Contest
Statistical learning per Data Science

HOMEWORK

- **CLAUDIA LANDI**
- **MICHAEL CAVICCHIOLI**
- **VITTORIO SANI**

XGBoost vs Gradient Boosting Machine

	XGBoost	Gradient Boosting Machine
Iperparametri	Maggior numero di iperparametri che possono essere ottimizzati	Meno iperparametri: tuning più semplice ma meno flessibile
Gestione degli alberi	Gestione efficiente di dati sparsi e valori mancanti	Gestione meno efficiente
Regolarizzazione	Per evitare overfitting e penalizzare alberi complessi	Non include regolarizzazione nella funzione obiettivo
Velocità ed Efficienza	Ottimizzazione cache e parallelismo	Non ottimizzato per il calcolo in parallelo

Dataset utilizzato

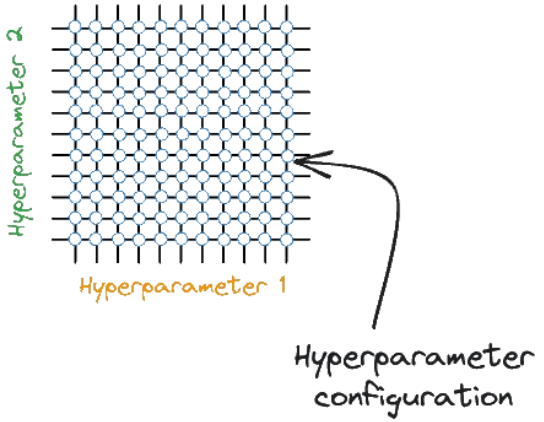
Airline Passengers Satisfaction

- **Variabili eterogenee** → Preprocessing e label encoding
- **Etichetta binaria** → satisfied (1), neutral or dissatisfied (0)
- **Grandi dimensioni** → 129487 righe, 25 colonne.

Tuning dei parametri

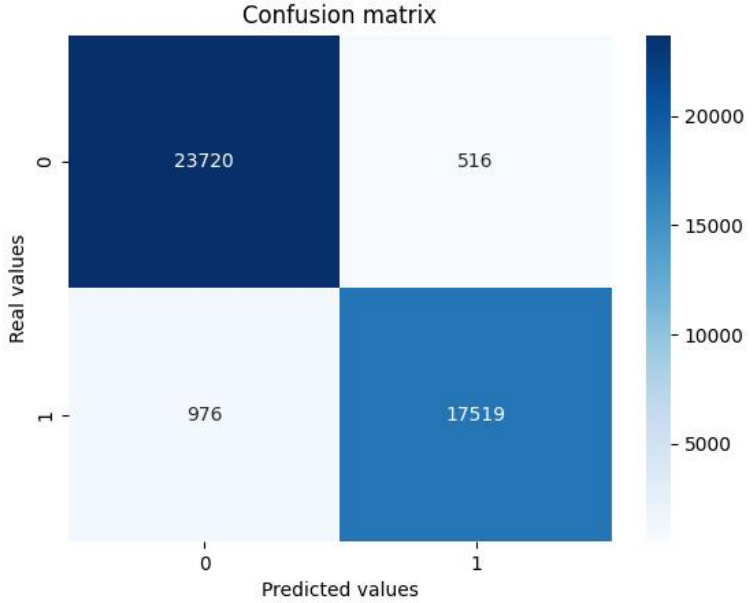
- Tuning sequenziale dei parametri:

- Tuning in parallelo con Grid Search:
  - Solo il parametro alpha è diverso,
  - L'accuracy è molto simile.

