

Personal Key Indicators of Heart Disease

Studente
Alessandro La Cava
247436

Sommario

- Exploratory Data Analysis
- Preprocessing
- Classificazione
- Ottimizzazione dei modelli
- Analisi dei risultati
- Conclusioni

EDA: dataset

Il dataset è costituito da circa 320 mila entry e 18 variabili: 9 booleane, 5 stringhe e 4 decimali.

Nel dataset non sono presenti valori mancanti; tuttavia, sono presenti entry duplicate.

Il dataset, inoltre, è fortemente sbilanciato.

EDA: attributi

- HeartDisease
- BMI
- Smoking
- AlcoholDrinking
- Stroke
- PhysicalHealth
- MentalHealth
- DiffWalking
- Sex
- AgeCategory
- Race
- Diabetic
- PhysicalActivity
- GenHealth
- SleepTime
- Asthma
- KidneyDisease
- SkinCancer

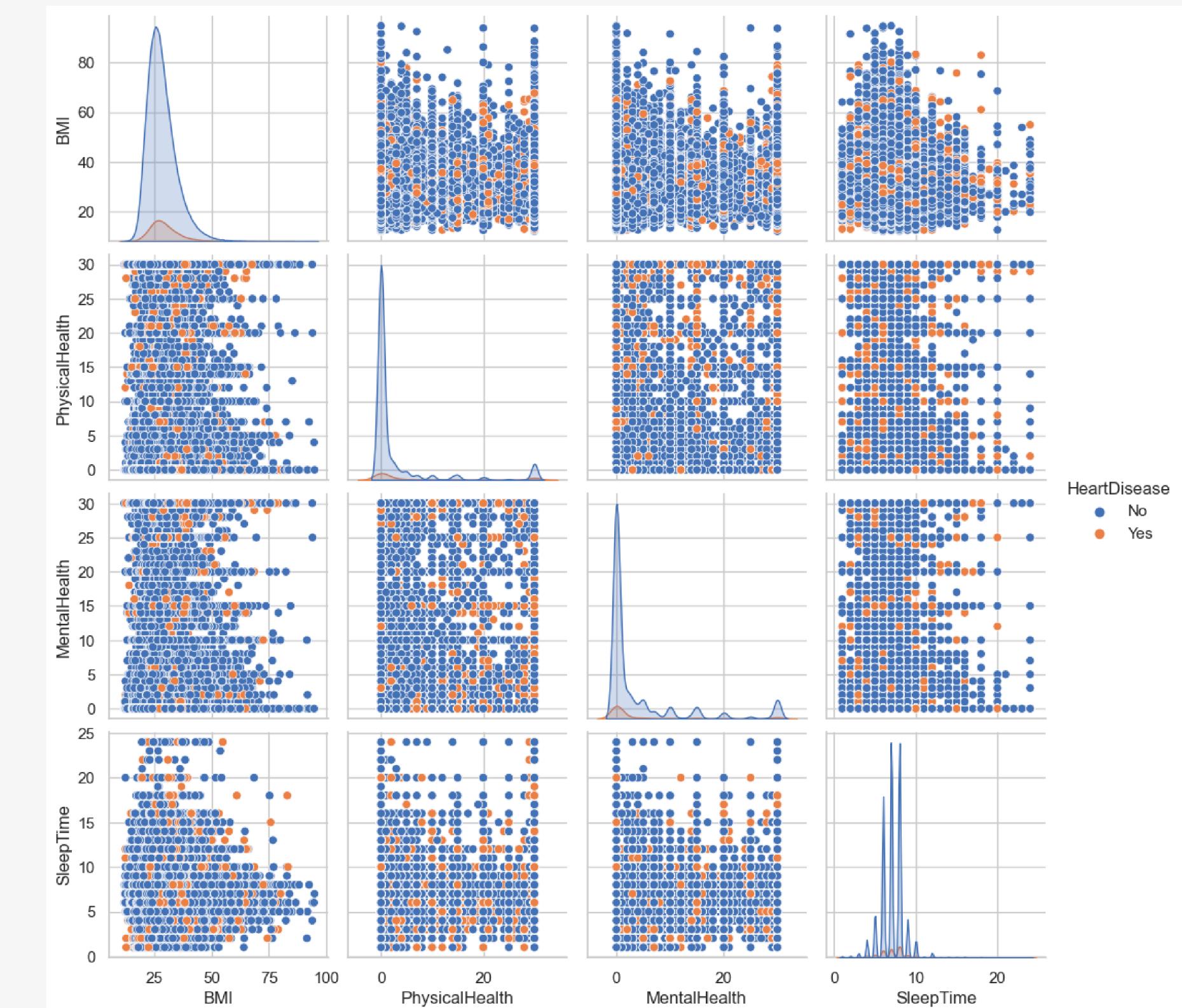
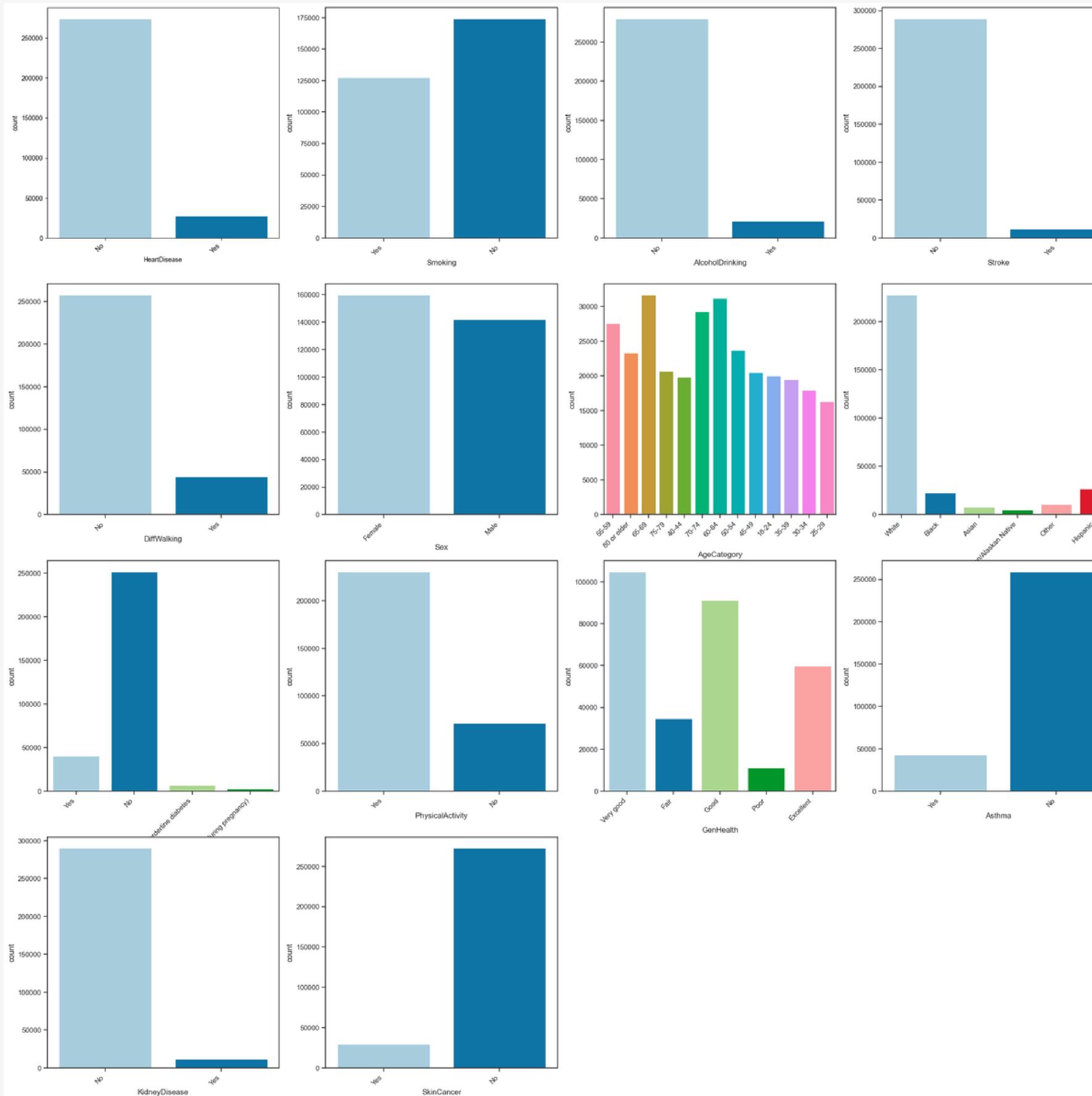
EDA: analisi

