Indice

[Introduzione 2](#_Toc108188093)

[Data Analysis e Data Visualization 2](#_Toc108188094)

[Matrice di correlazione 3](#_Toc108188095)

[Features Importances con RandomForestClassifier e ExtraTreesClassifier 3](#_Toc108188096)

[Classificazione 4](#_Toc108188097)

[Conclusioni 4](#_Toc108188098)

Indice delle figure

[Figura 1:CountPlot su feature "sex” 4](#_Toc108190095)

[Figura 3: Matrice di correlazione 5](#_Toc108190096)

[Figura 4: Confronto RandomForestClassifier e ExtraTreesClassifier 6](#_Toc108190097)

**Heart Diseases**

# 

# Introduzione

Lo scopo di questo progetto è quello di predire attraverso tecniche di Machine Learning quali individui sono più esposti e soggetti ad avere delle malattie cardiache. L’attenzione per le malattie cardiache è data dal fatto che sono una delle principali cause di morte per gli uomini. Ad esempio, negli Stati Uniti un quarto dei decessi è riconducibile a qualche tipo di malattia cardiaca. Il datasets che è utilizzato per lo studio delle malattie cardiche è “Heart Disease Dataset (Comprehensive)” di IEEE Dataport ([Heart Disease Dataset (Comprehensive) | IEEE DataPort (ieee-dataport.org)](https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive)). Questo dataset sulle malattie cardiache è ottenuto combinando 5 dataset popolari sulle malattie cardiache già disponibili in modo indipendente ma non combinati prima aiutare con l’obiettivo di far progredire la ricerca sull'apprendimento automatico relativo al CAD (Coronary artery disease) e sugli algoritmi di data mining e per far progredire la diagnosi clinica. I cinque set di dati utilizzati per la sua curatela sono:

1. Cleveland
2. Hungarian
3. Switzerland
4. Long Beach VA
5. Statlog (Heart) Data Set.

Il dataset contiene 12 attributi e 1190 istanze. I principali attributi presenti all’interno di questo dataset che contribuiscono ad un’analisi di questo tipo di malattia sono:

* ***age***
* ***sex***
* ***chest pain type***
* ***resting blood pressure***
* ***serum cholesterol***
* ***fasting blood sugar***
* ***resting electrocardiogram results***
* ***maximum heart rate achieved***
* ***exercise induced angina***
* ***oldpeak =ST***
* ***the slope of the peak exercise ST segment***
* ***class (target)***

Una descrizione più esaustiva degli attributi è reperibile presso il sito di IEEE Dataport.

## Tecnologie Utilizzate

Il progetto è sviluppato in linguaggio Python, tale scelta dipende principalmente dalle librerie che il linguaggio offre per il Machine Learning. Tra queste librerie sono state utilizzate:

* Pandas: per la gestione e la visualizzazione del dataset
* NumPy: per operare più facilmente su matrici e array di grandi dimensioni
* Scikit-learn: per gli algoritmi di Machine Learning
* MatPlotLib e Seaborn: per la visualizzazione grafica dei dati

# Data Analysis e Data Visualization

Prima di analizzare il dataset è necessario suddividere le 12 colonne tra features e target. Le prime 11 colonne fanno riferimento alle features messe a disposizione nel dataset, mentre l’ultima è relativa al target (etichetta classe di appartenenza).

## Data Visualization

Nella fase di Data Visualization sono state rappresentate le istanze del dataset sulla base delle diverse features disponibili. Questa fase è utile per vedere quali sono i casi relativi al Heart Diseases al variare delle features. I vari tipi di grafico che sono stati utilizzati per visualizzare i dati sono: Countplot, Displot e Boxplot. Ciascuna di queste tipologie di grafico consente di mettere in risalto diversi aspetti di ciascuna features. Ad esempio, il Countplot per la feature “sex” permette di capire quanti uomini rispetto alle donne soffrono di malattie cardiache:

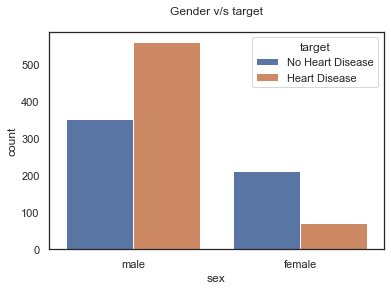
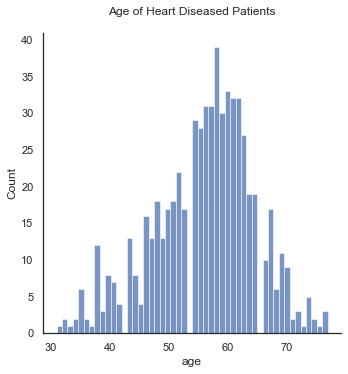
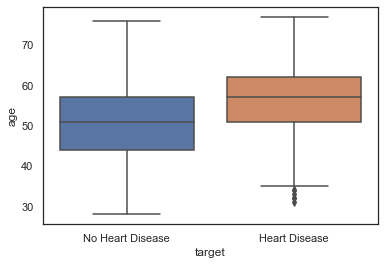


Figura :CountPlot su feature "sex”

Ad esempio, il Displot per la feature “age” consente di visualizzare quali sono le fasce di età che sono più soggette a malattie cardiache:



Ad esempio, il Boxplot per la feature “age” consente di visualizzare quale è il range di età degli individui che rischiano di avere malattie cardiache:



## Matrice di correlazione

Il primo passo effettuato nell’ambito della Data Analysis è la creazione della matrice di correlazione per individuare quali features sono più o meno correlate tra di loro. I valori all’interno della matrice di correlazione possono variare da -1 a +1. Come possiamo vedere nella matrice ottenuta, le features che sono più correlate con il target sono: ***chest pain type, exercise angina, oldpeak, ST slope, max hearth rate.***

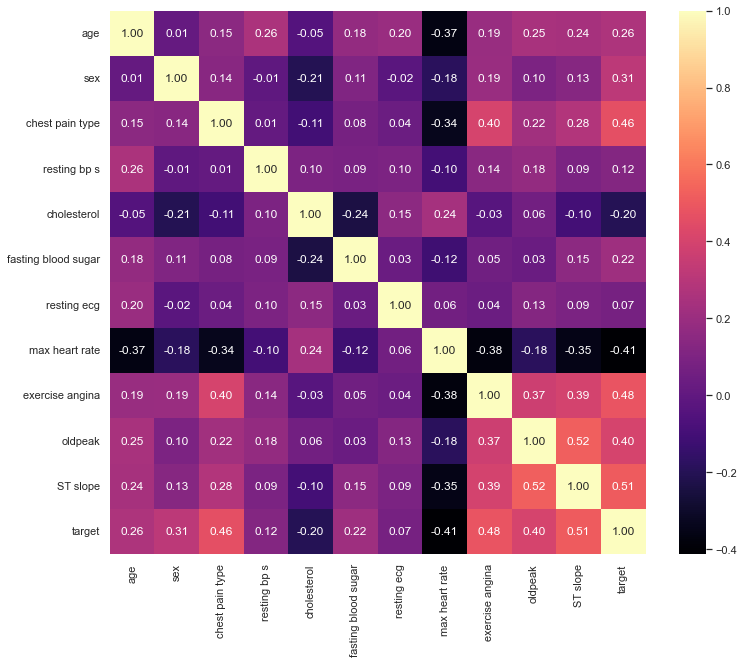
******

Figura : Matrice di correlazione

## 

## Features Importances con RandomForestClassifier e ExtraTreesClassifier

Per misurare l’importanza di ogni features sono stati utilizzati due classificatori: RandomForestClassifier e ExtraTreesClassifier. È difficile dire in anticipo se un RandomForestClassifier funzionerà meglio o peggio di un ExtraTreesClassifier. Generalmente, l'unico modo per sapere è provare entrambi e confrontarli. RandomForestClassifier è il più conveniente e il più ottimizzato tra i Decision Trees. L’ExtraTreesClassifier è molto più veloce per l’addestramento rispetto al Random Forest e utilizza le stesse API. Questi Random Forests permettono di misurare features importances in maniera semplice. Scikit-Learn misura l'importanza di una features osservando quanto i nodi dell'albero che utilizzano quella features riducono l’errore di classificazione.

Utilizzando le due diverse tecniche sul dataset in questione sono stati ottenuti i seguenti risultati:

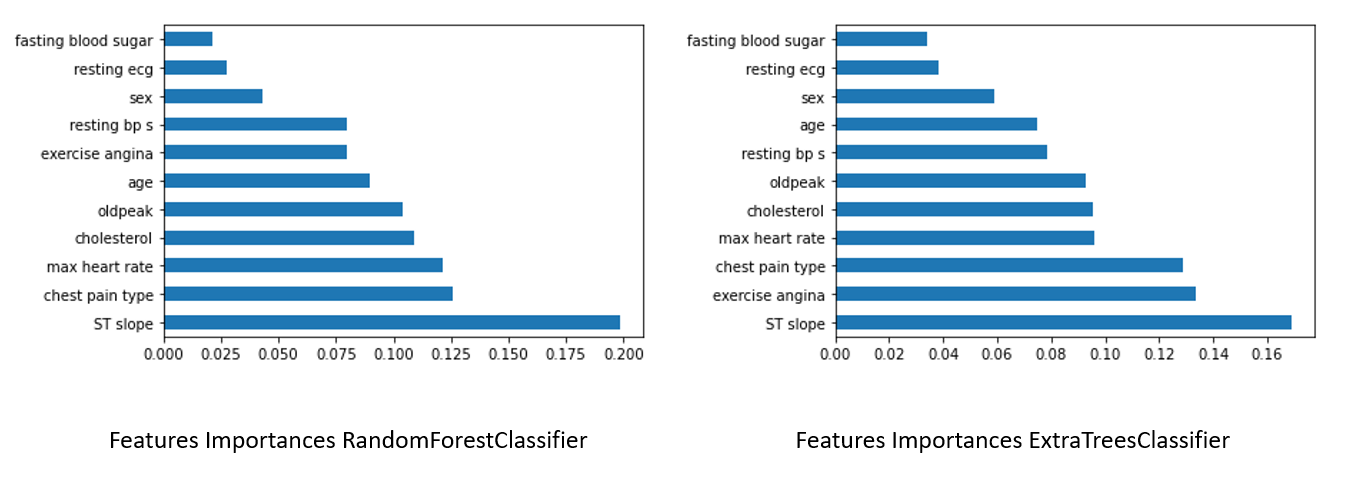


Figura : Confronto RandomForestClassifier e ExtraTreesClassifier

In conclusione, si può vedere che con tutti i metodi utilizzati in questa fase di analisi dei dati le features più discriminanti ai fini della classificazione sono le stesse ma con leggere differenze (ad esempio, la features exercise angina in RandomForestClassifier è molto meno importante rispetto a ExtraTreeClassifier)

# Classificazione

# Conclusioni