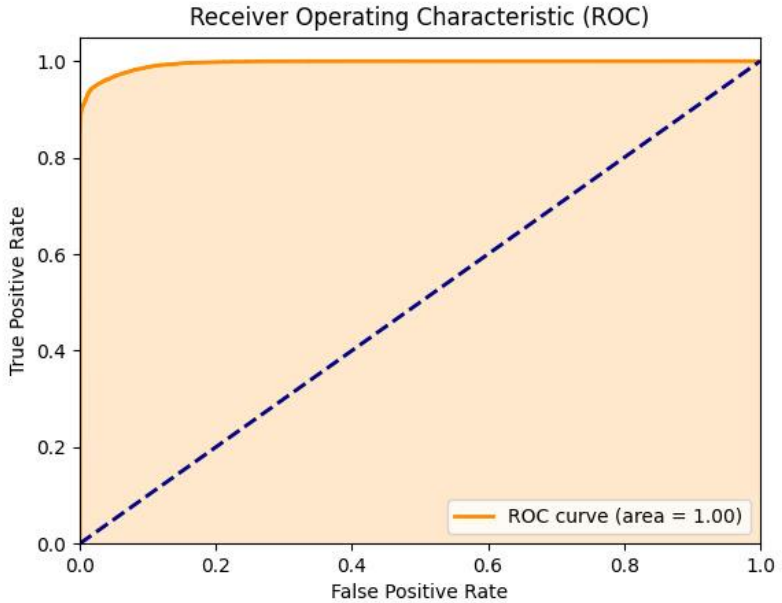


Parametri della ricerca sequenziale

Matrice di confusione



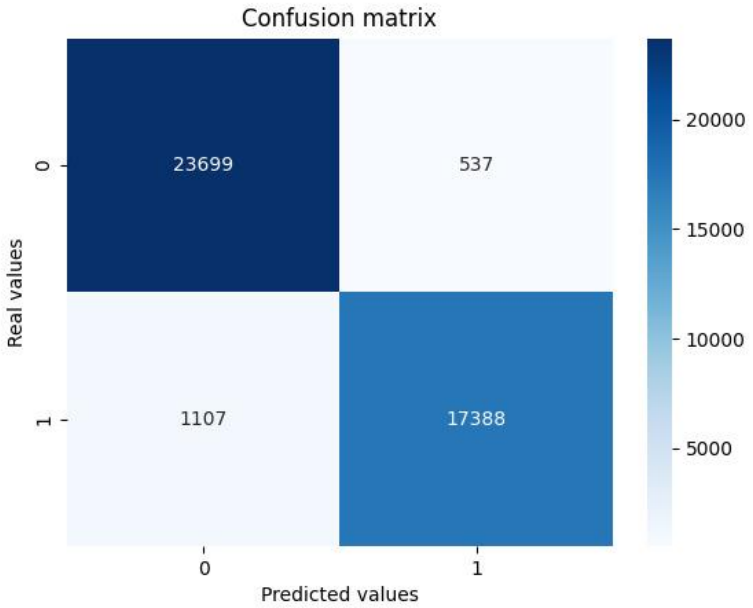
ROC - AUC



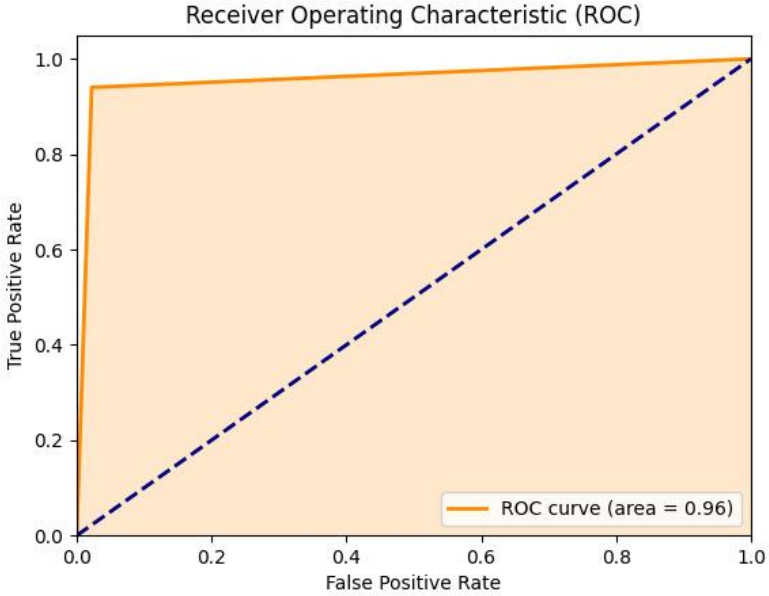
Accuracy: 0.965

Parametri ottenuti con Grid Search

Matrice di confusione



ROC - AUC



Accuracy: 0.959

Grazie per l’attenzione.