ANALISI E VALUTAZIONE CORONARY ARTERY DISEASE : CASO STUDIO CON APPROFONDIMENTO

CARMELA PIA SENATORE
UNIVERSITA' DEGLI STUDI DI SALERNO
MATRICOLA:0522501721

CORSO DI ADVANCED STATISTICAL LEARNING II

Abstract

Questo studio condotto nell'ambito del corso di advanced statistical learning II, comprende un approccio multifase. L'obiettivo principale, rispetto al fenomeno di studio, è individuare un modello con le migliori metriche di accuratezza che permetta la previsione della malattia CAD; la motivazione risiede nella possibilità di indirizzare i pazienti dello studio verso la migliore delle soluzioni. L'angiografia coronarica è considerata il gold standard per la diagnosi della CAD. Tuttavia, l'angiografia è una procedura invasiva. La possibilità di rilevare , in maniera non invasiva, la malattia delle arterie coronarie nei pazienti ad alto rischio basandosi su algoritmi di apprendimento automatico potrebbe essere utile per ridurre il carico medico e la grande perdita di vite umane. Questo può avvenire non solo selezionando un modello che permette di comprendere la relazione tra fattori medici e biologici ma anche attraverso una scelta mirata.

Nella prima parte, è stata condotta un'analisi esplorativa delle variabili per individuare associazioni significative. Questo comprende ACM , feature Engeneering, Test del Chi quadro e EDA.

Nella seconda parte, sono stati sviluppati modelli di classificazione al fine di scegliere il modello , le cui performance, si adattano meglio al caso studio. Scelto il miglior modello, è stato utilizzato il metodo SHAP per l'interpretabilità dei fattori.

INDICE

INDICE

Coronary Artery Disease

Carmela Pia Senatore

Introduzione

La malattia ischemica del cuore (IHD) rappresenta una delle principali cause di morte in tutto il mondo. La IHD si riferisce a un tipo di malattia cardiaca che manifesta ischemia miocardica, ipossia o necrosi causata da un restringimento o occlusione delle arterie coronarie, che è provocato dall'aterosclerosi.

Secondo la struttura anatomica delle arterie coronarie, ci sono tre arterie principali che forniscono sangue al cuore, ovvero, la LAD (arteria coronaria discendente anteriore sinistra), la LCX (arteria coronaria circonflessa sinistra) e la RCA (arteria coronaria destra).

La malattia delle arterie coronarie (CAD) si verifica quando il lume di una qualsiasi delle tre arterie coronarie è ristretto del 50% o di più. Poiché la CAD è una malattia altamente letale, la tempestiva individuazione e diagnosi sono cruciali per salvare vite e migliorare la prognosi.

L'angiografia coronarica è considerata il gold standard per la diagnosi della CAD. Tuttavia, l'angiografia è una procedura invasiva. La possibilità di rilevare , in maniera non invasiva, la malattia delle arterie coronarie nei pazienti ad alto rischio basandosi su algoritmi di apprendimento automatico potrebbe essere utile per ridurre il carico medico e la grande perdita di vite umane. Questo può avvenire non solo selezionando un modello che permette di comprendere la relazione tra fattori medici e biologici oltre che scelta mirata.

Lo studio è stato condotto in questo modo, per cui ogni sezione ha uno scopo specifico nell'analisi :

1. **Descrizione del Dataset**: Questa sezione fornisce una comprensione dettagliata di ciascuna variabile nel dataset, oltre che ricerca di paper per una migliore comprensione del dataset.
2. **Analisi Esplorativa dei Dati (EDA)**: vengono esplorate le strutture e le caratteristiche dei dati attraverso vari metodi grafici e statistici. Questo include la pulizia dei dati, analisi univariata, l'analisi bivariata, l'analisi multivariata, l'analisi della correlazione e dell'informazione mutua oltre che analisi delle corridpondenze multiple e features engeeniring. Queste analisi mirano a scoprire le relazioni potenziali tra le variabili e il loro impatto collettivo sulla previsione di CAD.
3. **Analisi Predittiva**: Questa parte comporta la pre-elaborazione dei dati, la divisione in set di addestramento e di test. Vengono, poi, sviluppati modelli predittivi, tra cui boosting, logistic regression, rf, XGBoost, LightGBM. I modelli vengono valutati utilizzando metriche come accuratezza, precisione, recall, punteggio F1 e punteggio roc_auc.
4. **SHAP**: Il modello selezionato viene interpretato utilizzando i valori SHAP (SHapley Additive exPlanations). Vengono utilizzati il plot dell'importanza delle feature e il plot di dipendenza per comprendere l'interpretabilità locale, attraverso la selezione di specifici pazienti dal dataset.
5. **Conclusioni**: Questa sezione riassume l'analisi e fornisce approfondimenti ottenuti dai processi di esplorazione e modellazione. Chiaramente si rifletterà sui limiti e i problemi riscontrati durante l'analisi.

1. Descrizione del dataset

Il dataset **Z-Alizadeh Sani** è disponibile pubblicamente nel repository UCI Machine Learning. Il dataset contiene informazioni circa pazienti che si sono presentati all'ospedale Shaheed Rajaei (in Iran) per via di dolori al petto. Per ciascun paziente sono incluse 55 caratteristiche appartenenti a quattro categorie.

Le quattro categorie sono:

1. Caratteristiche demografiche
2. Sintomi ed esame fisico
3. ECG
4. Caratteristiche dell'ecocardiografia

Questi campioni appartengono a due classi, ovvero, la classe **CAD** e la classe **normale.** Quando la stenosi del lume delle arterie coronarie di un campione raggiunge o supera il 50%, questo campione viene classificato come classe CAD; altrimenti appartiene alla classe normale.

Di conseguenza, su 303 campioni, 216 istanze corrispondenti al 71,29% sono della classe CAD, mentre 87 istanze corrispondenti al 28,71% sono della classe normale.

2. Definizione obiettivo

Nel corso dell' analisi, si mira, a definire gli obiettivi tramite l'utilizzo del dataset, con particolare attenzione su due aspetti principali.

- In primo luogo, verrà condotta un'analisi esplorativa e di model investigation. La prima permetterà di ottenere una visione più dettagliata dei pattern e delle relazioni presenti nei dati; la seconda individuerà il miglior modello utile alla risoluzione del problema.
- In secondo luogo, ci si concentrerà sull'interpretazione dei modelli stessi. Si utilizzeranno tecniche di visualizzazione avanzate per comprendere meglio quali feature influenzano le previsioni del modello e come avviene la decisione di classificazione.

Nel presente elaborato ci si propone lo studio della CAD soprattutto nell'ottica dell'individuazione di fattori che possano influenzare la variabile dipendente per la comprensione dell'impatto su dati medici, oltre che sull'individuazione di un modello che sia in grado di fare quanti meno errori possibili. Lo scopo sarà pertanto quello di distinguere aspetti determinanti e significativi per la diagnosi dei pazienti oltre che utilizzare questi per sviluppare una corretta classificazione della malattia utilizzando features numeriche e categoriali.
Il principale obiettivo è, però, verificare se le reti neurali sono in grado di ottenere risultati migliori rispetto ai modelli non black-box.

2.1 Introduzione

Seppur il campo medico è un campo noto, non è esente da problemi statistici. I problemi statistici relativi all'utilizzo di dati medici sono molteplici, di seguito i principali.

Il primo problema che può presentare il dataset è la presenza di missing values. Molto spesso per le p variabili raccolte non sono presenti valori per tutte le osservazioni prese in analisi. I dati mancanti sono un problema perché possono causare una perdita di potere statistico e una distorsione dei risultati, oltre a comportare molto spesso una selezione del campione non rappresentativa, che espone al rischio di una sovrastima o sottostima dell'effetto del predittore sui dati mancanti.

Il secondo problema è relativo alla correlazione presente fra le variabili indipendenti che porta al problema denominato come multicollinearità. Nell'ambito medico tale problema è di particolare rilevanza. I dati medici mostrano un elevata multicollinearità rispetto ad altri tipi di dati. Ciò è dovuto al fatto che alcune condizioni cliniche dei pazienti e misure biologiche adottate sono correlate fra di loro. Tali situazioni si riflettono in ambito statistico da una stretta correlazione fra le variabili.

Un terzo problema riguarda la natura stessa dei dati che sono caratterizzati dalla presenza massiccia di rumore di tipo biologico e tecnico, come il tipo di tecnologia applicata o il metodo utilizzato per la tac/pet eseguita.

Un quarto problema, concernente l'analisi della prevalenza, riguarda l'esplorazione delle tabelle di contingenza contenenti i dati relativi allo stato del paziente.

I problemi appena menzionati si concentrano principalmente nell'ambito statistico. È importante, però, sottolineare che ci sono diverse conseguenze a livello biologico e medico che non possono essere ignorate. La prima conseguenza è la corrispondenza tra la rilevanza biologica e quella statistica di un gene. La rilevanza biologica di un gene può fornire informazioni preziose per la scoperta di funzioni specifiche del gene, la determinazione di gruppi di geni che contribuiscono alla malattia. Queste informazioni possono influenzare le decisioni cliniche e la cura dei pazienti, rendendo la corretta valutazione della rilevanza biologica un processo fondamentale nell'ambito medico.

La seconda conseguenza è derivabile dal problema di inquinamento o contaminazione del campione: questa situazione si verifica quando il campione biologico viene accidentalmente contaminato da batteri, virus o da altri microrganismi durante il processo di prelievo o manipolazione.

3. Descrizione del dataset

Il dataset utilizzato, disponibile pubblicamente su UCI, si compone di 303 osservazioni (pazienti) su 55 variabili. Esso riporta diverse informazioni genetiche e mediche relative ai pazienti adulti di età compresa fra 30 e 86 anni, affetti o non da CAD, dell'ospedale di riferimento in Iran. I dati raccolti ricoprono un arco temporale non specificato. Di seguito vengono riportate le variabili di cui si compone il dataset.

Variabile	**Descrizione**
Age	Intervallo di età dei pazienti, compreso tra 30 e 86 anni. Questa variabile indica l'età cronologica del paziente al momento della raccolta dei dati.
Weight	Peso corporeo dei pazienti, espresso in chilogrammi (48-120 kg).
Sex	Sesso biologico dei pazienti (Maschio, Femmina).
BMI	Indice di Massa Corporea (BMI) calcolato come peso (kg) diviso per altezza al quadrato (m^2), valori tra 18 e 41.
DM	Presenza di diabete mellito (Sì, No).
HTN	Presenza di ipertensione (Sì, No).
Current Smoker	Indica se il paziente fuma attualmente (Sì, No).
Ex-Smoker	Indica se il paziente ha smesso di fumare (Sì, No).
FH	Presenza di una storia familiare di malattie cardiache (Sì, No).
Obesity	Presenza di obesità (Sì, se BMI > 25, No altrimenti).
CRF	Presenza di insufficienza renale cronica (Sì, No).
CVA	Presenza di un ictus pregresso (Sì, No).
Airway Disease	Presenza di malattie delle vie aeree (Sì, No).

Variabile	Descrizione
Thyroid Disease	Presenza di malattie della tiroide (Sì, No).
CHF	Presenza di insufficienza cardiaca congestizia (Sì, No).
DLP	Presenza di dislipidemia (Sì, No).
BP (blood pressure)	Pressione sanguigna misurata in mmHg (90-190).
PR (pulse rate)	Frequenza del polso misurata in battiti per minuto (50-110).
Edema	Presenza di gonfiore nei tessuti (Sì, No).
Weak peripheral pulse	Polso periferico debole (Sì, No).
Lung rales	Rumori anormali nei polmoni (Sì, No).
Systolic murmur	Presenza di soffio sistolico (Sì, No).
Diastolic murmur	Presenza di soffio diastolico (Sì, No).
Typical Chest Pain	Dolore toracico tipico (Sì, No).
Dyspnea	Difficoltà respiratoria (Sì, No).
Function class	Classe funzionale dei sintomi (1, 2, 3, 4).
Atypical	Dolore toracico atipico (Sì, No).
Nonanginal CP	Dolore toracico non anginoso (Sì, No).
Exertional CP	Dolore toracico da sforzo (Sì, No).
Low Th Ang	Angina a bassa soglia (Sì, No).
Rhythm	Tipo di ritmo cardiaco (Sin, AF).
Q Wave	Presenza di onde Q anomale (Sì, No).
ST Elevation	Presenza di elevazione del segmento ST (Sì, No).
ST Depression	Presenza di depressione del segmento ST (Sì, No).
T Inversion	Presenza di inversione dell'onda T (Sì, No).
LVH	Presenza di ipertrofia ventricolare sinistra (Sì, No).
Poor R progression	Scarsa progressione dell'onda R (Sì, No).
FBS	Glicemia a digiuno (mg/dl) (62-400).
Cr	Livello di creatinina (mg/dl) (0.5-2.2).
TG	Livello di trigliceridi (mg/dl) (37-1050).
LDL	Livello di lipoproteine a bassa densità (mg/dl) (18-232).
HDL	Livello di lipoproteine ad alta densità (mg/dl) (15-111).
BUN	Livello di azoto ureico (mg/dl) (6-52).
ESR	Tasso di sedimentazione degli eritrociti (mm/h) (1-90).
HB	Livello di emoglobina (g/dl) (8.9-17.6).
K	Livello di potassio (mEq/lit) (3.0-6.6).
Na	Livello di sodio (mEq/lit) (128-156).
WBC	Numero di globuli bianchi (cells/ml) (3700-18,000).
Lymph	Percentuale di linfociti (%) (7-60).
Neut	Percentuale di neutrofili (%) (32-89).
PLT	Numero di piastrine (1000/ml) (25-742).
EF	Frazione di eiezione (%) (15-60).
Region with RWMA	Numero di regioni con anomalie del movimento della parete (0-4).
VHD	Gravità della malattia delle valvole cardiache (Normale, Lieve, Moderata, Grave)

4. Data WRANGLING

Il "data wrangling," spesso chiamato anche "data munging" rappresenta un passaggio cruciale nell'analisi dei dati. Molto spesso si hanno a disposizione un insieme di dati grezzi provenienti da diverse fonti. Questi dati possono essere disorganizzati, contenere errori, dati mancanti o informazioni in formati diversi. Senza una preparazione adeguata, l'analisi e la modellazione dei dati sarebbero difficili, se non impossibili.

Il data wrangling è il processo di trasformazione dei dati disordinati e sporchi in un formato coerente e adatto all'analisi. Il passaggio comporta molteplici attività, tra cui la pulizia dei dati per correggere errori e rimuovere duplicati, la standardizzazione delle unità di misura, la trasformazione dei dati categorici in forme numeriche comprensibili, e la creazione di nuove variabili o caratteristiche quando necessario.

L'obiettivo finale del data wrangling è creare un dataset pulito, coerente e pronto per essere analizzato, riducendo così i potenziali errori e garantendo che i risultati dell'analisi siano accurati e significativi.

4.1 Import del dataset

4.2 Panoramica dei dati

Si ottiene una visulizzazione dell eprime 6 osservazioni per comprendere la struttura del dataset.

```
## # A tibble: 6 x 55
##     Age Weight Length Sex     BMI    DM   HTN `Current Smoker` `EX-Smoker`    FH
##   <dbl>  <dbl>  <dbl> <chr> <dbl> <dbl> <dbl>            <dbl>       <dbl> <dbl>
## 1    53     90    175 Male   29.4     0     1                1           0     0
## 2    67     70    157 Fmale  28.4     0     1                0           0     0
## 3    54     54    164 Male   20.1     0     0                1           0     0
## 4    66     67    158 Fmale  26.8     0     1                0           0     0
## 5    50     87    153 Fmale  37.2     0     1                0           0     0
## 6    50     75    175 Male   24.5     0     0                1           0     0
## # i 45 more variables: Obesity <chr>, CRF <chr>, CVA <chr>,
## #   `Airway disease` <chr>, `Thyroid Disease` <chr>, CHF <chr>, DLP <chr>,
## #   BP <dbl>, PR <dbl>, Edema <dbl>, `Weak Peripheral Pulse` <chr>,
## #   `Lung rales` <chr>, `Systolic Murmur` <chr>, `Diastolic Murmur` <chr>,
## #   `Typical Chest Pain` <dbl>, Dyspnea <chr>, `Function Class` <dbl>,
## #   Atypical <chr>, Nonanginal <chr>, `Exertional CP` <chr>, `LowTH Ang` <chr>,
## #   `Q Wave` <dbl>, `St Elevation` <dbl>, `St Depression` <dbl>, ...
```

Sono quindi presenti 303 osservazioni relative ai pazienti dell'indagine e 55 variabili di descrizione.

```
## [1] 303  55
```

4.3 Data Cleaning

I dati grezzi sono spesso disordinati e formattati male. Inoltre, potrebbero mancare definizioni appropriate che tengano conto della scala di misurazione utilizzata.

Per cui la pulizia dei dati consiste nel procedimento mediante il quale si esaminano e si migliorano i dati contenuti nel dataset, con l'obiettivo di assicurare che siano di alta qualità e validi per l'analisi statistica. Le procedure che caratterizzano questo passaggio sono le seguenti:

- **Analisi dei missing values.** Si verifica la presenza/assenza dei valori mancanti (NA) che può notevolmente influire sull'analisi. Si procede , quindi, decidendo di eliminarli (*na.omit*) sostituendoli con valori reali (esempio: imputazione di media o mediana).

- **Individuazione di valori anomali.** Si esamina attentamente il dataset per individuare eventuali valori inesatti o insoliti, e si prendono decisioni su come trattarli. Queste decisioni possono includere l'eliminazione dei valori problematici o la loro sostituzione con valori appropriati.
- **Trasformazione delle variabili.** Si verifica che tutte le variabili abbiano la classe appropriata. (Double o Integer per i numeri, factor per le variabili categoriali, ordered per le variabili categoriali ordinate).
- **Implementazione di nuove variabili.** Sulla base delle variabili già presenti nel dataset si possono creare delle nuove variabili, ad esempio facendo operazioni matematiche tra due variabili o creando una variabile multilivello sulla base di una variabile numerica.

4.3.1 Analisi dei missing values

Si verifica la presenza degli NA nel dataset. Nel dataset non sono presenti NA.

```
## [1] 0
```

4.3.2 Trasformazione delle variabili

Dalla tabella delle tipologie di formattazione dei dati, si osserva che molte variabili categoriche sono formattate come numeriche o caratteri (chr). Per garantire una corretta analisi, si procede con la conversione appropriata di queste variabili in fattori.

Inoltre, poiché alcune variabili presentano un numero esiguo di osservazioni per alcune classi di categorie, si è deciso di eliminarle dal dataset poiché non risultano rilevanti per lo studio.

4.3.3 Individuazione di valori anomali

La funzione in output fornisce un resoconto automatico delle statistiche di base per ciascuna variabile nel dataset. Queste statistiche comprendono il valore minimo, il valore massimo, la media, la mediana e la deviazione standard. Se la variabile è numerica, vengono calcolati anche i quartili e il range interquartile. Nel caso in cui la variabile sia di tipo carattere o un factor, la funzione restituirà il conteggio delle osservazioni per ciascun livello o valore univoco presente nella variabile stessa.

L'età dei pazienti varia da un minimo di 30 anni a un massimo di 86 anni, con una media di circa 58,9 anni. Il peso varia da 48 kg a 120 kg, con una media di 73,83 kg. L'altezza varia da 140 cm a 188 cm, con una media di 164,7 cm. L'indice di massa corporea (BMI) varia da 18,12 a 40,90, con una media di 27,25. I 303 campioni sono classificati in due classi principali: CAD (malattia coronarica) e normale. Un campione è classificato come CAD se la stenosi delle arterie coronarie raggiunge o supera il 50%, altrimenti appartiene alla classe normale. Dei 303 campioni, 216 (71,29%) appartengono alla classe CAD e 87 (28,71%) alla classe normale.

Variabile	Min.	1st Qu.	Mediana	Media	3rd Qu.	Max.	Frequenze
Age (anni)	30.0	51.0	58.0	58.9	66.0	86.0	
Weight (kg)	48.0	65.0	74.0	73.83	81.0	120.0	
Length (cm)	140.0	158.0	165.0	164.7	171.0	188.0	
Sex							F: 127, M: 176
BMI (Kg/m²)	18.12	24.51	26.78	27.25	29.41	40.90	

Variabile	Min.	1st Qu.	Mediana	Media	3rd Qu.	Max.	Frequenze
DM (Diabetes)							0: 213, 1: 90
HTN (Hypertension)							0: 124, 1: 179
Current Smoker	0.0	0.0	0.0	0.2079	0.0	1.0	
Ex-Smoker	0.0	0.0	0.0	0.0	0.0	1.0	0: 293, 1: 10
FH (Family History)							0: 255, 1: 48
Obesity							N: 92, Y: 211
DLP (Dyslipidemia)							N: 191, Y: 112
BP (mmHg)	90.0	120.0	130.0	129.6	140.0	190.0	
PR (Pulse Rate)	50.0	70.0	70.0	75.14	80.0	110.0	
Edema							N: 262, Y: 41
Systolic Murmur							N: 262, Y: 41
Typical Chest Pain							0: 139, 1: 164
Dyspnea							N: 169, Y: 134
Function Class	0	0	0	0	1	4	0: 211, 1: 1, 2: 73, 3: 18
Atypical							N: 210, Y: 93
Q Wave							0: 287, 1: 16
ST Elevation							0: 289, 1: 14
ST Depression							0: 232, 1: 71
T Inversion							0: 213, 1: 90
LVH (Left Ventricular Hypertrophy)							N: 283, Y: 20
FBS (mg/dl)	62.0	88.5	98.0	119.2	130.0	400.0	
Cr (mg/dl)	0.5	0.9	1.0	1.056	1.2	2.2	
TG (mg/dl)	37.0	90.0	122.0	150.3	177.0	1050.0	

Variabile	Min.	1st Qu.	Mediana	Media	3rd Qu.	Max.	Frequenze
LDL (mg/dl)	18.0	80.0	100.0	104.6	122.0	232.0	
HDL (mg/dl)	15.9	33.5	39.0	40.23	45.5	111.0	
BUN (mg/dl)	6.0	13.0	16.0	17.5	20.0	52.0	
ESR (mm/h)	1.0	9.0	15.0	19.46	26.0	90.0	
HB (g/dl)	8.9	12.2	13.2	13.15	14.2	17.6	
K (mEq/lit)	3.0	3.9	4.2	4.231	4.5	6.6	
Na (mEq/lit)	128	139	141	141	143	156	
WBC (cells/ml)	3700	5800	7100	7562	8800	18000	
Lymph (%)	7.0	26.0	32.0	32.4	39.0	60.0	
Neut (%)	32.0	52.5	60.0	60.15	67.0	89.0	
PLT (1000/ml)	25.0	183.5	210.0	221.5	250.0	742.0	
EF-TTE (%)	15.0	45.0	50.0	47.23	55.0	60.0	
Region RWMA	0.0	0.0	0.0	0.6205	1.0	4.0	
VHD (Valvular Heart Disease)							Mild: 149, Moderate: 27, Normal: 116, Severe: 11
Cath							Cad:216, Normal:87

Questi indicatori forniscono una panoramica completa delle condizioni di vita tenendo conto ,tuttavia, delle differenze che caratterizzano ciascun paziente. L'analisi dettagliata di ciascuna variabile verrà aggiunta in *appendice.*

5. Analisi delle associazioni fra features

Nell'ambito dell'analisi è stato condotto uno studio per esaminare le associazioni tra diverse variabili di interesse. Per valutare il legame e il grado di associazione si ricorre all'analisi delle corrispondenze multiple e al test del chi-quadro.

5.1 Test del Chi-Quadro

Tramite il test del chi quadrato di Pearson, si intende controllare se l'associazione fra due variabili sia statisticamente significativa. Il test del chi quadrato confronta i valori osservati di frequenza in una tabella di contingenza con i valori attesi, che rappresentano l'ipotesi di indipendenza tra le variabili, l'ipotesi nulla

(H_0) del test afferma che le due variabili sono indipendenti l'una dall'altra, mentre l'ipotesi alternativa (H_1) afferma che le due variabili sono associate in qualche modo. Le caratteristiche demografiche, di laboratorio e cliniche sono state confrontate sulla base dello status del paziente, tenendo conto dei risultati del test. Poiché il test viene effettuato solo su variabili categoriche, si è ritenuto necessario stabilire una soglia per la variabile BMI numerica poiché considerata fondamentale nello studio. La soglia ottimale è stata individuata tenendo conto dei risultati dell'analisi effettuata dallo Studio Danone sul sovrappeso e obesità (BMI>25) (Obesity) [1]. Inoltre, si ottengono delle variabili categoriali fattorizzate per le informazioni riguardanti il battito cardiaco oltre che si mantengono le informazioni circa l'ecografia, poichè secondo [2] in uno studio con pazienti malati di CAD, sono presenti evidenze di studi randomizzati che supportano il ruolo dell'ecocardiografia come guida processo decisionale clinico. Dallo studio ORBITA (Objective Randomized Blinded Investigation With Optimal Medical Therapy of Angioplasty in Stable Angina), un risultato secondario era una maggiore riduzione del punteggio tra i pazienti con CAD trattati con intervento coronarico percutaneo (PCI) rispetto con placebo (P0,0001).

Stabilito il valore di α=0.05,

- La probabilità che l'associazione sia dovuta al caso è prossima allo zero per le variabili: DM, HTN, TYPICAL CHEST PAIN, DYSPNEA, ATYPICAL, QWAVE, ST ELEVATION, ST DEPRESSION,TINVERSION E VHD. L'ipotesi nulla viene respinta e l'associazione è statisticamente significativa. Il valore di associazione maggiore si registra con la variabile TYPICAL CHEST PAIN;
- Diversamente per altre variabili , il valore del p-value porta a non rifiutare l'ipotesi nulla stabilendo l'indipendenza fra le coppie di variabili.

La Tabella riporta il valore associato del p-value per ciascuna categoria oltre che la partizione utile per descrivere le caratteristiche principali della popolazione. In totale sono stati osservati 303 pazienti , di cui 127 (41.91%) sono maschi e 176 (58.1%) sono femmine, l'età mediana riscontrata è 58 anni.