È stata effettuata un'analisi preliminare della struttura e della distribuzione degli attributi all'interno del dataset.

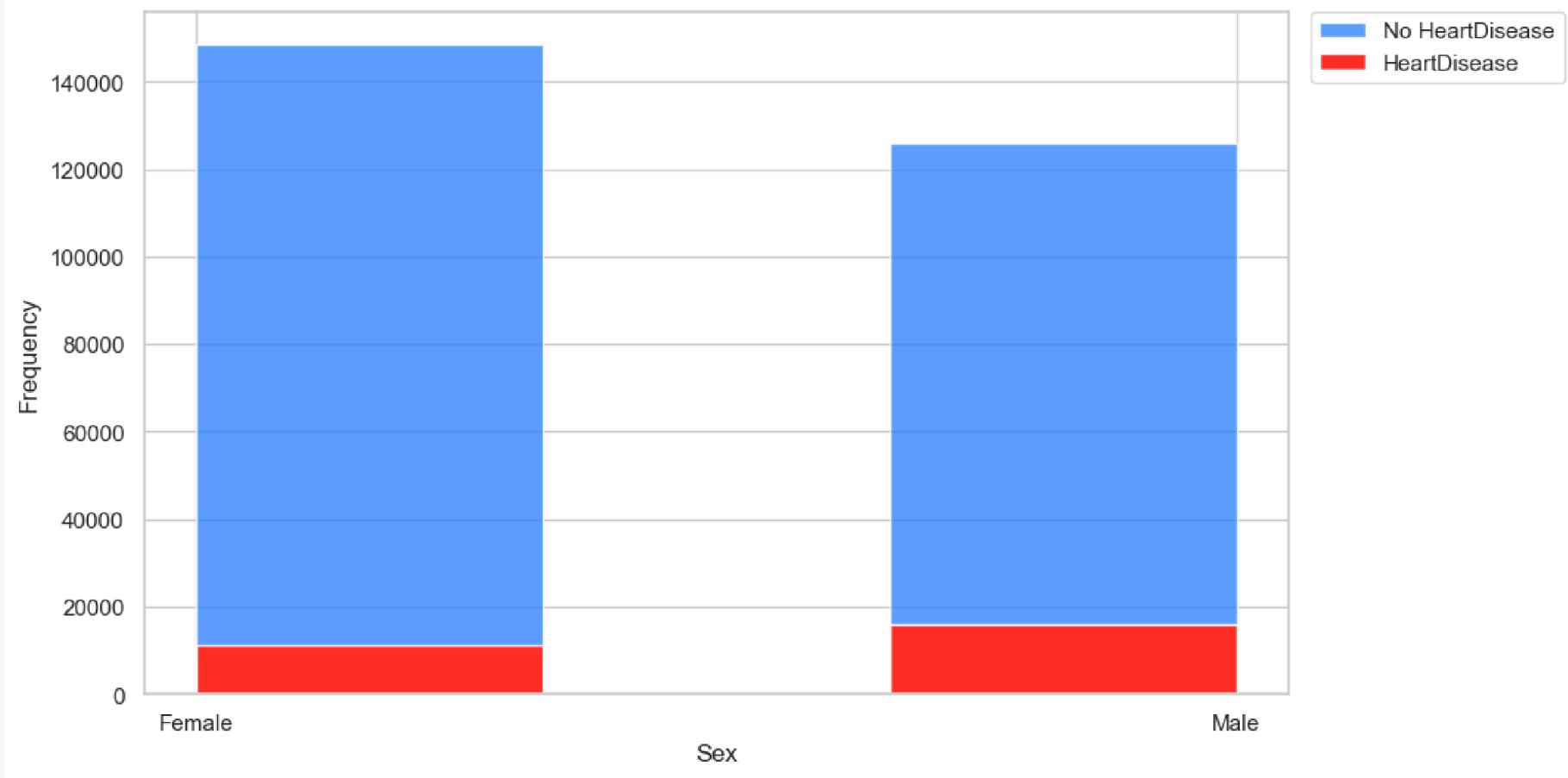
In seguito ci si è concentrati sull'analisi delle eventuali relazioni tra di esso, in modo da individuare quelli che potrebbero influenzare maggiormente l'insorgere di patologie cardiache.

In particolare, ci si è concentrati sull'analizzare le possibili relazioni tra l'attributo target HeartDisease e gli attributi Race, Sex, GenHealth, DiffWalking, Smoking e AgeCategory.

