**

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

*DIPARTIMENTO DI INFORMATICA*



Progetto di Fondamenti di Intelligenza Artificiale

Link Repository

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

Partecipanti:

Amendola Alfredo   
Di Tella Nazaro  
Fazio Alessandro  
Xu Xin Yu

***Sommario***

[*Capitolo 1* 3](#_Toc157974224)

[1.1 Definizione del Contesto 3](#_Toc157974225)

[1.2 Obiettivi 3](#_Toc157974226)

[1.3 Contesto Applicativo 4](#_Toc157974227)

[*Capitolo 2* 5](#_Toc157974228)

[2.1 CRISP – DM 5](#_Toc157974229)

[2.2 Specifiche P.E.A.S. 6](#_Toc157974230)

[2.2.1 Performance 6](#_Toc157974231)

[2.2.2 Environment 6](#_Toc157974232)

[2.2.3 Actuators 6](#_Toc157974233)

[2.2.4 Sensors 6](#_Toc157974234)

[2.3 Business Understanding 7](#_Toc157974235)

[2.4 Data Understanding 8](#_Toc157974236)

[2.4.1 Acquisizione dei Dati 8](#_Toc157974237)

[2.4.2 Analisi dei Dati 8](#_Toc157974238)

[2.4.3 Esplorazione dei Dati 9](#_Toc157974239)

[Grafico a Griglia 9](#_Toc157974240)

[Grafico Duplicati 10](#_Toc157974241)

[Grafico Correlazione Valori 11](#_Toc157974242)

[Grafico Distribuzione Totale 12](#_Toc157974243)

[2.4.4 Qualità dei Dati 12](#_Toc157974244)

[*Capitolo 3* 13](#_Toc157974245)