La prevalenza di CAD nell'intera coorte è del 73%. Dall'analisi comparativa della malattia in relazione alle caratteristiche demografiche, è emerso che i pazienti con il diabete mellito hanno una maggiore probabilità di ricevere diagnosi di CAD rispetto a quelli che non lo possiedono (88% vs 66%). Inoltre, è stato osservato che i pazienti con ipertensione presentano una maggiore frequenza in osservazioni per CAD che rispetto lo stato normale(147 vs 69). Al 93% dei pazienti , con dolore toracico tipico, è stato diagnosticata CAD mentre solo il 44% di coloro che non presentano il dolore è stata diagnosticata CAD. Inoltre, contrariamente a quanto emerso per il dolore toracico atipico, per il dolore toracico atipico il maggior numero di pazienti a cui viene diagnosticata la CAD non presentano questo sintomo.
Un altro dato significativo riguarda i pazienti che presentano difficoltà respiratoria. Questo gruppo ha una percentuale di prognosi del 64%, suggerendo che la presenza di questo sintomo è associata a una maggiore probabilità di diagnosi della malattia.

Per quel che concerne le variabili relative alle informazioni e assunzioni derivate dall'elettrocardiogramma, la non presenza di onde Q anomale ha un supporto di definizione maggiore: a 200 pazienti in questa categorie è stata diagnostica CAD mentre a 87 pazienti no. Lo stesso vale per la non presenza di elevazione del segmento. A differenza di queste , la presenza di depressione del segmento invece ha visto l'assegnazione della malattia all'83% dei pazienti mentre solo il 67% che non lo presentavano hanno avuto la stessa diagnosi.

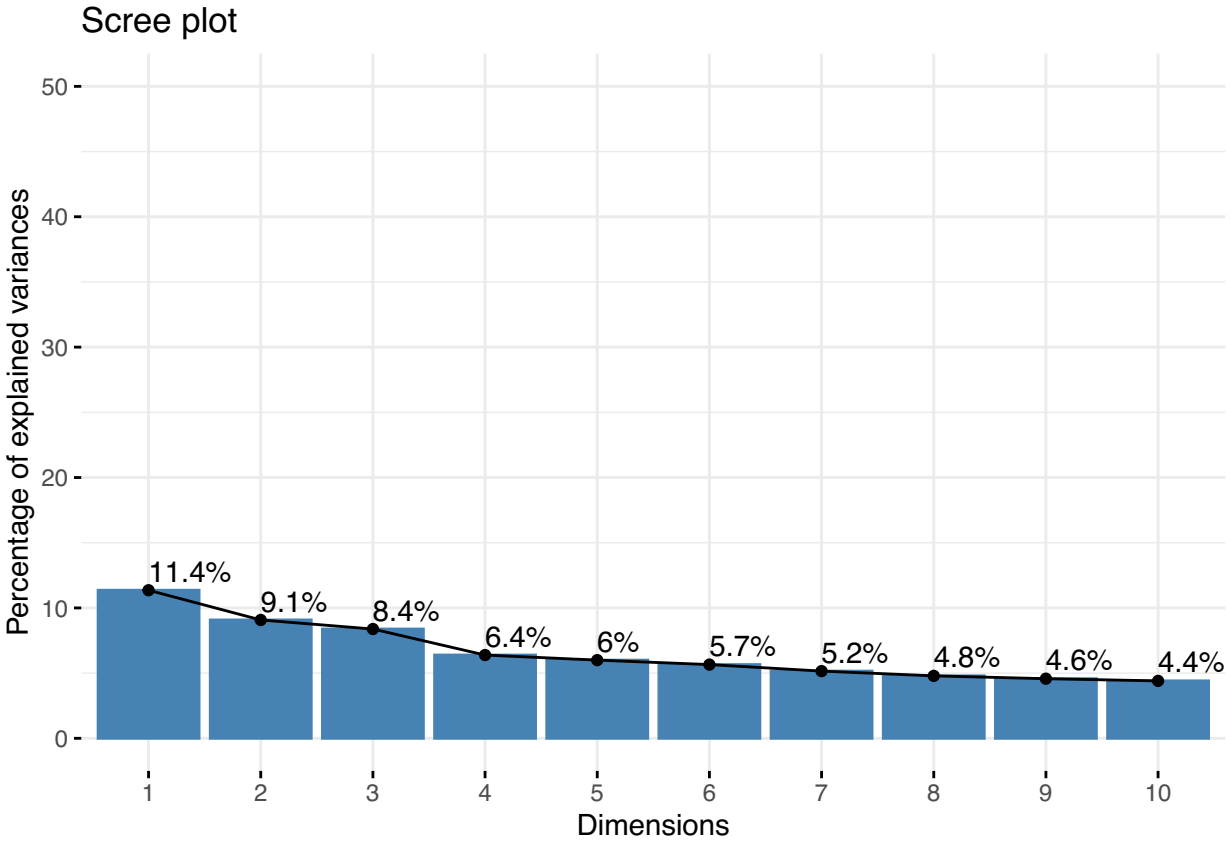
La categoria della gravità delle malattie delle valvole cardiache che presenta il maggior numero di pazienti a cui è sta diagnosticata la malattia nella coorte è quella mild, con una percentuale del 50%%, mentre solo il 22% dei pazienti in questa categoria non ha registrato la malattia.

Variable	Livelli	Totale	Target Positive	Target Negative	P.Value	V.Cramer
Sex	Fmale	127	86 (67.72%)	41 (32.28%)	0.243	0.067
	Male	176	130 (73.86%)	46 (26.14%)	0.243*	0.067
DM	0	213	136 (63.85%)	77 (36.15%)	0.00001	0.253

Variable	Livelli	Totale	Target Positive	Target Negative	P.Value	V.Cramer
	1	90	80 (88.89%)	10 (11.11%)	0.00001*	0.253
HTN	0	124	69 (55.65%)	55 (44.35%)	0.00000***	0.288
	1	179	147 (82.12%)	32 (17.88%)	0.00000***	0.288
EX-Smoker	0	293	208 (70.99%)	85 (29.01%)	0.53572	0.036
	1	10	8 (80.00%)	2 (20.00%)	0.53572	0.036
FH	0	255	180 (70.59%)	75 (29.41%)	0.53540	0.036
	1	48	36 (75.00%)	12 (25.00%)	0.53540	0.036
Obesity	N	92	67 (72.83%)	25 (27.17%)	0.69581	0.02246
	Y	211	149 (70.62%)	62 (29.38%)	0.69581	0.02246
DLP	N	191	137 (71.73%)	54 (28.27%)	0.82479	0.01272
	Y	112	79 (70.54%)	33 (29.46%)	0.82479	0.01272
Systolic Murmur	N	262	187 (71.37%)	75 (28.63%)	0.93263	0.00486
	Y	41	29 (70.73%)	12 (29.27%)	0.93263	0.00486
Typical Chest Pain	0	139	62 (44.60%)	77 (55.40%)	3.34442e-21***	0.543
	1	164	154 (93.90%)	10 (6.10%)	3.34442e-21***	0.543
Dyspnea	N	169	129 (76.33%)	40 (23.67%)	0.02929*	0.12521
	Y	134	87 (64.93%)	47 (35.07%)	0.02929*	0.12521
Function Class	0	211	145 (68.72%)	66 (31.28%)	0.13938	0.13457
	1	1	0 (0.00%)	1 (100.00%)	0.13938	0.13457
	2	73	56 (76.71%)	17 (23.29%)	0.13938	0.13457
	3	18	15 (83.33%)	3 (16.67%)	0.13938	0.13457
Atypical	N	210	176 (83.81%)	34 (16.19%)	4.48991e-13***	0.41592
	Y	93	40 (43.01%)	53 (56.99%)	4.48991e-13***	0.41592
Q Wave	0	287	200 (69.69%)	87 (30.31%)	0.00910***	0.14985
	1	16	16 (100.00%)	0 (0.00%)	0.00910***	0.14985
St Elevation	0	289	202 (69.90%)	87 (30.10%)	0.01504**	0.13968
	1	14	14 (100.00%)	0 (0.00%)	0.01504**	0.13968
St Depression	0	232	157 (67.67%)	75 (32.33%)	0.01194	0.14443
	1	71	59 (83.10%)	12 (16.90%)	0.01194**	0.14443
Tinversion	0	213	137 (64.32%)	76 (35.68%)	3.71919e-05***	0.23693
	1	90	79 (87.78%)	11 (12.22%)	3.71919e-05***	0.23693
LVH	N	283	200 (70.67%)	83 (29.33%)	0.37284	0.05120
	Y	20	16 (80.00%)	4 (20.00%)	0.37284	0.05120
VHD	mild	149	115 (77.18%)	34 (22.82%)	0.00103***	0.23121
	Moderate	27	22 (81.48%)	5 (18.52%)	0.00103***	0.23121
	N	116	76 (65.52%)	40 (34.48%)	0.00103***	0.23121
	Severe	11	3 (27.27%)	8 (72.73%)	0.00103***	0.23121

5.2 ACM

In seguito, è stato applicato l'analisi delle corrispondenze multiple il cui fine ultimo risiede nell'esplorazione e visualizzazione delle associazioni tra le categorie delle variabili prese in esame. La figura mostra lo screeplot delle dimensioni ottenute in seguito all'applicazione della analisi delle corrispondenze multiple. Sull'asse delle ordinate viene rappresentata la varianza spiegata da ciascuna componente, mentre sull'asse delle ascisse viene rappresentato il numero di componenti. La visualizzazione rende chiaro che la varianza e l'inerzia totale spiegata da ciascun fattore diminuisce a mano a mano che queste diventano maggiori in numero. Il punto in cui la curva inizia a livellarsi è definito "elbow point" indicando il numero di dimensioni significative, questo potrebbe essere individuato tra la componente numero 4 e la componente numero 5. Solitamente queste componenti riescono a spiegare le principali associazioni o pattern nei dati.

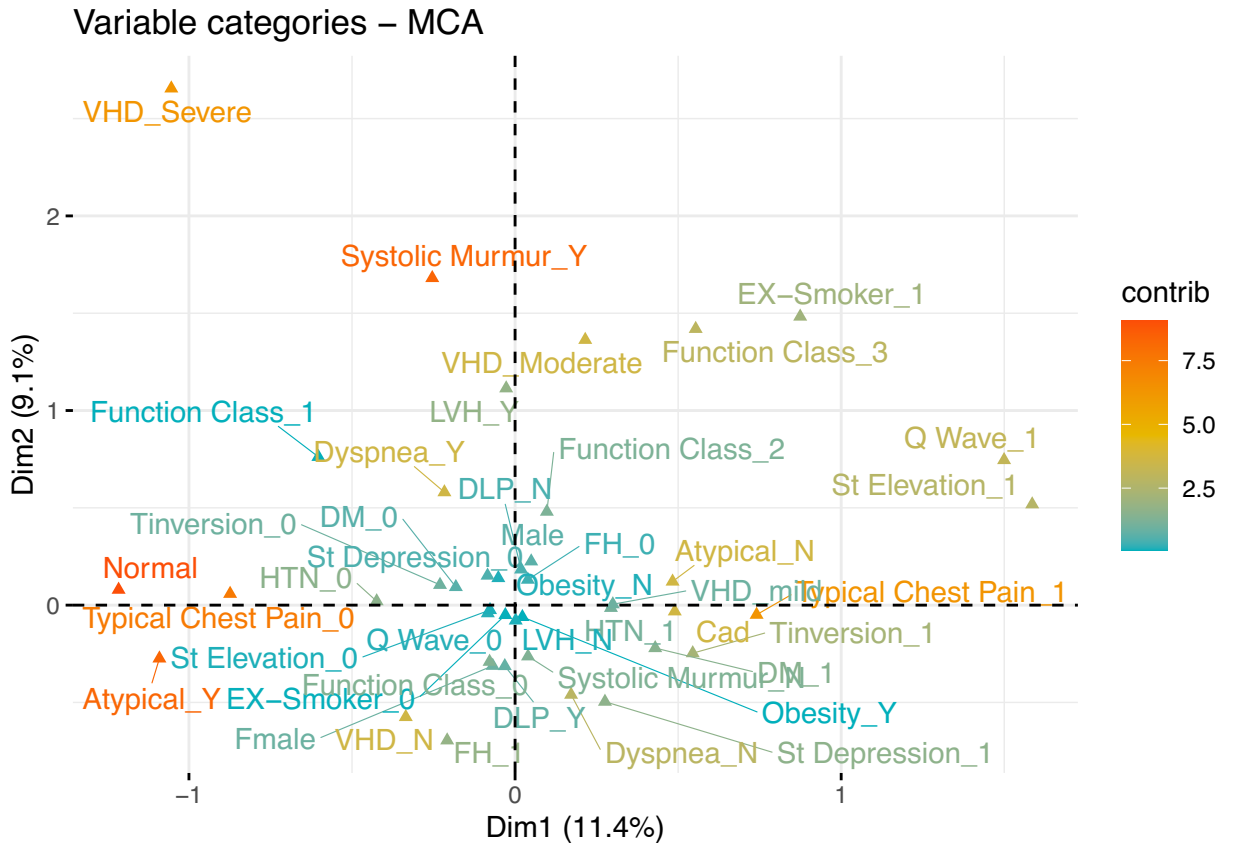


La visualizzazione bidimensionale degli assi principali è presentata nella Figura sottostante. Il grafico dell'analisi delle corrispondenze multiple viene creato proiettando le variabili categoriche su un piano bidimensionale in modo da visualizzare le loro relazioni. La rappresentazione è composta da un piano cartesiano con due assi principali. L'asse x rappresenta la prima componente principale dell'analisi delle corrispondenze multiple, mentre l'asse y rappresenta la seconda componente principale. L'inerzia totale spiegata dal grafico dell'analisi delle corrispondenze multiple basato sui primi due assi fattoriali, ovvero la misura del grado di dispersione del profilo attorno al profilo medio, è del circa 20%. Ogni variabile categoriale viene rappresentata come un punto sul grafico, e la posizione del punto indica la relazione tra le categorie di quella variabile e le altre variabili nel set di dati. Le variabili che sono più vicine tra loro sul grafico MCA sono quelle correlate positivamente, mentre le variabili che sono lontane l'una dall'altra sono correlate negativamente. Inoltre, le variabili vengono designate in base al loro contributo che viene definito sulla scala di colori al fianco del grafico. Ciò permette di distinguere i punti lontani dall'origine degli assi (colorati in rosso), i quali rappresentano le variabili che contribuiscono maggiormente all'analisi delle corrispondenze, dai punti vicini agli assi (colorati di azzurro), espressione di una minore contribuzione delle stesse. Le righe e le colonne che sono simili in base ai loro valori di variabile, tendono ad essere posizionate vicine tra loro nello

spazio, mentre quelle che sono dissimili tendono a essere posizionate lontano l'una dall'altra. Ciò consente di individuare facilmente le relazioni tra le righe e le colonne della tabella di dati, facilitando l'analisi e la comprensione. La presenza di linee tratteggiate che collegano le categorie di diverse variabili, esprime un'associazione significativa.

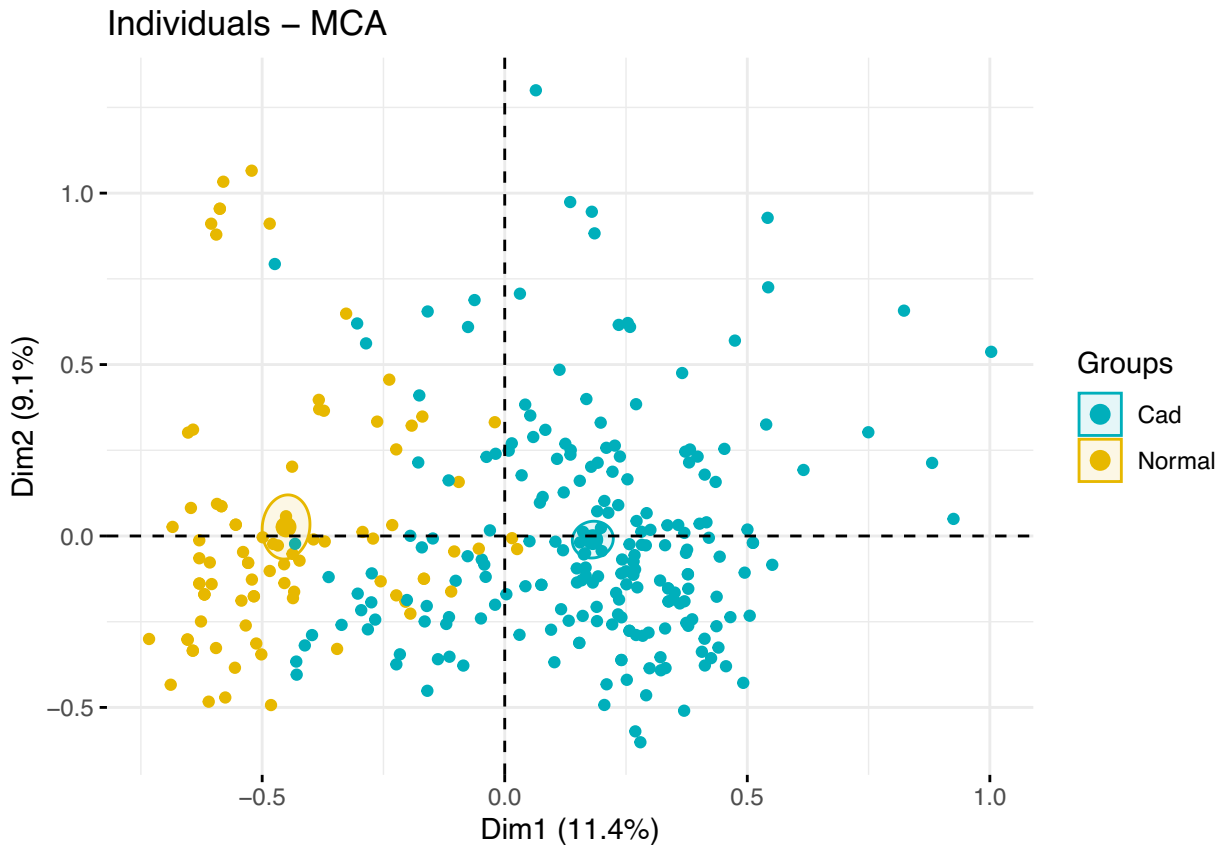
Di seguito quelle essenziali e da prendere in considerazione:

1) La relazione tra valvulopatia severa e presenza di soffio sistolico (VHD_Severe (Valvulopatia cardiaca severa) e Systolic Murmur_Y (Presenza di soffio sistolico)). Queste due categorie sono vicine sul grafico, indicando che i pazienti con valvulopatia cardiaca severa spesso hanno anche un soffio sistolico;
2) L'elevato contributo della relazione tra pazienti che mostrano sintomi di dolore toracico tipico e presenza di onde anomale nell'ECG;
3) La vicinanza tra presenza di obesità e diabete mellito, merita una nota di approfondimento. La vicinanza tra queste categorie suggerisce una frequente co-occorrenza di obesità e diabete mellito nei pazienti;
4) Function Class_3. Questa categoria è distante dal centro e colorata in arancione, indicando una contribuzione alta e specifica alla variabilità nei dati.Spesso si associa con condizioni severe come "VHD_Severe" e "Systolic Murmur_Y";
5) Gli ex-fumatori mostrano relazioni con variabili di malattia cardiaca come "Typical Chest Pain_1" e "LVH_N" (assenza di ipertrofia ventricolare sinistra).
6) Presenza di difficoltà respiratoria e CHF_1 (Presenza di insufficienza cardiaca): la loro vicinanza suggerisce che la difficoltà respiratoria è comunemente associata a insufficienza cardiaca congestizia.



Si consideri l'analisi particolare di diagnosi come punto di partenza nella figura successiva, è possibile evidenziare sul biplot la disposizione degli individui. I punti si diffondono intorno agli assi principali. Il primo asse fattoriale distingue situazioni relative a pazienti a cui è stata diagnosticata la malattia a destra, da

pazienti senza CAD a sinistra. I primi fortemente caratterizzati da situazioni gravose in termini medici, si fa riferimento a pazienti on classe di sintomi 3, obesi, con dolori tipici, presenza di depressione del segmento ecg,ex fumatori, con un BMI>25 ; I secondi, avvantaggiati in termini medici, non mostrano inversione dell T nell ECG, e non ottengono la diagnosi di CAD. I risultati riportati dall'analisi delle corrispondenze multiple potrebbero evidenziare andamenti particolari nella popolazione fornendo un punto di partenza per l'analisi esplorativa.



5.4 EDA

Per quel che concerne l'analisi univariata delle variabili e eventuali approfondimenti sulle singole variabili, è presente nell'appendice una dettagliata analisi dei dati. Per cui ci si concentrerà sul'analisi bivariata e multivariata.

Analisi Bivariata

L'analisi bivariata permette di lavorare su ogni unità statistica e rilevare *congiuntamente* i due caratteri statistici XeY, generando la rilevazione doppia (X, Y). Si può trattare di due caratteri qualitativi, due caratteri quantitativi o un carattere qualitativo e uno quantitativo.

Prima di analizzare le relazioni fra variabili si rende necessaria un'analisi preliminare degli indici e misure che caratterizzano tali rapporti.

Covarianza

Il coefficiente è il seguente:

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}((x_i - \overline{x})(y_i - \overline{y}))$$

- se $cov = 0$, XeY sono incorrelate, non esiste alcun legame lineare fra di loro;
- se $cov > 0$, XeY sono correlate positivamente; a variazioni positive (negative) di una variabile corrispondono variazioni positive(negative) dell'altra variabile;
- se $cov < 0$, XeY sono correlate negativamente; a variazioni positive di una variabile, corrispondono in media, variazioni negative dell'altra variabile e viceversa;

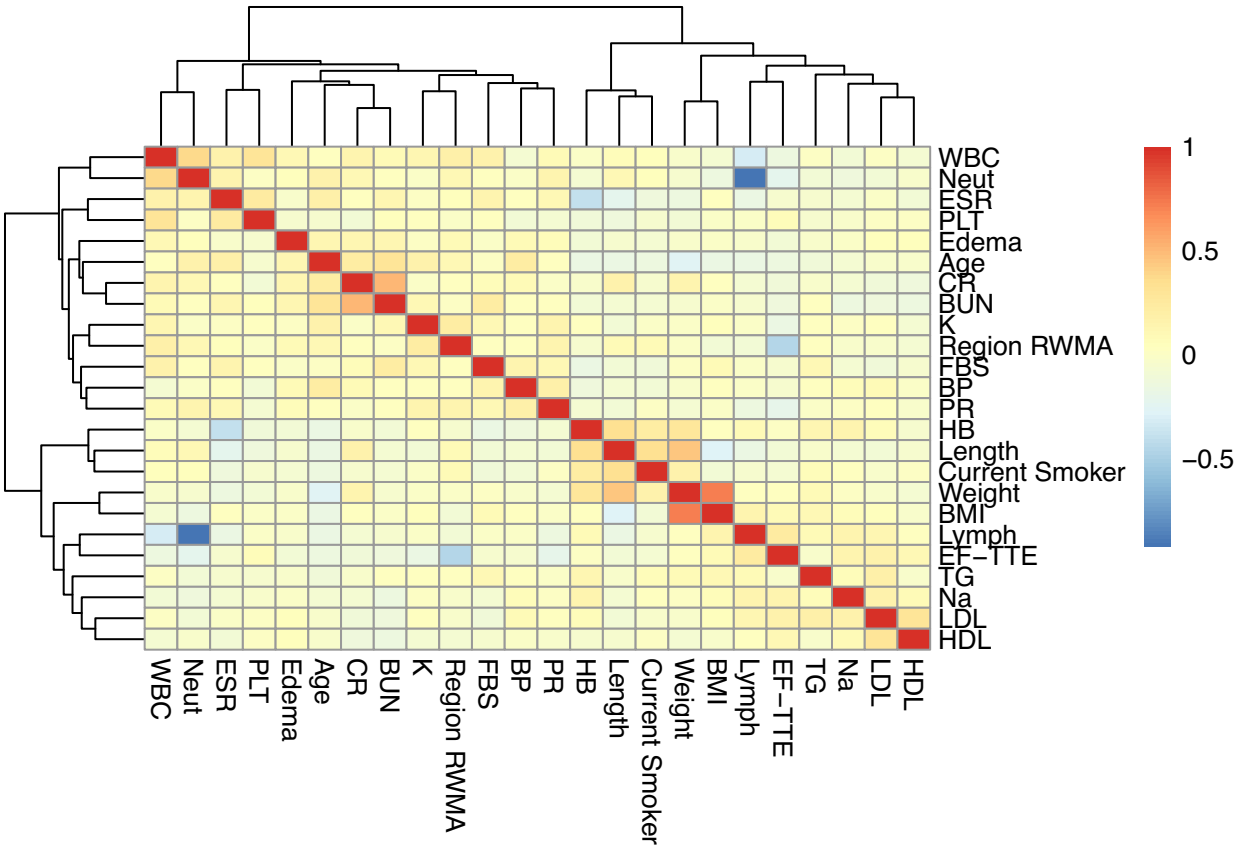
Nonostante la grandezza del dataframe, si è cercato di cogliere i tratti essenziali dalla matrice di covarianza:

- Per le relazioni positive significative:
    1. TG e WBC: C'è una forte covarianza positiva tra i livelli di trigliceridi e il numero di globuli bianchi, suggerendo che alti livelli di trigliceridi possono essere associati a un aumento del numero di globuli bianchi.
    2. WBC e Neut (neutrofili): C'è una forte covarianza positiva tra il numero di globuli bianchi e la percentuale di neutrofili, il che è atteso, poiché i neutrofili sono una sottoclasse dei globuli bianchi.
    3. TG e LDL (lipoproteine a bassa densità): La covarianza positiva significativa indica che alti livelli di trigliceridi tendono a essere associati ad alti livelli di LDL, che sono entrambi fattori di rischio per le malattie cardiovascolari.
- Per le relazioni negative significative:
    1. TG e Age (età): C'è una covarianza negativa significativa tra i livelli di trigliceridi e l'età, suggerendo che i livelli di trigliceridi tendono a diminuire con l'aumentare dell'età dei pazienti.
    2. TG e HDL (lipoproteine ad alta densità): C'è una covarianza negativa significativa tra i livelli di trigliceridi e i livelli di HDL, il che è atteso poiché HDL è considerato il "colesterolo buono" e spesso ha un comportamento inverso rispetto ai trigliceridi.
    3. FBS (glicemia a digiuno) e TG: Anche se questo è positivo, è importante notarlo perché un alto livello di glicemia a digiuno è associato a un alto livello di trigliceridi, suggerendo una correlazione con il diabete mellito.

