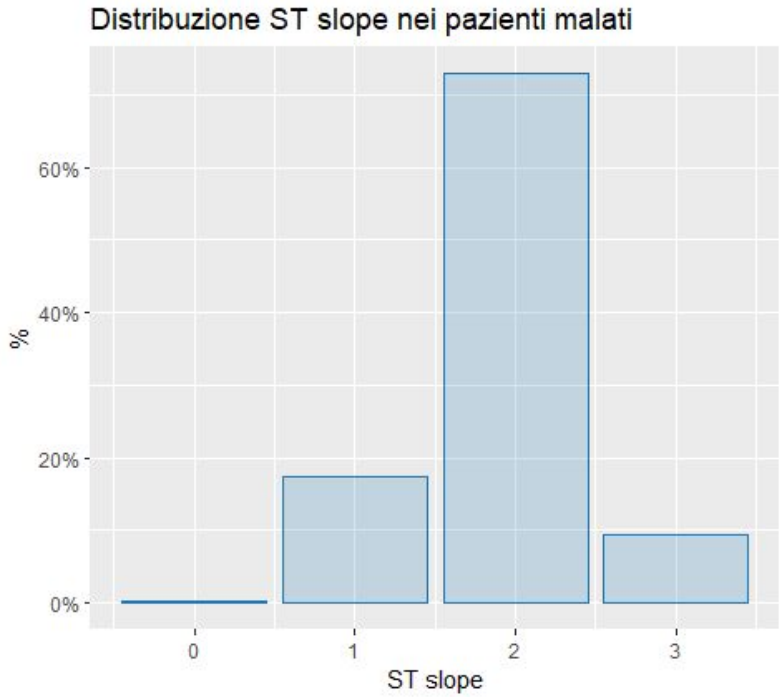
Analisi dei dati: Variabili categoriche



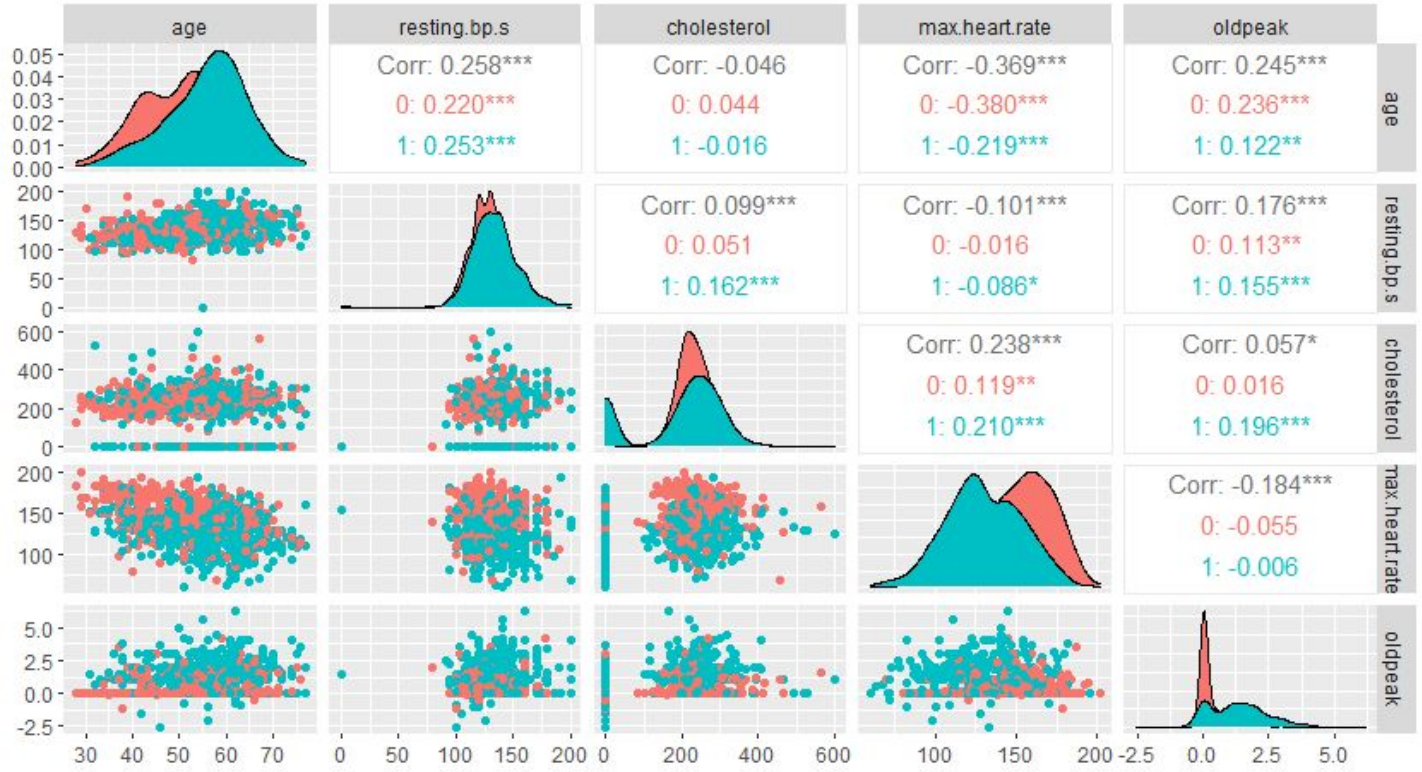


Analisi dei dati: Variabili categoriche





Analisi dei dati: Variabili numeriche



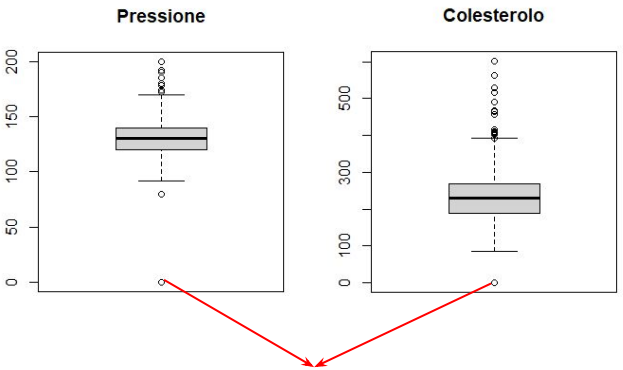
Analisi dei dati

- Individuati dei valori anomali nel dataset → necessaria gestione degli outliers
- Non si è ritenuto necessario applicare tecniche di feature reduction:
    - Lo spazio delle features è già abbastanza ridotto
    - In ogni caso applicare la PCA al dataset non funzionerebbe poichè contiene anche variabili