# *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.

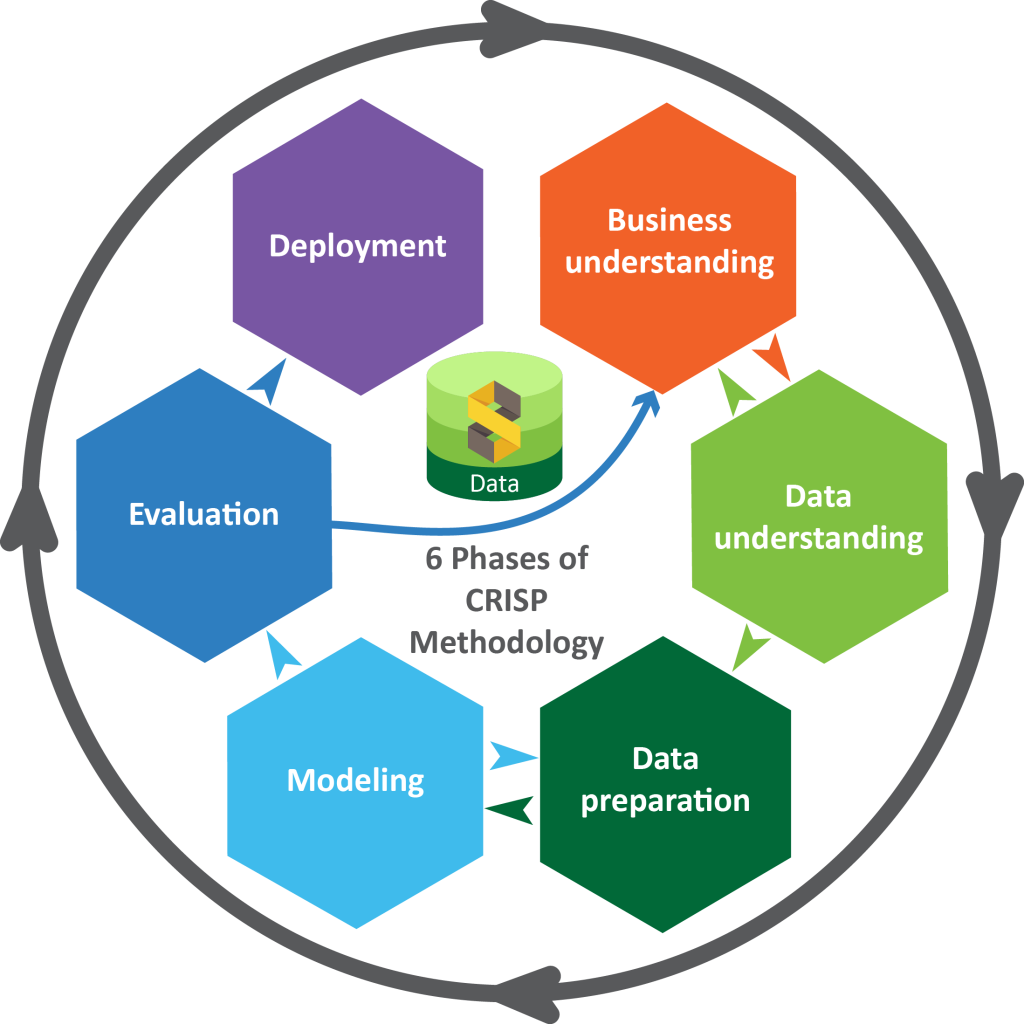
# *Capitolo 2*

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

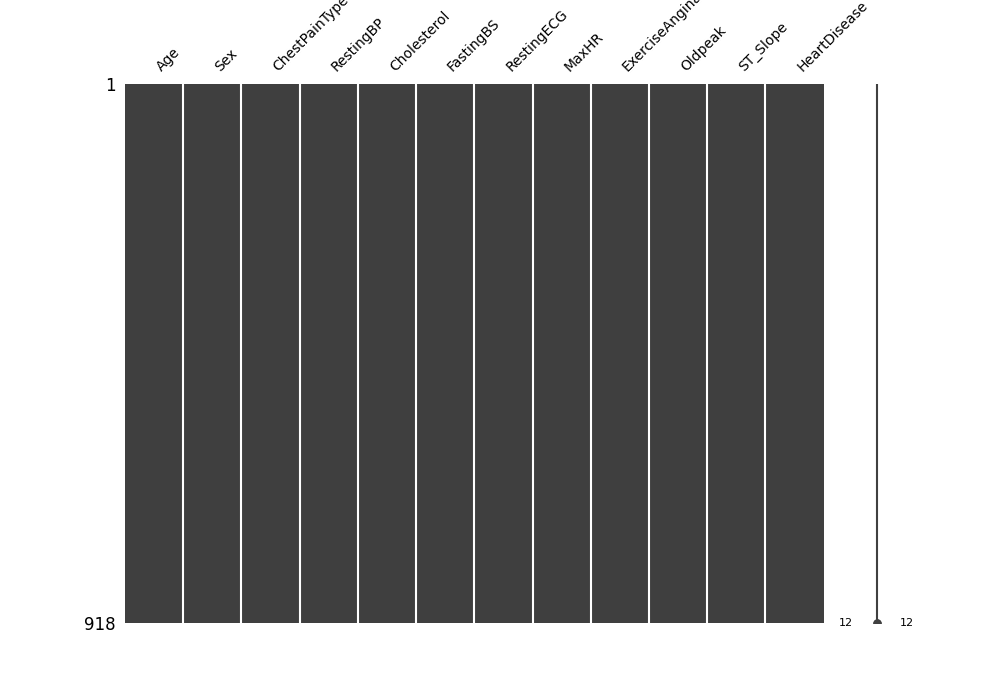
### 2.4.3 Esplorazione dei Dati

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.

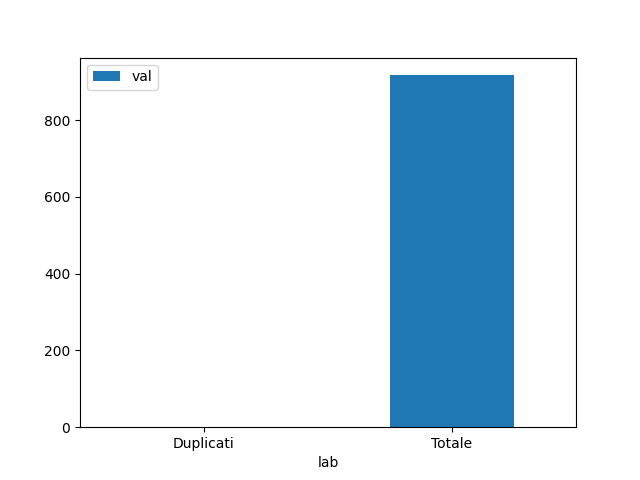


Nel nostro caso non ci sono valori nulli.