Heatmap per la Correlazione

La correlazione può essere positiva, quando le variazioni delle due variabili vanno nella stessa direzione, o negativa, quando le variazioni delle due variabili vanno in direzioni opposte. L'espressione del coefficiente di correlazione avviene attraverso un valore, che varia da -1 a +1. Un coefficiente di correlazione di 0 indica l'assenza di correlazione, mentre un valore di +1 o -1 indica una correlazione perfetta tra le due variabili positivamente o negativamente. E' possibile quindi procedere alla visualizzazione del suddetto coefficiente attraverso un grafico pheatmap, utilizzato per creare mappe di calore, ovvero grafici in cui le celle di una tabella sono colorate in base al loro valore, al fine di evidenziare modelli o tendenze nei dati.Le righe e le colonne del grafico rappresentano le variabili numeriche, come prima descritte. La scala di valori rappresentata dalla legenda posizionata a destra è espressione del grado di correlazione fra variabili. Per cui, variabili che presentano un grado di associazione molto elevato , positivo, avranno in corrispondenza del match riga-colonna un colore tendente al rosso/arancione. Una casella colorata di giallo indica un'assenza di correlazione; una casella colorata celeste, tendente al blu scuro, indica una debole o forte associazione negativa. Risulta essenziale denotare che è una matrice simmetrica, la diagonale principale è colorata di rosso scuro poiché riflette un ovvia associazione pari a 1 della variabile con sé stessa, per cui tutto ciò che si ripete nella parte superiore della diagonale si ripete nella parte inferiore della diagonale.

Degna di nota la forte correlazione negativa tra Nut e Lymph.

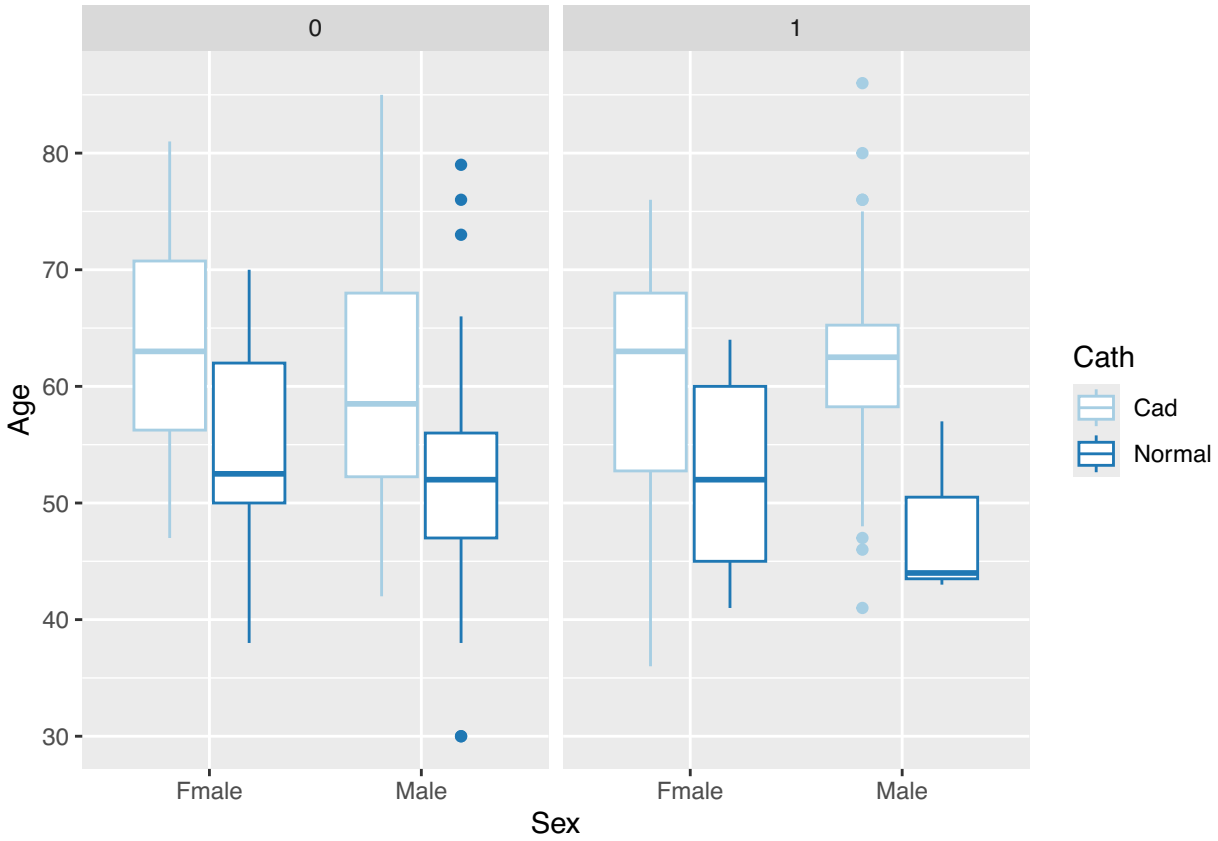


EDA

Successivamente, è stata condotta un'analisi esplorativa dei dati per ottenere una comprensione approfondita delle informazioni. Il criterio di analisi ha tenuto conto dei risultati ottenuti precedentemente.

La coorte inclusa nello studio è costituita da pazienti con età compresa tra 30 e 86 anni. Nel seguente boxplot viene effettuata una panoramica dell'età dei pazienti rispetto al genere, prognosi e diabete mellito.

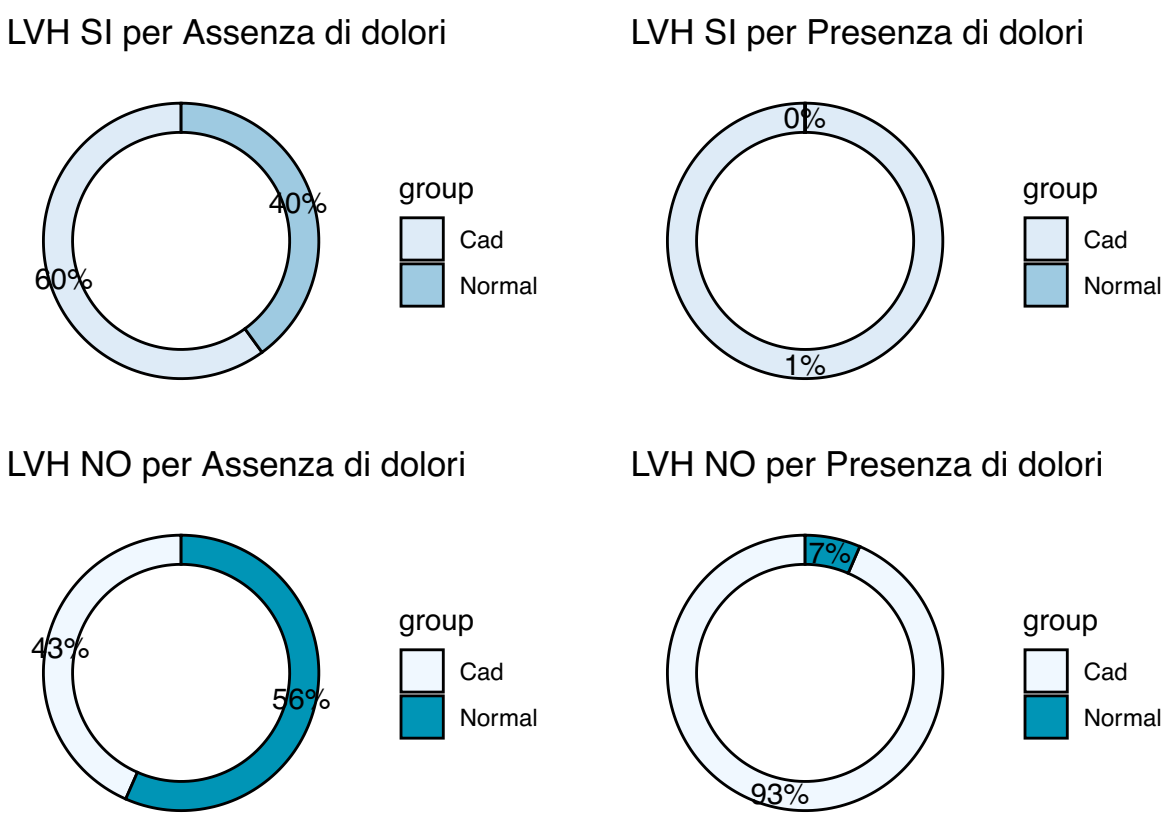
Dalla Figura emerge una predominanza significativa di pazienti che non presentano il diabete mellito . Questa manifestazione è supportata da boxplot più ampi, indicando una maggiore variabilità dei dati. Inoltre, i valori delle mediane differiscono in base al genere e alla presenza di diabete. Per il gruppo di pazienti che non ha il diabete: l'età mediana per le donne che non hanno CAD è intorno ai 65 anni con una distribuzione che va dai 55 ai 75 anni circa; per la stessa categoria ma con diagnosi di malattia l'età mediana è simile, ma la distribuzione è più ampia con più outlier verso gli 80 anni. Per i maschi con diagnosi di malattia normale l'età mediana è intorno ai 55 anni, con una distribuzione più ampia rispetto alle femmine. Per il gruppo di pazienti che sviluppa il diabete: la mediana delle donne , con CAD, è simile ai pazienti senza diabete, ma la distribuzione è più stretta. Mentre per gli uomini l'età mediana è la più bassa tra tutti i gruppi maschili, con una distribuzione ampia e alcuni outlier. Per cui, i pazienti con malattia coronarica, tendono ad essere più anziani rispetto a quelli senza questa patologia. La presenza di diabete mellito sembra influenzare ulteriormente l'età dei pazienti con malattia coronarica, soprattutto tra i maschi. Inoltre, le donne con malattia coronarica tendono ad essere più anziane rispetto agli uomini con la stessa condizione.



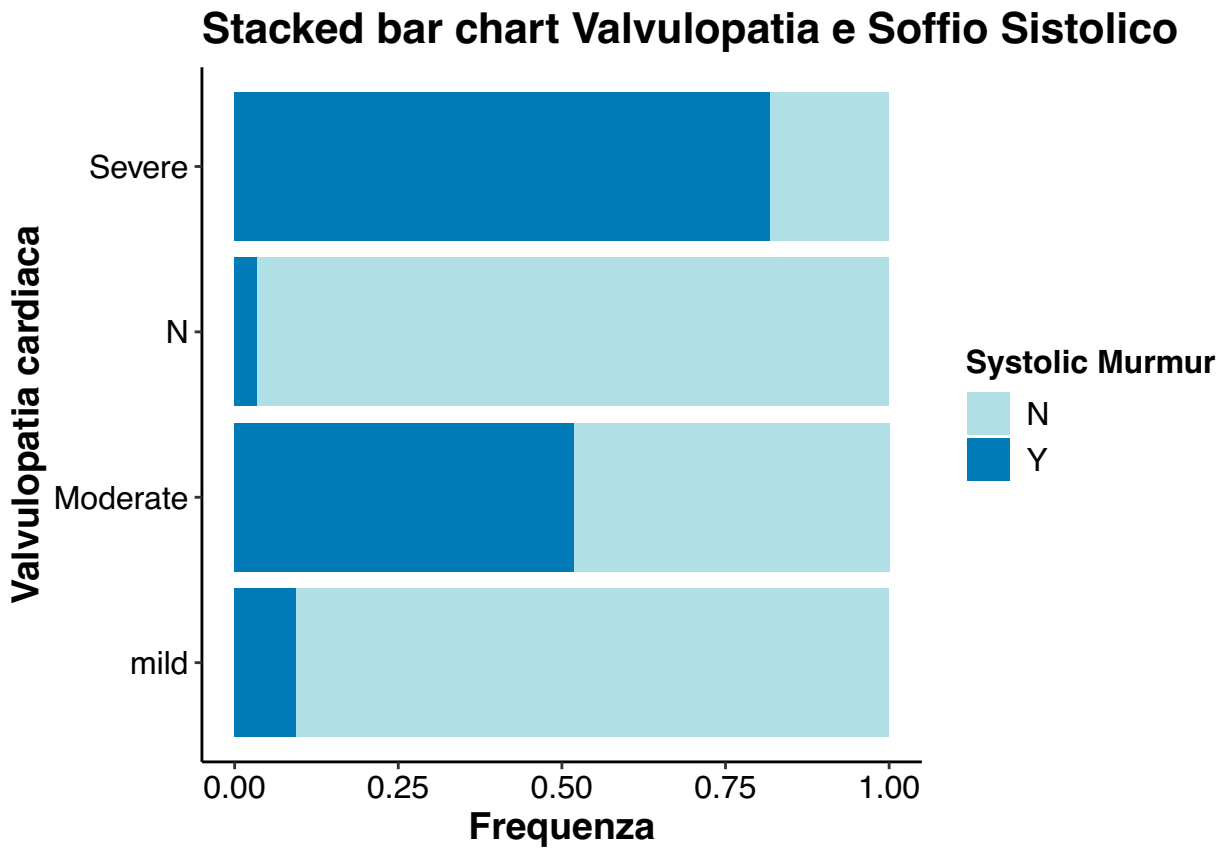
Come da evidenza riportata dall'analisi delle corrispondenze viene investigata la relazione tra pazienti che mostrano sintomi di dolori tipici e presenza di ipertrofia ventricolare. La prevalenza di soggetti che sperimentano dolori tipici è del 54%. I risultati riportano che, la presenza di ipertrofia ventricolare sinistra, implica che la percentuale di pazienti che ottengono la diagnosi di CAD è maggiore nel caso in cui si hanno sintomi di dolori tipici, rispetto ai casi in cui non sono presenti sintomi (1% vs 60%). Di conseguenza, emerge una connessione significativa tra a presenza di sintomi e iperventilazione. Nel migliore dei casi, quando i pazienti non hanno ipertrofia ventricolare e non hanno sintomi, la probabilità di non avere la diagnosi aumenta.

I risultati non sono per nulla in contrasto con la letteratura, l LVH è infatti riconosciuto come un marcatore significativo. Questa condizione è associata ad un aumento del rischio di CAD perché l'ipertrofia ventricolare sinistra può essere una risposta all'ipertensione, che è un noto fattore di rischio per CAD; oltrechè, l'aumento della massa ventricolare richiede più ossigeno e sangue, ma le arterie coronarie potrebbero non essere in grado di fornire abbastanza sangue, specialmente se sono già parzialmente ostruite.[3]

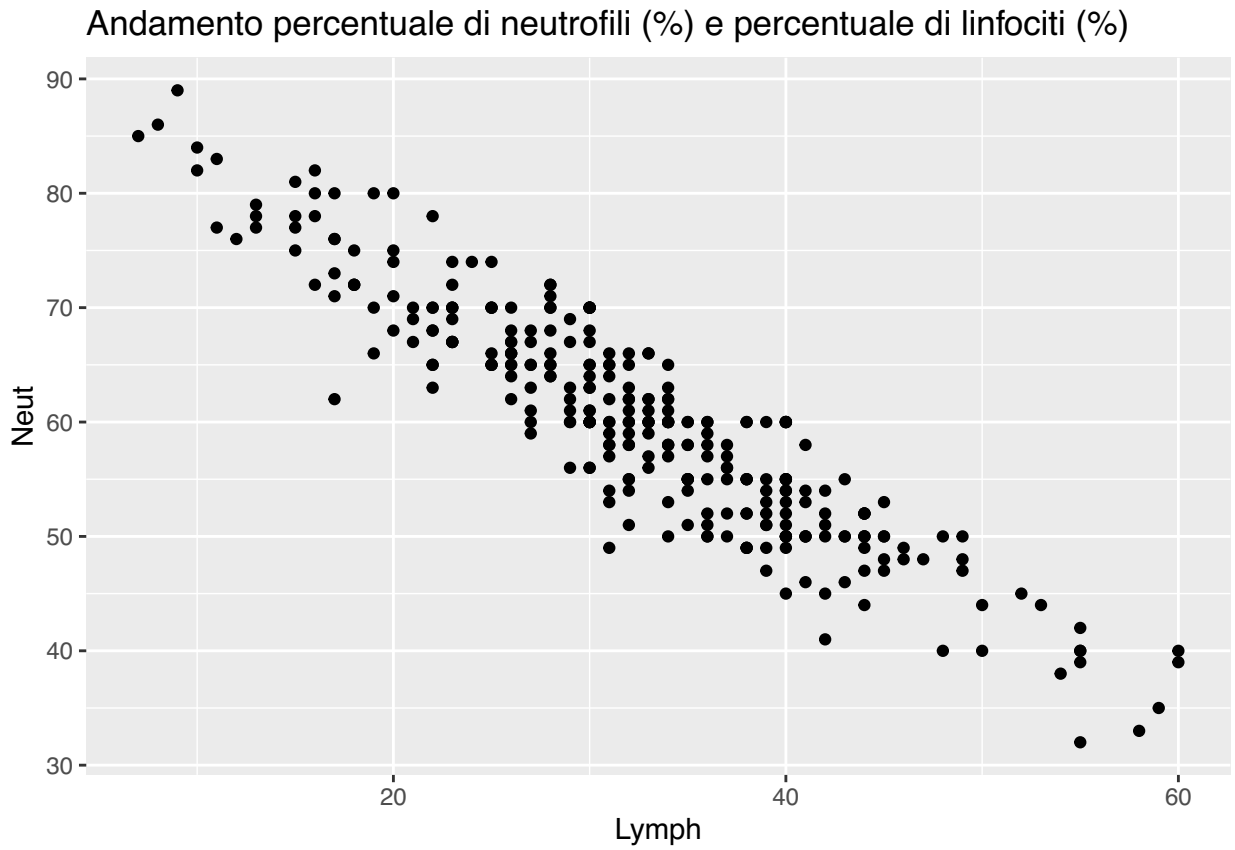
Ancora, i sintomi di dolori toracici tipici, spesso descritti come un'oppressione o dolore al petto che può irradiarsi a braccia, collo o schiena, sono un segno classico di ischemia cardiaca, che è spesso causata da CAD. La presenza di questi sintomi aumenta significativamente la probabilità di diagnosi di CAD perché i sintomi indicano che il cuore potrebbe non ricevere abbastanza ossigeno, suggerendo una possibile ostruzione nelle arterie coronarie.[4]



Ancora, la Figura fornisce un grafico a barre sovrapposte che visualizza le frequenze di soffio sistolico rispetto alle diverse categorie di valvulopatia cardiaca. Le categorie di valvulopatia sono rappresentati come righe orizzontali nel grafico. La lunghezza totale di ogni riga rappresenta il 100% dei pazienti coinvolti nello studio. Le differenze nella lunghezza dei segmenti colorati all'interno di ogni riga riflettono le diverse frequenze associate alle diverse categorie. La categoria *Severe* presenta una frequenza di soffio sistolico più elevata, apri all'80% della stessa. Al contrario, la frequenza più bassa è mostrata nella categoria *N*, con quasi il 93% di pazienti senza soffio sistolico.



La forte correlazione negativa tra neutrofili e linfociti è giustificabile biologicamente, in quanto quando c'è un aumento della percentuale di neutrofili, spesso si osserva una diminuzione relativa della percentuale di linfociti, e viceversa. Questo effetto è parzialmente dovuto al fatto che la formula leucocitaria totale è costante, quindi un aumento di un tipo di cellula implica una diminuzione relativa degli altri tipi.

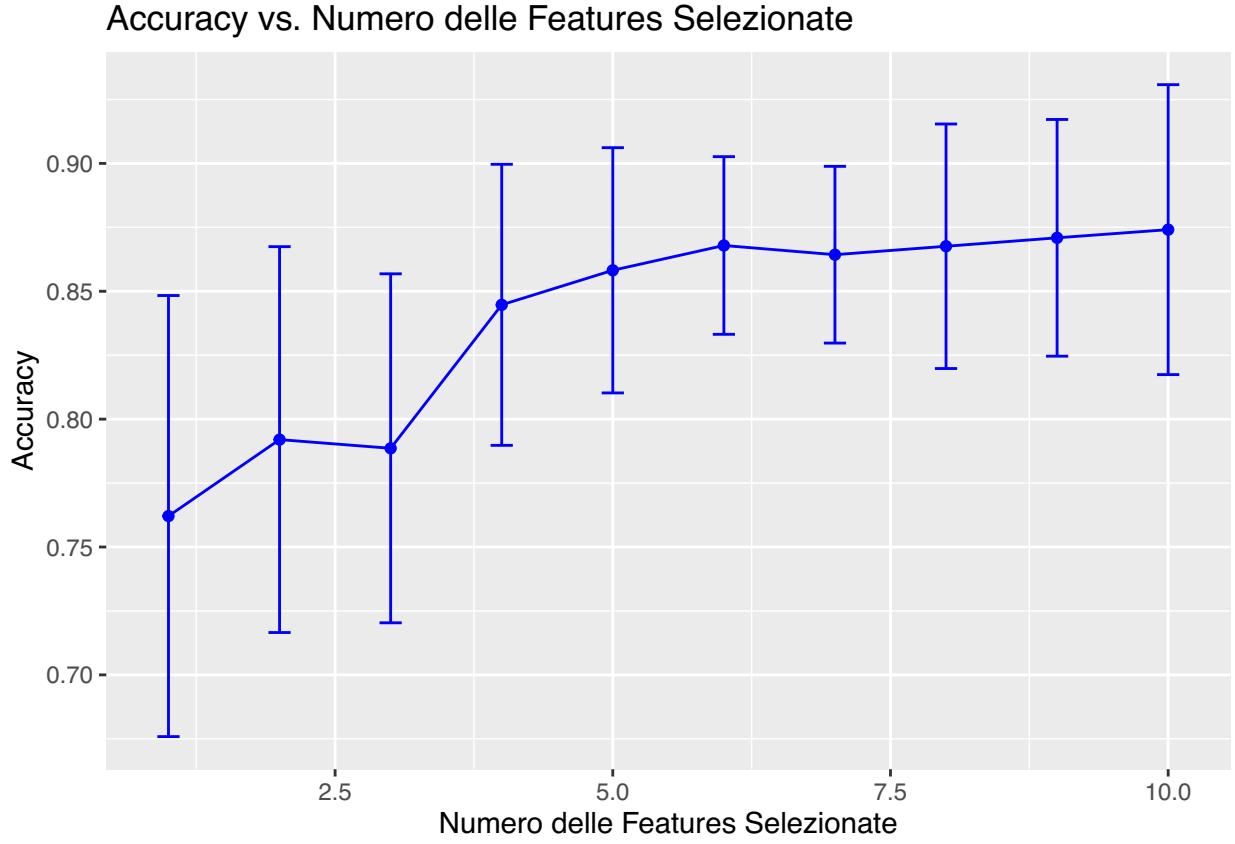


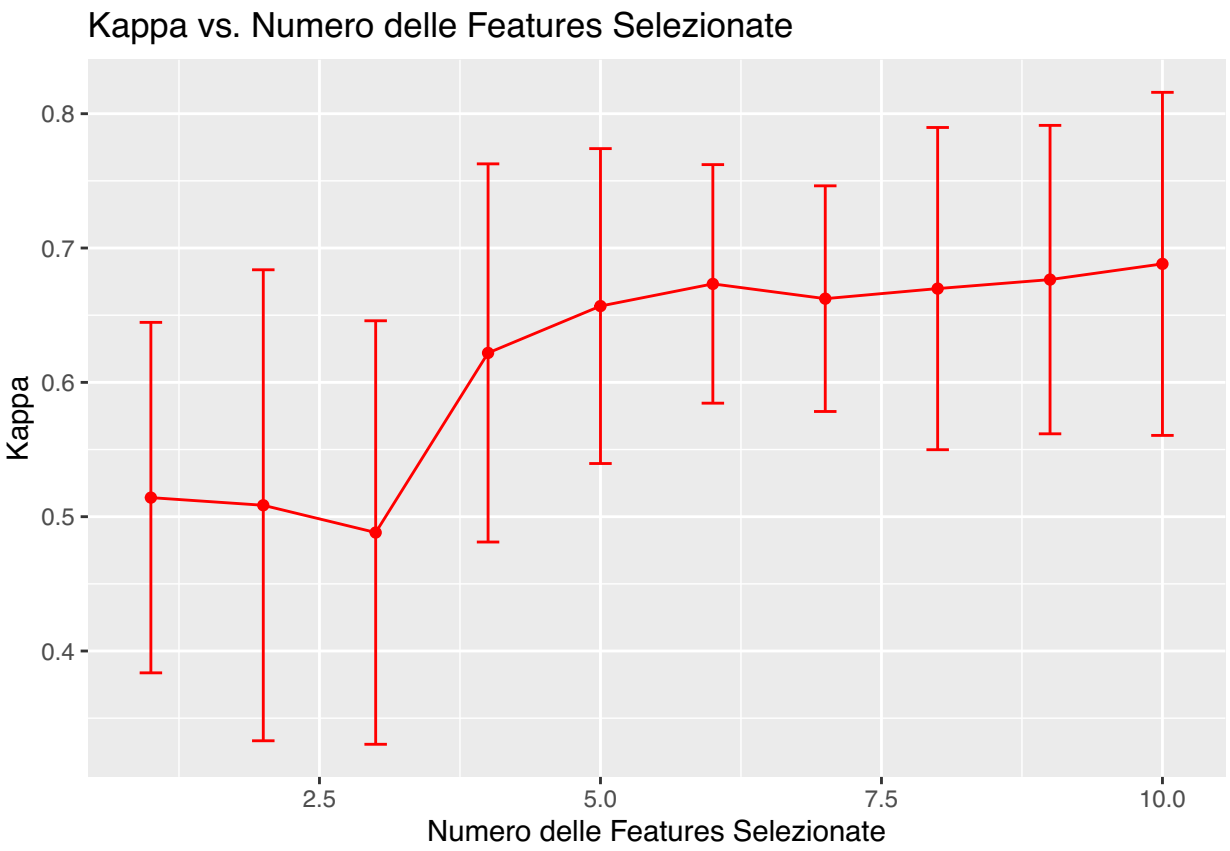
6. Feature Engeeniring: Recursive Feature Elimination

La Recursive Feature Elimination è una tecnica di selezione delle caratteristiche che mira a identificare le caratteristiche più rilevanti per un modello di machine learning. RFE funziona rimuovendo ricorsivamente le caratteristiche meno importanti e costruendo il modello sulle caratteristiche rimanenti. Il processo continua fino a quando non viene raggiunto il numero desiderato di caratteristiche. Viene quindi effettuata una valutazione della rilevanza delle caratteristiche rispetto alla variabile dipendente tramite RFE. I valori di Accuracy e Kappa nella tabella possono essere interpretati come misure della bontà delle caratteristiche selezionate. Più alto è il valore, più le caratteristiche selezionate possono essere rilevanti o informative. Questi sembrano aumentare all'aumentare del numero di caratteristiche selezionate, fino a stabilizzarsi a un certo punto nel processo di selezione delle caratteristiche. La deviazione standard dell'accuratezza e del coefficiente di kappa tende a diminuire all'aumentare del numero di caratteristiche selezionate, indicando una maggiore coerenza nei risultati.

```
##
## Recursive feature selection
##
## Outer resampling method: Cross-Validated (10 fold)
##
## Resampling performance over subset size:
##
##  Variables Accuracy  Kappa AccuracySD KappaSD Selected
##          1   0.7625 0.5098    0.08828  0.1660
##          2   0.7830 0.5162    0.09908  0.1795
```

```
##          3   0.8289 0.5772    0.05095  0.1144
##          4   0.8481 0.6338    0.04634  0.1075
##          5   0.8552 0.6421    0.06948  0.1680
##          6   0.8520 0.6343    0.06410  0.1604
##          7   0.8487 0.6228    0.05451  0.1327
##          8   0.8616 0.6610    0.04972  0.1166
##          9   0.8619 0.6568    0.05133  0.1191
##         10   0.8717 0.6804    0.04836  0.1107
##         42   0.8754 0.6697    0.04865  0.1316        *
##
## The top 5 variables (out of 42):
##    Typical Chest Pain, Age, Atypical, Region RWMA, EF-TTE
```





7. Data Pre-processing

Le reti neurali tendono a ottenere prestazioni migliori quando le variabili sono standardizzate. Questo perché la standardizzazione dei dati consente di portare tutte le variabili su una scala comune, eliminando così eventuali disparità di scala tra di esse. Inoltre, la standardizzazione accelera il processo di addestramento e favorisce una convergenza più rapida durante l'ottimizzazione dei parametri della rete.