categoriche (andrebbero codificate, oppure andrebbe trovata una tecnica alternativa)

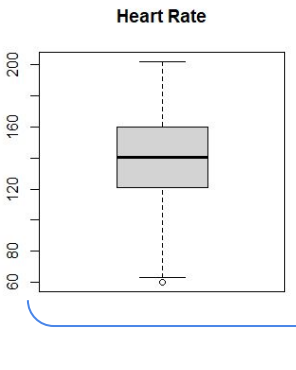
Gestione degli outliers

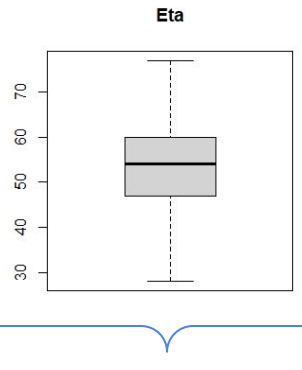
Dato Q1 = 1° quartile, Q3 = 3° quartile e IQR = Q3-Q1, allora un outlier è un punto che si trova sotto [Q1 - k*IQR] oppure sopra [Q3+k*IQR] (dove k regola l'ampiezza dell'intervallo, nel nostro caso k=1.5).

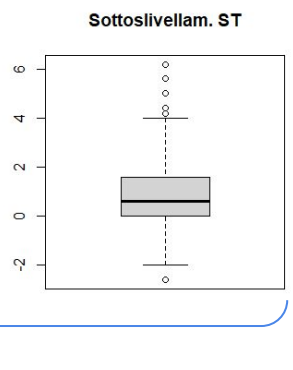
Non sempre è giusto eliminare gli outliers dal dataset. Quando però è evidente che questi sono errori, allora è giusto rimuoverli.



