**

*Università degli Studi di* ***S****alerno*

*DIPARTIMENTO DI INFORMATICA*



Progetto di Fondamenti di Intelligenza Artificiale

Anno Accademico *2023/2024*

Link Repository

<https://github.com/OddlyHod/HTH>

Partecipanti:

Amendola Alfredo   
*Di Tella Nazaro*  
Fazio Alessandro  
Xu Xin Yu

***Sommario***

[*Capitolo 1* 4](#_Toc158244024)

[1.1 Definizione del Contesto 4](#_Toc158244025)

[1.2 Obiettivi 4](#_Toc158244026)

[1.3 Contesto Applicativo 5](#_Toc158244027)

[*Capitolo 2* 6](#_Toc158244028)

[2.1 CRISP – DM 6](#_Toc158244029)

[2.2 Specifiche P.E.A.S. 7](#_Toc158244030)

[2.2.1 Performance 7](#_Toc158244031)

[2.2.2 Environment 7](#_Toc158244032)

[2.2.3 Actuators 7](#_Toc158244033)

[2.2.4 Sensors 7](#_Toc158244034)

[2.3 Business Understanding 8](#_Toc158244035)

[2.4 Data Understanding 9](#_Toc158244036)

[2.4.1 Acquisizione dei Dati 9](#_Toc158244037)

[2.4.2 Analisi dei Dati 9](#_Toc158244038)

[2.4.3 Esplorazione dei Dati 10](#_Toc158244039)

[Grafico a Griglia 10](#_Toc158244040)

[Grafico Duplicati 11](#_Toc158244041)

[Grafico Correlazione Valori 12](#_Toc158244042)

[Grafico Distribuzione Totale 13](#_Toc158244043)

[Distribuzione dei dati 14](#_Toc158244044)

[Features Categoriche 14](#_Toc158244045)

[Features Numeriche 15](#_Toc158244046)

[Distribuzione dei Dati rispetto Variabile Target 16](#_Toc158244047)

[Features Categoriche 16](#_Toc158244048)

[Features Numeriche 16](#_Toc158244049)

[2.4.4 Qualità dei Dati 17](#_Toc158244050)

[*Capitolo 3* 18](#_Toc158244051)

[3.1 Data Preparation 18](#_Toc158244052)

[3.1.1 Data Cleaning 18](#_Toc158244053)

[3.1.2 Feature Scaling 18](#_Toc158244054)

[3.2 Feature Selection 19](#_Toc158244055)

[3.2.1 Features Categoriche 19](#_Toc158244056)

[3.2.2 Features Numeriche 19](#_Toc158244057)

[3.3 Data Balancing 20](#_Toc158244058)

[*Capitolo 4* 21](#_Toc158244059)

# *Capitolo 1*

Introduzione al Contesto

## Definizione del Contesto

L’insufficienza cardiaca o scompenso cardiaco è una condizione per cui il cuore non riesce a pompare sangue in quantità sufficiente da soddisfare le esigenze dell’organismo. L’insufficienza cardiaca non si manifesta all’improvviso ma si sviluppa lentamente, spesso nell’arco di anni. L’insufficienza cardiaca è una patologia molto diffusa: colpisce infatti circa 14 milioni di europei.   
In Italia, lo scompenso riguarda il 2% della popolazione, circa 1.200.000 di pazienti con una crescita media del 2,3% nei prossimi 10 anni.

Sia l’Insufficienza cardiaca acuta che quello cronica sono associate ad una elevata mortalità e al rischio di andare incontro a frequenti ospedalizzazioni ed inoltre ha un effetto negativo sulla qualità della vita.

L’insufficienza cardiaca si accompagna a *sintomi* caratteristici:

* Dispnea (mancanza di fiato);
* Ortopnea (difficoltà a respirare quando si è distesi);
* Tosse frequente;
* Gonfiore (edema) di piedi, caviglie e gambe;
* Debolezza generale, affaticamento o stanchezza;
* Perdita di appetito;
* Senso di ripienezza o tensione addominale.

## Obiettivi

L’obiettivo che il progetto HTH si pone è quello di ridurre al minimo l’errore umano creando e sviluppando un modello di intelligenza artificiale per predire uno scompenso cardiaco.

Tramite lo sviluppo di questo modello si prova ad automatizzare la diagnosi di uno scompenso cardiaco lasciando al medico più tempo per concentrarsi sul trattamento.

Il sistema proposto utilizzerà una varia gamma di attributi numerici e categorici, variabili che spaziano dall’età ai valori del colesterolo spaziando per il tipo di dolore che si accusa al petto.  
Utilizzando un set di variabili più eterogenee si riduce la probabilità di un falso positivo/negativo.

Uno dei focus, se non ***il*** focus, di questo progetto sarà la fase di validazione del modello, verranno valutate affidabilità e precisione e questi valori verranno messi a confronto con strumenti di diagnosi tradizionali.

Introduzione al Contesto

## Contesto Applicativo

Questo progetto non solo si pone come scopo la creazione di un modello fatto e finito per il contesto clinico, ma vede, in una sua integrazione ed implementazione futura per il possibile utilizzo diretto da parte del pubblico, integrandolo all’interno di un applicativo web.

Questo potrebbe portare a vantaggi esponenziali, diagnosticando una patologia precocemente, velocemente e soprattutto remotamente.

