**progetto machine learning a.a. 2022/2023**

heart desease prediction

**elisa catena 547007 valentina fontanarosa 547333**

Immagine che contiene invertebrato

Descrizione generata automaticamente

Sommario

[Contesto 3](#_Toc140601643)

[Dataset 3](#_Toc140601644)

[Dataset source 4](#_Toc140601645)

[Preparazione dei dati 5](#_Toc140601646)

[Data preprocessing 5](#_Toc140601647)

[Data visualization 5](#_Toc140601648)

[Ricerca outliers 6](#_Toc140601649)

[Train-Test Split 7](#_Toc140601650)

[Training del modello 7](#_Toc140601651)

[Training con LASSO regression 7](#_Toc140601652)

[Training con Ridge regression 8](#_Toc140601653)

[Training con Logistic regression 9](#_Toc140601654)

[Training con Decision Trees 10](#_Toc140601655)

[Conclusioni 12](#_Toc140601656)

# Contesto

Le malattie cardiovascolari sono una delle principali cause di mortalità nel mondo. La rilevazione precoce e l'intervento tempestivo sono fondamentali per prevenire le malattie cardiache e ridurre il rischio di infarto. Un approccio per identificare individui ad alto rischio di infarto è sviluppare modelli predittivi utilizzando dati clinici e demografici. In questo notebook, esploreremo un dataset contenente informazioni su pazienti con sospetta malattia cardiaca.

# Dataset

Riguardo il dataset, ci sono 14 attributi:

1. Età: età del paziente (in anni)
2. Sesso: genere
3. Dolore al petto: tipo di dolore al petto (valori da 1 a 4)
4. Pressione sanguigna a riposo: pressione sanguigna a riposo (in mmHg) all'ammissione in ospedale
5. Colesterolo sierico: colesterolo sierico in mg/dL
6. Glicemia a digiuno: zucchero nel sangue a digiuno > 120 mg/dL (probabile essere diabetico) 1 = vero, 0 = falso
7. Elettrocardiogramma a riposo: risultati dell'elettrocardiogramma a riposo

* Valore 0: normale
* Valore 1: anomalia dell'onda ST-T (inversioni dell'onda T e/o elevazione o depressione dell'onda ST di> 0.05 mV)
* Valore 2: probabile o definita ipertrofia ventricolare sinistra secondo i criteri di Estes

1. Massimo battito cardiaco: il numero più alto di battiti al minuto che il cuore può raggiungere durante l'esercizio fisico
2. Angina indotta dall'esercizio: (1 = sì; 0 = no)
3. Depressione ST indotta dall'esercizio: depressione ST indotta dall'esercizio rispetto al riposo (in mm, ottenuta sottraendo i punti più bassi del segmento ST durante l'esercizio e il riposo)
4. Pendenza: la pendenza del segmento ST all'esercizio fisico massimo, le anomalie ST-T sono considerate un indicatore cruciale per identificare la presenza di ischemia

* Valore 1: in salita
* Valore 2: piatto
* Valore 3: in discesa

1. Numero di vasi principali: numero di vasi principali (0-3) colorati con fluoroscopia.
2. Tallio: può provocare abbassamento della pressione arteriosa (ipotensione) e dei battiti del cuore (bradicardia), seguito da pressione alta (ipertensione) con battiti cardiaci veloci (tachicardia) e da aritmie e infarto
3. Heart desease: presenza o meno di malattia cardiaca

## Dataset source

<https://www.kaggle.com/code/miladistic/heart-disease-prediction-using-lasso-regression/input?select=Heart_Disease_Prediction.csv>

# Preparazione dei dati

## Data preprocessing

Per semplificare l’analisi dei dati, si mappa la colonna ‘Heart desease’ (i cui valori erano ‘presence’ e ‘absence’) in una nuova colonna ‘Heart desease codes’ con valori binari (1 = ‘presence’ e 0 = ‘absence’).

Inoltre, si verifica la presenza di valori mancanti o duplicati (che erano entrambi assenti).

## Data visualization

Verifichiamo se i dati sono sbilanciati:

Immagine che contiene testo, schermata, diagramma, Rettangolo

Descrizione generata automaticamente Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

Si può notare che i dati a disposizione sono piuttosto sbilanciati.