Outliers evidentemente errati (172 osservazioni)







Outliers assenti o plausibili

Naive Bayes

Naive Bayes è un modello di apprendimento supervisionato in grado di effettuare classificazioni basate sul Teorema di Bayes.
Il Teorema di Bayes permette di calcolare per ogni istanza la probabilità di appartenenza ad una classe. Nello specifico, calcola la probabilità di un evento A condizionata a un altro evento B.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Per ogni caratteristica vengono dunque calcolate le probabilità semplici e le probabilità condizionate (in questo caso B assume i possibili valori di target).
La predizione su una nuova istanza verrà effettuata semplicemente effettuando il prodotto tra le probabilità precedentemente calcolate.

Naive Bayes: perchè sceglierlo

PRO

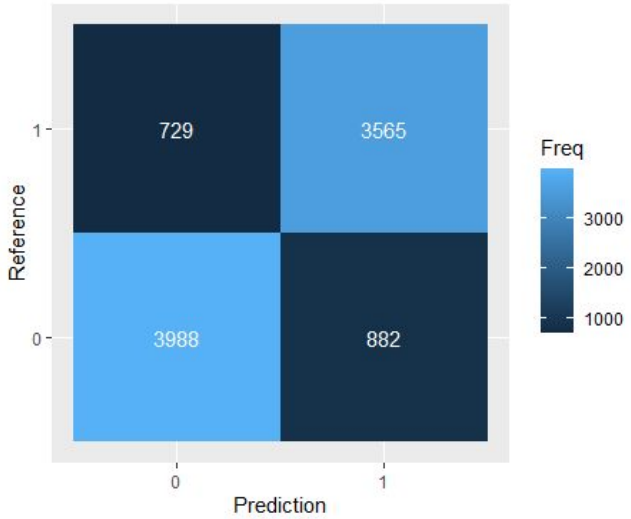
- Semplicità di utilizzo: si adatta bene sia a dati numerici che categorici e non ha iperparametri da selezionare
- Semplicità computazionale: garantisce alta velocità sia in fase di training che di predizione
- Robustezza rispetto al rumore
- Scalabilità

CONTRO

- Si basa sull'assunzione che le variabili siano tra loro indipendenti
- No regressione

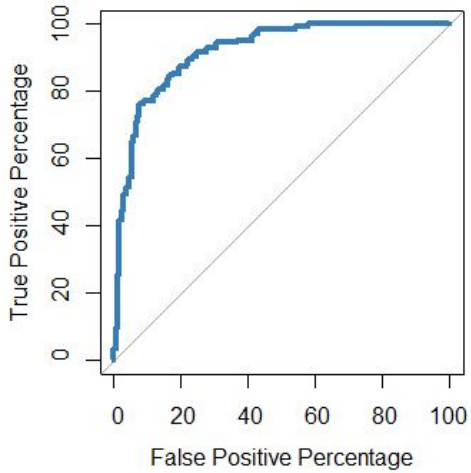
Naive Bayes: performance (10-fold cross evaluation)

Matrice di confusione



Accuracy = 82.4%
Precision = 80.1%
Recall = 83%
F-measure = 81.6%
95% CI : (0.78, 0.87)

Curva ROC



AUC = 91.8%

Decision Tree (CART)

L’albero di decisione è un modello di apprendimento supervisionato che cerca di effettuare una classificazione costruendo un albero decisionale in cui:

- Ogni nodo interno rappresenta una variabile
- Un arco verso un nodo figlio rappresenta un possibile valore per la variabile del nodo genitore
- Una foglia corrisponde al valore predetto per la classe, ottenuto a partire dai valori dei nodi che la precedono

