

Breast Cancer Prediction

Chiara Coscarelli
a.a 2023/2024

Introduzione

Il cancro al seno (BC) è uno dei tumori più comuni tra le donne in tutto il mondo e, secondo le statistiche globali, rappresenta la maggior parte dei nuovi casi di cancro e dei decessi correlati al cancro, rendendolo un problema di salute pubblica significativo nella società odierna.



Obiettivi

L'obiettivo è classificare se il cancro al seno è benigno o maligno.



Obiettivi

- l'analisi approfondita dei dati estrapolati da un dataset.
- l'identificazione di feature associate alle diagnosi.
- l'implementazione di un modello di apprendimento in grado di calcolare la probabilità che un tumore al seno sia benigno o maligno.

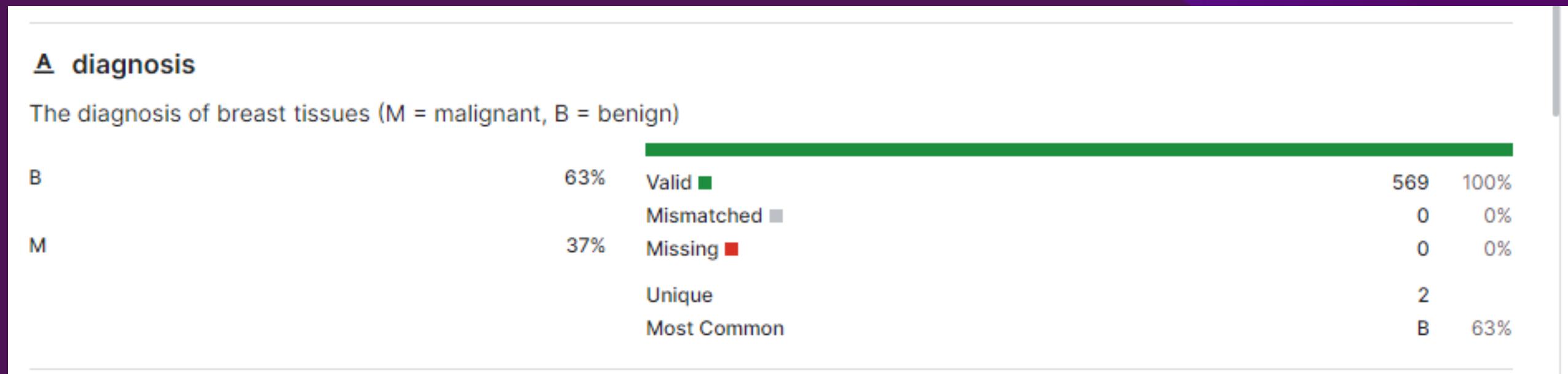
Analisi dei dati

- Acquisizione dei dati
- Esplorazione dei dati
- Analisi della qualità dei dati.