Analisi di Business e dei Dati

#### Grafico Duplicati

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

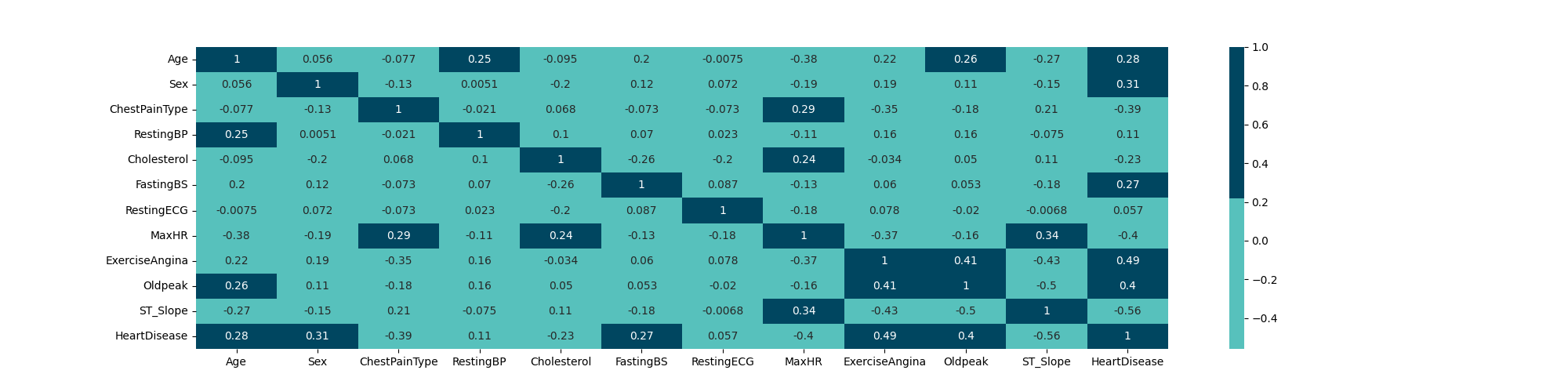


Non sono presenti duplicati.