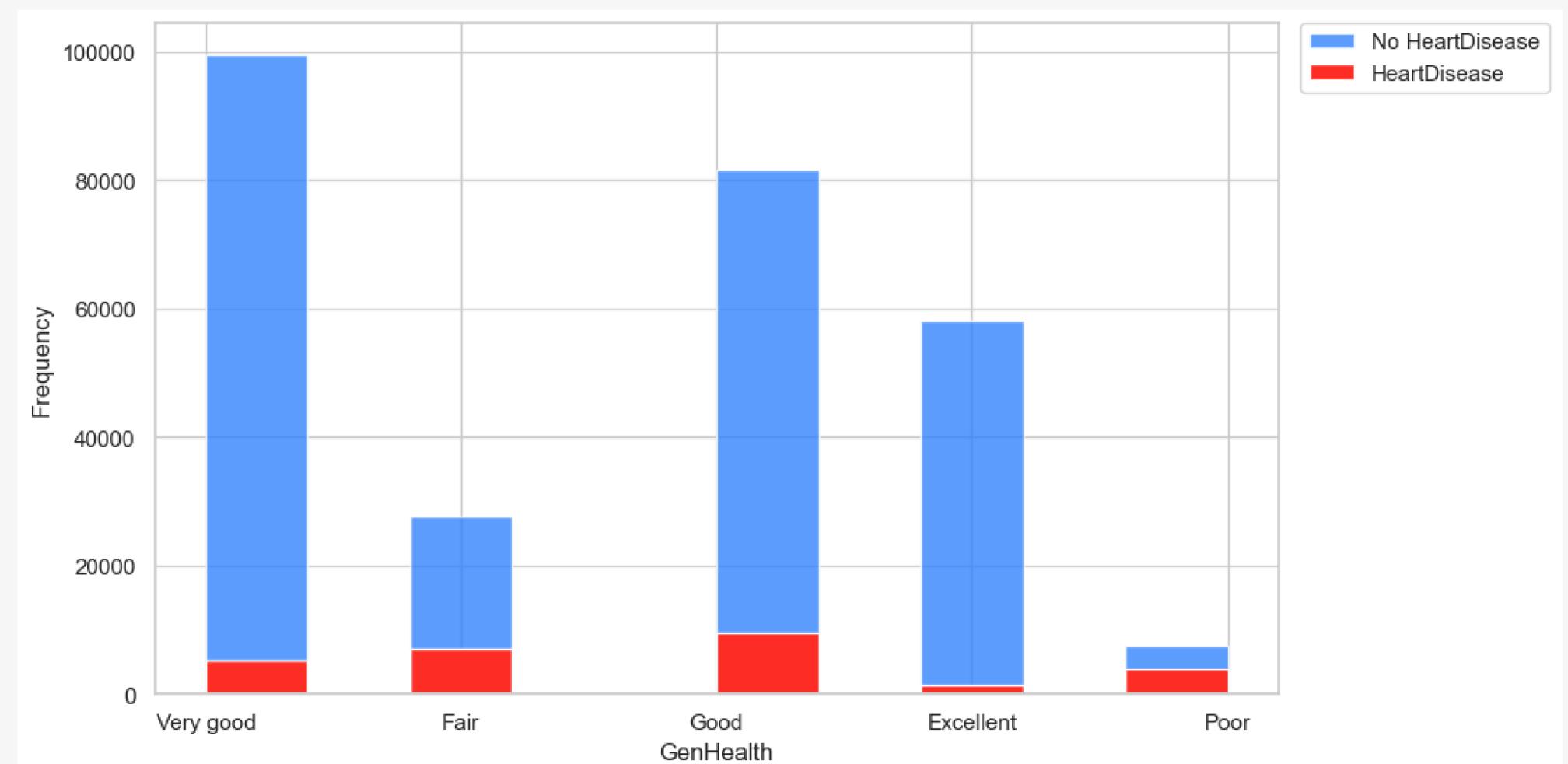
Distribuzione degli attributi



HeartDisease
● No
● Yes

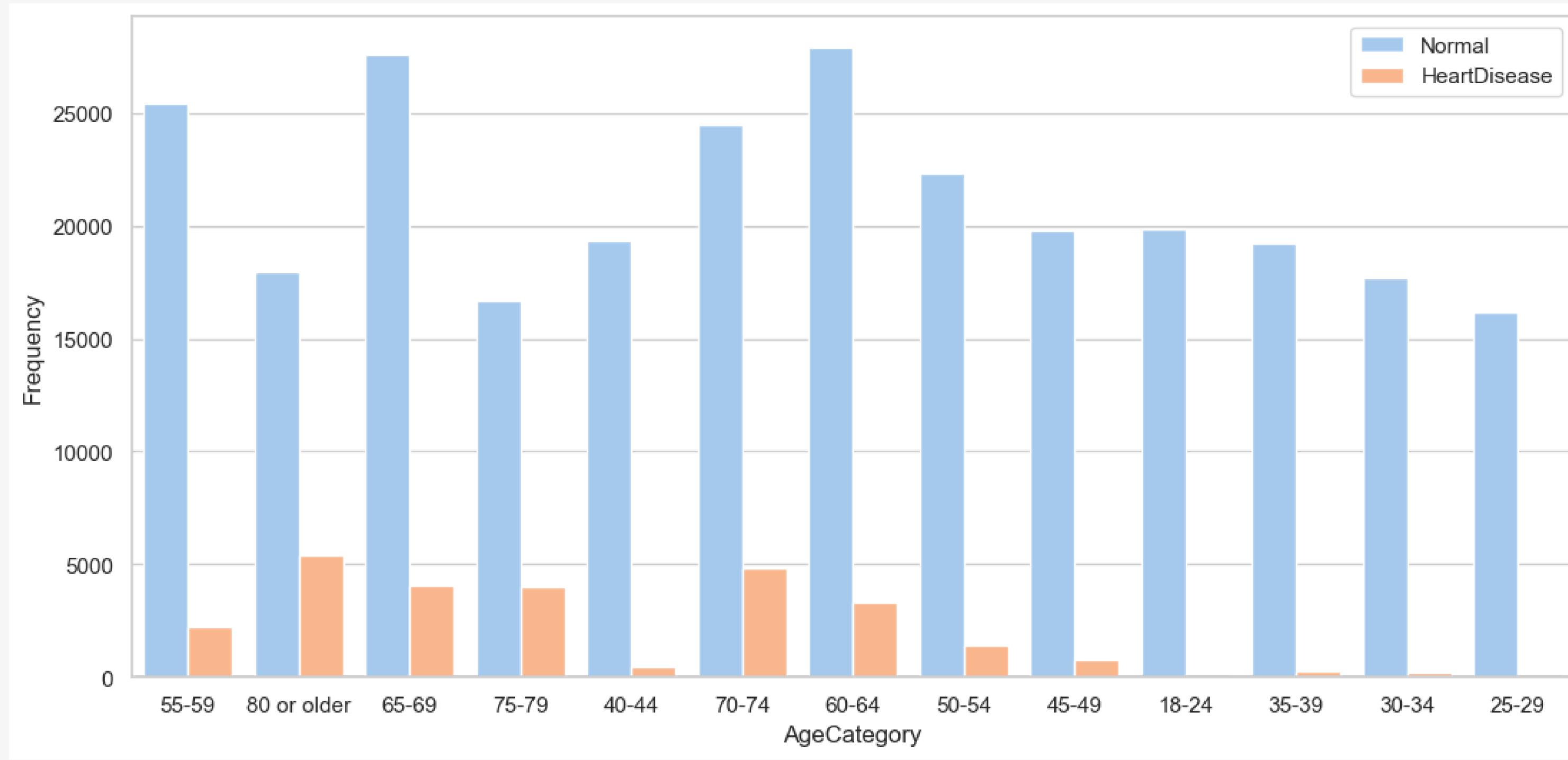


Distribuzione della
feature **Sex**

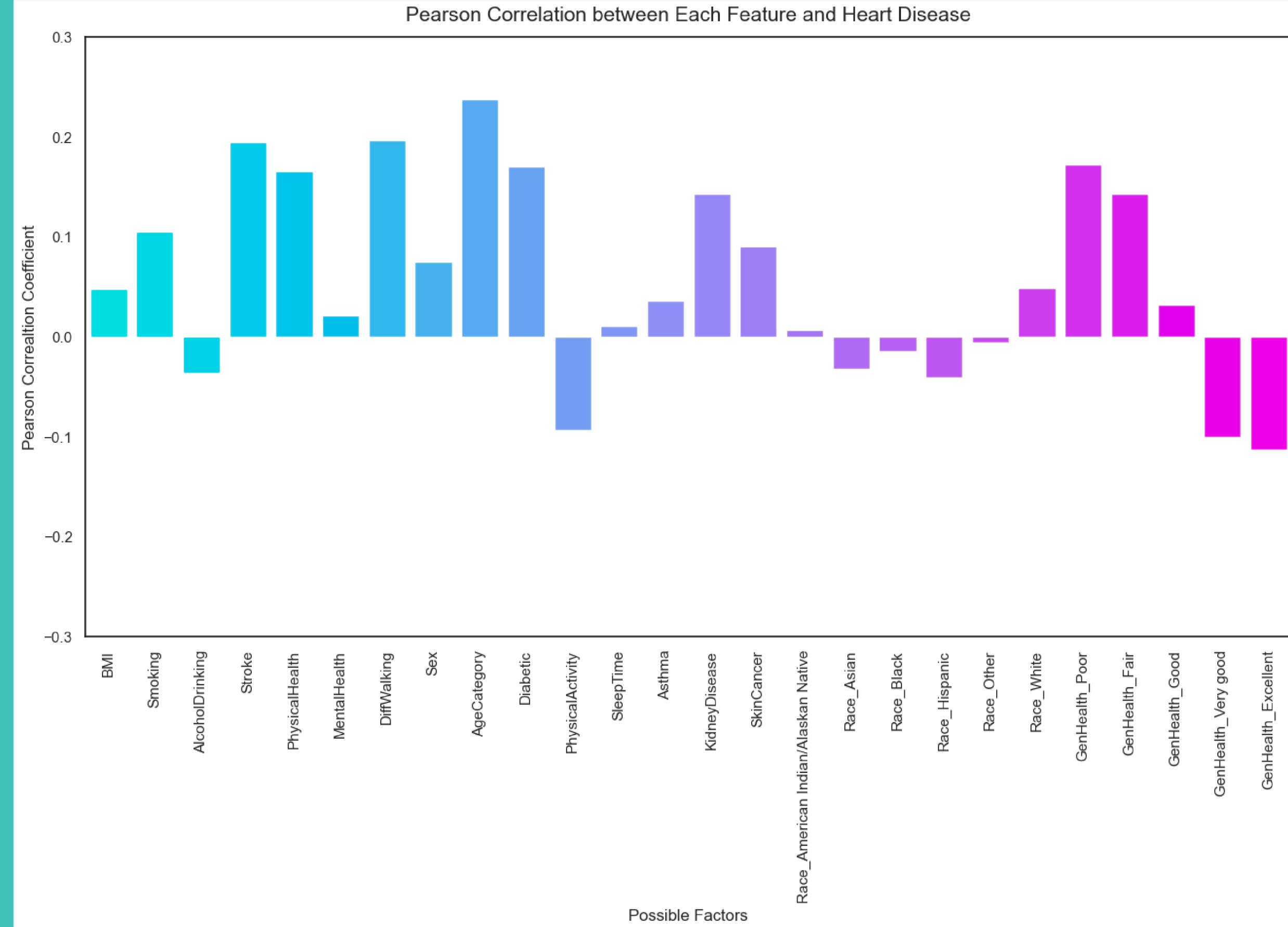


Distribuzione della
feature **GenHealth**

Distribuzione della feature **AgeCategory**



Correlazione tra ogni feature e la feature target **HeartDisease**



Informazioni ottenute

Risultato 1

L'etnia potrebbe influenzare l'insorgere di malattie cardiache.

Risultato 2

I pazienti di sesso maschile sono maggiormente soggetti all'insorgere di malattie cardiache

Risultato 3

Lo stato di salute generale influenza notevolmente l'insorgere di malattie cardiache.

Risultato 4

I pazienti che presentano difficoltà nel camminare o nel salire le scale sono più soggetti all'insorgere di malattie cardiache.

Risultato 5

Fumare aumenta l'insorgere di malattie cardiache.

Risultato 6

L'età è il fattore che influenza maggiormente l'insorgere di malattie cardiache.

Informazioni ottenute