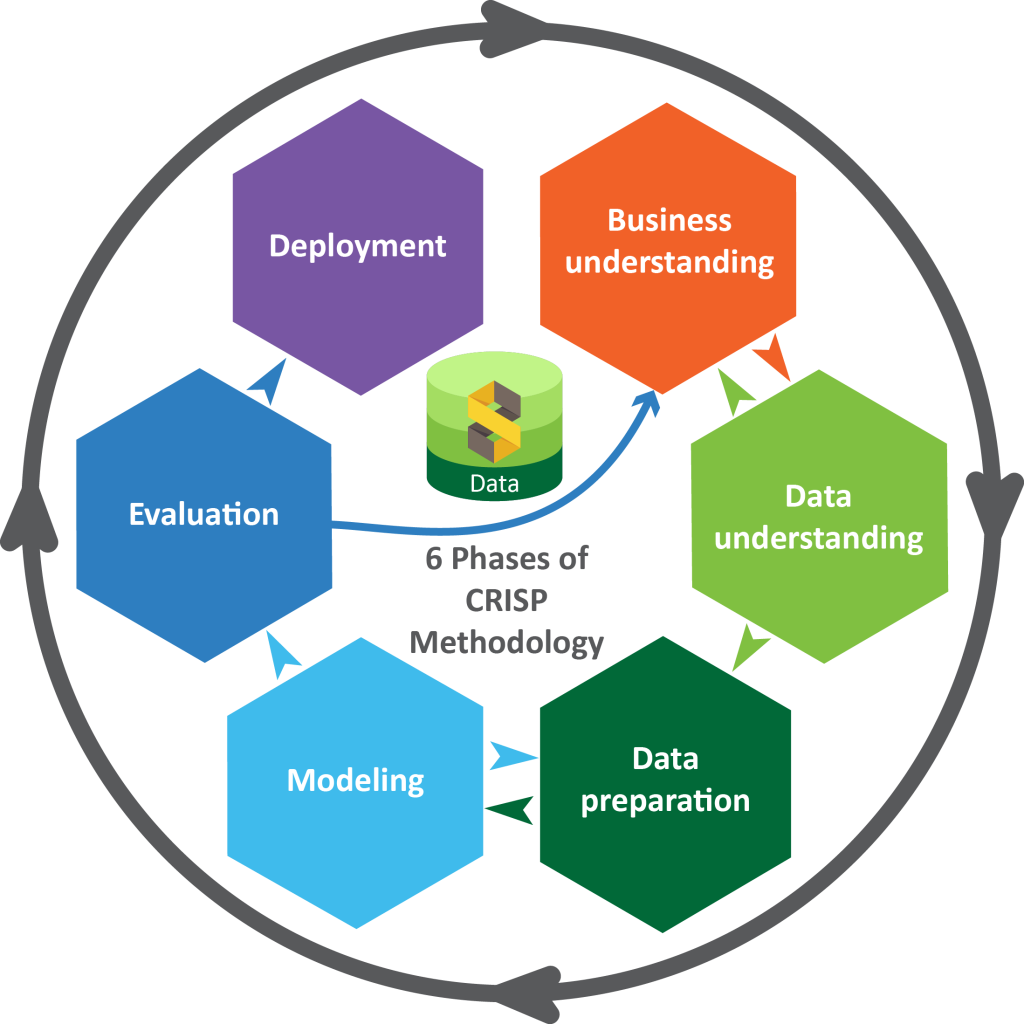
# *Capitolo 2*

Analisi di Business e dei Dati

## 2.1 CRISP – DM

CRISP-DM è l'acronimo di Cross-Industry Standard Process for Data Mining, è un process model che mette a disposizione un approccio strutturato a progetti di data mining.

Il modello si compone di 6 diverse fasi, ognuna specializzata per un ambito e tutte le fasi possono essere eseguite in maniera scollegata, ovvero, è possibile seguire un determinato percorso tra fasi ma è anche possibile effettuare il backtracking e tornare ad una fase precedente.



Analisi di Business e dei Dati

## 2.2 Specifiche P.E.A.S.

PEAS è l’acronimo in Inglese di Performance Environment Actuators Sensors. È utilizzato per raggruppare in un unico termine le caratteristiche dell’ambiente operativo.

### 2.2.1 Performance

Misura di prestazione adottata per valutare l’operato del modello.  
 Nel nostro caso le misure di prestazione sono i valori di ***precision***, ***accuracy***, ***recall*** ed ***f1-score***.

### 2.2.2 Environment

L’ambiente in cui opera il modello.  
 Nel nostro caso, il modello opera in un contesto clinico ovvero nell’insieme di tutti gli EHS1, le cartelle elettroniche dei pazienti.

Le caratteristiche dell’Environment sono:

* Completamente Osservabile (Poiché conosco tutte le informazioni riguardo all’EHS)
* Stocastico (Poiché lo stato successivo è influenzato da quelli precedenti)
* Episodico (Poiché ogni previsione è a sé stante)
* Discreto (Poiché il risultato è o affermativo o negativo)
* Singolo (Poiché il modello non è multi-agente)

### 2.2.3 Actuators

Gli attuatori disponibili all’agente per intraprendere le azioni.  
Nel nostro caso, i risultati della valutazione.

### 2.2.4 Sensors

I sensori attraverso i quali l’ambiente riceve gli input percettivi.  
Che nel nostro caso sono i valori predittivi del modello, ovvero i valori sui quali il modello effettuerà le sue predizioni.

Analisi di Business e dei Dati

## 2.3 Business Understanding

La fase iniziale del CRISP – DM è fondamentale per la raccolta dei requisiti e la definizione degli obiettivi di business che si intende raggiungere.