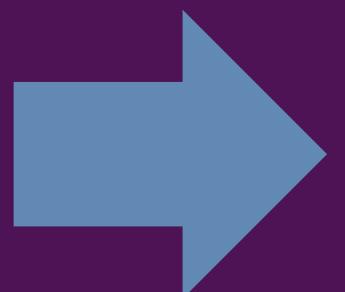
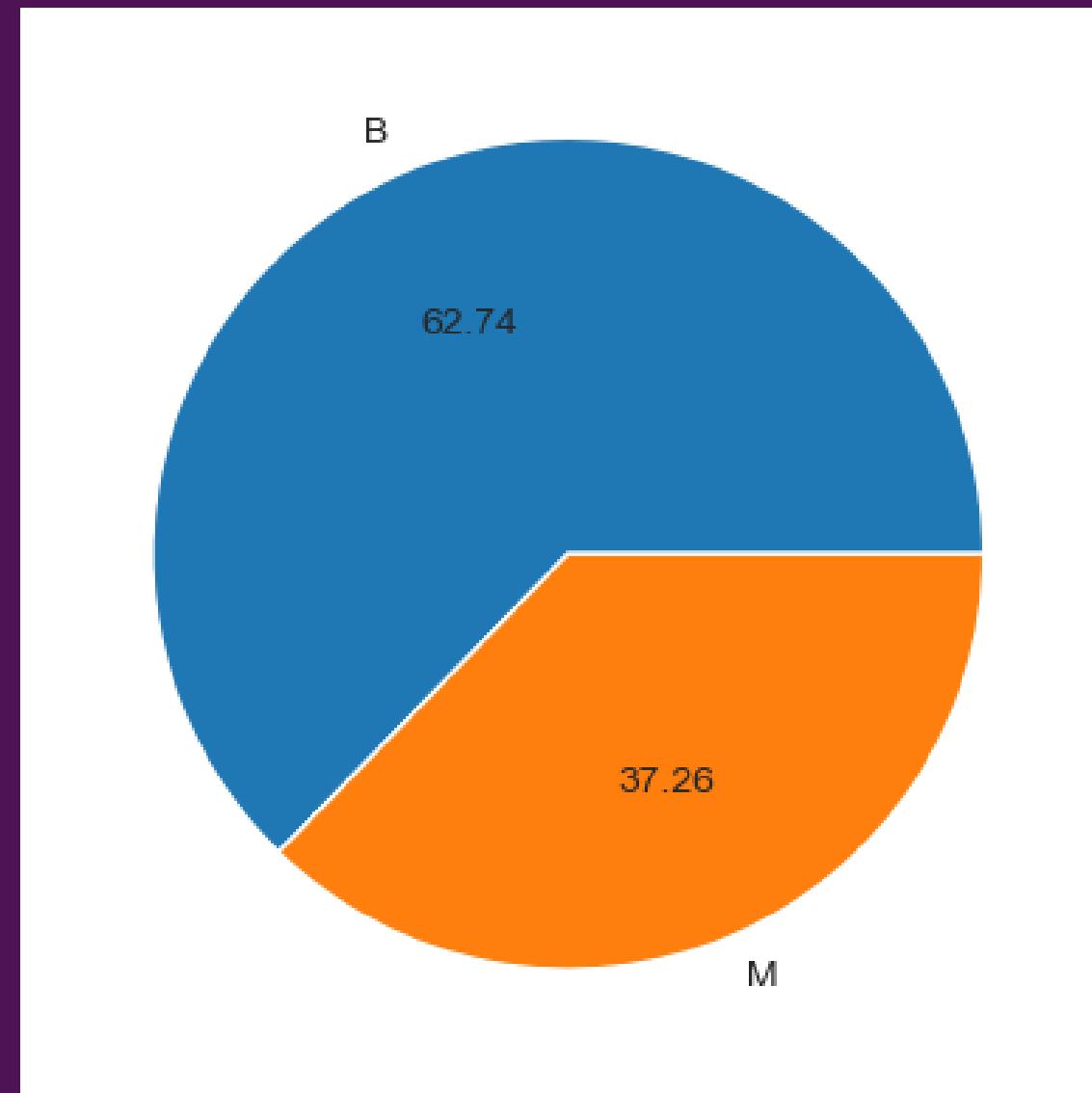
Analisi dei dati

- il dataset in esame contiene 569 righe e 32 colonne.
- La colonna che usiamo per predire è ‘Diagnosi’, e indica se il cancro al seno è maligno o benigno



Analisi dei dati

- 357 B (benigne)
- 212 M (maligne)



Problema: sbilanciamento dei dati!

Data Preparation

Data preparation

Data Cleaning

1. Pulizia dei dati
2. Rimozione colonne vuote
3. Trasformazione M=1 e B=0
4. Eliminazione valori nulli e duplicati

Feature Scaling

1. Split del dataset
2. Normalizzazione

Feature Selection

1. Visualizzazione dei dati attraverso la matrice di correlazione
2. Eliminazione variabili altamente correlate

Data Balancing

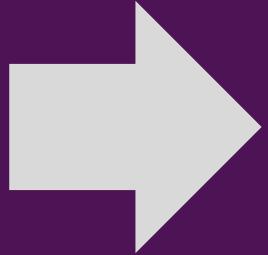
1. Bilancio dataset sbilanciato
2. Applicazione RandomUndersempler