Inoltre, è stato possibile notare come alcuni fattori, come MentalHealth e SleepTime, non influenzino molto l'insorgere di malattie cardiache (si è scelto di mantenerle all'interno del dataset).

Per quanto riguarda gli altri fattori, essi influenzano in maniera più o meno moderata l'insorgere di malattie cardiache; in particolare, Stroke, PhysicalHealth e Diabetic sembrano essere i fattori più influenti.

Preprocessing

GenHealth ed AgeCategory sono state rappresentate come delle variabili ordinali, ovvero si è scelto di introdurre un ordinamento all'interno dei possibili valori assunti da tali attributi. Nel caso di AgeCategory è stato utilizzato l'ordinamento lessicografico, mentre nel caso di GenHealth è stato imposto il seguente ordinamento:

["Poor" < "Fair" < "Good" < "Very good" < "Excellent"]

```
GenHealth_category=[ "Poor", "Fair", "Good", "Very good", "Excellent"]
df.GenHealth = df.GenHealth.astype(CategoricalDtype(ordered=True,
categories=GenHealth_category))
df.AgeCategory = df.AgeCategory.astype(CategoricalDtype(ordered=True))
```

Preprocessing

L'attributo Diabetic può assumere 4 diversi valori: "Yes", "No", "No, borderline diabetes" e "Yes (during pregnancy)".