L’albero così descritto può essere realizzato, a partire dal dataset fornito, utilizzando diversi criteri.
Nel caso dell’algoritmo CART la metrica utilizzata è l’indice di Gini

$$I = 1 - \sum_{i=1}^{k} fi^2$$

dove f_i sono le frequenze relative alle k modalità di X

Decision Tree: perchè sceglierlo

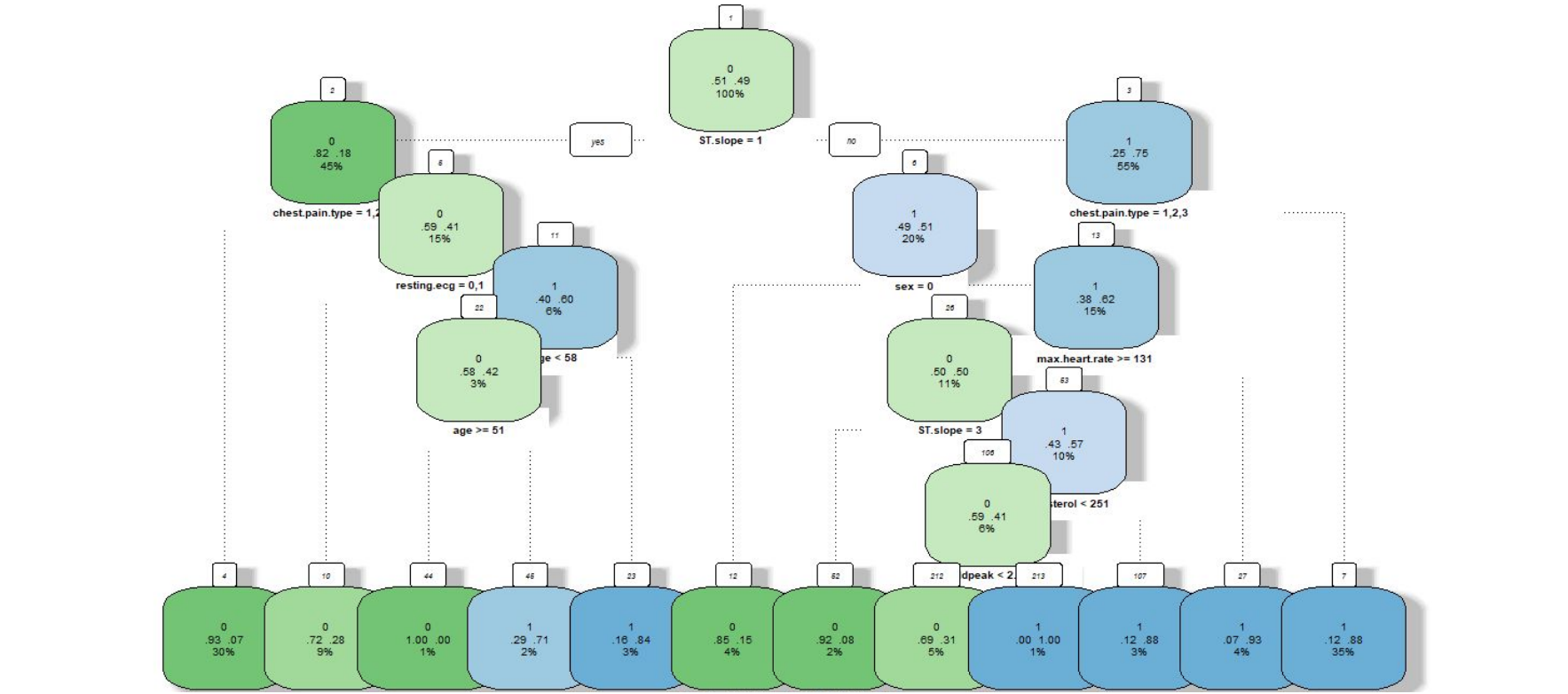
PRO

- Semplicità di utilizzo: si adatta bene sia a dati numerici che categorici e non ha iperparametri da selezionare
- Robustezza rispetto al rumore
- Alta interpretabilità

CONTRO

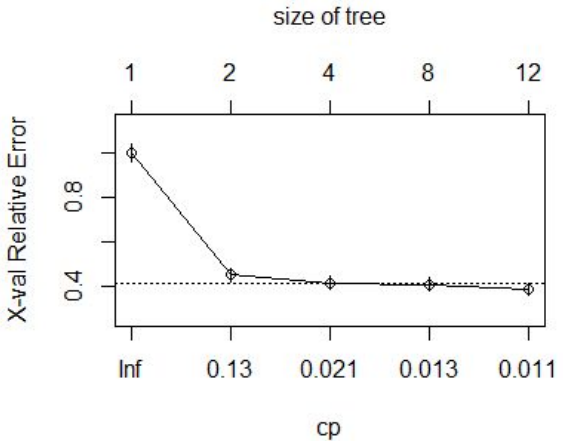
- Rischio di overfitting
- Meno performanti rispetto ad altri algoritmi