La fase di business prevede la definizione dei business success criteria, ovvero i criteri secondo i quelli potremo accertare che il sistema è costruito in linea agli obiettivi di business.

In questa fase vengono anche selezionate le tecnologie ed i tool necessari al raggiungimento dei business success criteria.

* ***Obiettivo di Business***

L’obiettivo di business è quello di stimare se un paziente, a partire dai suoi dati clinici, è o non è affetto da scompenso cardiaco

* ***Risorse***

Per creare ed addestrare il nostro modello abbiamo bisogno di un dataset, che prenderemo dal sito Keggle2, che mette disposizione vari dataset. Nel nostro caso utilizzeremo un dataset che mette tratta cartelle cliniche di oltre mille pazienti. (Per il trattamento dei dati cliccare qui)

* ***Rischi***

Uno dei rischi principali è la poca accuratezza del modello che sarebbe causata da un dataset che presenta poca eterogeneità, nel nostro caso questo problema verrà analizzato ed eventualmente trattato nelle fasi successive.

* ***Tecnologie***

Per l’analisi, la modellazione, l’addestramento e la visualizzazione grafica dei dati e del modello verranno utilizzate varie tecnologie, come ad esempio *Python* in concomitanza di varie librerie, come pandas3, numpy4, pyplot5 e seaborn6 per le informazioni sui dati ed sklearn6 per la fase di *feature engineering* e la fase di *modeling*.

Analisi di Business e dei Dati

## 2.4 Data Understanding

La seconda fase del CRISP - DM consiste nell’identificazione, collezione ed analisi dei dataset. Innanzitutto, quindi, vengono *acquisiti i dati* necessari al raggiungimento degli obiettivi di business e tecnici. I dati verranno poi caricati in un tool di *analisi dei dati*, quindi documentati ed esaminati.   
Successivamente si passa alla fase di *esplorazione dei dati*, durante la quale vengono visualizzati ed infine la fase di *qualità dei dati*, dove vengono identificati eventuali problemi di qualità (come ad esempio dati mancanti).

### 2.4.1 Acquisizione dei Dati

Il dataset per l’addestramento (e la valutazione) del modello è stato reperito in formato csv da keggle. Tutti i dati sono cartelle elettroniche di pazienti (EHS) e quindi contengono i risultati delle analisi effettuate per paziente (ECG ed analisi del sangue).

### 2.4.2 Analisi dei Dati

Il dataset, presenta circa 1000 cartelle cliniche e quindi 1000 pazienti e le feature per ogni paziente sono le seguenti:

* Age: Età del paziente; [Il numero di anni]
* Sex: Sesso del paziente; [M: Male, F: Female]
* ChestPainType: Tipo di dolore al petto; [TA, ATA, NAP, ASY]7
* RestingBP: Pressione sanguigna a riposo; [in mm/Hg]
* Cholesterol: Valore del colesterolo; [in mm/dl]
* FastingBS: Livello di zucchero nel sangue a digiuno; [1 se è maggiore di 120, 0 altrimenti]
* RestingECG: Valori dell’ECG a riposo; [Normal, ST, LVH]8
* MaxHR: Il valore massimo della freq. Cardiaca; [Valore tra 60 e 202]
* ExerciseAngina: Angina indotta da esercizi; [Y: Sì, N: No]
* Oldpeak: Sottolivellamento del tratto ST; [Valore Numerico]
* ST\_Slope: Pendenza dal picco dell’ST; [Up: upsloping, Flat: flat, Down: downsloping]
* HeartDisease: Classe di Output; [1: Scompenso, 0: Normale]

Note:

7: Differenti tipi di dolore al petto Typical Angina, Atypical Angina, Non – Anginal Pain e Asymptomatic.

8: Differenti valori dell’ECG a riposo ST (anormalità onda ST-T) e LVH (probabile o sicura ipertrofia ventricolare).

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Sex | CPT | RestingBP | Cholesterol | FastingBS | RestECG | MaxHR | ExcAng | OldPeak | OPS | HD |
| 62 | F | TA | 160 | 193 | 0 | Normal | 116 | N | 0 | Up | 0 |
| … | … | … | … | … | … | … | … | … | … | … | … |