Per questioni di semplicità, è stato scelto di semplificare il possibile range di valori e considerare "No, borderline diabetes" come un semplice "No" e "Yes (during pregnancy)" come un "Yes".

```
encode_Diabetic = {'Yes': True, 'No': False, 'No, borderline diabetes': False, 'Yes (during pregnancy)':True}
df['Diabetic'] = df['Diabetic'].apply(lambda x: encode_Diabetic[x])
df['Diabetic'] = df['Diabetic'].astype('bool')
```

Classificazione

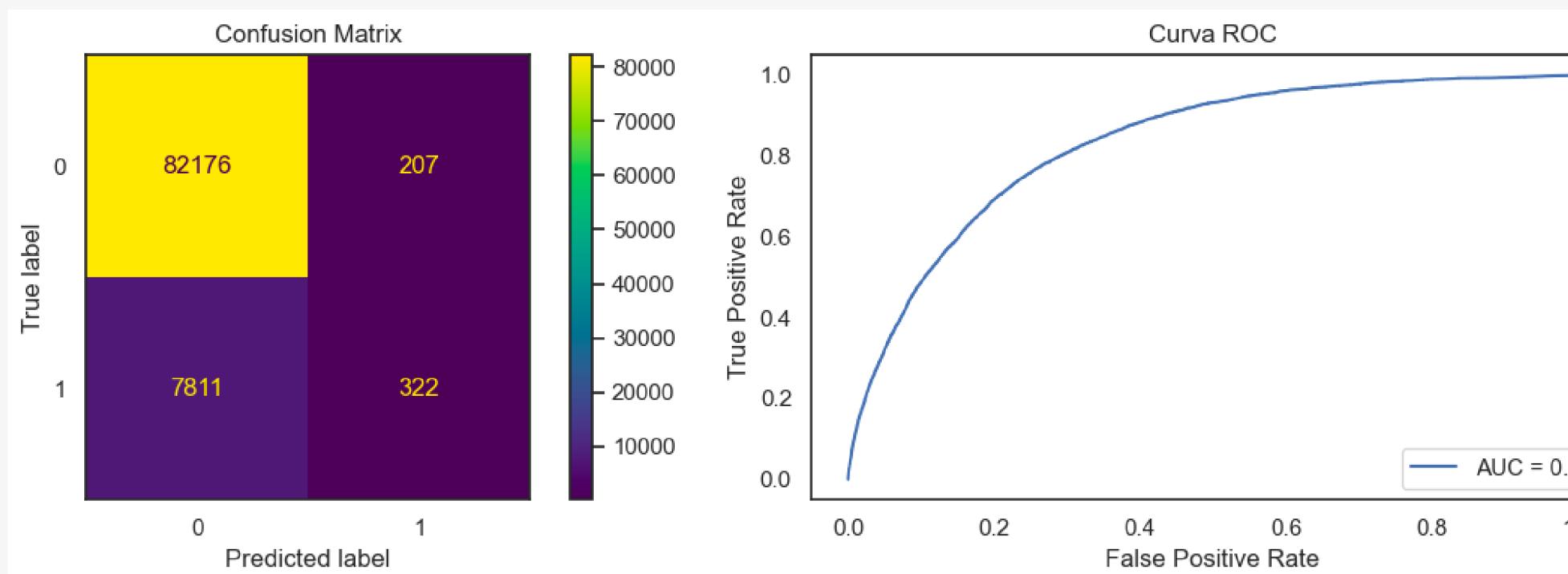
Bilanciamento del dataset

È stato implementato e addestrato un classificatore RandomForest per mostrare come un dataset fortemente sbilanciato possa degradare notevolmente l'accuratezza e le performance del modello.

Successivamente il dataset è bilanciato sia tramite oversampling che tramite undersampling e il modello RandomForest è stato nuovamente addestrato su dataset bilanciato.

Per i successivi modelli è stato utilizzato il dataset bilanciato mediante oversampling e normalizzato mediante feature scaling.

Random Forest con dataset sbilanciato

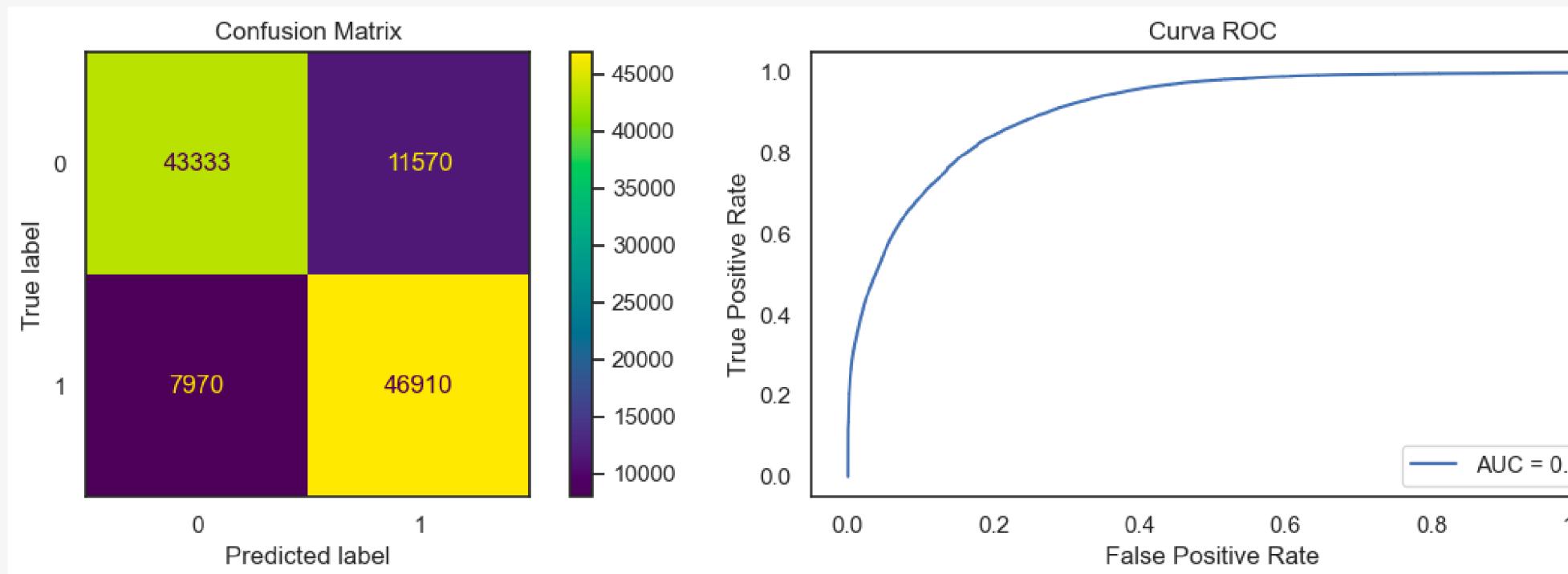


Random Forest Classifier (Imbalanced Dataset)

Classification Report:

	precision	recall	f1-score	support
0	0.91	1.00	0.95	82383
1	0.61	0.04	0.07	8133
accuracy				0.91
macro avg	0.76	0.52	0.51	90516
weighted avg	0.89	0.91	0.87	90516