Data Modeling

Data Modelling

- Scelta dell'algoritmo

apprendimento
supervisionato

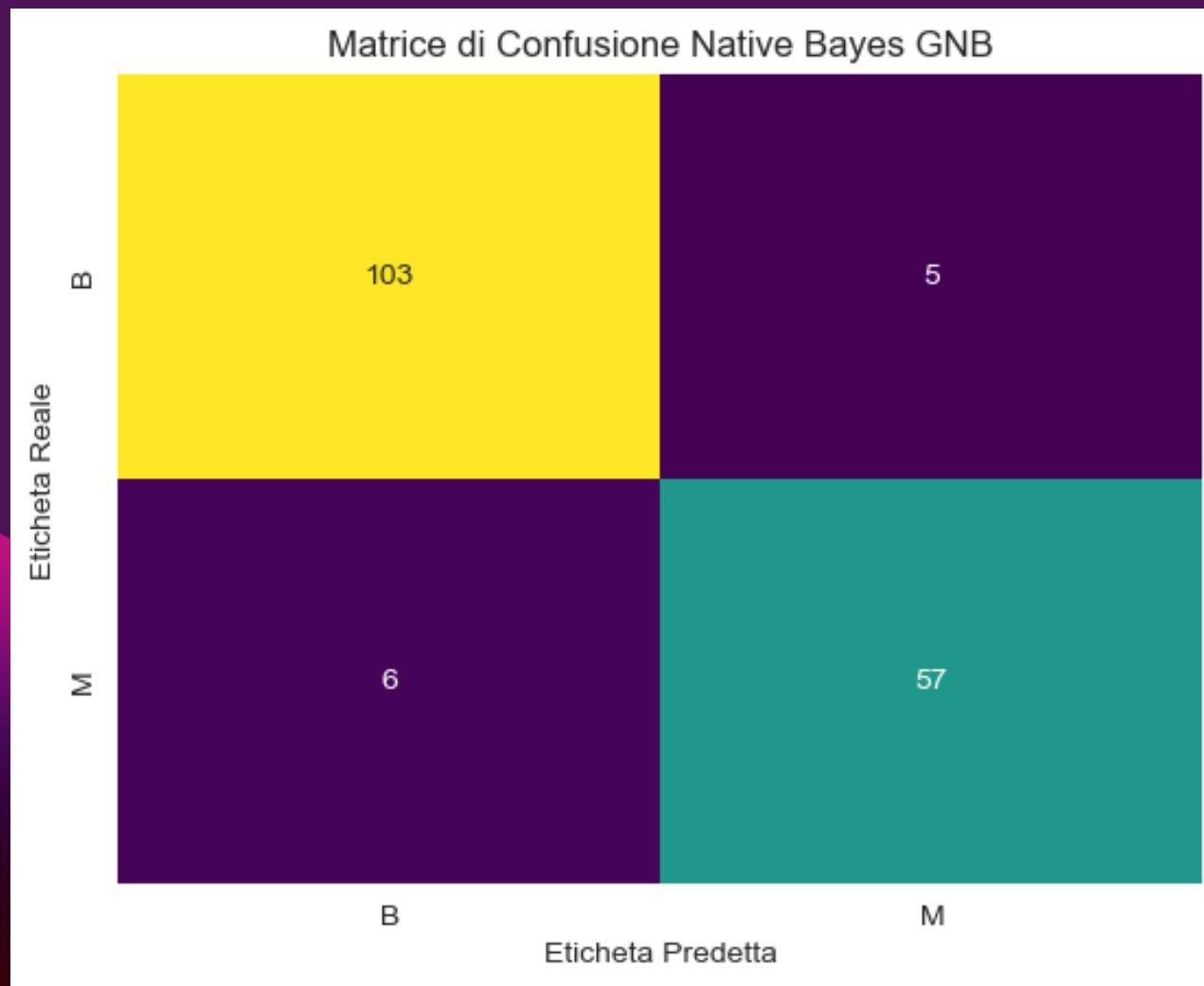


classificazione

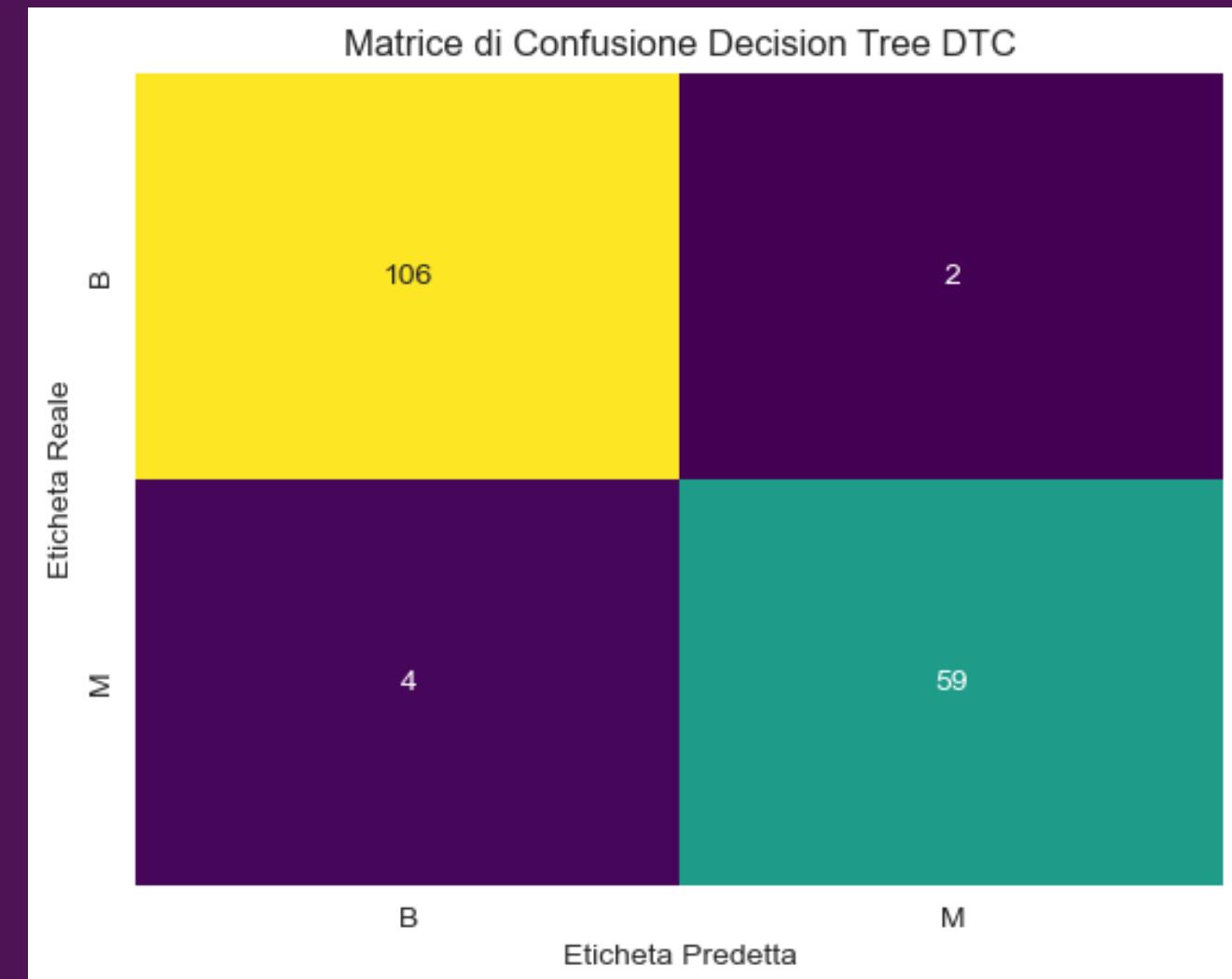
Data Modelling

- Addestramento

Native Bayes

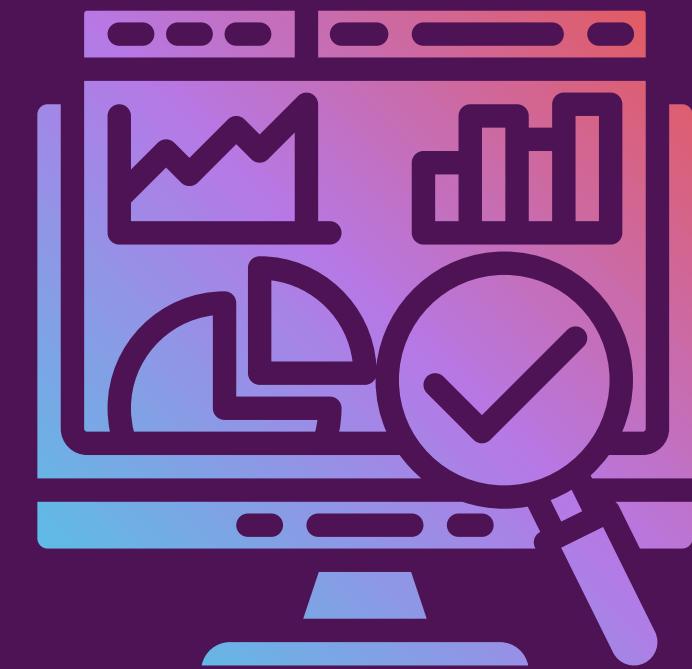


Decision Tree



Data Modelling