Un esempio di paziente in forma tabellare

Analisi di Business e dei Dati

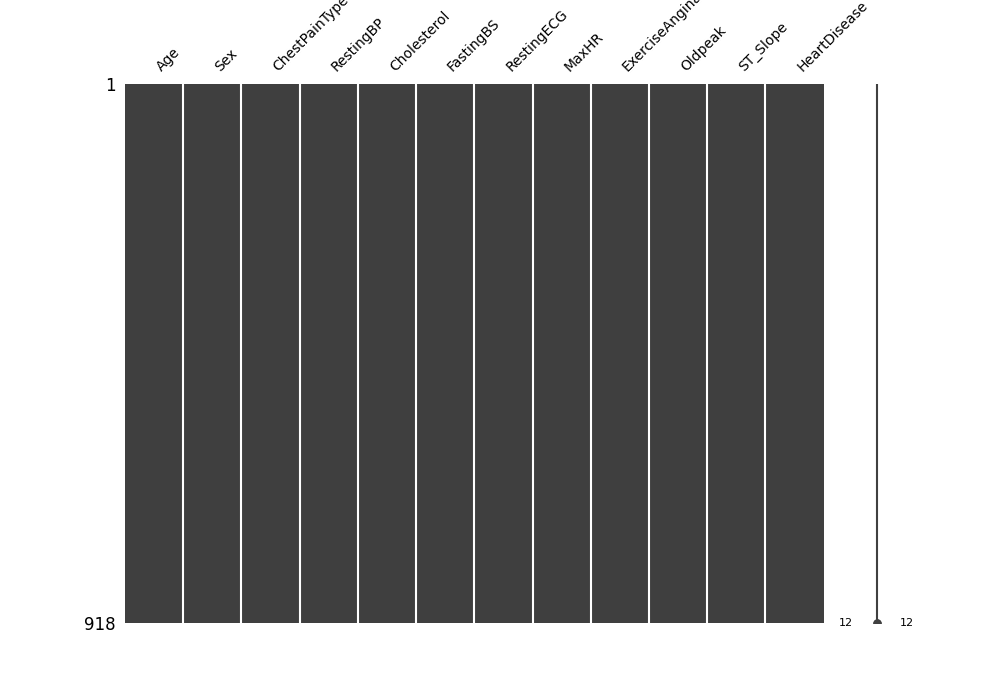
### 2.4.3 Esplorazione dei Dati

In questa fase, verrà effettuata la visualizzazione dei dati per trarre conclusioni sulla completezza, sulla distribuzione e sulla correlazione dei dati.

In primis, ci conviene esplorare il dataset alla ricerca di possibili valori nulli.   
(Per la fase di esplorazione ci avvarremo dell’utilizzo di Python)

#### Grafico a Griglia

Con questo grafico a griglia, ci viene evidenziata la presenza o meno di valori nulli.

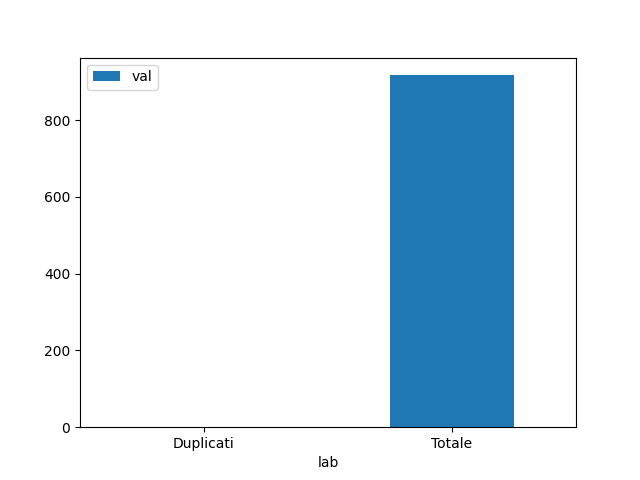


Nel nostro caso non ci sono valori nulli.

Analisi di Business e dei Dati

#### Grafico Duplicati

Da una semplice analisi utilizzando la funzione duplicate() di pandas, ci accorgiamo che:

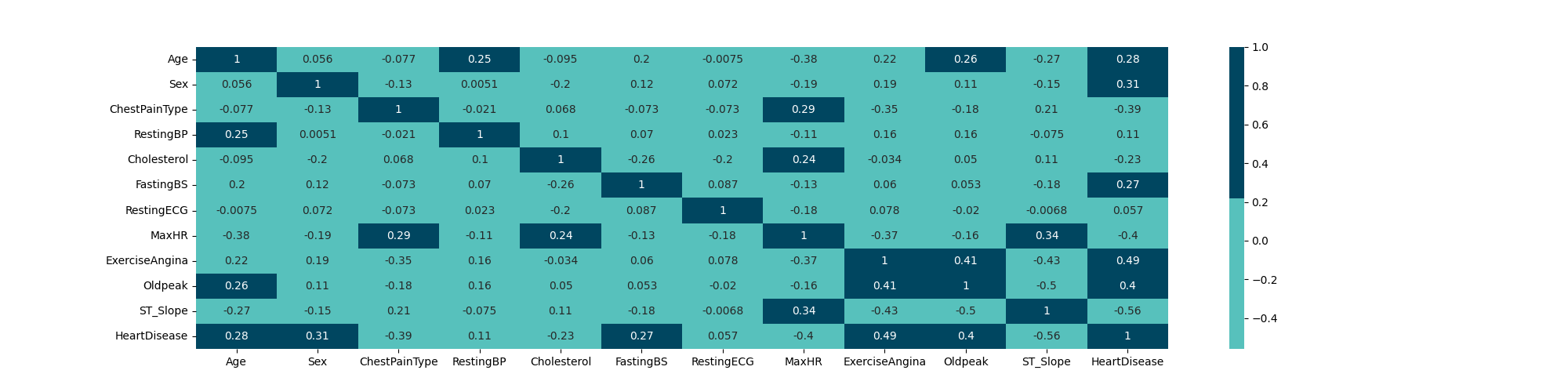


Non sono presenti duplicati.

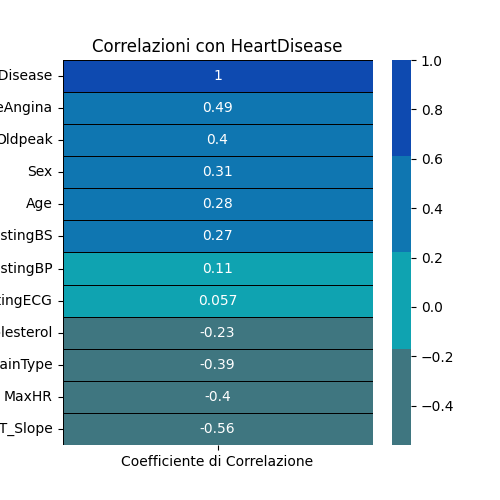
Analisi di Business e dei Dati

#### Grafico Correlazione Valori

Con questa serie di grafici analizzeremo le correlazioni, se presenti, tra Features e Variabile di Target:



Di questa matrice, selezioniamo solo i valori che sono in correlazione che HeartDisease:

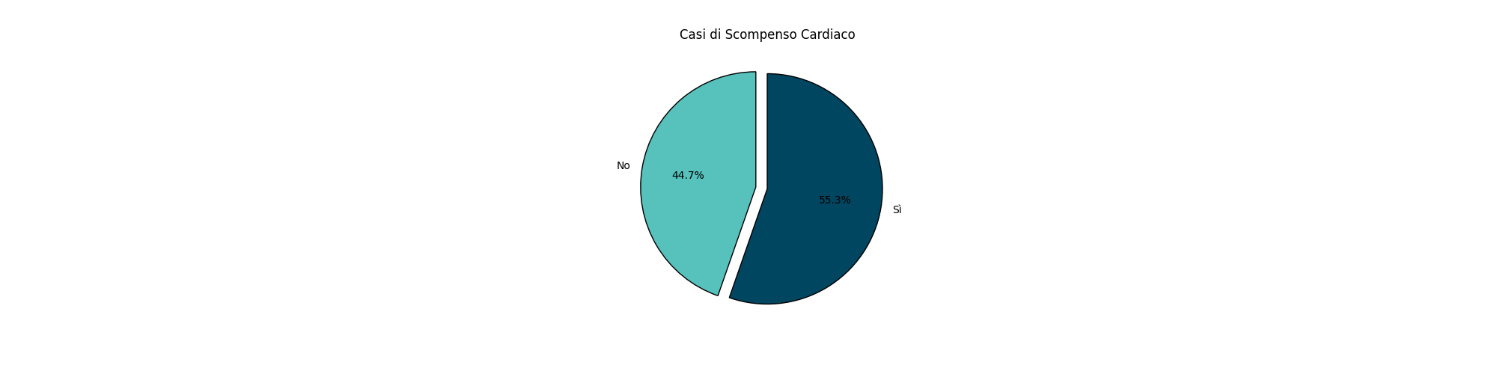


Si nota quindi che la maggior parte delle variabili di feature sono in correlazione con la variabile di target, in particolare, ExcerciseAngina.

Analisi di Business e dei Dati

#### Grafico Distribuzione Totale

Con questo grafico a torta, andremo a valutare se il dataset è bilanciato o meno.