Per standardizzare le variabili, si utilizza la funzione scale(), che ridimensiona i dati in modo che abbiano una media zero e una deviazione standard unitaria. Questo processo assicura che le variabili abbiano una distribuzione simile, il che semplifica il processo di apprendimento per la rete neurale. Oltre ciò è necessario formattare diversamente le variabili che hanno lo spazio nel nome.

8. Model fitting

Nella parte di Model Fitting, verranno processati i dati e creati dei modelli predittivi con l'annessa valutazione degli iperparametri. Verranno costruiti i seguenti modelli : reti neurali, random forest, boosting, modello logistico, knn. Sussessivamente verrà effettuato un confronto sulle metriche e verrà scelto il miglior modello da analizzare nella parte successiva.

L'obiettivo principale dei modelli parametrici è ottenere delle stime significative e valide che contribuiscano alla comprensione delle relazioni causa-effetto specifiche nella popolazione. La perseguibilità dell'obiettivo tiene conto dell'individuazione di risultati non necessariamente generalizzabili ma piuttosto validi per un sottoinsieme ristretto della popolazione. La variabile dipendente Carth assume valore CAD se il paziente ha ottenuto diagnosi di malattie coronariche, Normal altrimenti. È stato effettuato uno split del dataset originale

in training set, contenente l' 80% delle osservazioni, e test set, contenente il restante 20%. La motivazione alla base di questa scelta è che la distinzione in due set di dati consente di valutare le prestazioni del modello in modo imparziale e oggettivo. Dopo aver addestrato il modello sul training set, esso viene testato sul test set per valutare come si comporta su dati non osservati durante l'addestramento, fornendo una stima realistica delle prestazioni del modello. L'obiettivo principale è sviluppare un modello che sia in grado di generalizzare bene, ovvero, fare previsioni accurate su nuovi dati non ancora osservati. Inoltre, questo approccio aiuta a selezionare il modello che offre le migliori prestazioni e che sarà più generalizzabile per i dati futuri.

8.1 Modello logistico con stepwise selection

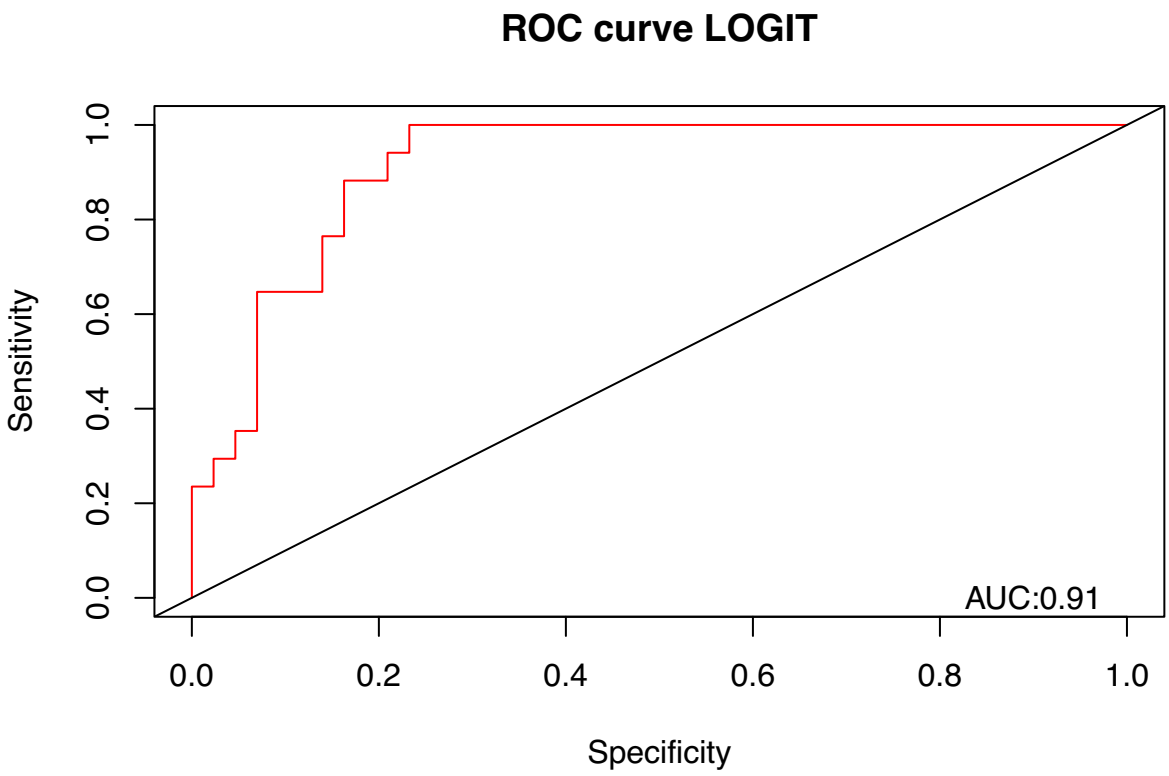
Il primo modello analizzato è il logistico con variable selection implementata con il metodo della stepwise forward. Questo metodo ha selezionato, tra le variabili disponibili, la presenza/assenza di dolori tipici, le regioni con anomalie, l'età, la presenza/assenza di inversione della T nell'ecocardiografia, diabete, presenza pregessa di malattie cardiache familiari, il sesso biologico, il livello di tigliceridi, frequenza del polso, livello di azoto, presenza/assenza di onda nell'elettrocardiogramma e ipertensione. Per la valutazione dell'importanza e della significatività dei coefficienti stimati del modello viene effettuato il test di Wald sui singoli parametri, confrontando l'ipotesi H_0: il coefficiente associato a una variabile sia statisticamente diverso da 0, contro l'alternativa H_1: il coefficiente è statisticamente diverso da 0. Il valore del test viene quindi espresso dal p-value, fissata una soglia critica di livello α=0.05, risultano essere statisticamente significative tutte le variabili , a eccezione di Q wave e BUN, permettendo il rifiuto dell'ipotesi H_0. Tutte le metriche di riferimento verranno raccolte nella sezione utile al confronto tra modelli. Il valore di AUC è 0.91.

```
##
## Call:
## glm(formula = Step5.reg$formula, family = binomial(link = "logit"),
##     data = dat_train, x = TRUE)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.777e+01  4.816e+00   3.689 0.000225 ***
## Typical_Chest_Pain1 -4.189e+00  7.584e-01  -5.523 3.33e-08 ***
## Region_RWMA         -1.895e+00  6.615e-01  -2.865 0.004172 **
## Age                 -1.589e-01  3.586e-02  -4.431 9.38e-06 ***
## Tinversion1         -3.076e+00  8.679e-01  -3.544 0.000394 ***
## DM1                 -2.889e+00  9.135e-01  -3.162 0.001567 **
## FH1                 -4.119e+00  1.093e+00  -3.769 0.000164 ***
## PR                  -1.265e-01  4.073e-02  -3.105 0.001901 **
## TG                  -1.115e-02  4.091e-03  -2.726 0.006408 **
## SexMale             -2.459e+00  8.049e-01  -3.055 0.002253 **
## HB                   4.666e-01  2.196e-01   2.125 0.033617 *
## Q_Wave1             -1.606e+01  1.242e+03  -0.013 0.989677
## HTN1                -1.092e+00  5.775e-01  -1.890 0.058717 .
## BUN                  7.508e-02  4.940e-02   1.520 0.128547
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 291.800  on 242  degrees of freedom
## Residual deviance:  93.717  on 229  degrees of freedom
## AIC: 121.72
##
## Number of Fisher Scoring iterations: 17
```

```
##   McFadden Nagelkerke   CoxSnell
##  0.6788300  0.7974054  0.5574287
```

```
## [1] 121.7173
```

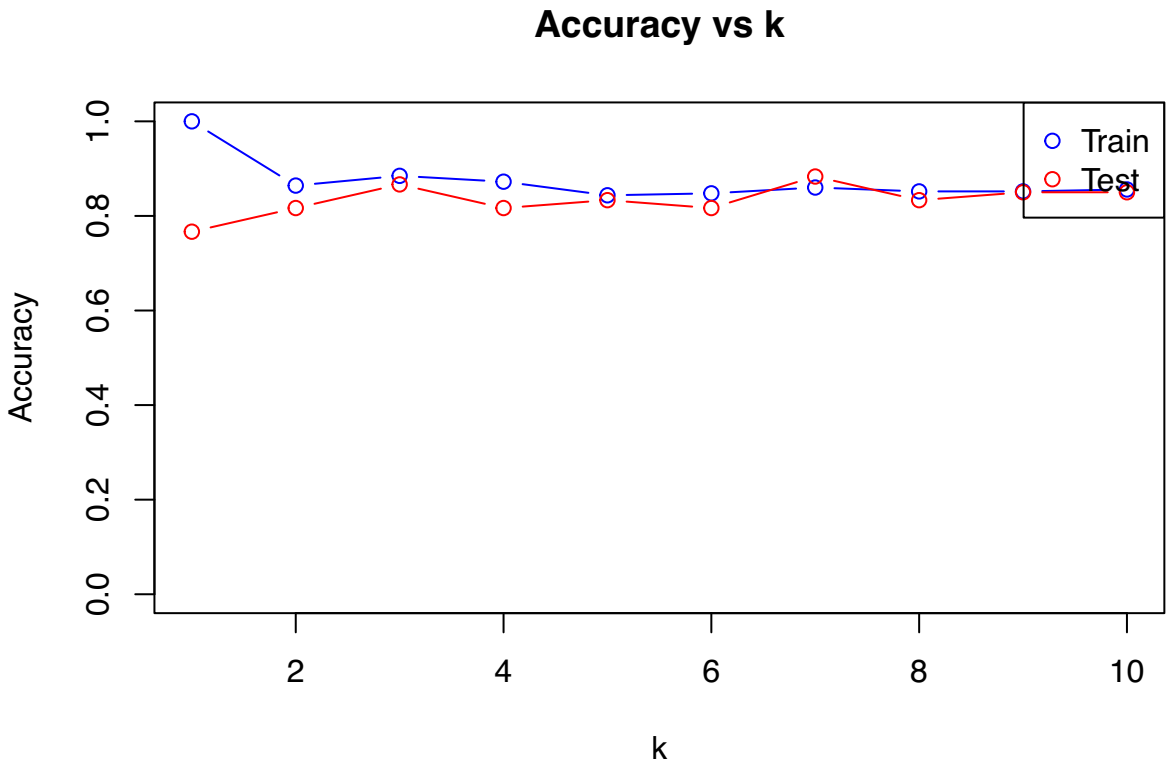
```
## [1] 170.6201
```

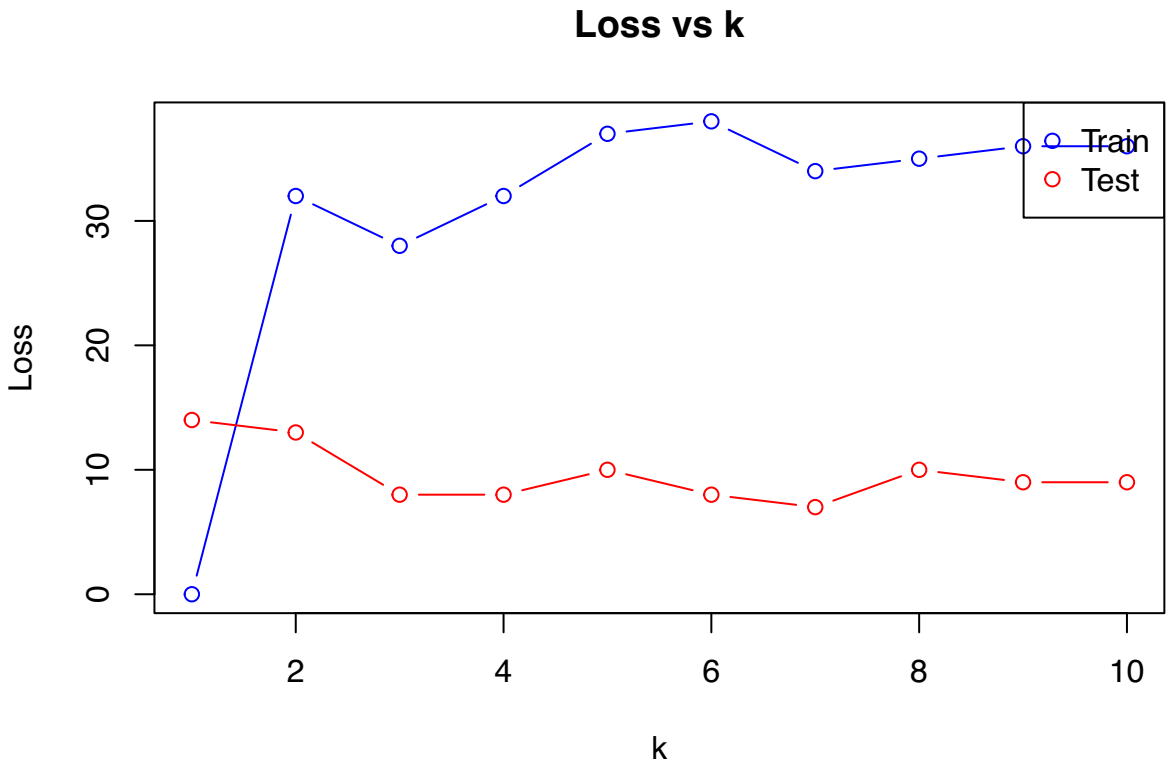


8.2 KNN

Il modello KNN è stato valutato utilizzando diversi valori di k, variando da 1 a 10. Le prestazioni del modello sono state valutate sia sul set di addestramento che sul set di test.

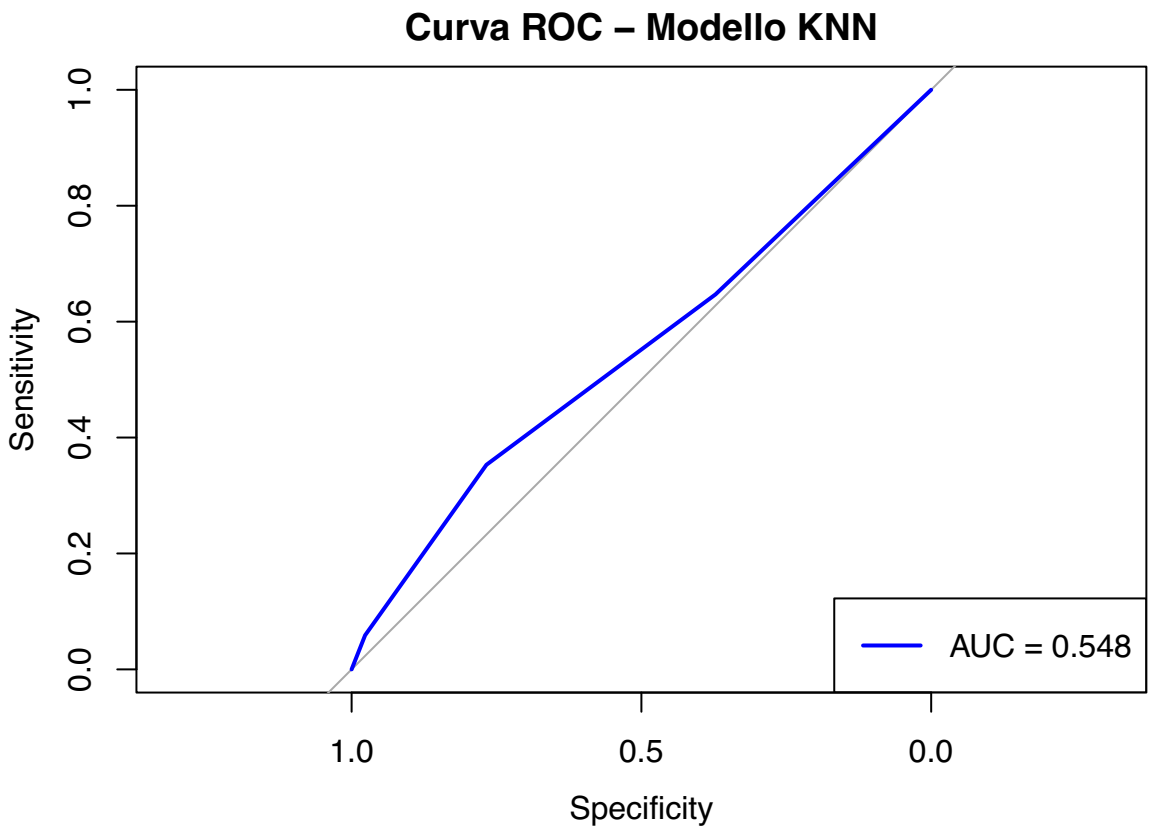
Si nota che il modello presenta un'accuratezza di addestramento più alta rispetto all'accuratezza di test per la maggior parte dei valori di k considerati. Questa discrepanza potrebbe suggerire un problema di overfitting, dove il modello si adatta troppo ai dati di addestramento e non generalizza bene ai nuovi dati. Tuttavia, si osserva che l'accuratezza di test raggiunge il suo valore massimo quando k è pari a 7, il che potrebbe indicare una buona capacità di generalizzazione del modello a questo valore di k. Seppur dal grafico dell'accuratezza sembra esserci una situazione di parità e convergenza, dal grafico della loss sul train e validation set sembra chiaro e evidente che all'aumentare del numero di K si presenti una situazione di overfitting. Seppur la loss sul training diminuisce convergendo sino a mantenersi stabile, mentre la loss sul test set continua a aumentare sintomo di evidenza per cui il modello si sta adattando ai dati troppo e non è in grado di generalizzare.


Accuracy vs k
Train
Test
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
k



Il valore di accuratezza , 0.65%, è il più basso in assoluto.

L' AUC è leggermente migliore rispetto a un classificatore random.



8.3 RF

Il modello Random Forest è stato addestrato su un dataset composto da 243 campioni e 42 variabili predittive, con l'obiettivo di classificare due classi: 'Cad' e 'Normal'. Il modello è stato validato utilizzando una tecnica di cross-validazione a 5 fold, garantendo una robusta valutazione delle sue prestazioni.

Durante la fase di tuning, sono stati valutati diversi valori per il parametro mtry, che rappresenta il numero di variabili selezionate casualmente per la costruzione di ciascun albero nel modello Random Forest. La metrica dell'accuratezza è stata utilizzata per selezionare il modello ottimale, con il valore più alto riscontrato per mtry = 24. Questo valore è stato quindi utilizzato per il modello finale.

```
## Random Forest
##
## 243 samples
##  42 predictor
##   2 classes: 'Cad', 'Normal'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 194, 194, 195, 195, 194
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.7941327  0.3887645
##   24    0.8761905  0.6892750
```

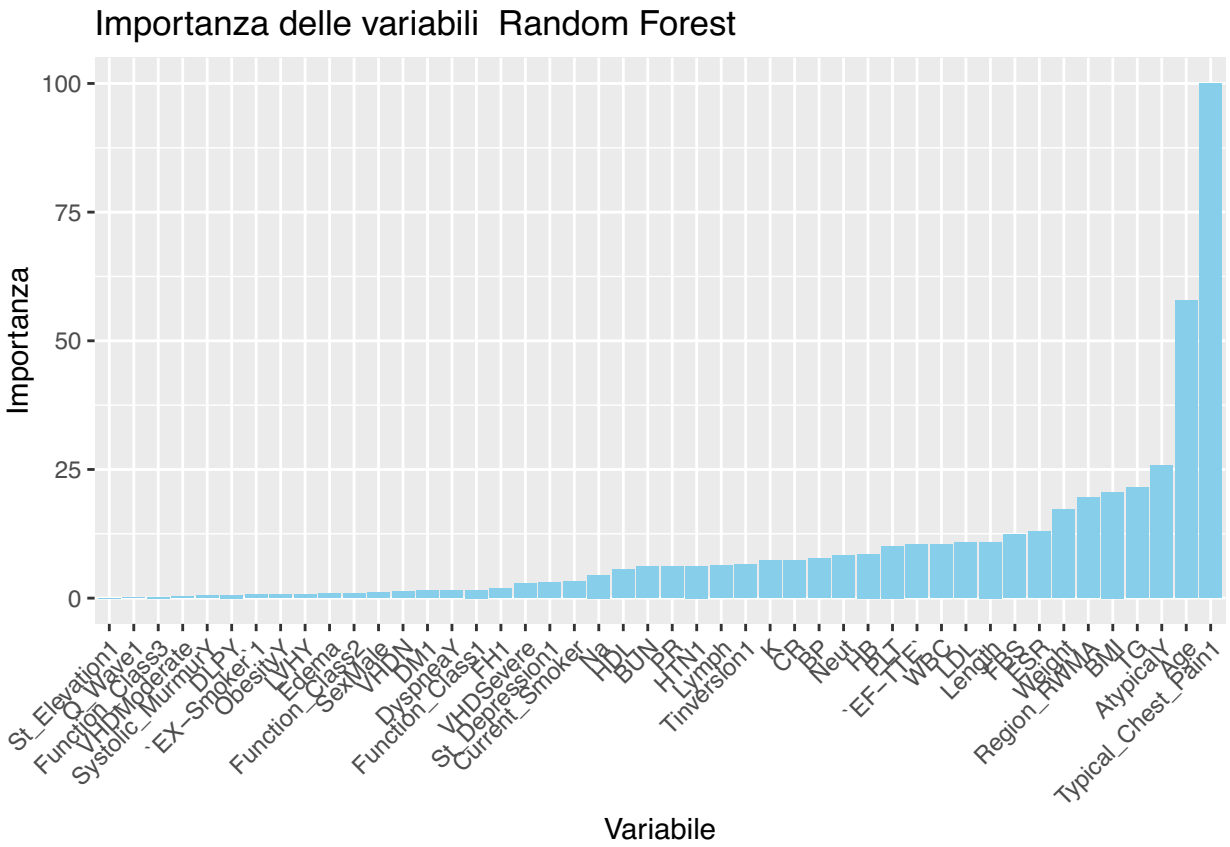
```
##   46    0.8680272  0.6710360
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 24.
```

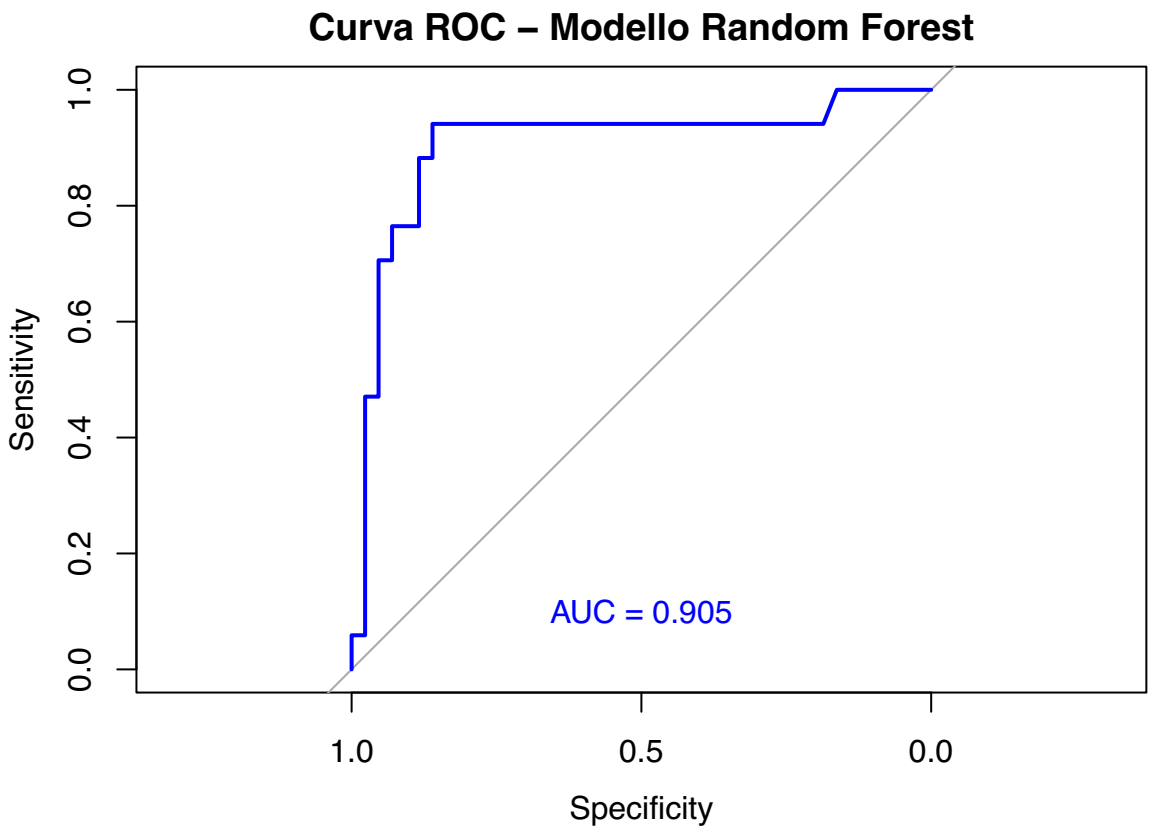
L'accuratezza del modello (0.85) è significativamente superiore al tasso di informazione nullo (No Information Rate) di 0.7167, con un p-value associato di 0.01221, indicando che il modello è statisticamente significativo.