Decision tree: training

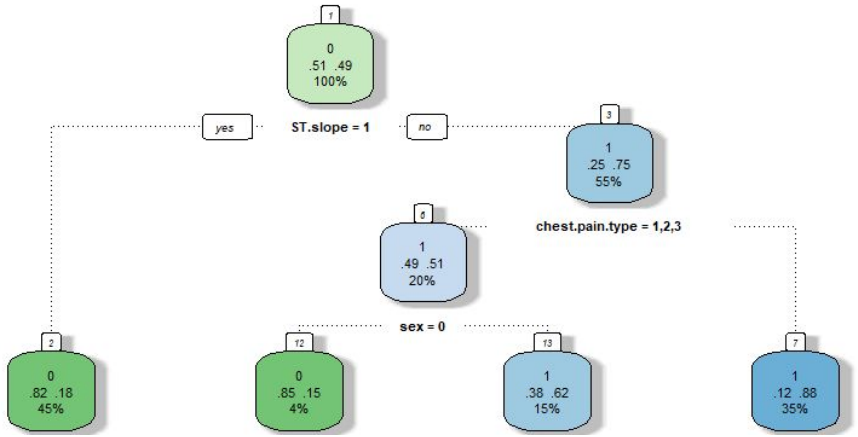


Decision tree: pruning dell'albero

	CP	nsplit	rel error	xerror	xstd
1	0.554017	0	1.00000	1.00000	0.037519
2	0.031856	1	0.44598	0.44875	0.031124
3	0.013850	3	0.38227	0.41274	0.030187
4	0.012465	7	0.31302	0.40720	0.030035
5	0.010000	11	0.26039	0.38504	0.029404