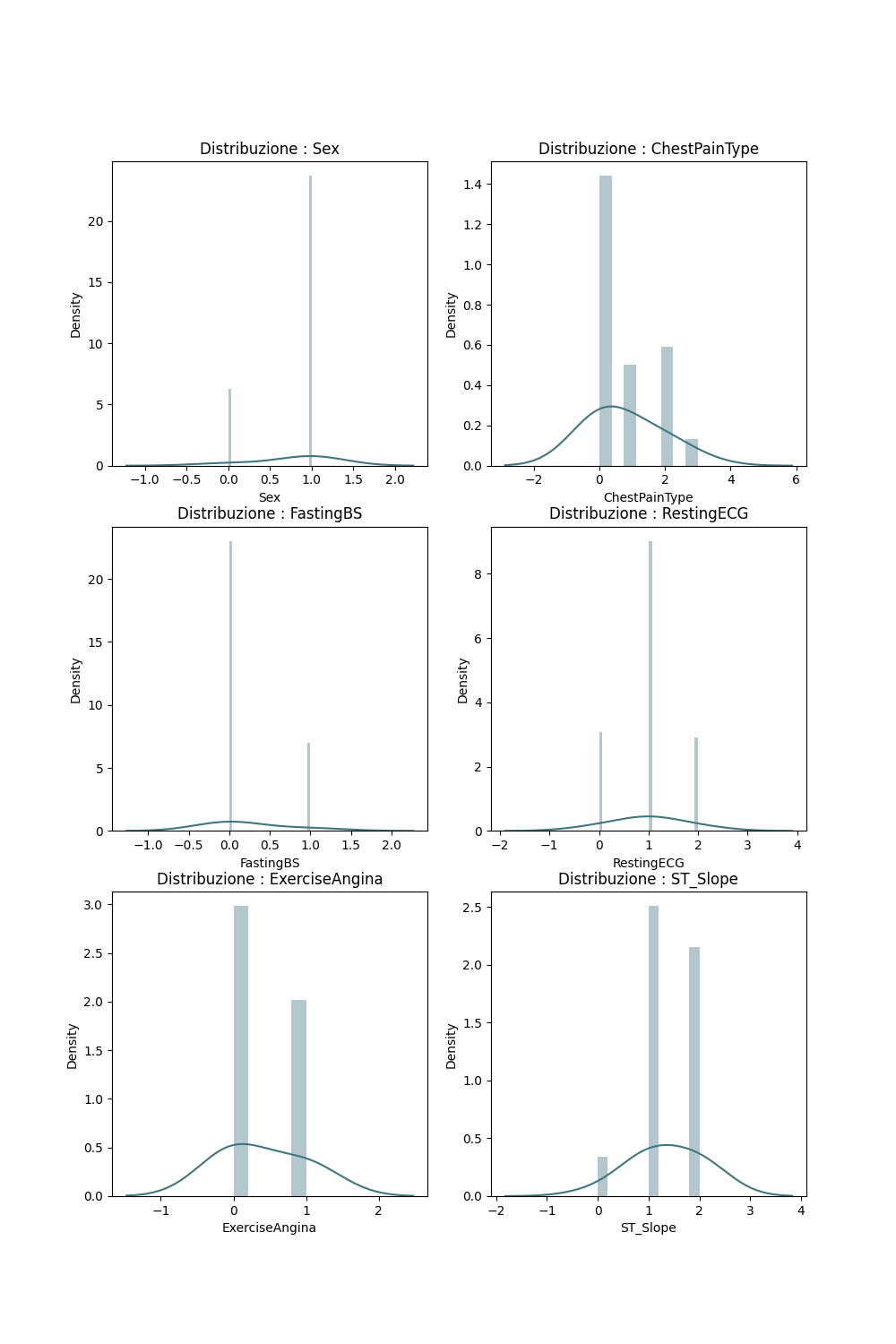


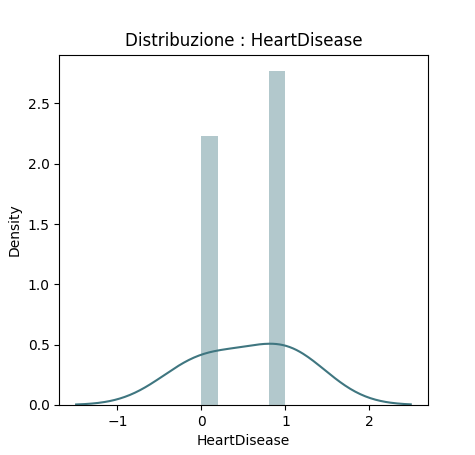
Da come si può notare, i casi si bilanciano, quindi il dataset è bilanciato.

#### Distribuzione dei dati

Adesso andremo a visualizzare le varie distribuzioni dei dati, in modo tale da visualizzare eventuali scompensi in termini di valori, che dovranno essere poi corretti in fase di Data Balancing.

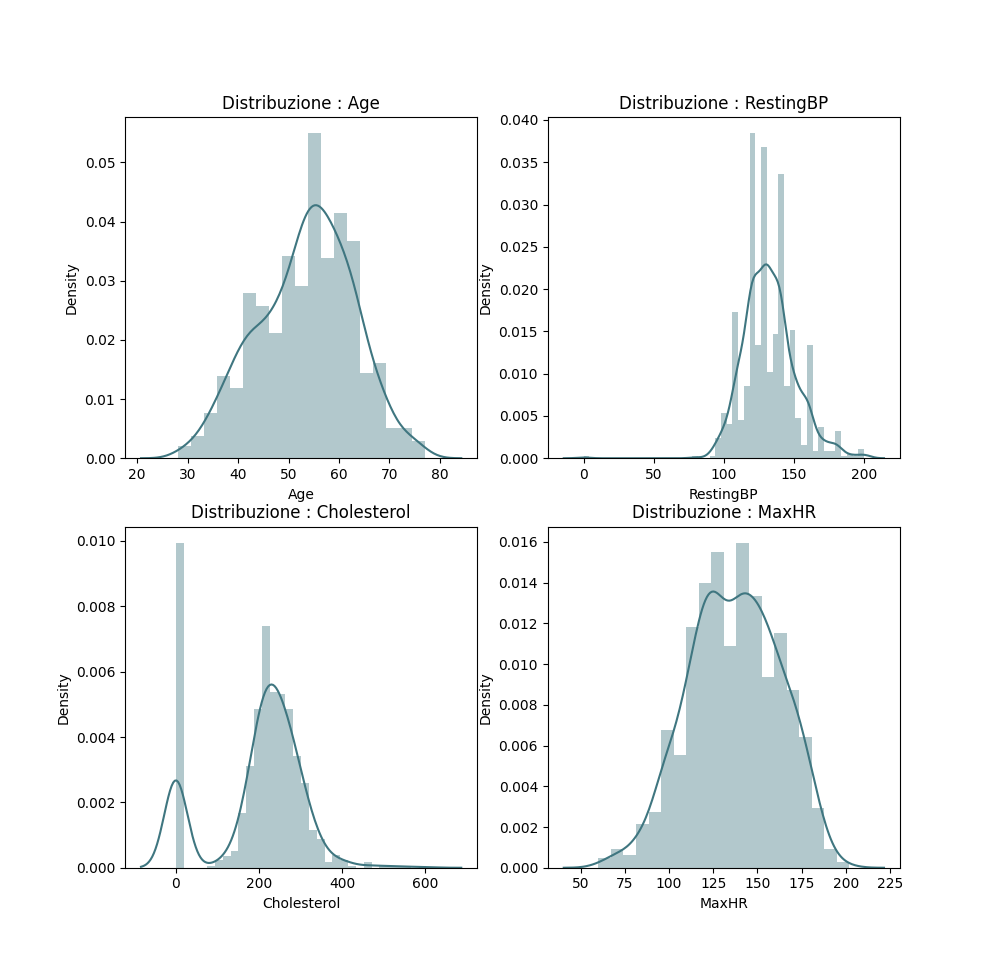
##### Features Categoriche

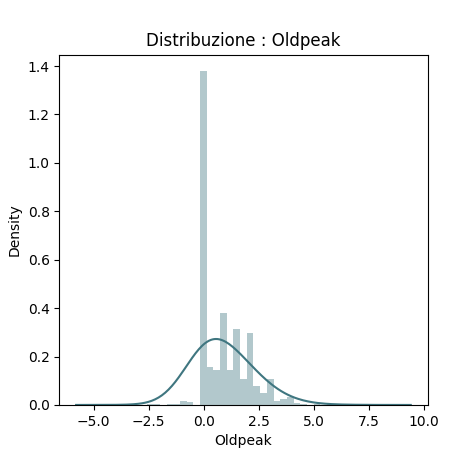




Tutte le features categoriche sono ***normalmente distribuite***.