L'importanza delle variabili in un modello Random Forest è una misura che indica quanto ciascuna variabile predittiva contribuisce alla capacità predittiva del modello. In altre parole, è un modo per quantificare l'impatto che ogni variabile ha sulle decisioni prese dal modello. Questa misura è particolarmente utile per comprendere quali fattori sono più influenti nel determinare l'output del modello.

Queste le variabili più importanti:

```
## rf variable importance
##
##   only 20 most important variables shown (out of 46)
##
##                     Overall
## Typical_Chest_Pain1 100.000
## Age                  57.769
## AtypicalY            25.773
## TG                   21.425
## BMI                  20.623
## Region_RWMA          19.620
## Weight               17.156
## ESR                  12.879
## FBS                  12.422
## Length               10.885
## LDL                  10.829
## WBC                  10.430
## 'EF-TTE'             10.401
## PLT                  10.130
## HB                    8.553
## Neut                  8.277
## BP                    7.672
## CR                    7.316
## K                     7.280
## Tinversion1           6.603
```

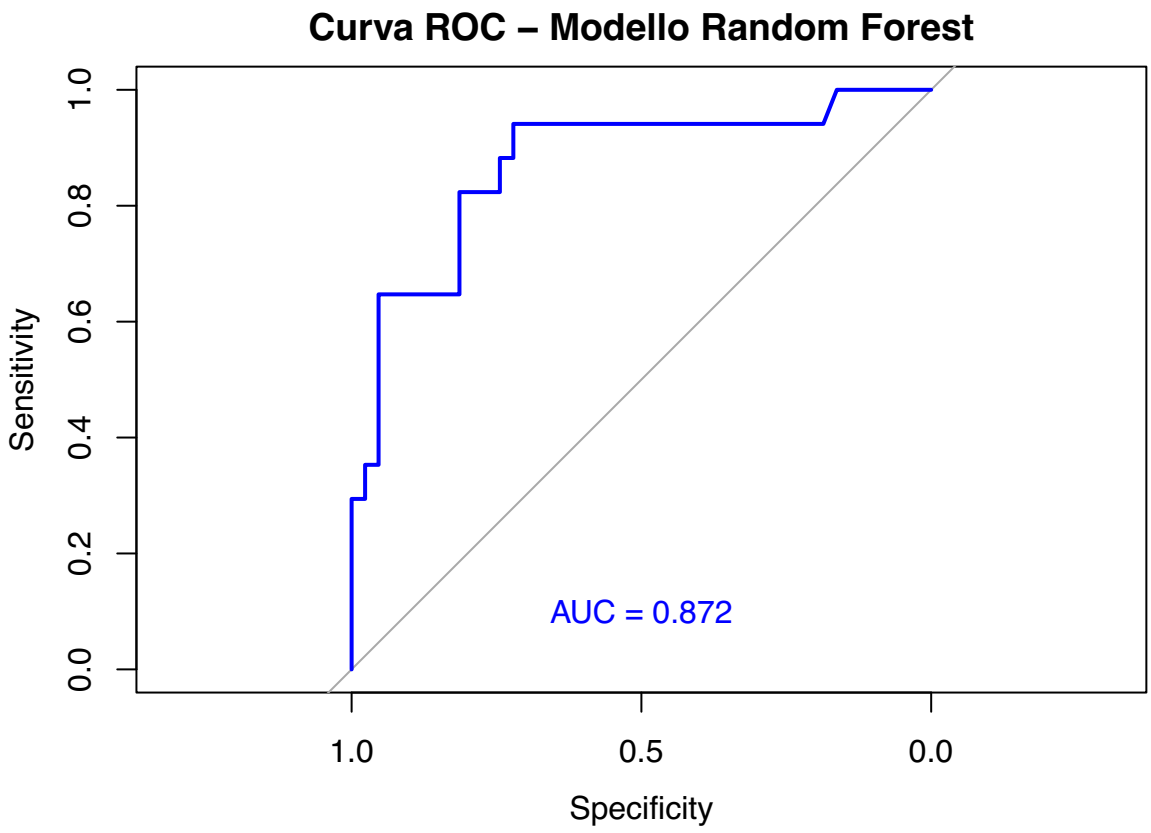

Importanza delle variabili Random Forest
Importanza
100
75
50
25
0
St_Elevation1
Q_Wave1
Function_Class3
VHDModerate
Systolic_Murmur1
DLP1
`EX-Smoker.1
Obesity1
LVH1
Edema
Function_Class2
SexMale
VHDN
DM1
Dyspnea1
Function_Class1
FH1
VHDSevere
St_Depression1
Current_Smoker1
Na
HDL
BUN
PR
HTN1
Lymph
Tinversion1
K
CPK
BP
Neut
HB
PLT
`EF-TTE`
WBC
LDL
Length
FBS
ESR
Weight
Region_RWMA
BMI
TG
Atypical1
Age
Typical_Chest_Pain1
Variabile



8.4 Modello RF 2 : dopo variable importance

Si considerano quindi le variabili più importanti e si costruisce un nuovo modello per comprenderne la capacità di predizione. Utilizzando solo 9 variabili il modello mostra un'accuratezza simile al precedente modello:

```
## Random Forest
##
## 243 samples
##  10 predictor
##   2 classes: 'Cad', 'Normal'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 194, 195, 194, 194, 195
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.8396259  0.5808740
##    6    0.8602041  0.6467625
##   10    0.8396259  0.6009354
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 6.
```



8.5 BOOSTING

Il modello viene valutato tramite cross validation con 5 fold. Durante il processo di tuning, sono stati considerati i seguenti parametri:

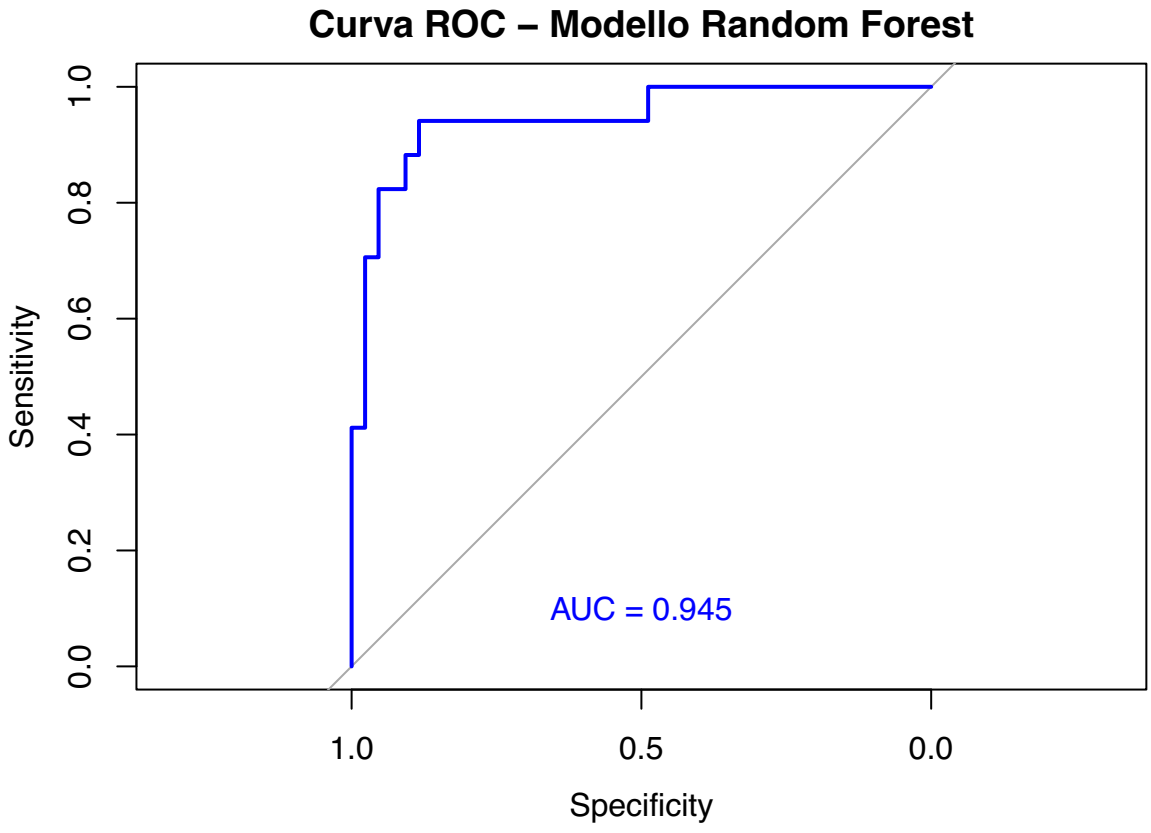
-interaction.depth: Profondità massima di ciascun albero di decisione. -n.trees: Numero totale di alberi nel modello. -shrinkage: Tasso di apprendimento, fissato a 0.1. -n.minobsinnode: Numero minimo di osservazioni richieste in un nodo terminale, fissato a 10.

I risultati della cross-validation con i parametri di tuning mostrano che il miglior modello è stato ottenuto con n.trees = 100 e interaction.depth = 1.

Il modello Stochastic Gradient Boosting ha mostrato un'ottima performance nel prevedere CAD. L'accuratezza del modello sul set di test è risultata essere 0.9, con un intervallo di confidenza al 95% che va da 0.7949 a 0.9624, indicando una performance robusta e stabile. La Sensitivity del modello è estremamente alta (0.9535), suggerendo che il modello è molto efficace nel rilevare i casi di CAD (vero positivo). La Specificity è più bassa (0.7647), ma ancora accettabile, indicando che il modello ha una discreta capacità di identificare correttamente i casi normali (vero negativo). Il Positive Predictive Value di 0.9111 e il Negative Predictive Value di 0.8667 indicano che il modello ha un'elevata precisione nelle sue predizioni, riducendo al minimo i falsi positivi e i falsi negativi.

```
## Stochastic Gradient Boosting
##
## 243 samples
##  42 predictor
##   2 classes: 'Cad', 'Normal'
```

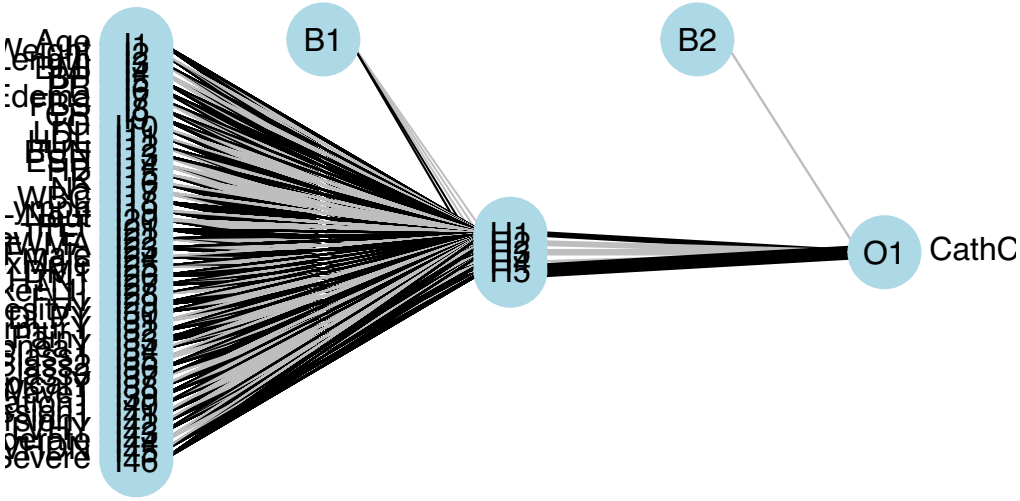
```
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 195, 194, 195, 194, 194
## Resampling results across tuning parameters:
##
##   interaction.depth  n.trees  Accuracy   Kappa
##   1                   50      0.8392857  0.6077324
##   1                  100      0.8473639  0.6283043
##   1                  150      0.8309524  0.5934659
##   2                   50      0.8350340  0.5960971
##   2                  100      0.8225340  0.5771146
##   2                  150      0.8144558  0.5438045
##   3                   50      0.8349490  0.6054328
##   3                  100      0.8227041  0.5650035
##   3                  150      0.8392007  0.6032484
##
## Tuning parameter 'shrinkage' was held constant at a value of 0.1
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were n.trees = 100, interaction.depth =
##  1, shrinkage = 0.1 and n.minobsinnode = 10.
```

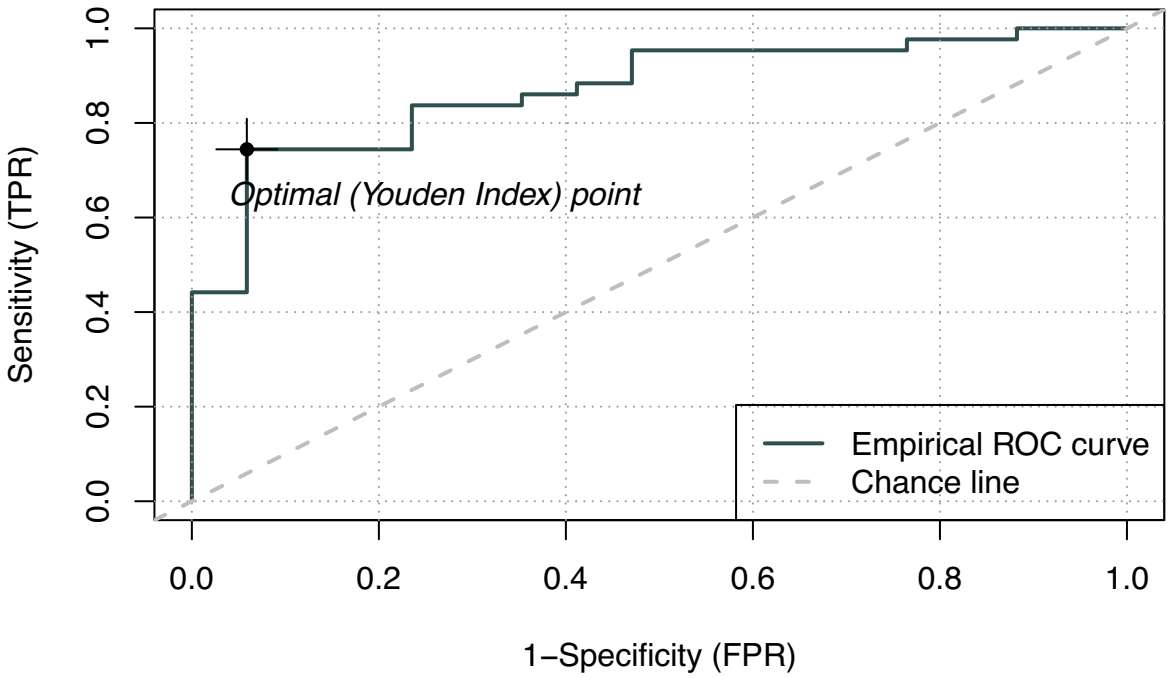


8.6 Rete neurale

La rete neurale utilizzata è una configurazione 30-5-1 con un totale di 150 pesi. Questo significa che ci sono 30 neuroni di input, 5 neuroni nel livello nascosto e 1 neurone di output. La configurazione è quindi una rete neurale single layer feed forward, la funzione utilizzata alla creazione utilizza come metodo di ottimizzazione BFGS.

Il modello è stato configurato con una funzione di perdita di entropia e un parametro di regolarizzazione (decay) impostato a 0.1. Il parametro di decay di 0.1 applica una penalizzazione sui pesi della rete neurale, riducendo la magnitudine dei pesi durante l'addestramento. Questo aiuta a prevenire l'overfitting, migliorando la generalizzazione del modello sui dati di test



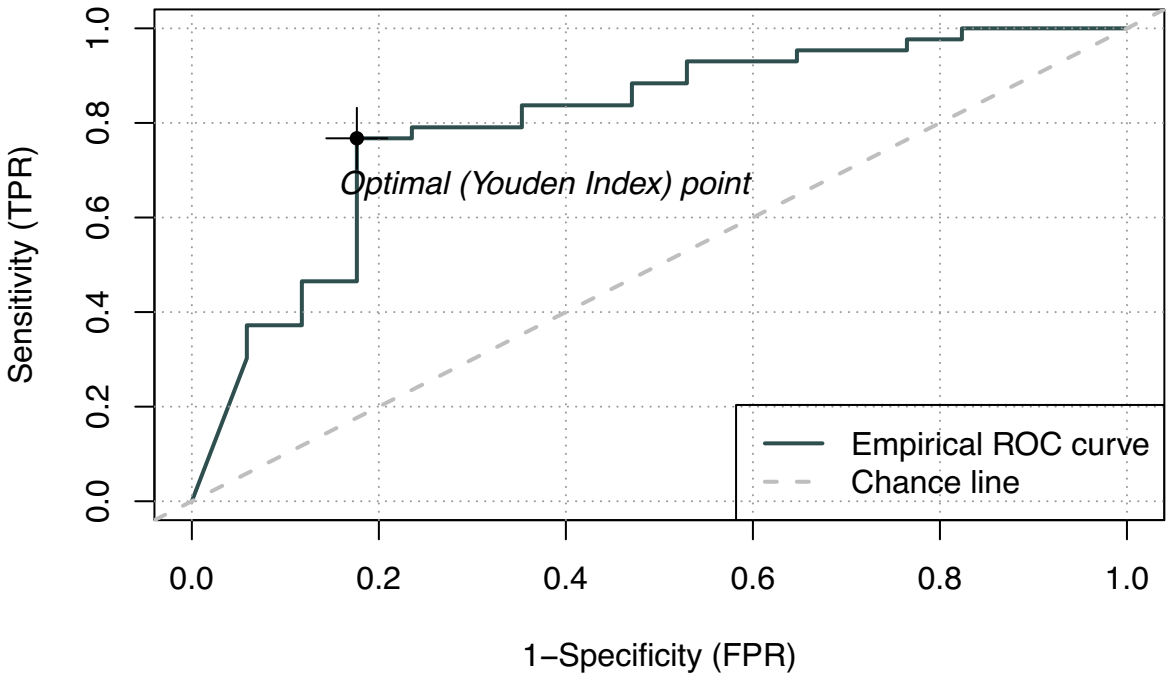


Youden's Index è una misura utilizzata frequentemente in combinazione con l'analisi della curva ROC (Receiver Operating Characteristic) per valutare l'efficacia di un test diagnostico. L'indice varia tra -1 e 1, con un valore massimo di 1 che indica una perfetta capacità discriminatoria del test (cioè, il test identifica correttamente tutti i positivi e tutti i negativi). Il valore massimo di Youden's Index lungo la curva ROC viene utilizzato per selezionare il cut-off ottimale del test diagnostico. Questo punto ottimizza contemporaneamente la sensibilità e la specificità, massimizzando la capacità del test di discriminare tra i casi positivi e negativi.

```
##
##  Method used: empirical
##  Number of positive(s): 43
##  Number of negative(s): 17
##  Area under curve: 0.8714
```

8.7 Rete Neurale con decay 0.01

Viene addestrata la stessa rete con 5 neuroni al livello intermedio, per verificare quanto il modello ha necessità di essere complesso per prevedere bene; in tal caso il decay è stato fissato più basso : 0.01. I risultati, in termini di accuratezza, sono nettamente migliori rispetto alla rete neurale con lo stesso numero di neuroni ma con una penalizzazione più alta.



8.8 Rete Neurale con CV

Si può pensare di migliorare le performance della rete neurale con l'utilizzo della K-fold Cross-Validationcon k=5, in cui si prendono in considerazione i seguenti iperparametri:

- Numero di neuroni nello strato nascosto: Consideriamo una sequenza di neuroni da 1 a 10
- Decadimento dei pesi: si considera 0.5, 0.1, 0.01, 0.001, 0.02
- Numero di epoche di apprendimento: si considerano 100,200 e 500

Per ogni combinazione degli iperparametri considerati, viene stimata una rete neurale. Al termine del processo, si otterranno k metriche di performance, ciascuna relativa ai k passaggi effettuati. Le k metriche di performance vengono aggregate per ottenere una singola metrica finale. Alla fine, viene selezionata la combinazione di iperparametri che garantisce la migliore accuratezza nella cross-validation.

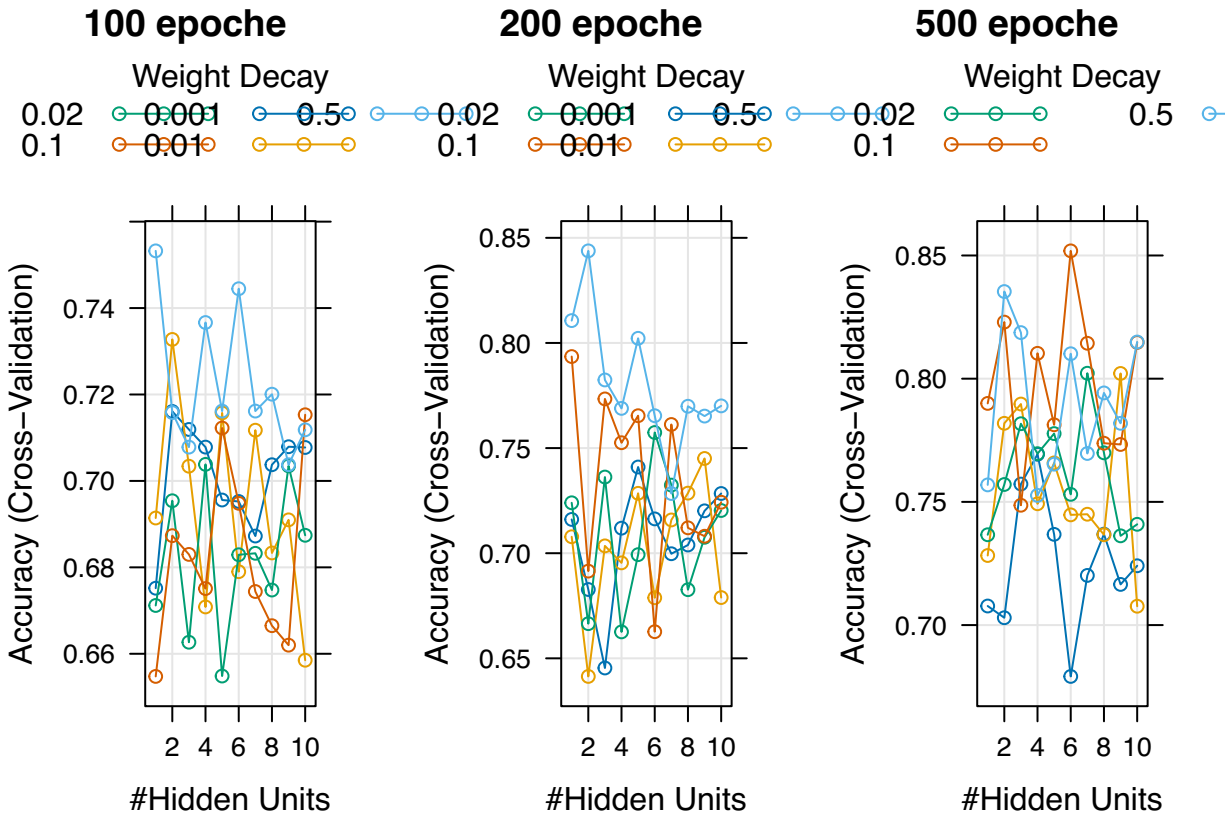
Il grafico mostra l'accuratezza della cross-validation per la rete neurale in funzione del numero di unità nascoste, del decadimento del peso e del numero di epoche. Ci sono tre pannelli, ciascuno rappresenta un numero diverso di epoche: 100, 200 e 500. All'interno di ogni pannello, sono tracciate diverse curve che rappresentano differenti valori di decadimento del peso: 0.001, 0.01, 0.02, 0.1 e 0.5.

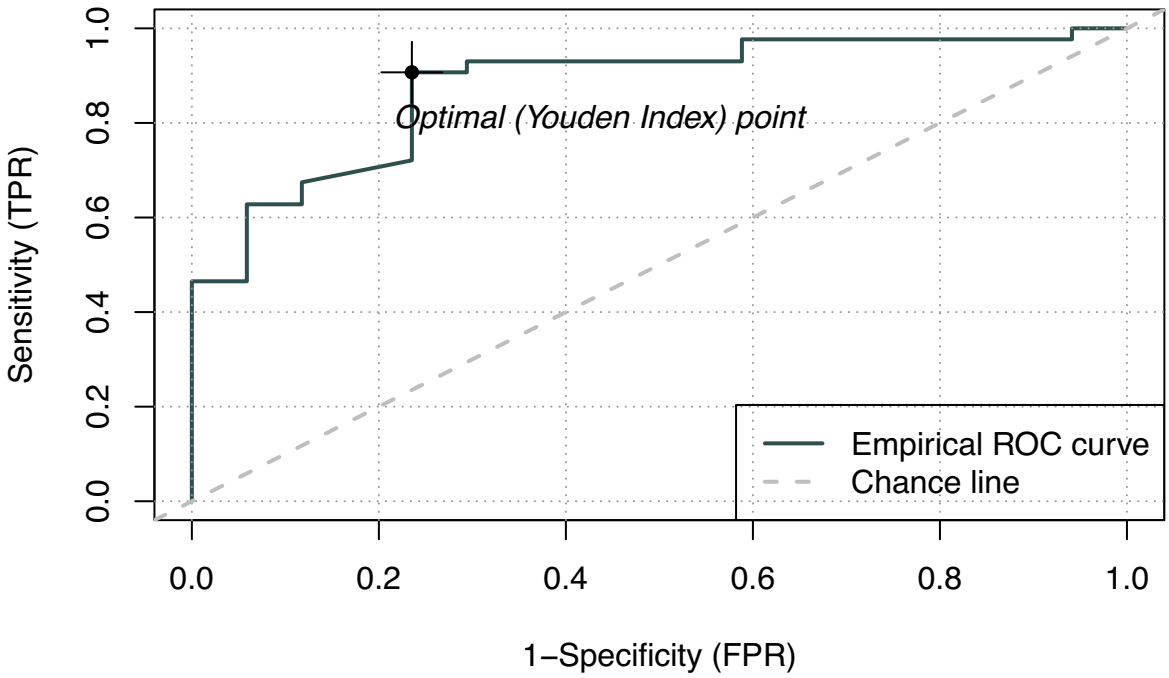
Osservando il primo pannello, relativo alle 100 epoche, si nota una considerevole variabilità nell'accuratezza a seconda del numero di unità nascoste e del valore di decadimento del peso. Non emerge una chiara tendenza, sebbene i valori di decadimento più bassi (0.001 e 0.01) sembrano avere performance relativamente migliori rispetto agli altri.