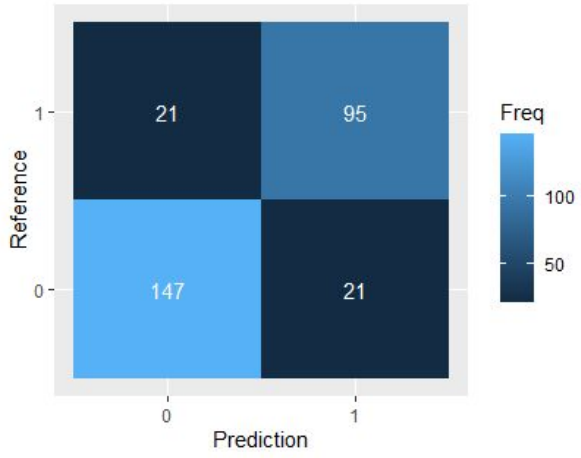


CP = 0.02



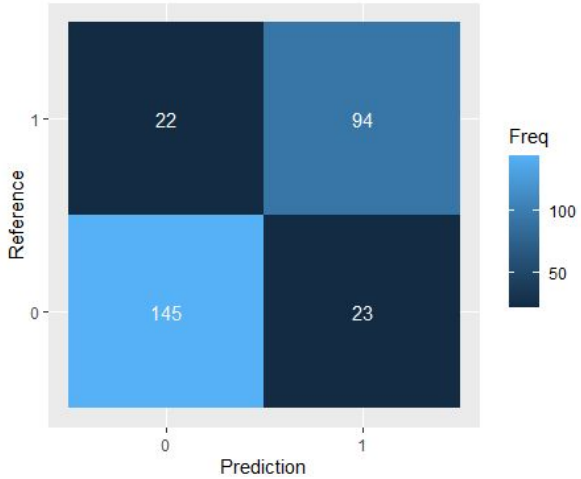
Decision tree: pruning dell'albero

Pre pruning



Accuracy = 85.2%
Precision = 81.9%
Recall = 81.9%

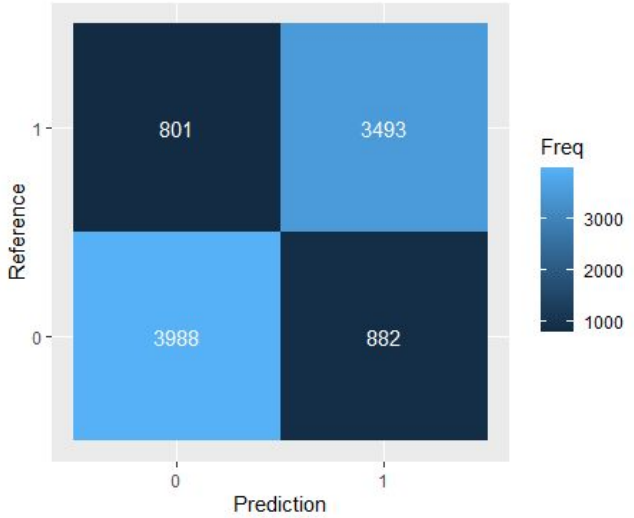
Post pruning



Accuracy = 84.2%
Precision = 80.3%
Recall = 81%

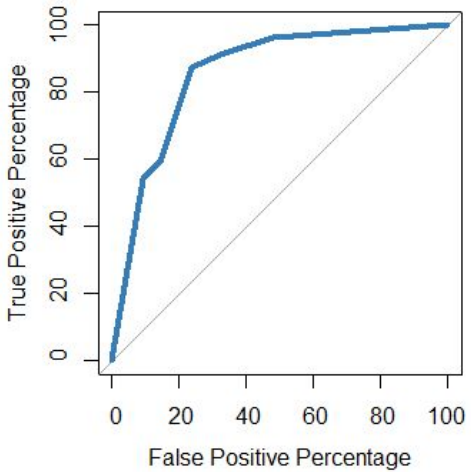
Decision Tree: performance (10-fold cross evaluation)

Matrice di confusione



Accuracy = 81.5%
Precision = 80%
Recall = 80.7%
F-measure = 80.4%
95% CI : (0.79, 0.88)

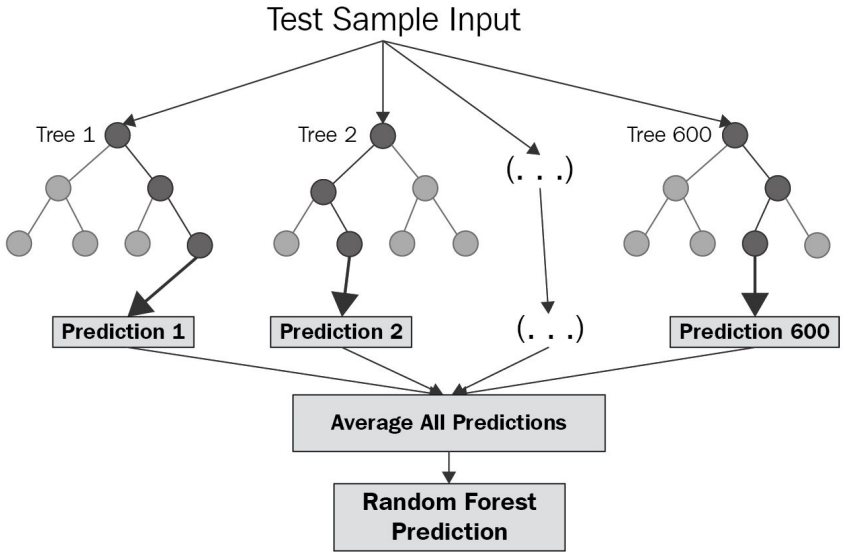
Curva ROC



AUC = 85.6%

Random Forest

Random Forest appartiene alla classe dei metodi di apprendimento d'insieme, ovvero quei metodi che usano modelli multipli per ottenere una migliore prestazione predittiva rispetto ai modelli singoli da cui è costituito. In particolare Random Forest viene ottenuto tramite bagging di alberi di decisione: l'algoritmo dunque addestra diversi alberi di decisione a partire da subset del trainset ed effettua predizioni interrogando tutti gli alberi e scegliendo la risposta più popolare (o facendo la media delle risposte).