Random Forest con dataset bilanciato tramite oversampling



Random Forest Classifier (Oversampling)

Classification Report:

	precision	recall	f1-score	support
0	0.84	0.79	0.82	54903
1	0.80	0.85	0.83	54880
accuracy				0.82
macro avg	0.82	0.82	0.82	109783
weighted avg	0.82	0.82	0.82	109783

Modelli utilizzati

I modelli che sono stati presi in considerazione sono:

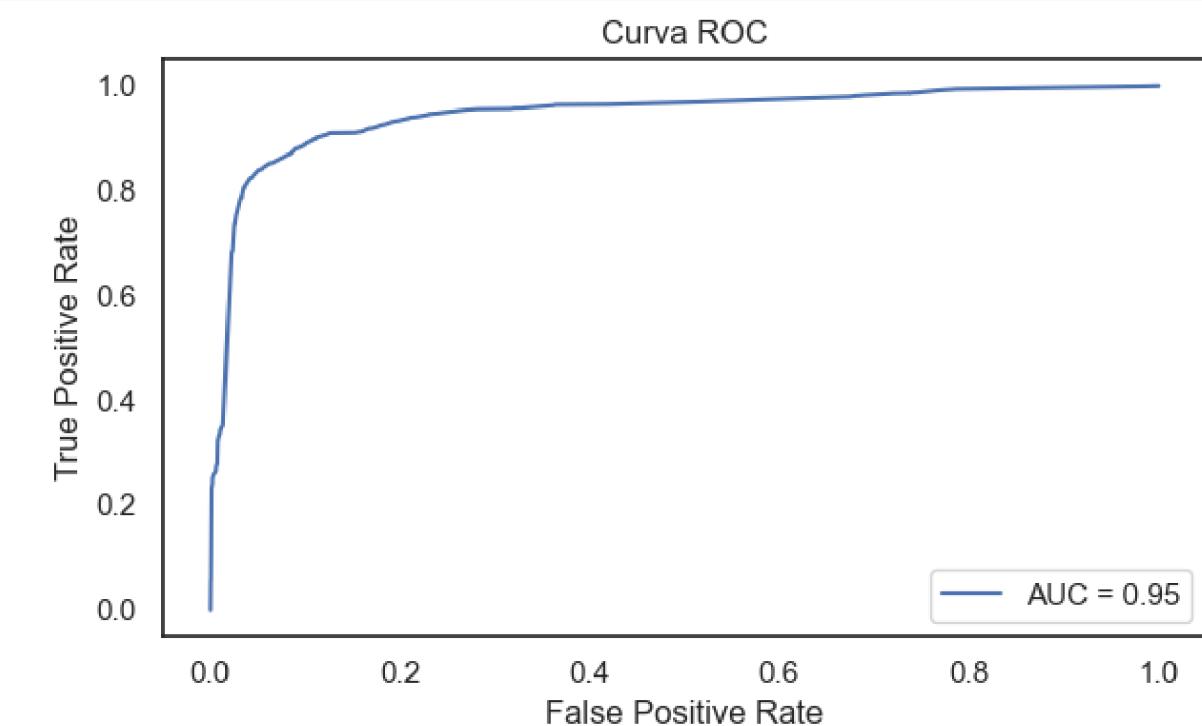
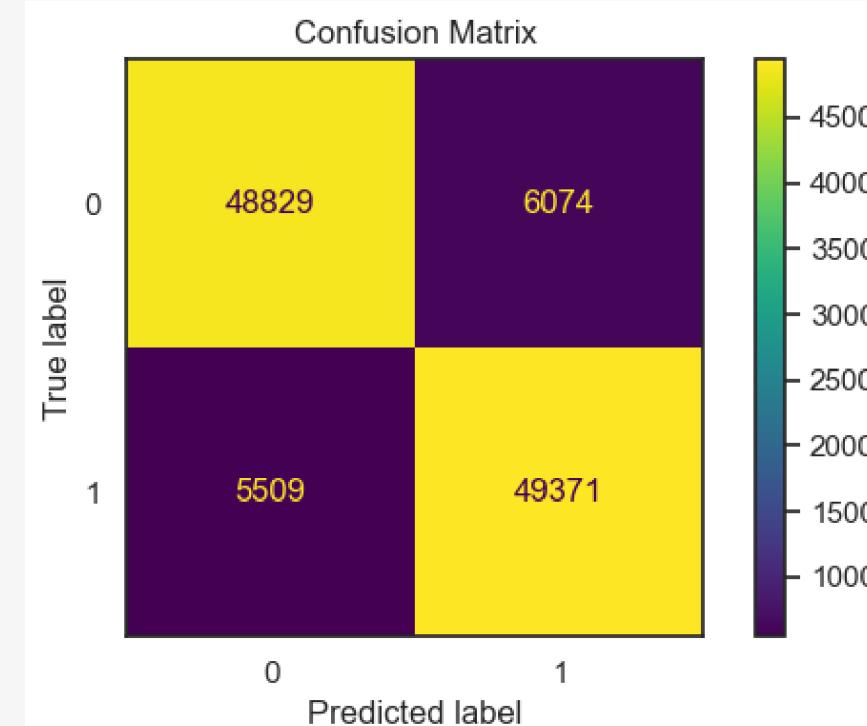
- SGDClassifier
- K-Nearest Neighbors Classifier
- Decision Tree Classifier
- Logistic Regression
- Random Forest Classifier
- ExtraTrees Classifier
- Naive Bayes Classifier
- XGBoost Classifier
- AdaBoost Classifier
- Voting Classifier
- Bagging Classifier
- Rete neurale multi-livello

Modelli utilizzati

Per ogni modello sono state analizzate le principali metriche di valutazione, in particolare accuracy e F-measure, e sono state visualizzate matrice di confusione e curva ROC.