Nel secondo pannello, per 200 epoche, l'accuratezza mostra una maggiore stabilità e tende a essere più alta rispetto al pannello precedente. Anche qui, i valori di decadimento più bassi continuano a performare meglio, in particolare 0.001 e 0.01, che raggiungono picchi di accuratezza più elevati.

Infine, nel pannello delle 500 epoche, si osserva una situazione simile, con una performance complessiva che si stabilizza ulteriormente. Tuttavia, la variabilità è ancora presente, suggerendo che l'aumento del numero di epoche non elimina completamente l'influenza del numero di unità nascoste e del valore del decadimento del peso sull'accuratezza. Ancora una volta, i valori di decadimento del peso più bassi sembrano offrire le migliori performance in termini di accuratezza.

Il miglior modello , in termini di accuratezza sembra essere size=2 e decay=0.01, n di epoche 200.






```
##
##  Method used: empirical
##  Number of positive(s): 43
##  Number of negative(s): 17
##  Area under curve: 0.8769
```

9. Confronto fra modelli

L'obiettivo principale del seguente paragrafo è la valutazione delle prestazioni e delle caratteristiche dei modelli implementati precedentemente, tramite confronto, per l'identificazione del modello più adatto allo scopo della ricerca. La definizione del modello tiene conto delle premesse fatte inizialmente, ovvero la possibilità che le previsioni siano valide per i soggetti della popolazione di riferimento e non necessariamente generalizzabili. L'utilità del modello è principalmente proiettata per la pratica medica e la previsione della malattia per i pazienti (diagnosi, stima prognostica e supporto decisionale). Per effettuare un confronto accurato, sono stati presi in considerazione diversi criteri, tra cui l'adeguatezza del modello alle specifiche della popolazione di studio, la precisione delle stime dei parametri, e la capacità predittiva tramite l'utilizzo di indici.

L'efficacia dei modelli e l'accuratezza delle previsioni sono valutate attraverso l'utilizzo di metriche quali accuratezza (acc), precisione, recall, f1 score e Area Under the Curve (AUC). Dati i risultati, emerge che il modello di Boosting con cross validation a 5 fold è il più performante, con un'accuratezza (acc) del 93% e una AUC di 0.94. Questo indica una capacità eccellente di distinguere tra le classi e un'elevata accuratezza nelle previsioni. Il modello dimostra una precisione e un recall entrambe pari a 0.93 e 0.93 rispettivamente, risultando in un f1 score di 0.92. Il Random Forest con importanza delle variabili mostra anche esso buone prestazioni, con una recall del 93% e una AUC di 0.90, suggerendo un buon equilibrio tra rilevamento dei veri

positivi e minimizzazione dei falsi negativi. La regressione logistica con selezione stepwise e Random Forest con validazione incrociata a 5 fold presentano un'accuratezza simile del 85%, ma differiscono leggermente nelle altre metriche. La regressione logistica ha una precisione particolarmente alta (0.94), indicando un basso numero di falsi positivi, mentre il Random Forest con CV5 mostra una recall più alta (0.90), suggerendo una maggiore capacità di identificare i veri positivi. Il modello KNN si posiziona come il meno performante, con un'accuratezza del 65% e una AUC di 0.54, indicando difficoltà nel distinguere correttamente tra le classi.

Per quanto riguarda i modelli di Rete Neurale, il modello con la validazione incrociata a 5 fold risulta il più performante tra essi, con un'accuratezza del 87% e una AUC di 0.90. Gli altri modelli di Rete Neurale mostrano prestazioni inferiori, con accuratezze tra 78% e 80% e valori di AUC compresi tra 0.66 e 0.78.

In conclusione, il modello di Boosting con CV a 5 fold è il più efficace in termini di bilanciamento tra accuratezza, precisione, recall e AUC, dimostrando una capacità superiore di classificare correttamente i dati rispetto agli altri modelli analizzati, per cui qualora si avesse necessità di spiegare le interazioni tra le variabili è il modello più adatto.

Tuttavia, sebbene non sia il modello con le prestazioni complessive più elevate in termini di accuratezza e AUC, il modello di Rete Neurale con size 5 e decay 0. 1 è particolarmente adatto in scenari dove è fondamentale minimizzare i falsi negativi per garantire una diagnosi tempestiva della malattia. Questo approccio è preferito quando l'obiettivo principale è garantire che il maggior numero possibile di veri positivi sia identificato, migliorando così le possibilità di intervento precoce e trattamento efficace.

Performance dei Modelli

Model	Approach	acc	precision	recall	f1	auc
Logistic Regression	Stepwise Selection	0.8500	0.9400	0.8300	0.8800	0.9100
KNN	CV 5 Folds	0.6500	0.7500	0.7600	0.7500	0.5400
Random Forest	CV 5 Folds	0.8100	0.8400	0.9000	0.8700	0.9000
Random Forest	Variable Importance	0.8500	0.8900	0.9300	0.8900	0.8700
Boosting	CV 5 Folds	0.9300	0.9300	0.9300	0.9200	0.9400
Rete Neurale	Size 5, Decay 0.1	0.8000	0.7000	0.9400	0.7200	0.7800
Rete Neurale	Size 5, Decay 0.01	0.7800	0.5800	0.8200	0.6800	0.6600
Rete Neurale con CV	CV 5 Folds	0.8700	0.6744	0.7700	0.7700	0.9000

10. SHAP (SHapley Additive exPlanations)

SHAP sta per SHpley Additive ExPlanations. È un modo per calcolare l'impatto di una funzionalità sul valore della variabile target. L'idea è di considerare ogni caratteristica come un giocatore e il set di dati come una squadra. Ogni giocatore dà il proprio contributo al risultato della squadra. La somma di questi contributi dà il valore della variabile target dati alcuni valori delle caratteristiche (cioè dato un particolare record).

Il concetto principale è che l'impatto di una caratteristica non si basa solo sulla singola caratteristica, ma sull'intero insieme di caratteristiche del set di dati. Pertanto, SHAP calcola l'impatto di ogni caratteristica sulla variabile target (chiamata valore shap) utilizzando il calcolo combinatorio e riqualificando il modello su tutta la combinazione di caratteristiche che contiene quella che stiamo considerando. Il valore medio assoluto dell'impatto di una caratteristica rispetto a una variabile target può essere utilizzato come misura della sua importanza.

Il grafico degli shap aiuta a visualizzare l'importanza relativa delle variabili nel modello. Ogni barra nel grafico rappresenta l'importanza della variabile corrispondente nei valori SHAP, per cui come ciascuna caratteristica influisce sulla predizione del modello, evidenziando sia la direzione sia l'entità del contributo di ciascuna variabile.. Le barre più lunghe indicano una maggiore importanza della variabile nel modello.

L'intercetta è pari a 0.71. Questo valore rappresenta il punto di partenza della predizione del modello, ovvero il valore di base a cui vengono aggiunti (o sottratti) i contributi delle altre variabili per arrivare alla predizione finale.

PLT : - Valore: 742 - Contributo: +0.25 (Bastone verde) - Un valore di PLT pari a 742 contribuisce positivamente alla predizione del modello con un incremento di 0.25. Il bastone verde indica che questa variabile ha un impatto significativo nel senso positivo, suggerendo che alti valori di PLT sono associati a un aumento nella predizione del modello.

ST Depression: - Valore: 1 - Contributo: +0.029 (Bastone verde) - Un valore di ST depression pari a 1 contribuisce con un incremento di 0.029 alla predizione del modello. Anche in questo caso, il bastone verde indica un impatto positivo, suggerendo che la presenza di depressione del segmento ST ha una correlazione positiva con la variabile target del modello.

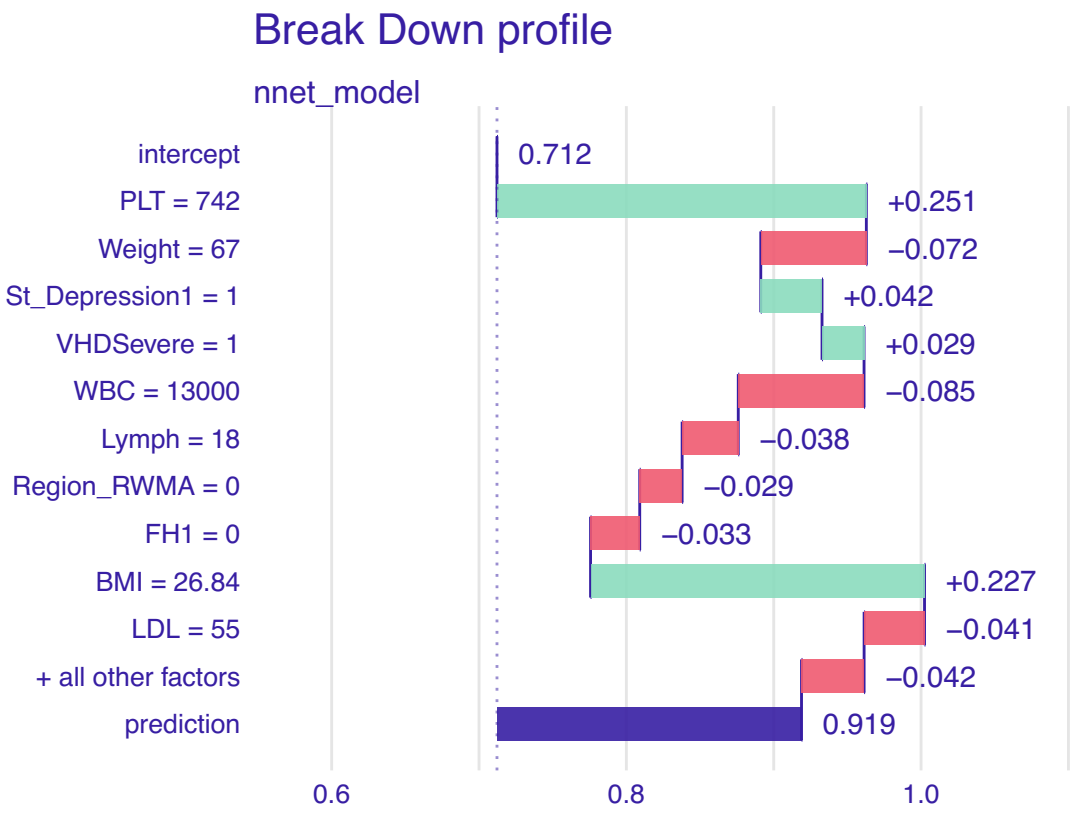
Weight: - Valore: 67 - Contributo: -0.072 (Bastone rosso) - Un peso di 67 contribuisce negativamente alla predizione del modello con una diminuzione di 0.072. Il bastone rosso indica che questa variabile ha un impatto negativo, suggerendo che un peso maggiore è associato a una diminuzione nella predizione del modello.

VHD Severe : - Contributo: +0.029 (Bastone verde) - La presenza di VHD Severo contribuisce positivamente alla predizione del modello con un incremento di 0.029. Questo suggerisce che la presenza di una grave malattia valvolare cardiaca è associata a un aumento nella predizione del modello.

Lymph: - Valore: 18 - Contributo: -0.03 (Bastone rosso) - Un livello di linfociti pari a 18 contribuisce negativamente alla predizione del modello con una diminuzione di 0.03. Il bastone rosso indica un impatto negativo, suggerendo che un basso numero di linfociti è associato a una diminuzione nella predizione del modello.

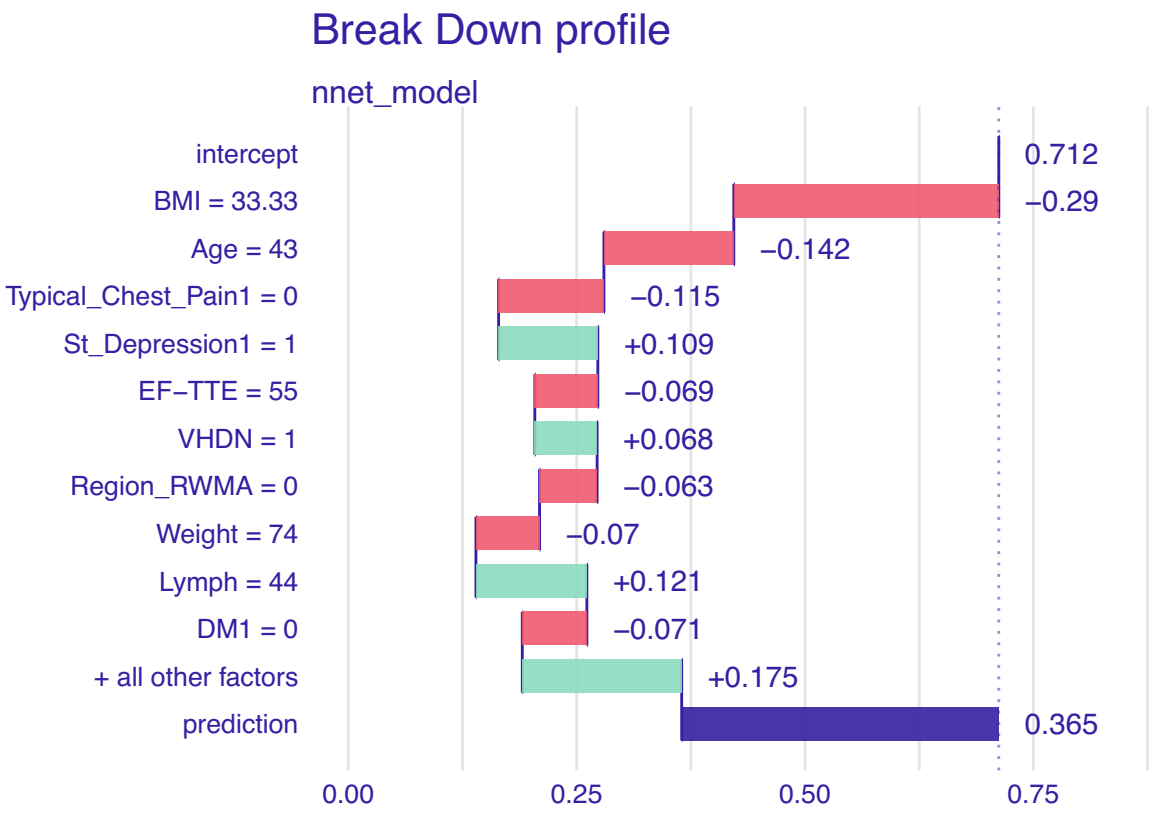
I risultati sono importanti perchè permettono di ottenere un certo grado di certezza circa la previsione oltre che dare consigli su come procedere. In particolare, nel caso specifico, se procedere con angiografia invasiva o ripetere stress test. Per il paziente in questione , sulla base della rete neurale, è possibile dare un'analisi dettagliata:

Anche se presenta VHD allo stato severo, non ha precedenze in famiglia circa malattie coronarie , ma è un paziente obeso (BMI>25), i valori del ECG mostrano segni di ischemia (st depression=1) , LDL=25 i globuli rossi sono alti (130000) al limite. Il modello prevede la probabilità di malattia coronarica pari al 91.9%. Si pensa che il paziente dovrebbe sottoporsi ad un'angiografia coronarica come strategia invasiva precoce.



Prendendo come esempio il paziente 8, la situazione è completamente diversa:

Anche se nel ECG sono presenti segni di ischemia (st depression), è obeso BMI>25, ma la malattia delle valvole è normale, non mostra dolori atipici e non ha il diabete, il modello ha previsto la probabilità di malattia coronarica pari al 36%. Si pensa che il paziente non dovrebbe sottoporsi ad un'angiografia coronarica ora ma provare a ripetere stress test per vedere se ci sono segni di ischemia.



11. Conclusioni

L' obiettivo era costruire un modello di previsione della malattia coronarica, confrontare le prestazioni di vari algoritmi di apprendimento automatico e analizzare la spiegabilità dell'algoritmo di apprendimento automatico selezionato con la recall più bassa.

Nella parte Descrizione del set di dati, si è cercato di comprendere ciascuna funzionalità in modo elaborato per avere una maggiore comprensione del dominio.

Nella parte Analisi esplorativa dei dati, ci si è fatta un'idea della distribuzione dei dati ed esaminato le proprietà statistiche di base . Ciò includeva la comprensione dei tipi di variabili, il controllo dei valori mancanti e la visualizzazione della distribuzione di varie funzionalità . Oltre tutto, grazie all'analisi delle corrispondenze multiple e tecniche di feature engeeniring si è cercato di comprendere l'impatto e le associazioni fra le variabili.

Nella parte Analisi predittiva, i dati sono stati preprocessati, sono stati sviluppati modelli predittivi con ottimizzazione degli iperparametri, spesse volte con cross-validation. Sono stati costruiti diversi modelli: logistic regression, random forest, boosting, knn e reti neurali. La rete neurale con la recall più alta è stata scelta per analizzare la spiegabilità.

In SHAP, si è cercato di interpretare il modello di Rete Neurale con il valore SHAP. Per comprendere l'interpretabilità locale del modello, sono stati scelti il paziente n. 1 e il paziente n. 8 dal set di dati di training, tramite l'analisi di shap si è cercato quindi di ottenere un parere medico e definire la probabilità di malattia.

Tuttavia, è importante considerare i seguenti limiti nella valutazione dei risultati e delle conclusioni di questo progetto :

1. Limitazioni dei dati: È possibile che i dati utilizzati per l'analisi siano limitati in termini di dimensione del campione, copertura temporale o completezza delle informazioni. Queste limitazioni possono influire sulla generalizzabilità dei risultati e sulla precisione delle stime ottenute.
2. Mancanza di variabili rilevanti: È possibile che alcune variabili potenzialmente influenti non siano state incluse nell'analisi a causa di limitazioni dei dati o di altre ragioni. La mancanza di queste variabili potrebbe limitare la capacità dei modelli di spiegare completamente la previsione.
3. Potenziale confondimento: Nonostante gli sforzi per controllare le variabili confondenti, potrebbe essere presente un potenziale confondimento non misurato o non considerato nei modelli. Ciò potrebbe influire sugli effetti stimati delle variabili di interesse e sulla validità delle conclusioni.
4. Limitazioni dell'interpretazione causale: nonostante gli sforzi per controllare le variabili confondenti, gli studi osservazionali come questo possono presentare limitazioni nella determinazione di relazioni causali tra le variabili di interesse.

In conclusione, la scelta del modello migliore dipenderà dal tipo di informazione che si vuole estrarre dallo stesso: è stata scelta la rete neurale il cui valore di recall è più alto, in quanto si vuole evitare il più possibile i falsi negativi; i falsi positivi possono essere considerati di minore importanza in ambito medico. In alternativa, qualora dal modello si ha necessità di ottenere informazioni più approfondite circa il legame dei fattori-malattia è possibile scegliere il boosting le cui misure di accuratezza e prestazione sono lo stesso elevate.

Riferimenti bibliografici

[1] https://www.sculati.it/media/sovrappeso_obesita.pdf (pagina 13) [2] https://www.ahajournals.org/doi/10.1161/CIR.0000000000001168#d1e2894 (Capitolo 3) [3] https://www.ahajournals.org/doi/10.1161/JAHA.118.010107#:/~:text=Left%20ventricular%20hypertrophy%20(LVH)%20is,for%20cardiovascular%20disease%20and%2
[4] https://www.thecardiologyadvisor.com/ddi/chest-pain-differential-diagnosis/#:/~:text=Chest%20pain%20differential%20

Appendice: analisi univariata del dataset

L'analisi univariata permette di esplorare una variabile alla volta. Le Statistiche descrittive sono strumenti cruciali per riassumere un gruppo di osservazioni nel modo più semplice possibile.

Per le variabili categoriali: 1. si calcolano le tabelle di frequenza 2. diagramma a barre 3. pie chart

Per le variabili quantitative: 1. si calcolano le misure di posizione o tendenza centrale come media, mediana e quartili 2. misure di dispersione come la deviazione standard, MAD, varianza etc.. 3. misure di forma come asimmtria e curtosi 4. boxplots 5. stime di densità Kernel 6. Normal probability plots

Età (variabile 1)

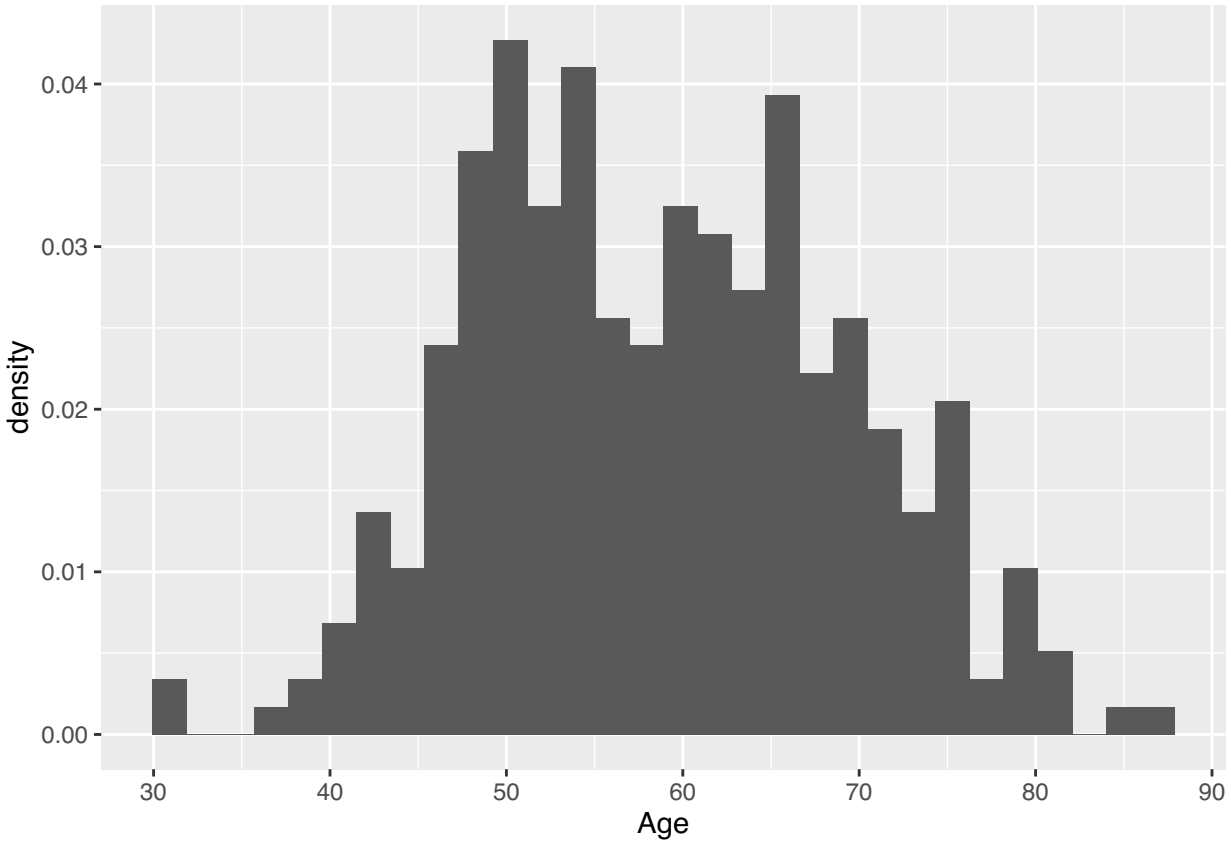
In media, si è osservato che l'età dei partecipanti è circa 58.9 anni. La mediana è di 58, per cui la metà dei partecipanti ha da 30 a 58 anni, mentre l'altra metà da 58 a 86. Il risultato suggerisce una distribuzione leggermente asimmetrica. Rispetto ai quartili, il primo quartile (Q1) è di 51%, il secondo quartile (Q2, che coincide con la mediana) è di 58 e il terzo quartile (Q3) è di 66. La simmetria è leggermente positiva, con un valore di circa 0.06, per cui la distribuzione tende ad essere leggermente spostata verso destra, indicando una maggiore concentrazione di pazienti la cui età è superiore a 58. La curtosi, con un valore di circa -0.5, indica una distribuzione con code poco pesanti.

```
##           Statistiche       Valori
## 1               Media  58.89768977
## 2             Mediana  58.00000000
```

```
## 3                  Q1  51.00000000
## 4                  Q2  58.00000000
## 5                  Q3  66.00000000
## 6            Varianza 107.99943173
## 7 Deviazione standard  10.39227750
## 8           Simmetria   0.06666667
## 9             Curtosi  -0.43636086
```

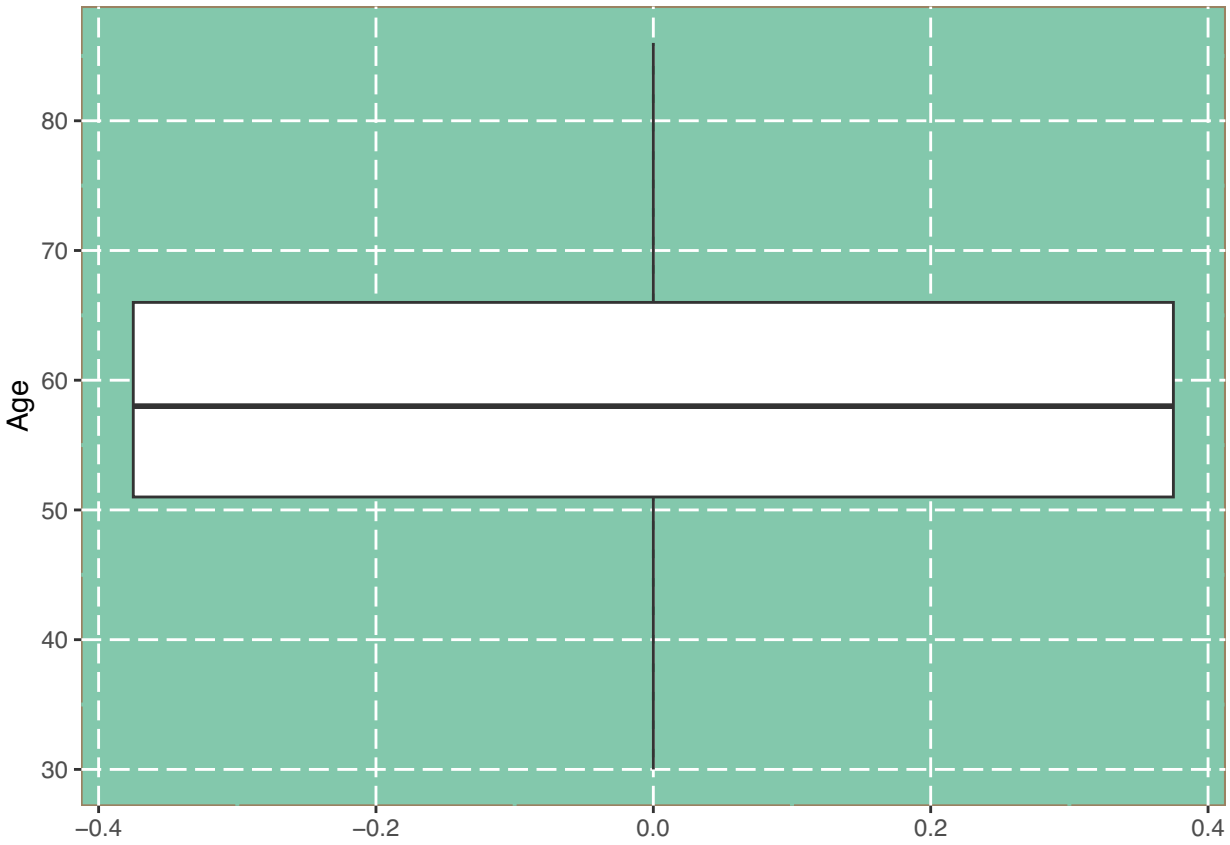
Istogramma

L'istogramma è una rappresentazione grafica per la visualizzazione della variabile numerica. Dal grafico non c'è evidenza di bimodalità.