Random Forest: perchè sceglierlo

PRO

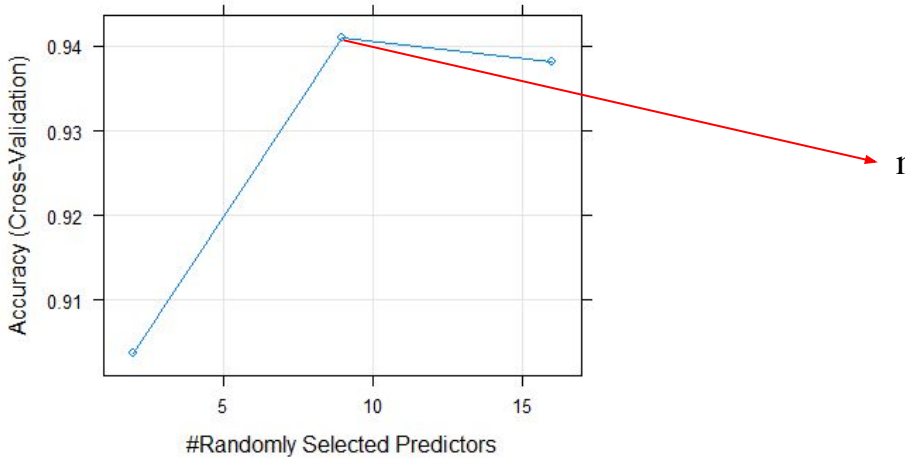
- Abbassa il rischio di overfitting degli alberi
- Performance migliori rispetto agli altri algoritmi di classificazione
- Robustezza rispetto al rumore

CONTRO

- Alta complessità computazionale
- Rinuncia all'interpretabilità tipica dei singoli alberi di decisione

Random Forest: training

Il training del modello richiede la selezione del parametro mtry, ovvero il numero di predittori che verranno usati per il training degli alberi (selezionati randomicamente).

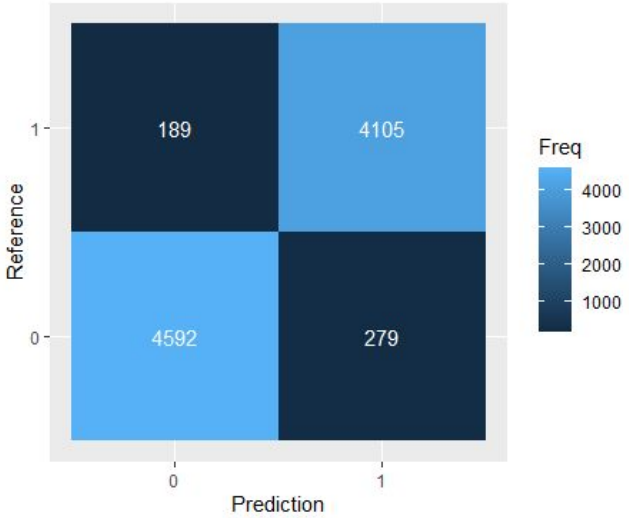


mtry = 9 è la soluzione ottimale

*Il numero di alberi addestrati, invece, è stato fissato a 500

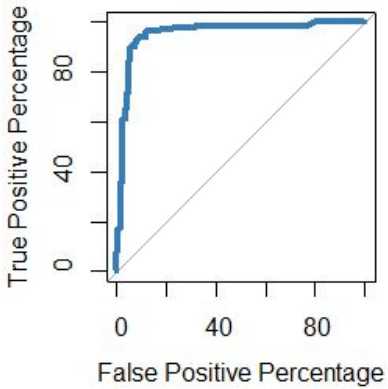
Random Forest: performance (10-fold cross evaluation)

Matrice di confusione



Accuracy = 94.9%
Precision = 93.6%
Recall = 95.6
F-measure = 94.6
95% CI : (0.89, 0.95)