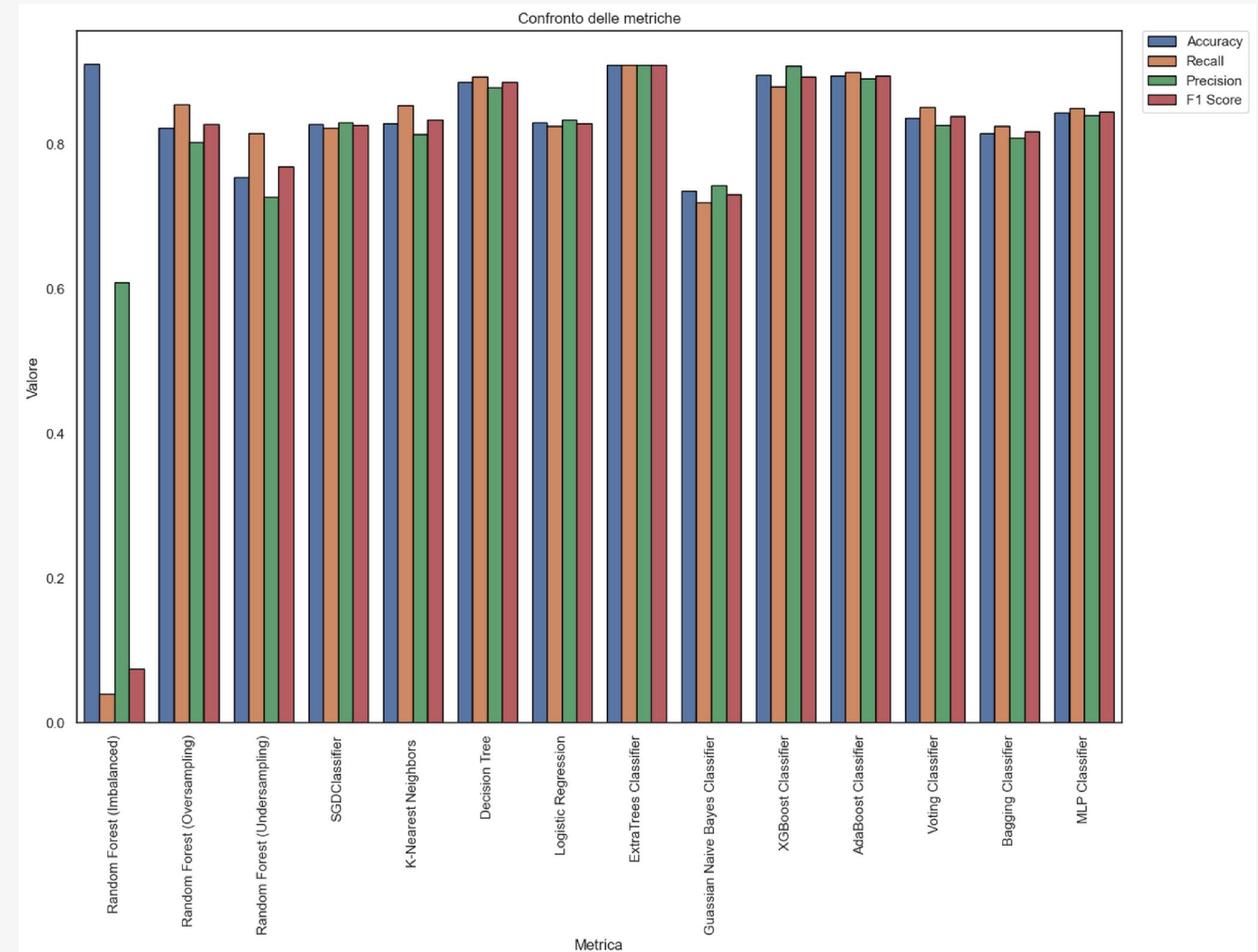
AdaBoost Classifier				
Classification Report:				
	precision	recall	f1-score	support
0	0.90	0.89	0.89	54903
1	0.89	0.90	0.90	54880
accuracy			0.89	109783
macro avg	0.89	0.89	0.89	109783
weighted avg	0.89	0.89	0.89	109783

Esempio AdaBoost Classifier



Ottimizzazione

Una volta addestrati i vari modelli, alcuni tra i migliori sono stati ottimizzati tramite GridSearch.



Ottimizzazione

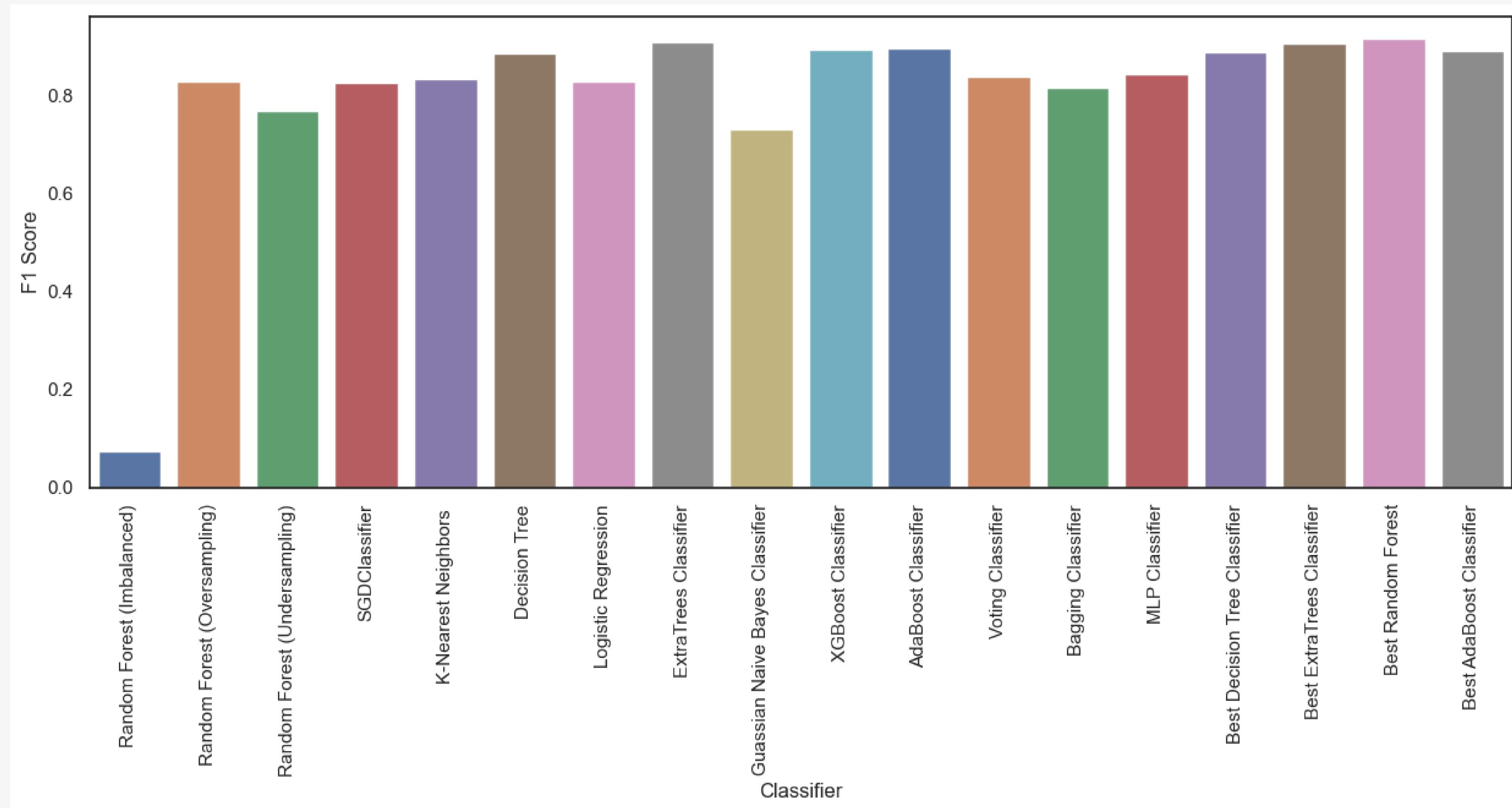
AdaBoost: {'base_estimator__criterion': 'entropy',
'base_estimator__splitter': 'best', 'n_estimators': 2}