##### Features Numeriche





Dai grafici, si nota che l’unico parametro a non avere una distribuzione normale è Oldpeak, che ha una distribuzione skewed, ovvero presenta una notevole asimmetria  
 centrale.

#### Distribuzione dei Dati rispetto Variabile Target

Infine, in questa fase viene messa in corrispondenza la correlazione tra features categoriche/numeriche con la variabile di target

##### Features Categoriche

Da questa serie di grafici è possibile notare che gli uomini hanno, probabilmente, una predisposizione agli scompensi cardiaci, che la maggioranza dei casi di scompenso cardiaco non comportano dolore al petto e che l’angina aumenta la probabilità di contrarre uno scompenso cardiaco.

##### Features Numeriche

Tutte le features numeriche sono distribuite normalmente rispetto ad HeartDisease.

### 2.4.4 Qualità dei Dati

Da come si è potuto notare dalle fasi precedenti, non vi sono valori nulli, né scompensi del dataset, l’unico problema che sorge, è la scala dei dati, che dovranno essere standardizzati e normalizzati nella fase seguente, la fase di Data Preparation.

# *Capitolo 3*

Data Preparation – Preparazione dei Dati

## 3.1 Data Preparation

L’obiettivo di questa fase è quello di preparare i dati in modo tale da poter essere utilizzati nelle successive fasi del processo. In primis si includono i processi di pulizia dei dati, poi si selezionano le features che hanno più potere predittivo ed infine i dati vengono

### 3.1.1 Data Cleaning

Poiché il dataset non presenza istanze vuote e/o valori nulli, questa fase non è necessaria per il corretto procedimento del processo.

### 3.1.2 Feature Scaling

Per parlare di feature selection, si passa prima per la fase di feature scaling, ovvero, dato che un modello di predizione non può prendere come input valori non numerici e che un modello di predizione tratta i dati senza curarsi della loro unità, è necessaria una fase di Standardizzazione/Normalizzazione.

Viene utilizzata la standardizzazione, per valori che sono normalmente distribuiti, il contrario invece per la normalizzazione.

Nel nostro caso normalizzeremo solo ed esclusivamente *Oldpeak.*

Andando ad utilizzare sklearn come libreria per standardizzare (StandardScaler) e normalizzare (MinMaxScaler), il risultato è il seguente:

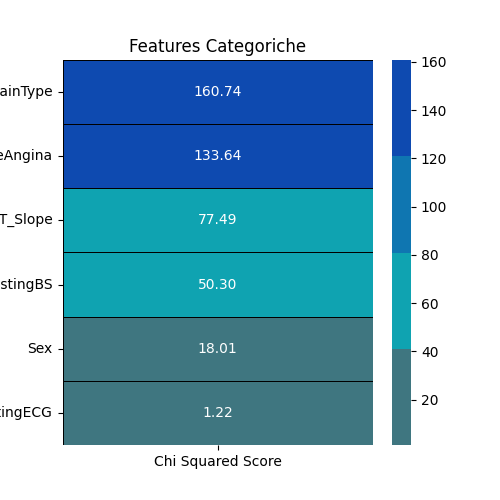
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Sex | CPT | RestingBP | Cholesterol | FastingBS | RestECG | MaxHR | ExcAng | OldPeak | OPS | HD |
| -1.43 | 1 | 1 | 0.41 | 0.83 | 0 | 1 | 1.38 | 0 | 0.30 | 2 | 0 |
| … | … | … | … | … | … | … | … | … | … | … | … |

Data Preparation – Preparazione dei Dati

## 3.2 Feature Selection

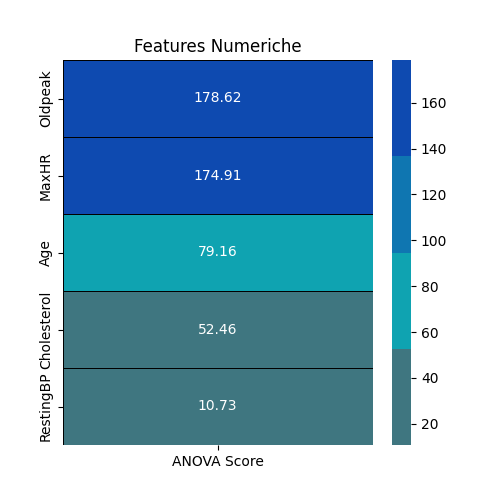
In questa fase utilizzeremo le matrici di correlazione precedentemente ricavate nella fase di Esplorazione dei Dati ed utilizzeremo anche il test del Chi Quadrato, che è un test statistico che serve per determinare se e quali features categoriche influenzano o meno la distribuzione dei dati. Mentre per quanto riguarda le features numeriche utilizzeremo l’ANOVA test, che è un altro test statistico.

### 3.2.1 Features Categoriche



Da come si può notare, tutte le features categoriche presentano un buon indice di indipendenza tranne per “RestingECG”, che quindi non verrà considerata per l’addestramento del modello.

### 3.2.2 Features Numeriche



Mentre nel caso delle features numeriche, l’unica che presenta un indice relativamente inferiore agli altri è “RestingBP” che quindi non verrà considerata per l’addestramento del modello.

Data Preparation – Preparazione dei Dati

## 3.3 Data Balancing

La fase di data balancing è necessaria quando si presentano scompensi di casi notevoli all’interno di un dataset, questo perché il modello di ML potrebbe lavorare peggio nella realtà se addestrato su dati prevalentemente di un determinato tipo.