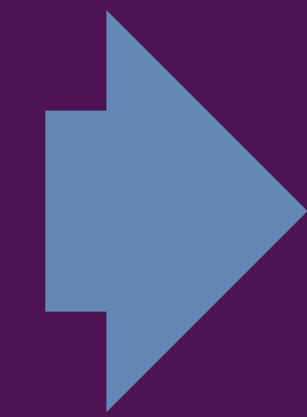
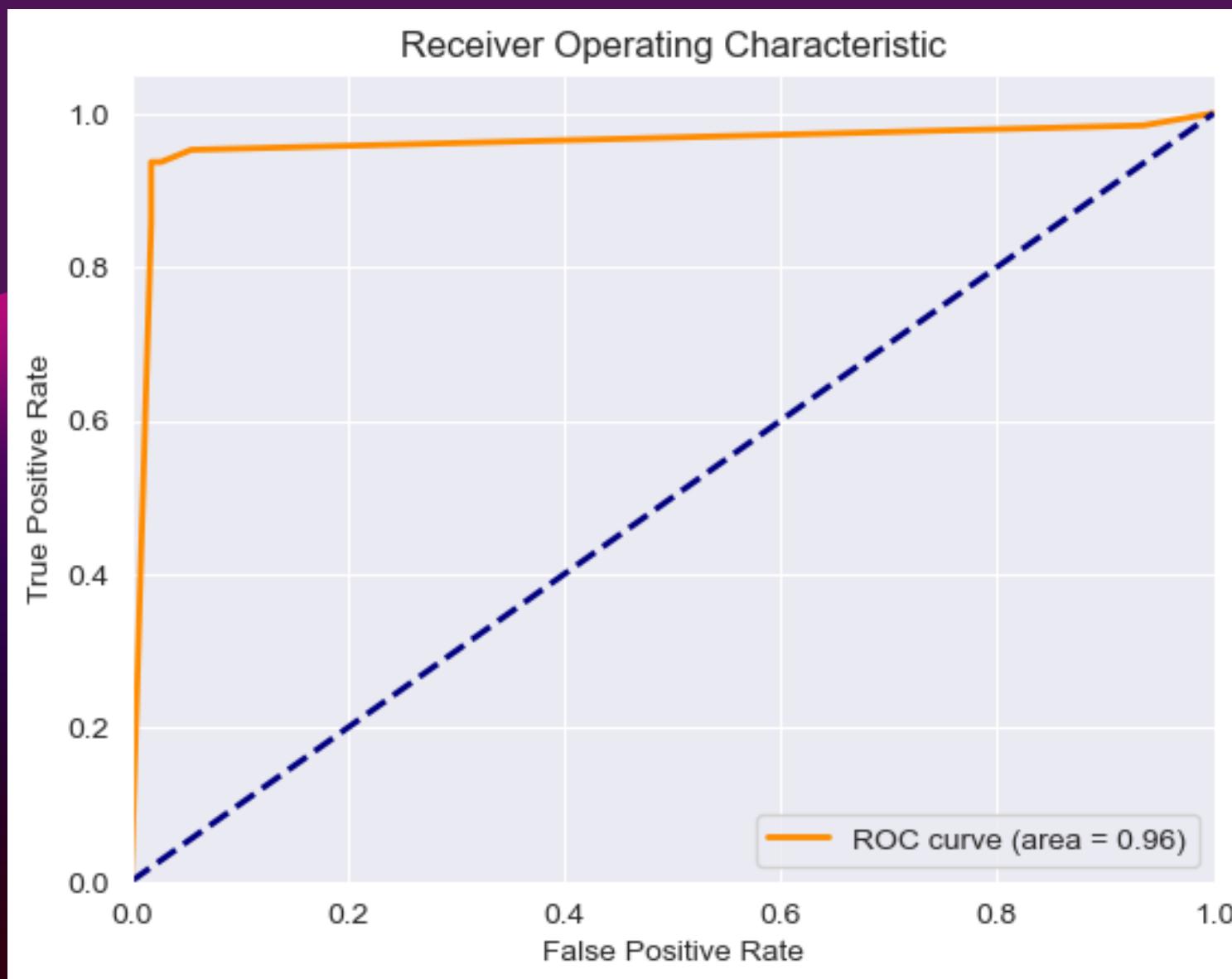
- Analisi dei risultati ottenuti
- A partire dai valori espressi dalla matrice di confusione ovvero TP, TN, FP, FN possiamo calcolare accuratezza, recall, precisione.



Variante	Precisione	Recall	Accuratezza
GNB	0.9193548387096774	0.9047619047619048	0.935672514619883
DTC	0.9672131147540983	0.9365079365079365	0.9649122807017544

Evaluation

- ROC-AUC curve



Il mio modello ha ottenuto un valore pari a 0.96, che è un ottimo risultato. Quindi posso considerare la costruzione del modello e in generale dell'approccio completa.

Grazie per l'attenzione