Random Forest: {'max_depth': None, 'min_samples_split':2,
'n_estimators': 30}

ExtraTrees: {'class_weight': 'balanced', 'max_depth':None,
'n_estimators': 10}

Decision Tree: {'criterion': 'entropy', 'max_depth': None, 'splitter':'best'}

Analisi dei risultati



Analisi dei risultati

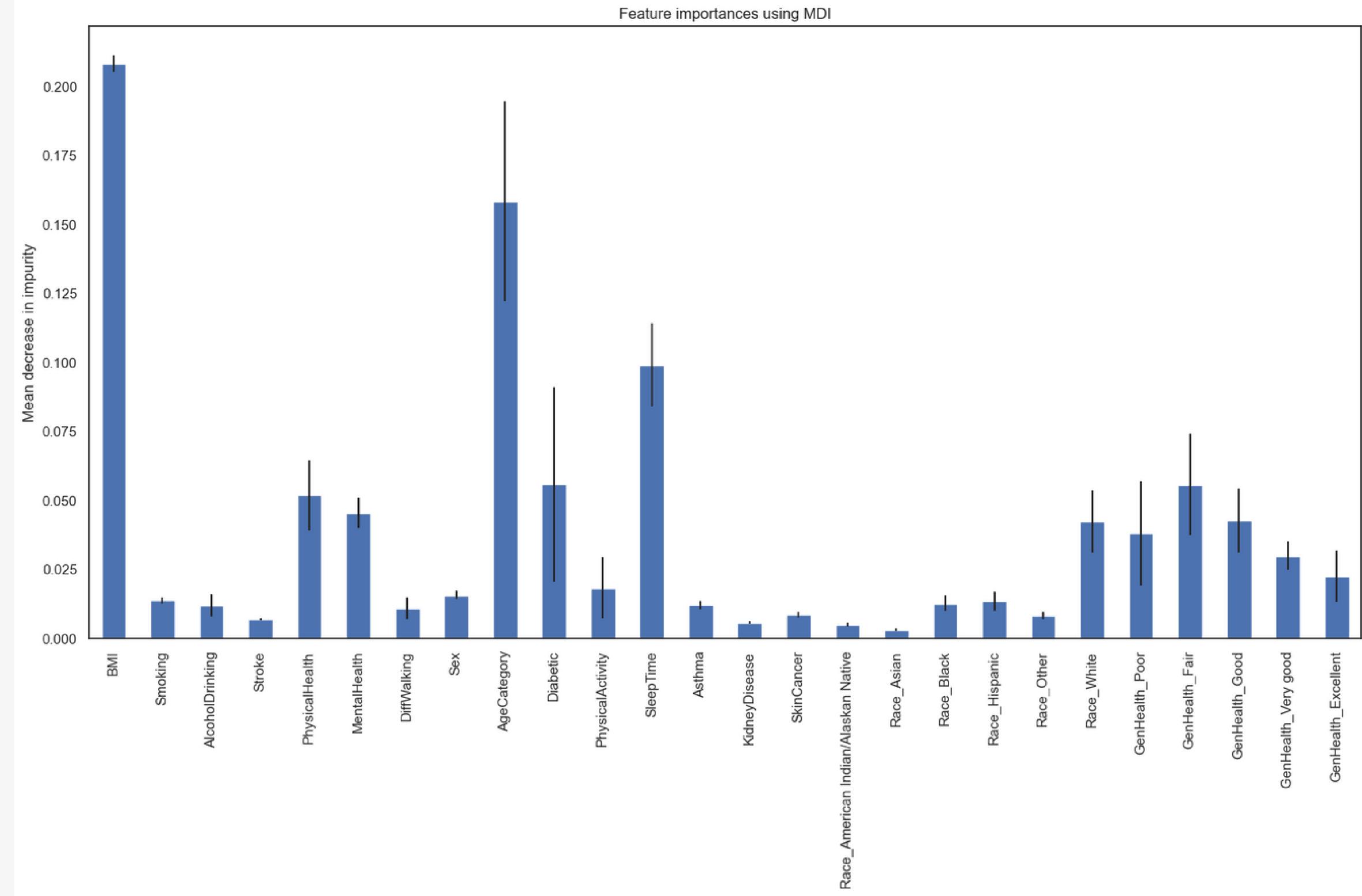
Come si può evincere dal grafico, il miglior modello è Random Forest ottimizzato tramite GridSearch.

Più in generale, è possibile notare come la maggior parte dei modelli più performanti siano modelli basati su alberi decisionali. In particolare, AdaBoost, ExtraTrees, XGBoost e Decision Tree risultano essere molto performanti sia nella versione normale che nella versione ottimizzata con GridSearch.

Analisi dei risultati

		Accuracy	Recall	Precision	F1 Score
	Best Random Forest	0.916153	0.910733	0.920681	0.915680
	ExtraTrees Classifier	0.909849	0.909785	0.909868	0.909826
	Best ExtraTrees Classifier	0.907344	0.900747	0.912755	0.906711
	AdaBoost Classifier	0.894492	0.899617	0.890450	0.895010
	XGBoost Classifier	0.895539	0.880175	0.908038	0.893889
	Best AdaBoost Classifier	0.896796	0.850091	0.937636	0.891720
	Best Decision Tree Classifier	0.887269	0.894825	0.881462	0.888093
	Decision Tree	0.885447	0.893786	0.879082	0.886373
	MLP Classifier	0.843792	0.849818	0.839644	0.844700
	Voting Classifier	0.836131	0.851531	0.826033	0.838588
	K-Nearest Neighbors	0.829272	0.853426	0.814044	0.833270
	Logistic Regression	0.829855	0.824526	0.833349	0.828914
	Random Forest (Oversampling)	0.822013	0.854774	0.802155	0.827629
	SGDClassifier	0.827022	0.823196	0.829484	0.826328
	Bagging Classifier	0.815154	0.825510	0.808697	0.817017
	Random Forest (Undersampling)	0.754058	0.815621	0.727050	0.768793
	Gaussian Naive Bayes Classifier	0.735396	0.720044	0.742768	0.731229
	Random Forest (Imbalanced)	0.911419	0.039592	0.608696	0.074348

Feature importances per il modello Random Forest ottimizzato



Conclusioni

Dall'analisi delle feature importances del modello Random Forest (Grid) è stato possibile notare come le caratteristiche che presentano maggior importanza siano:

- BMI
- AgeCategory
- SleepTime