Boxplot

Il boxplot è composto da una scatola che si estende da un quartile all'altro, cioè dal primo quartile (Q1) al terzo quartile (Q3). La lunghezza della scatola rappresenta l'intervallo interquartile (IQR) e indica dove si concentra la maggior parte dei dati. All'interno della scatola c'è una linea che rappresenta la mediana dei dati, che è il valore centrale nella distribuzione. I baffi si estendono dalla scatola verso l'alto e verso il basso. Possono rappresentare la dispersione dei dati e indicare quanto lontano si estendono i dati al di fuori dell'intervallo interquartile. I punti che cadono al di fuori dei baffi sono spesso considerati valori anomali o estremi. Questi punti rappresentano dati che si discostano significativamente dalla maggior parte delle osservazioni. La mediana è posizionata centrale rispetto osservazioni. Non sembrano esserci valori anomali e la distribuzione rispecchia la simmetria calcolata precedentemente.



BP (variabile 2)

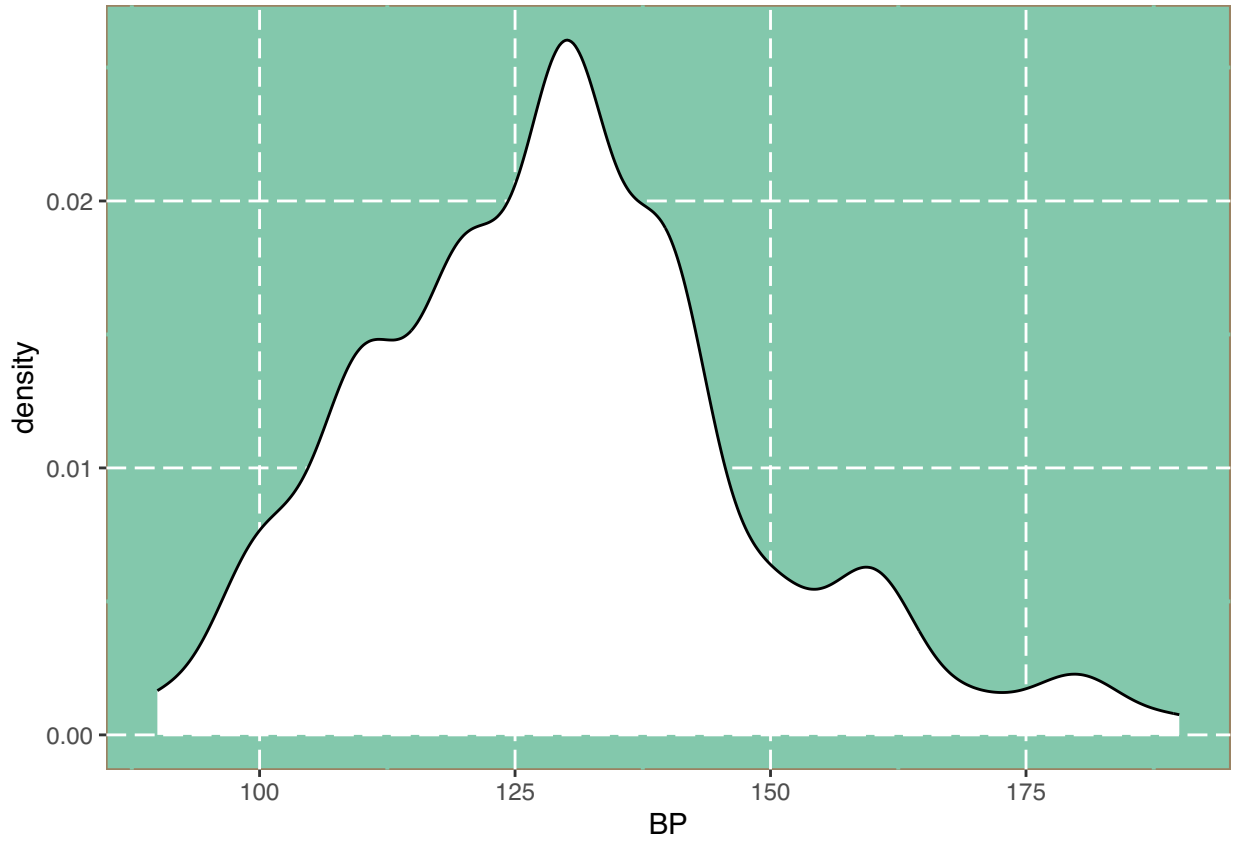
La pressione sanguigna è una variabile importante nel dataset. In media, gli individui registrano un valore pari a 130. La mediana è di 129.55, suggerisce una distribuzione relativamente bilanciata. Rispetto ai quartili, il primo quartile (Q1) è di 120, il secondo quartile (Q2, che coincide con la mediana) è di 130 e il terzo quartile (Q3) è di 140. In media la pressione sanguigna si si discosta di 18.93 dalla media. La simmetria è pari a 0. La curtosi è leggermente positiva con un valore di 0.43. Questo indica una distribuzione leggermente schiacciata rispetto a una distribuzione normale, suggerendo che ci siano meno valori estremi o "code" rispetto a una distribuzione più piatta.

```
##           Statistiche      Valori
## 1               Media 129.5544554
## 2             Mediana 130.0000000
## 3                  Q1 120.0000000
## 4                  Q2 130.0000000
## 5                  Q3 140.0000000
## 6            Varianza 358.6518261
## 7 Deviazione standard  18.9381051
## 8           Simmetria   0.0000000
## 9             Curtosi   0.4750052
```

Density

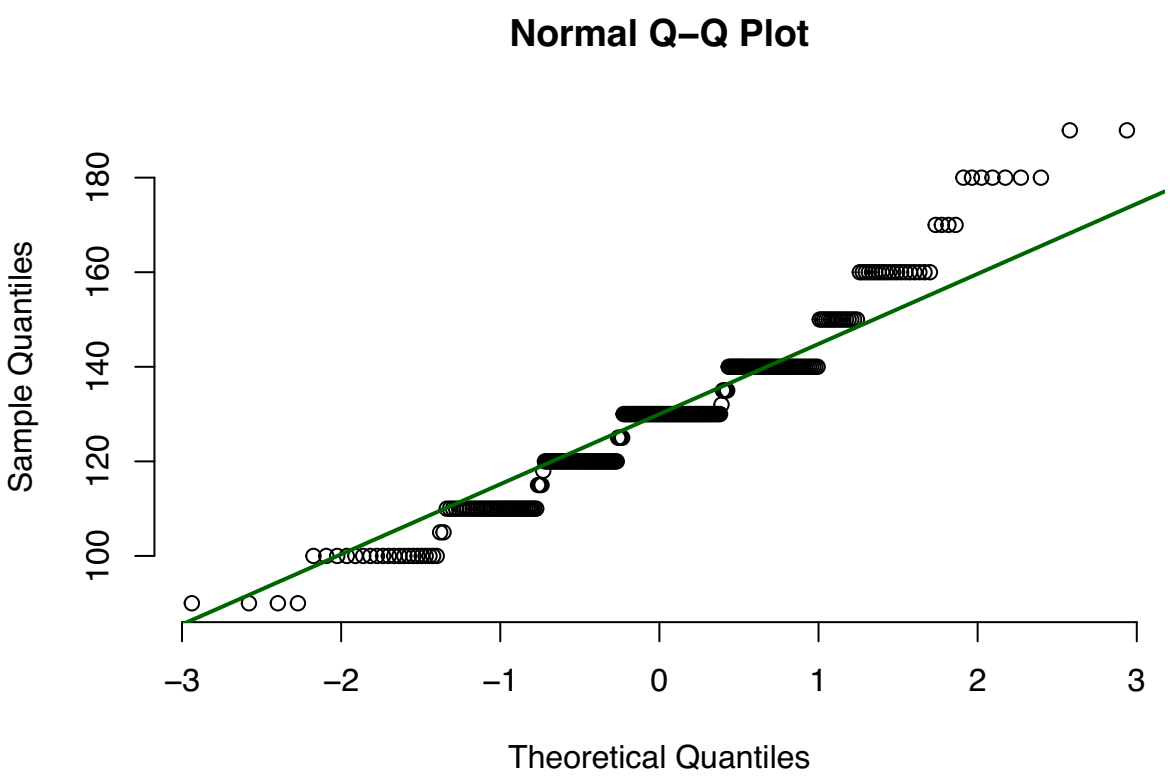
Il grafico di densità è un tipo di grafico che permette di visualizzare la forma della distribuzione di una variabile continua. Questa visualizzazione è particolarmente utile per comprendere la concentrazione e la

variabilità dei dati e per individuare eventuali modalità o tendenze all'interno della distribuzione. Nel grafico di densità, sull'asse delle ordinate è rappresentata la densità di probabilità, che rappresenta la probabilità di trovare un'osservazione in una determinata posizione lungo l'asse delle ascisse. Più la curva è alta in un punto, maggiore è la probabilità che le osservazioni cadano in quella posizione. La distribuzione mostra una curva che rispiecchia a valori appena calcolati non ci sono bimodalità ma sono comunque presenti picchi di distribuzioni che rendono la rappresentazione frastagliata.



Normal qqplot

Un QQ-plot rappresenta una comparazione tra i quantili di due distribuzioni: la distribuzione dei dati osservati e una distribuzione teorica di riferimento (in questo caso la distribuzione normale). L'asse orizzontale rappresenta i quantili teorici della normale. L'asse verticale rappresenta i quantili osservati delle spese per la casa rispetto alla normale. I punti sul grafico rappresentano le coppie di quantili. Ogni punto corrisponde a un valore nei dati osservati (asse y) e al suo corrispondente quantile teorico (asse x) secondo la normale. I dati osservati non seguono esattamente la distribuzione della normale infatti pochi punti si allineano sulla retta di riferimento evidenziando discrepanze dalla distribuzione.



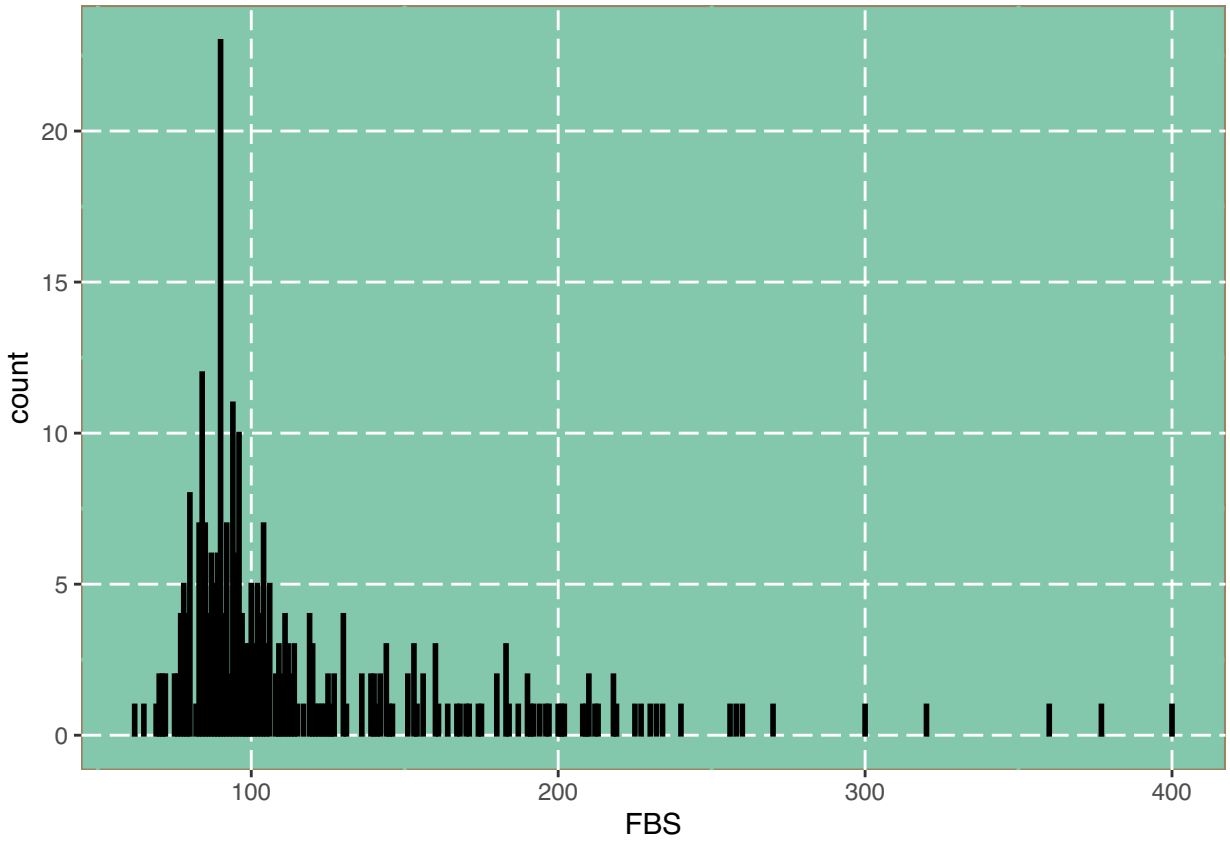
FBS (variabile 3)

FBS rappresenta la glicemia a digiuno, i pazienti in media registrano un valore prossimo a 119. La mediana è 98, per cui è espressione di asimmetria positiva. Il primo quartile (Q1) è 88,5 mg/dl, il secondo quartile (Q2, che è anche la mediana) è 98 mg/dl e il terzo quartile (Q3) è 130 mg/dl. La varianza è di circa 2712, il che indica che i dati sono poco vicini alla media. La deviazione standard è 52 indicando che la dispersione dei dati è moderata. Il valore della simmetria è 0,54 esprimendo una leggera tendenza verso la destra nella distribuzione, cioè alcuni pazienti potrebbero avere un valore più alto di glicemia rispetto alla media. La curtosi misura quanto le code della distribuzione differiscono da una distribuzione normale. Il valore di 6 indica che la distribuzione ha code più pesanti di una distribuzione normale.

```
##           Statistiche       Valori
## 1               Media  119.1848185
## 2             Mediana   98.0000000
## 3                  Q1   88.5000000
## 4                  Q2   98.0000000
## 5                  Q3  130.0000000
## 6            Varianza 2712.2902323
## 7 Deviazione standard   52.0796528
## 8           Simmetria    0.5421687
## 9             Curtosi    6.3107250
```

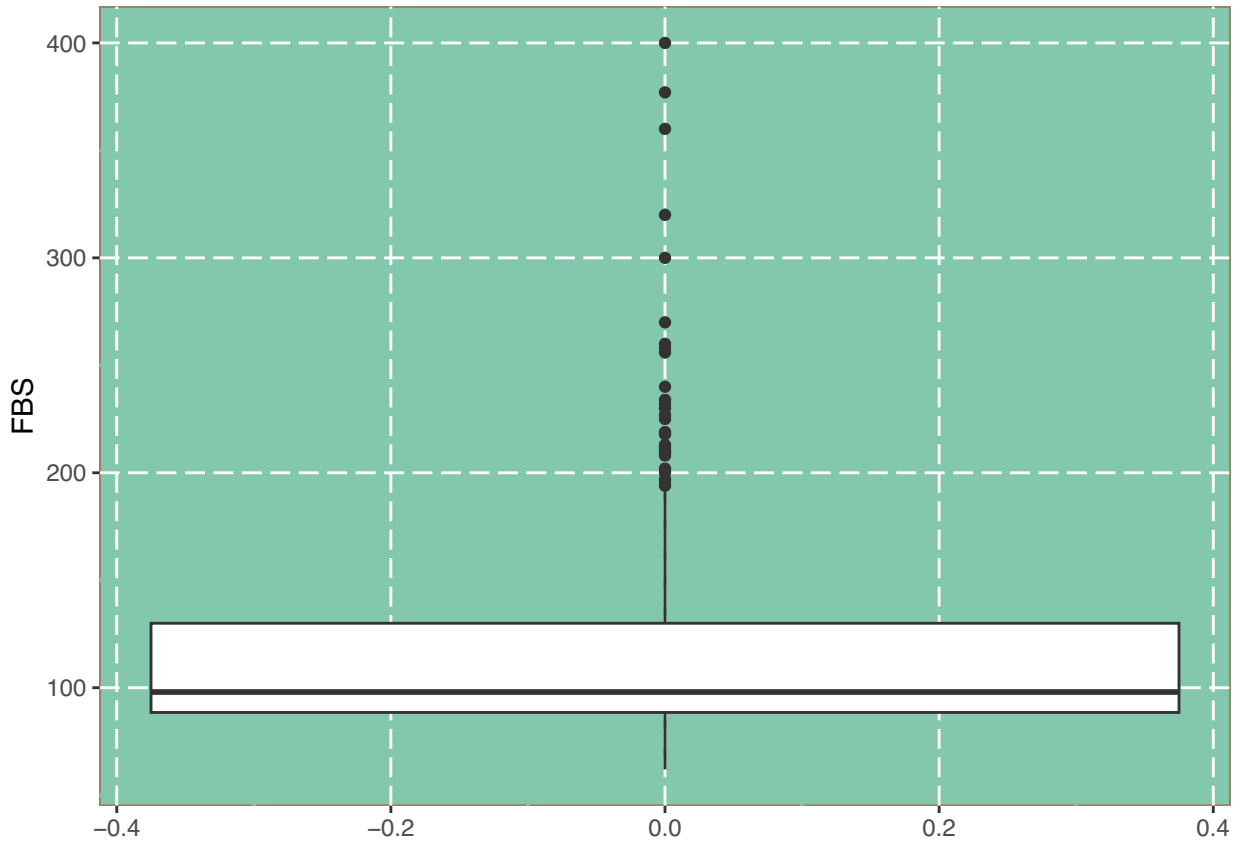
Istogramma

La stragrande maggioranza delle osservazioni si presenta con valori fra 30 e 130. Come già evidenziato dalla moda la maggior frequenze delle osservazioni si presenta nel valore di 98, questo si riflette nel grafico con presena di barra più alta. Non c'è evidenza di bimodalità ma presenza di leggera asimmetria negativa.



Boxplot

Il boxplot evidenza l'asimmetria positiva nei dati. Sono presenti possibili valori anomali.



TG- Trigliceridi (variabile 4)

In media i pazienti registrano valori di globuli rossi intorno a 150 mg/dl. La mediana , se pur di poco , di discosta di 30 mg/dl. Osservando i quartili, si nota che il primo quartile (Q1) è del 90 mg/dl, il secondo quartile (Q2, equivalente alla mediana) è 122 mg/dl, e il terzo quartile (Q3) è 177 mg/dl. La dispersione intorno alla media è circa 97 mg/dl.
La simmetria, con un valore vicino a zero (0,26), suggerisce che la distribuzione è approssimativamente simmetrica, senza una forte tendenza verso destra o sinistra. La curtosi, con un valore positivo di 25, indica una distribuzione con code più pesanti rispetto a una distribuzione normale.

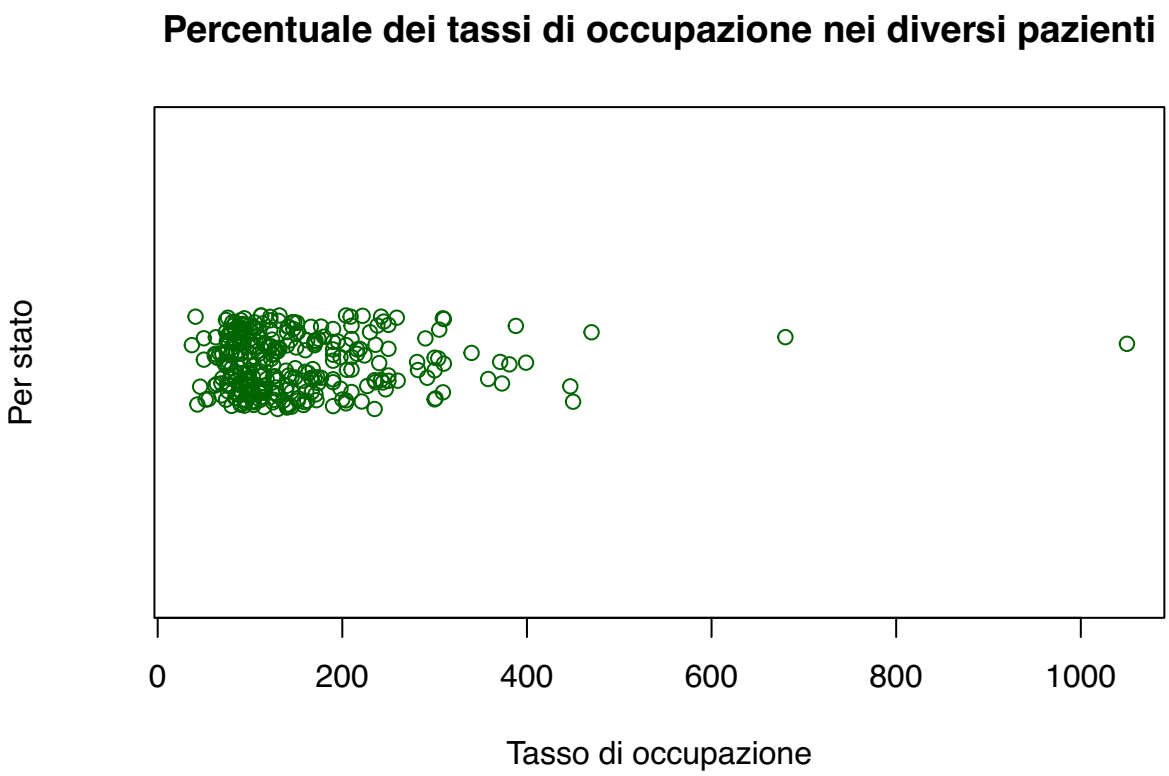
```
##           Statistiche       Valori
## 1               Media  150.3432343
## 2             Mediana  122.0000000
## 3                  Q1   90.0000000
## 4                  Q2  122.0000000
## 5                  Q3  177.0000000
## 6            Varianza 9596.0539855
## 7 Deviazione standard   97.9594507
## 8           Simmetria    0.2643678
## 9             Curtosi   25.3108249
```

Stripchart

Nel grafico a stripchart, i punti dati vengono disposti lungo un'unica linea orizzontale o verticale, in modo che ciascun punto rappresenti un'osservazione o un valore specifico. I punti possono sovrapporsi se ci sono molte

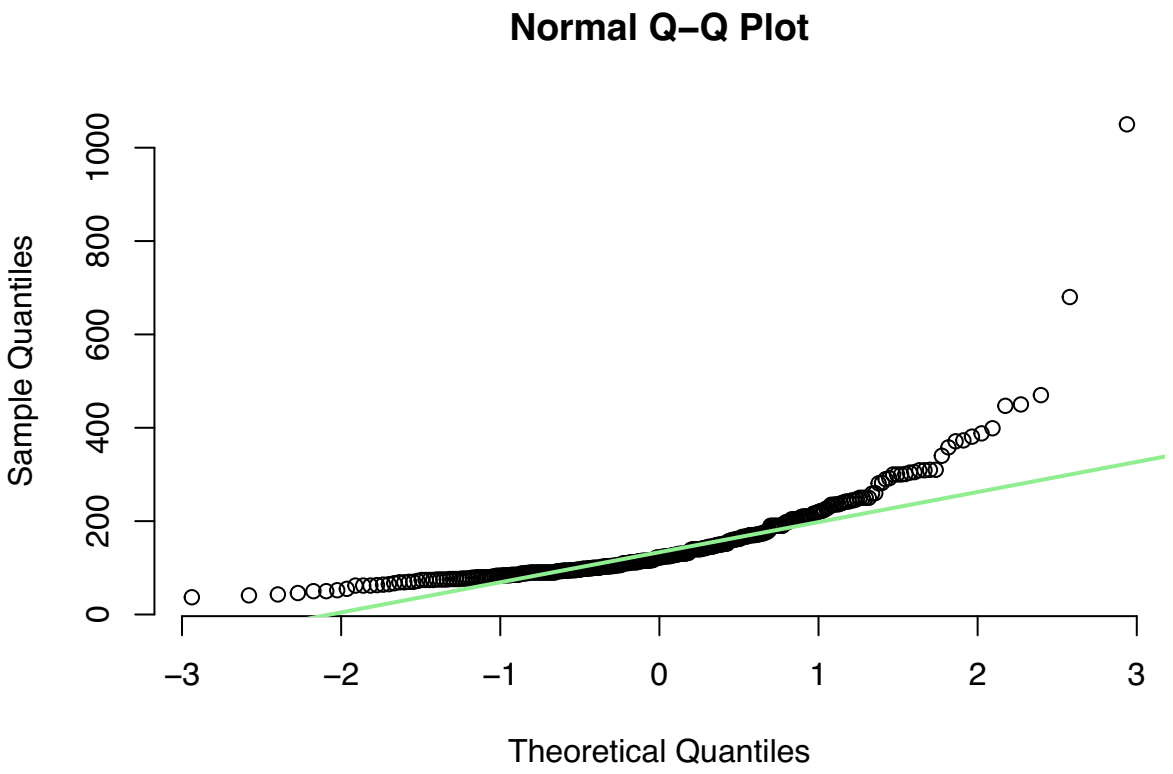
osservazioni con lo stesso valore. Questo tipo di grafico è spesso utilizzato per evidenziare la distribuzione dei dati, la concentrazione dei punti in determinate regioni e possibili outliers o valori anomali.