Curva ROC

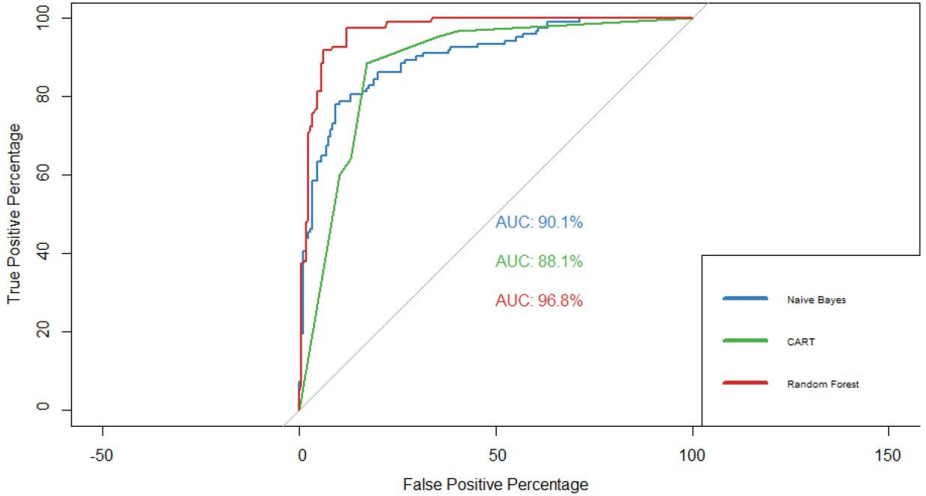


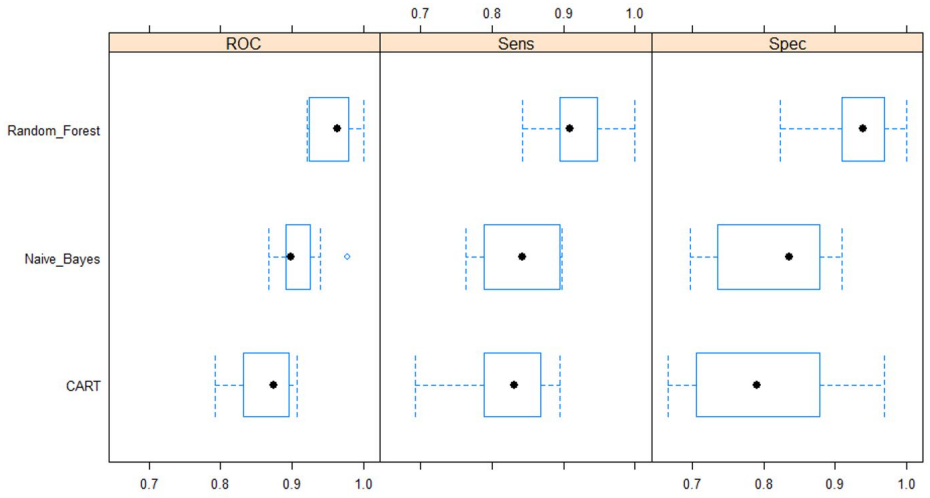
AUC = 95.2%

Confronto tra modelli

Naive Bayes	Decision Tree (CART)	Random Forest
Accuracy = 82.4%	Accuracy = 81.5%	Accuracy = 94.9%
Precision = 80.1%	Precision = 80%	Precision = 93.6%
Recall = 83%	Recall = 80.7%	Recall = 95.6
F-measure = 81.6%	F-measure = 80.4%	F-measure = 94.6
95% CI : (0.78, 0.87)	95% CI : (0.79, 0.88)	95% CI : (0.89, 0.95)

Confronto tra modelli: ROC





Confronto tra modelli: tempo

	Everything	Final Model	Prediction
Naive Bayes	3.57	0.01	NA
Dec. Tree (CART)	0.75	0.02	NA
Random Forest	14.75	0.42	NA

Conclusioni

Si ritiene che il modello più consigliabile fra i tre sia Random Forest che, nonostante la bassa interpretabilità e l'onerosità computazionale, ha ottenuto risultati nettamente superiori agli altri algoritmi.
Tra gli altri due modelli invece, la preferenza ricade su Decision Tree poiché quest'ultimo garantisce un'ottima interpretabilità e, allo stesso tempo, performance paragonabili a quelle di Naive Bayes.