A questo problema si risponde con l’undersampling o con l’oversampling:

* Undersampling: Rimuovere casualmente istanze all’interno della classe di maggioranza.
* Oversampling: Aggiungere copie di istanze all’interno della classe di minoranza.

Queste due tecniche però potrebbero presentare dei problemi, la prima potrebbe causare problemi di rimozione di istanze particolarmente importanti per l’apprendimento del modello; mentre la seconda potrebbe creare *overfitting* ovvero quando il modello risponde accuratamente per dati pregressi ma in modo errato per dati nuovi, questo perché ***ha imparato a “memoria”***.

Nel nostro caso però, avendo un dataset pressoché ***bilanciato*** (la differenza di casi è di un 5% trascurabile) non adopereremo nessuna delle due tecniche. In alternativa si potrebbe pensare di utilizzare l’undersampling poiché statisticamente non ci sono casi ***particolarmente*** rilevanti.

# *Capitolo 4*

Data Modeling – Modellazione dei Dati

Il prossimo step è quello di creare un modello per l’apprendimento

**3.1 Gaussian Naive Bayes**

L’ algoritmo Naive Bayes considera le caratteristiche di una nuova istanza, e calcola la probabilità che queste facciano parte di una classe tramite l’applicazione del teorema di Bayes, come suggerito dal nome questo è un algoritmo “ingenuo” perché assume che le varie caratteristiche non siano correlate tra loro, di conseguenza in fase di classificazione, l’algoritmo non andrà a considerare la potenziale utilità di una combinazione di più caratteristiche. L’ assunzione di indipendenza condizionale semplifica il calcolo delle probabilità condizionali. Sebbene questa assunzione possa non essere realistica in molti casi, Gaussian Naive Bayes è noto per la sua semplicità e velocità di addestramento e classificazione. L’algoritmo è ampiamente utilizzato in applicazioni come il riconoscimento del linguaggio naturale, la classificazione di documenti e altre situazioni in cui le caratteristiche di input possono essere approssimate da una distribuzione gaussiana. Tuttavia, è importante valutare attentamente le assunzioni dell’algoritmo e considerare se sono adatte al problema specifico prima di applicarlo.

**3.2 Decision Tree**

Un Decision Tree (albero decisionale) è un algoritmo di apprendimento automatico ampiamente utilizzato per problemi di classificazione e regressione. Prende il nome dalla sua struttura a forma di albero, composta da nodi interni che rappresentano decisioni o attributi e foglie che rappresentano le classi o i valori di output. Un Decision Tree `e quello di suddivide il dataset in modo ricorsivo in base agli attributi con rilevanza maggiore, al fine di ottenere una classificazione o una stima precisa. Inizialmente, il Decision Tree seleziona un attributo che meglio separa le istanze in base alla loro classe o valore di output. Successivamente, l’albero si ramifica in base alle possibili combinazioni di valori dell’ attributo selezionato, creando sotto-alberi corrispondenti ai rami. L’algoritmo continua iterativamente la divisione del dataset, fin quando o le classi sono ben definite, oppure si raggiunge un’altezza massima dell’albero, stabilita precedentemente. Tra i vantaggi di questo algoritmo ci sono: l’ interpretabilità, in quanto il processo decisionale, rappresentato dal percorso dell’albero è di facile comprensione; la possibilità di operare sia su dati numerici che categorici e la possibilità di combinarlo con altre tecniche e di gestire dati mancanti o rumorosi. Mentre il difetto del Decison Tree è rappresentato dall’ overfitting, soprattutto se abbiamo un albero molto profondo e complesso, per far fronte a questo problema possiamo avvalerci di tecniche come la potatura o dell’utilizzo di parametri per regolare la complessità.

**3.3 RANDOM FOREST**

Random Forest è un algoritmo di apprendimento automatico basato suensemble, che combina l’output di più alberi decisionali per ottenere una previsione più accurata. è ampiamente utilizzato per problemi di classificazione e regressione. L’idea principale di Random Forest `e quella di creare un insieme (ensemble) di alberi decisionali, ognuno dei quali viene addestrato su un sottoinsieme casuale e indipendente del dataset di addestramento. Durante la fase di addestramento, per ogni albero viene selezionato casualmente un sottoinsieme di dati e un sottoinsieme di attributi. Ciò introduce una variazione e una diversità nell’addestramento dei singoli alberi. Durante la fase di previsione, ciascun albero produce una previsione e la classe (o il valore) che ottiene la maggioranza delle previsioni degli alberi viene considerata come l’output finale del modello di Random Forest. Random Forest offre diversi vantaggi, tra cui la capacità di gestire grandi dataset con molte caratteristiche, la robustezza agli outlier e la riduzione dell’overfitting rispetto a un singolo albero decisionale. Inoltre, può gestire sia dati numerici che categorici senza la necessità di normalizzazione o codifica dei dati. Un altro aspetto interessante di Random Forest è la possibilità di ottenere una stima dell’importanza delle diverse caratteristiche utilizzate per la previsione. Questa informazione può essere utile per identificare le caratteristiche più influenti e comprenderne il ruolo nel processo decisionale. In conclusione, Random Forest `e un algoritmo di ensemble che sfrutta la diversità degli alberi decisionali per ottenere previsioni più accurate e stabili. La sua flessibilità, scalabilità e capacità di gestire diversi tipi di dati lo rendono una scelta popolare in diversi contesti di apprendimento automatico