Per la variabile che rappresenta i trigliceridi i dati si concentrano per lo più tra 0 e 200 con 2 valori che cadono fuori di 600 mg/dl e pochi sopra 400 mg/dl. In condizioni di osservazioni normali si potrebbe pensare a valori anomali ma in questo caso i valori giustificano la condizione dei pazienti, non è del tutto impensabile lavorare con questo tipo di dati.



Normal Qq plot

Per la variabile relativa ai trigliceridi la similarità della distribuzione con quella della variabile casuale normale è più evidente rispetto alla variabile analizzata precedente. I dati osservati seguono quasi esattamente la distribuzione teorica della normale, i punti ,infatti, si allineano non perfettamente alla retta di 45 gradi.



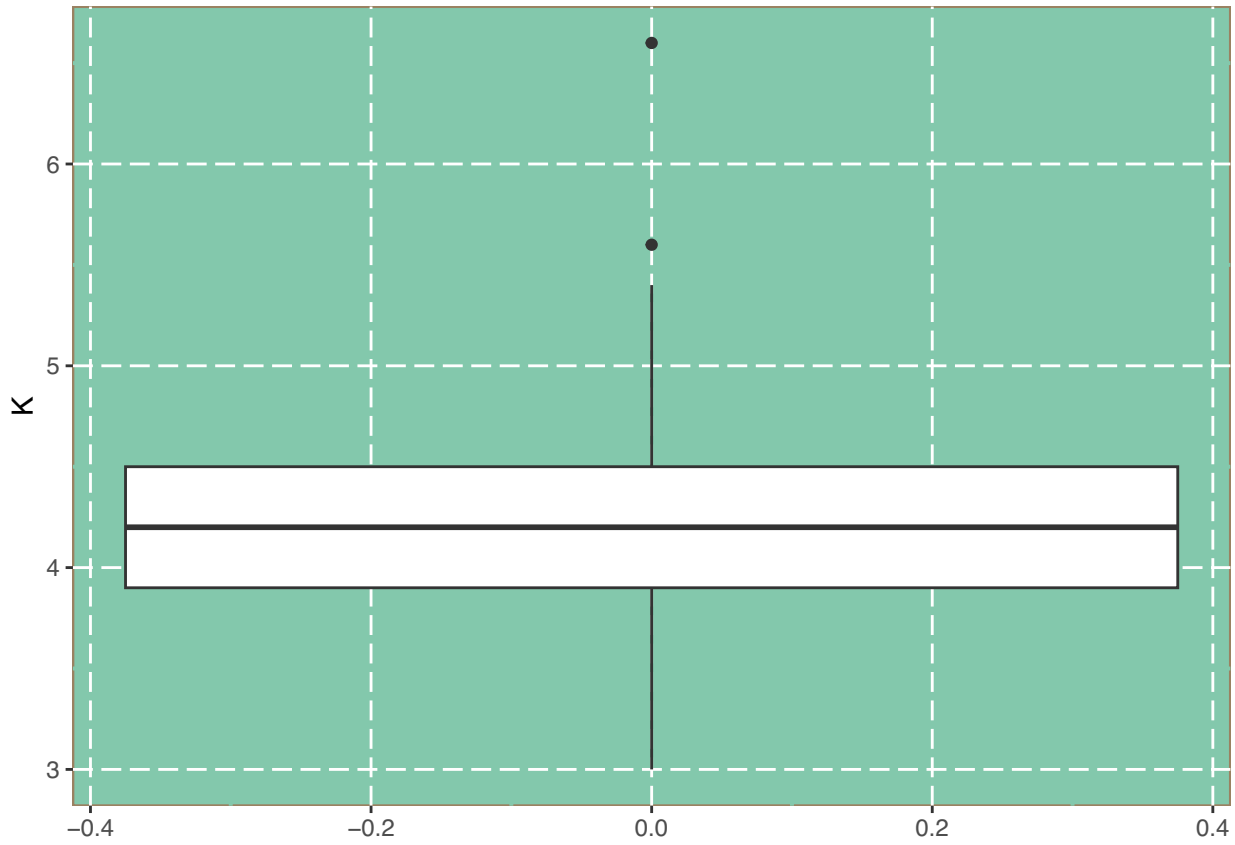
K- Potassio (Variabile 5)

Il livello medio di potassio è di 4.2307. La mediana, che è 4.2, indica che metà dei pazienti ha un livello di potassio inferiore a 4.2 e l'altra metà ha un livello superiore a questo valore. Il primo quartile (Q1), pari a 3.9, mostra che il 25% dei pazienti ha un livello di potassio inferiore a questo valore. Il secondo quartile (Q2), che coincide con la mediana, è 4.2. Il terzo quartile (Q3) è 4.5, indicando che il 75% dei pazienti ha un livello di potassio inferiore a questo valore, mentre il restante 25% ha valori superiori. Il coefficiente di simmetria è 0, suggerendo che la distribuzione dei valori di potassio è perfettamente simmetrica rispetto alla media. Infine, la curtosi è 2.08, indicando che la distribuzione dei livelli di potassio ha una forma relativamente piatta rispetto a una distribuzione normale.

```
##           Statistiche    Valori
## 1               Media 4.2306931
## 2             Mediana 4.2000000
## 3                  Q1 3.9000000
## 4                  Q2 4.2000000
## 5                  Q3 4.5000000
## 6            Varianza 0.2099489
## 7 Deviazione standard 0.4582018
## 8           Simmetria 0.0000000
## 9             Curtosi 2.0837864
```

Boxplot

Il boxplot non evidenza sintomi di asimmetria. Sono presenti solo due valori anomeli per la distribuzione.



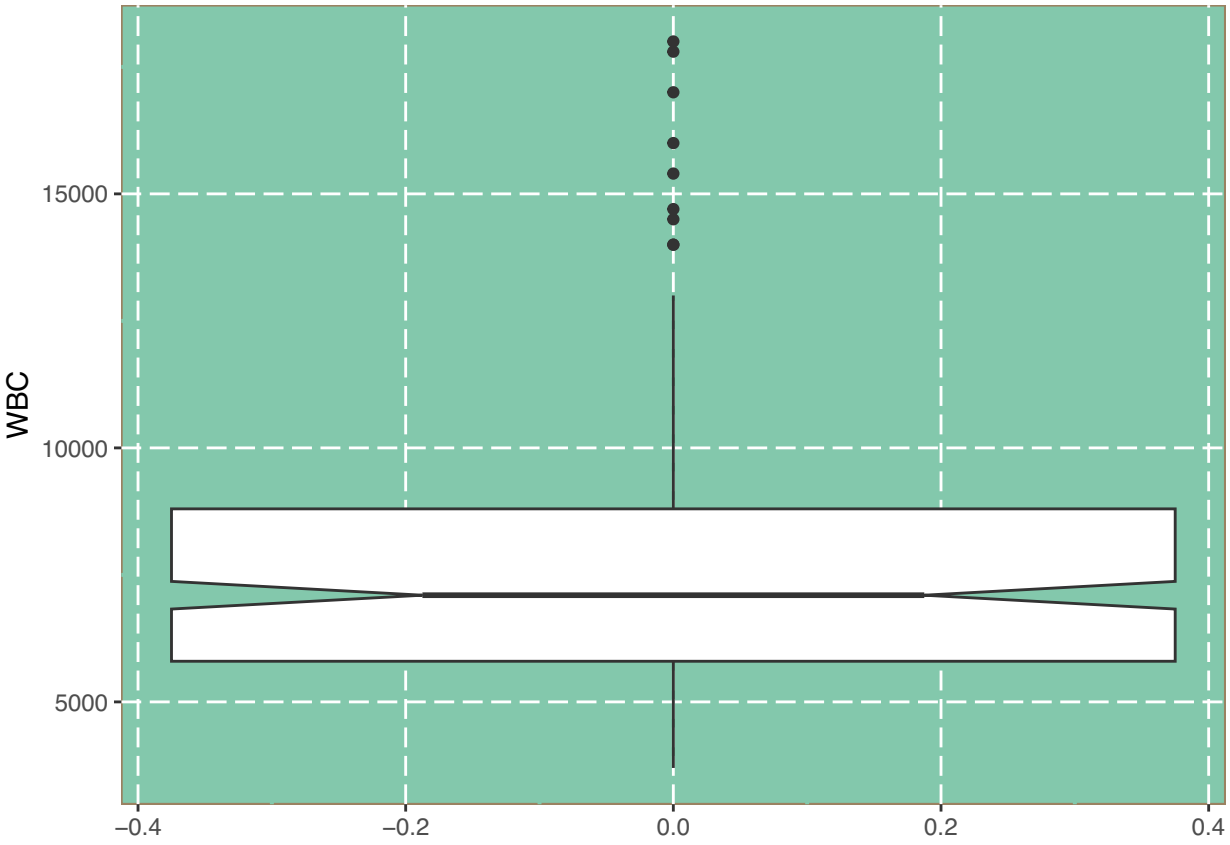
WBC- Globuli rossi (Variabile 6)

Il numero medio di globuli rossi (WBC) è di 7.562,046. La mediana del numero di globuli rossi è di 7.100, il che indica che metà dei pazienti ha un conteggio di globuli rossi inferiore a 7.100 e l'altra metà ha un conteggio superiore a questo valore. Il primo quartile (Q1), che è 5.800, mostra che il 25% dei pazienti ha un numero di globuli rossi inferiore a questo valore. Il secondo quartile (Q2), che coincide con la mediana, è 7.100. Il terzo quartile (Q3) è 8.800, indicando che il 75% dei pazienti ha un conteggio di globuli rossi inferiore a questo valore, mentre il restante 25% ha conteggi superiori. Il coefficiente di simmetria è 0,1333333, suggerendo che la distribuzione dei conteggi di globuli rossi è leggermente asimmetrica rispetto alla media. Infine, la curtosi è 2,793368, indicando che ha una forma relativamente normale ma leggermente più piatta rispetto a una distribuzione normale.

```
##           Statistiche       Valori
## 1               Media 7.562046e+03
## 2             Mediana 7.100000e+03
## 3                  Q1 5.800000e+03
## 4                  Q2 7.100000e+03
## 5                  Q3 8.800000e+03
## 6            Varianza 5.826138e+06
## 7 Deviazione standard 2.413739e+03
## 8           Simmetria 1.333333e-01
## 9             Curtosi 2.793368e+00
```

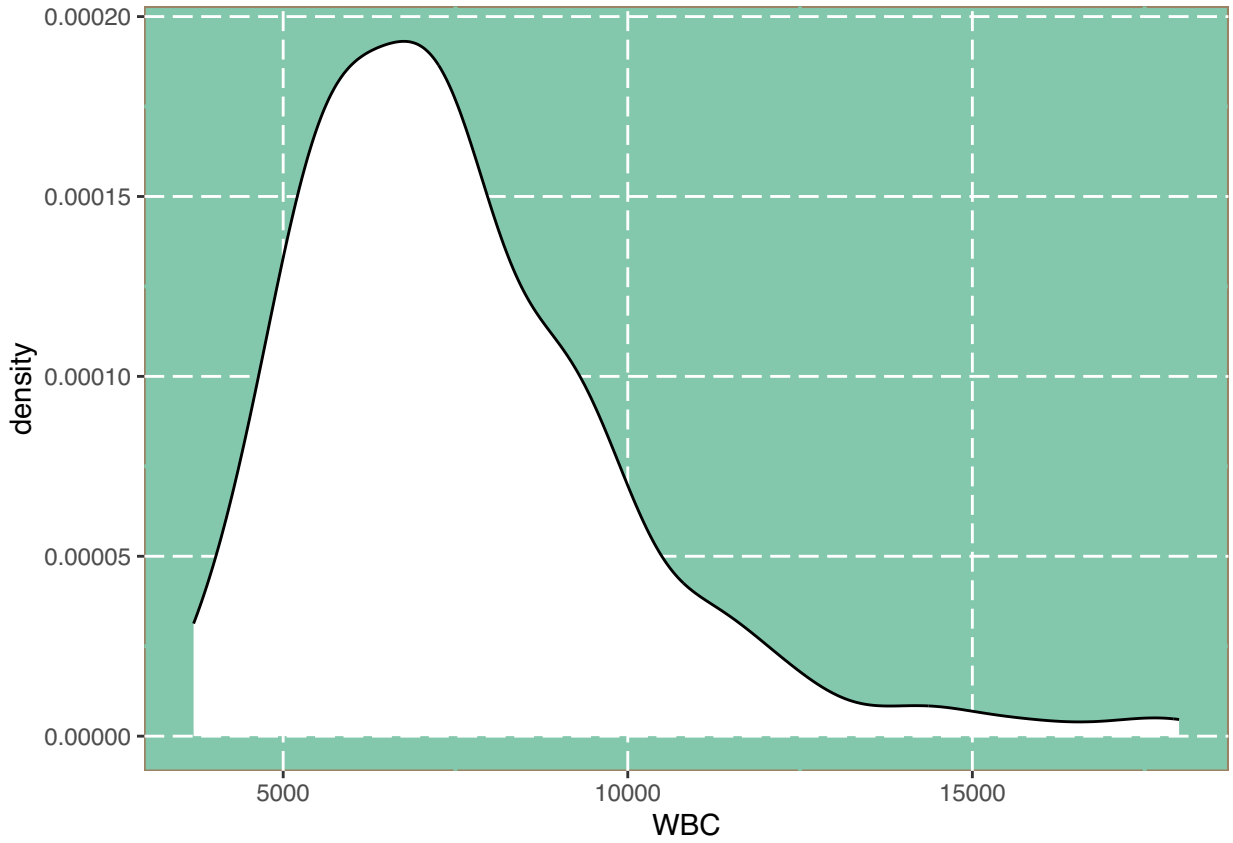
Box a Intaglio

Si possono utilizzare anche i boxplot ad intaglio. Questi sono una rappresentazione grafica dei boxplot ma con l'aggiunta dell'intervallo di confidenza. Con un grado di fiducia del 95%, l'intervallo di confidenza approssimato per i globuli rossii è (6829.418 7370.582).



Density

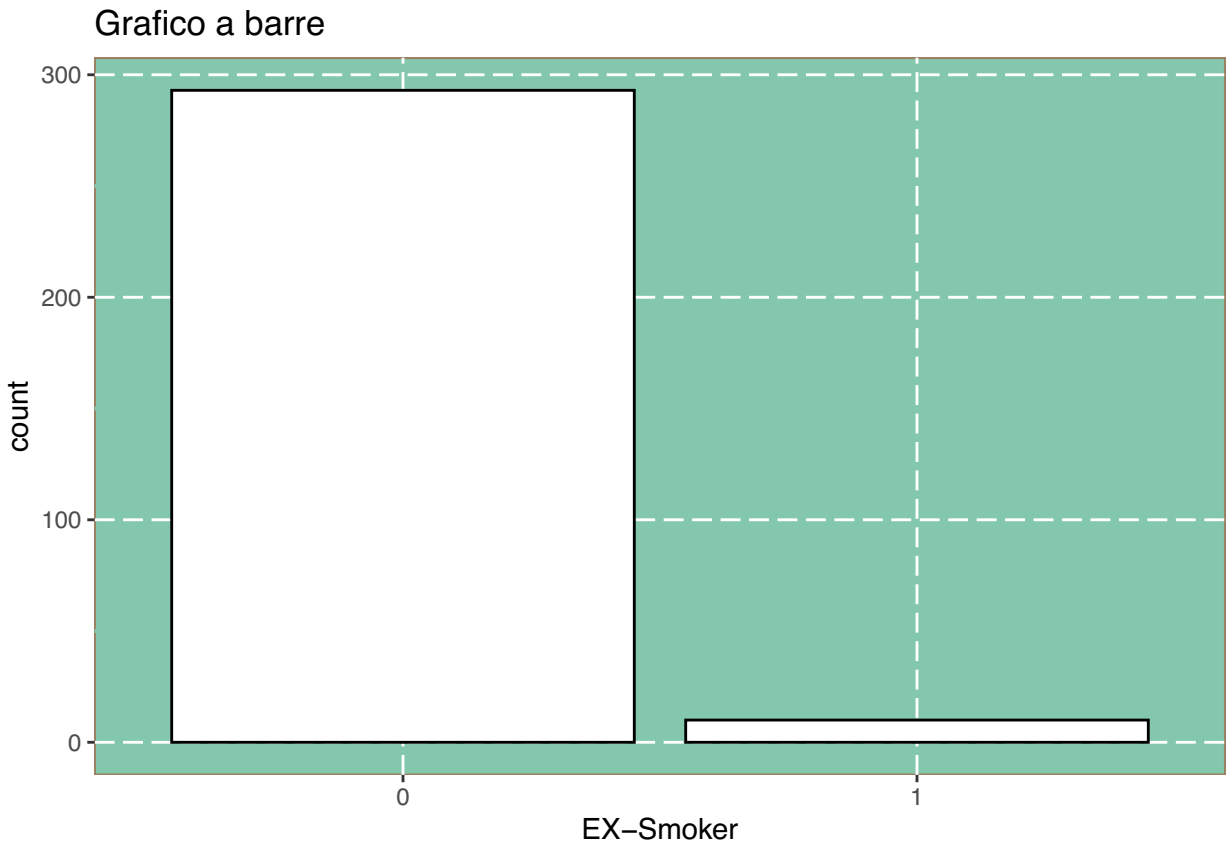
Dalla visualizzazione della distribuzione è chiaramente possibile assimilare la totale assenza di asimmetria della distribuzione.



Ex-Smoker (Variabile 1)

La variabile **EX-Smoker** rappresenta il numero di pazienti che sono stati fumatori in passato. Sono presenti 293 osservazioni con valore di ex-smoker 0 e 10 con valore 1.

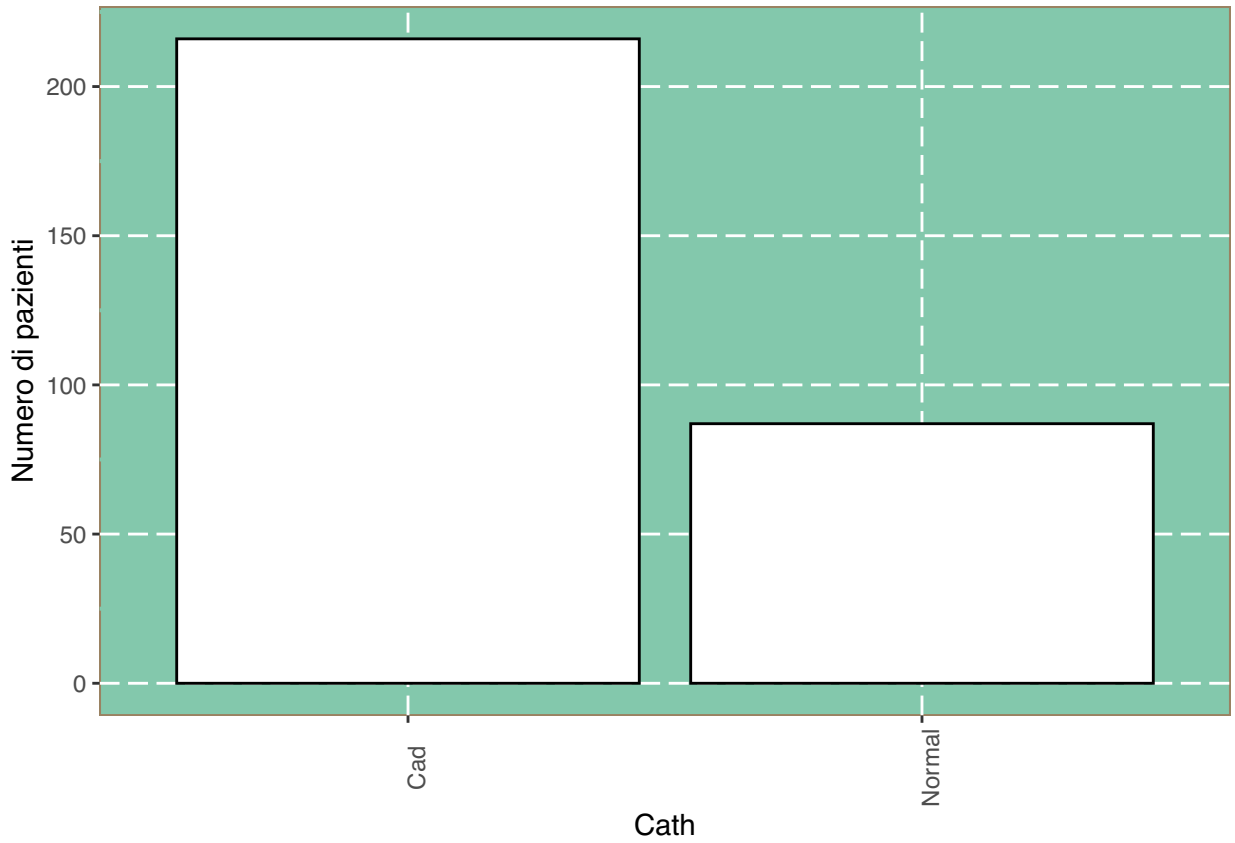
```
##   Frequenza Frequenza_percentuale
## 0       293              96.69967
## 1        10               3.30033
```



Cath - CAD o Normale (Variabile 2)

La variabile **Cath** rappresenta la variabile target dello studio. Dato il campione ristretto, il numero di osservazioni si riduce al 71% dei pazienti con prognosi di malattia coronaria e 28% in condizioni mediche normali.

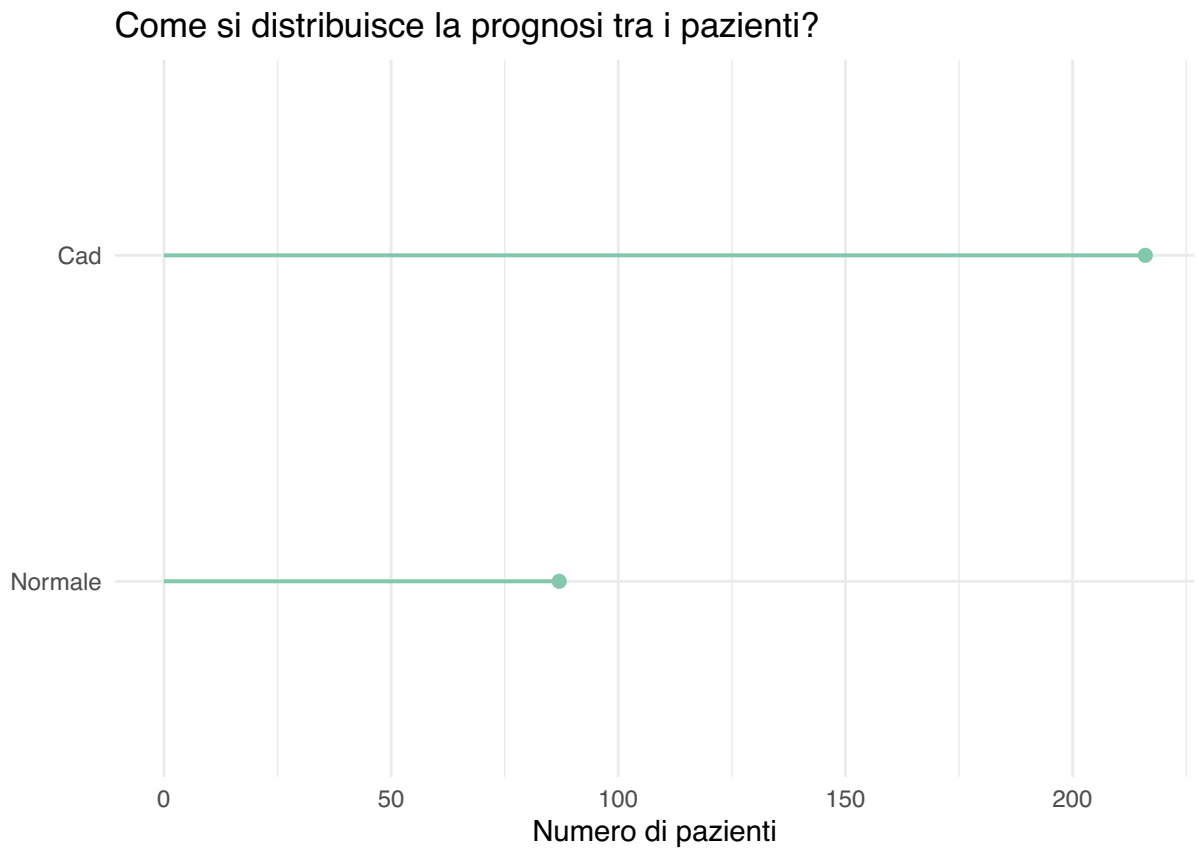
```
##        Frequenza Frequenza_percentuale
## Cad          216              71.28713
## Normal        87              28.71287
```



Lollipop Chart

Un metodo alternativo per la visualizzazione di una variabile categoriale è il lollipop chart. Un *Lollipop Chart* è un grafico che presenta punti di dati come cerchi o dischi ("lollipops") posti su un asse orizzontale, che rappresenta una variabile indipendente o categoria. Ogni lollipop rappresenta un singolo punto dati e la sua posizione sull'asse orizzontale indica il valore di quella variabile. Il "lollipop" è collegato a una linea verticale o "astina" che si estende verso sinistra da ciascun punto dato fino a una seconda scala. La seconda scala rappresenta la variabile dipendente o un valore di riferimento, come una media o una soglia.

Per cui, 16 stati hanno valore medio della soddisfazione dell'individuo al di sotto della media complessiva, mentre circa 22 stati hanno valore medio della soddisfazione al di sopra della media.



Il "Donut Chart" è un grafico circolare diviso in sezioni, o "fette," che rappresentano le diverse categorie di dati. Ogni fetta corrisponde a una categoria specifica e la dimensione di ciascuna fetta è proporzionale alla percentuale che quella categoria rappresenta rispetto al totale. E' un'alternativa al "Pie Chart" distinguendosi per il foro al centro, che crea un anello vuoto. Il "Donut Chart" è utile per visualizzare chiaramente come le diverse categorie contribuiscono a un totale, e la dimensione dell'anello interno rappresenta il totale complessivo.

Target CAD