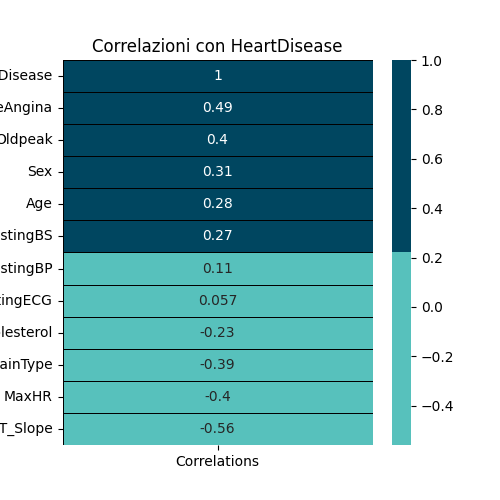
Analisi di Business e dei Dati

#### 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:

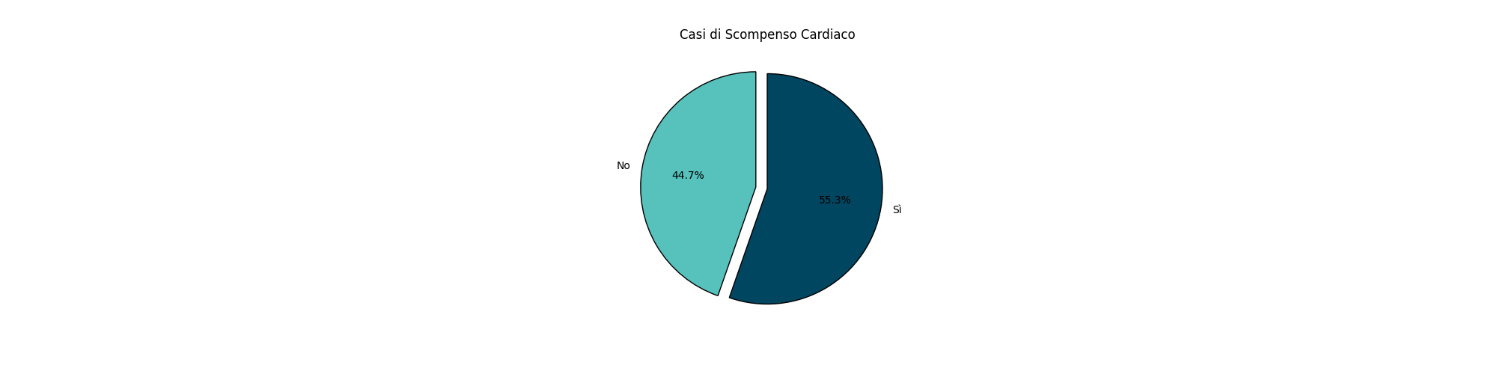


Si nota quindi che la maggior parte delle variabili di feature sono in correlazione con la variabile di target, in particolare, ExcerciseAngina, che quindi è la variabile con più potere predittivo.

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.

### 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

Nella fase di data preparation, si modellano i dati in modo da risolvere i problemi di qualità riscontrati nell’ analisi e mettere in evidenza i dati con maggiore rilevanza e quindi potere predittivo.

3.1 data cleaning

Data l’assenza di valori nulli, non è necessario eseguire operazioni di data cleaning.

3.2 feature selection

Durante l’analisi degli attributi abbiamo notato come ogni variabile di feature avesse una correlazione con la variabile di target, di conseguenza la decisione migliore è prendere in considerazione tutti i valori

3.3 feature scaling

Non è stato necessario applicare alcuna normalizzazione sui dati.

3.4 data balancing

Le tecniche di data balancing mirano a bilanciare le classi del dataset, quando una classe presenta un numero di esempi molto inferiore rispetto ad un’altra, per fare questo utiliziamo due tecniche:

-oversampling: attraverso questa tecnica si aumentano gli esempi nella classe minoritaria del dataset generando dati sintetici. Uno degli algoritmi di oversampling più utilizzati `e l’ADASYN (Adaptive Synthetic Sampling). ADASYN `e un metodo basato su campionamento sintetico che mira a generare nuovi esempi sintetici per la classe minoritaria in modo adattivo. L’algoritmo tiene conto della densità dei punti nel dataset e genera più esempi sintetici per le regioni meno dense, consentendo di affrontare meglio le aree più complesse e meno rappresentate della classe minoritaria. In questo modo, ADASYN cerca di bilanciare le distribuzioni delle classi nel dataset. Il rischio dell’ oversampling è quello di rendere la classe minoritaria troppo statica, aggiungendo esempi troppo simili, questo metterebbe in difficoltà gli algoritmi.

Undersampling: D’altra parte, l’undersampling `e una tecnica che punta a ridurre la presenza di esempi della classe maggioritaria nel dataset, eliminando alcuni dei dati di quella classe. RandomUnderSampler `e un metodo di undersampling comune ed efficace che opera selezionando casualmente un sottoinsieme dei dati della classe maggioritaria in modo che il numero di esempi delle classi sia bilanciato. Questo metodo `e semplice da implementare e pu`o essere efficace quando il dataset ha una dimensione sufficientemente grande. Il rischio in questo caso è quello di togliere troppi dati e ritrovarsi con classi troppo piccole.

In entrambi i casi, `e fondamentale valutare attentamente l’impatto delle tecniche di oversampling o undersampling sul modello di apprendimento automatico. E possibile utilizzare misure di valutazione, come l’accuratezza o ` la precisione per determinare se la tecnica di campionamento ha migliorato le prestazioni del modello rispetto alla situazione di partenza