La matrice di correlazione è una tabella che mostra le correlazioni tra le diverse variabili del dataset. Questa matrice è di particolare interesse poiché fornisce una panoramica delle relazioni lineari tra le feature. La correlazione misura la forza e la direzione della relazione tra due variabili.

La matrice di correlazione dei dati è la seguente:

Immagine che contiene schermata, diagramma, modello, testo

Descrizione generata automaticamente

## Ricerca outliers

Gli outliers sono osservazioni che si discostano significativamente dal resto dei dati in un dataset. Queste osservazioni anomale possono influenzare negativamente la qualità dei risultati delle analisi; pertanto, è importante rilevare e gestire gli outliers durante l'esplorazione e la pulizia dei dati.

Per la ricerca degli outliers sono stati utilizzati i boxplot (grafici utilizzati per visualizzare la distribuzione di un insieme di dati numerici per avere un'idea della centralità, della dispersione e della presenza di outliers nei dati), in particolare uno per ciascuna feature del dataset.

Immagine che contiene diagramma, Piano, Rettangolo, schematico

Descrizione generata automaticamente

Dopo aver localizzato gli outliers (8 nel dataset in esame), sono stati eliminati per rendere il dataset più preciso.

## Train-Test Split

I dati vengono suddivisi in due sottoinsiemi: un training set (dati per addestrare il modello) e un test set (dati per testare il modello), di dimensioni rispettivamente dell’80% e del 20% dei dati iniziali.

# Training del modello

## Training con LASSO Regression

È stata utilizzata la funzione LASSO con cross validation (con fattore di cv pari a 5) per selezionare automaticamente il parametro di regolarizzazione migliore per il modello.

Le prestazioni sono state valutate al variare di una soglia, impostata per ottenere risultati binari.

Immagine che contiene linea, Diagramma, diagramma, testo

Descrizione generata automaticamente

## Training con Ridge Regression

È stata utilizzata la funzione Ridge con cross validation (con fattore di cv pari a 5) per selezionare automaticamente il parametro di regolarizzazione migliore per il modello.

Le prestazioni sono state valutate al variare di una soglia, impostata per ottenere risultati binari.

Immagine che contiene linea, Diagramma, diagramma, testo

Descrizione generata automaticamente

## Training con Logistic Regression

Per ottenere prestazioni migliori del modello, si applica una normalizzazione delle features in modo che abbiano una media di zero e una deviazione standard di uno.

Si addestra il modello con la Logistic regression e si ottiene la seguente matrice di confusione:

Immagine che contiene schermata, testo, quadrato, Rettangolo

Descrizione generata automaticamente

Questo significa che sono stati rilevati

* 0 falsi negativi
* 4 falsi positivi
* 24 veri positivi
* 25 veri negativi

Il report mostra le seguenti metriche:

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

* Precision: la percentuale di istanze predette come positive che sono corrette (True Positive / (True Positive + False Positive)).
* Recall: la percentuale di istanze positive che sono state correttamente predette (True Positive / (True Positive + False Negative)).
* F1-score: la media armonica tra precisione e recall (2 \* Precision \* Recall / (Precision + Recall)).
* Support: il numero di istanze di ciascuna classe nel set di test.

## Training con Decision Trees

I Decision Trees sono un tipo di modello di apprendimento supervisionato che prende decisioni sequenziali per arrivare a una previsione finale.

Ogni nodo rappresenta una caratteristica (feature) del dataset, e ogni ramo rappresenta una possibile decisione o una divisione basata su quella caratteristica. I nodi foglia rappresentano le classi o i valori di output finali.

Immagine che contiene diagramma, linea, design

Descrizione generata automaticamente

Si ottiene la seguente matrice di confusione:

Immagine che contiene testo, schermata, diagramma, Rettangolo

Descrizione generata automaticamente

Il report mostra le seguenti metriche:

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

# Conclusioni

In conclusione, si può stabilire che il modello più preciso è quello addestrato con Logistic Regression con uno score di 0.92, leggermente superiore rispetto allo score di Ridge Regression e LASSO Regression. Il modello che offre le prestazioni minori, invece, è quello addestrato con i Decision Trees.