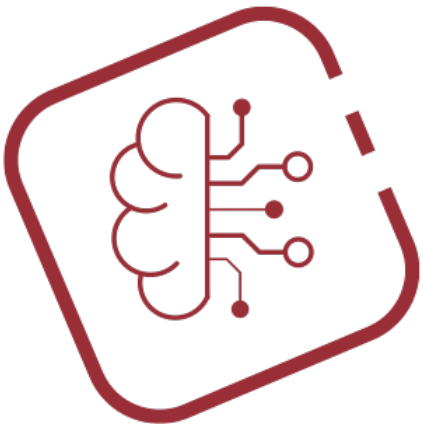
ITS Angelo Rizzoli
Corso ITS Machine Learning Specialist



Linguaggi di Programmazione
per il Machine Learning

Programmazione Base in R: Heart Disease

Lorenzo Buffolino
Jalex Follosco

Corso 2020 - 2022

Contenuti

1 Introduzione

Con l'analisi che segue in tale report, si è voluto analizzare una possibile correlazione tra il genere, l'età e i parametri cardiovascolari misurate in ogni osservazione (i.e. l'individuo sottoposto alle analisi) al fine di stabilire l'eventuale presenza di malattie cardiovascolari nell' individuo osservato . Per analizzare le possibili correlazioni, sulla base della traccia, e stata utilizzata una versione del dataset,che si compone di osservazioni registrate dalla V.A. Medical Center, Long Beach and Cleveland Clinic Foundation, donate alla UCI Machine Learning Repository nel 1988. [1] Attraverso la regressione lineare si è in grado di stabilire possibili correlazioni tra i diversi parametri numerici delle osservazioni misurate alla Clinica, e all' età del singolo individuo, anche al fine di analizzarne i trend delle incidenze (patterns) . Per finire, con la fase di training del dataset, con metrica basata sull'accuratezza, si restituiscono dei punteggi di probabilità della presenza di eventuali malattie cardiache e l'analisi della veridicità dei punteggi di probabilità tramite il metodo della matrice confusionale (Confusion Matrix)

[1] UCI Machine Learning Repository `http://archive.ics.uci.edu/ml/datasets/Heart+Disease`

2 | Dataset

Il dataset utilizzato è una versione del dataset,composto da osservazioni registrate dalla V.A. Medical Center, Long Beach and Cleveland Clinic Foundation, donate alla UCI Machine Learning Repository nel 1988. [1]. Il dataset si presenta con 17 attributi, analoghi all'età e sesso, e ai parametri cardiovascolari del paziente osservato alla Clinic Foundation. Nel paragrafo successivo si approfondiscono gli attributi del dataset interessato

2.1 Attributi

Il dataset è suddiviso in 15 attributi, i quali sono definiti secondo quanto segue:

- `x` è una variabile identificativa dell'osservazione.
- `age` indica l'età dell' individuo osservato.
- `sex` indica il sesso dell'individuo osservato.
- `cp` indica il tipo di dolore al petto presente nell'individuo osservato.Può essere indice della presenza dell' *Angina*[2]. Il valore attribuito può essere: `0` se l'osservato è asintomatico; `1` se l'osservato presenta dolori al petto anormali derivati dall' *Angina*; `2` se l'osservato presenta dolori al petto non derivati dall' Angina; `3` se l'osservato presenta dolori normali derivati dall' Angina.
- `trestbps` indica la pressione sanguinia, misurata in mm/Hg, presente a riposo nell'osservato, registrata all'inizio del ricovero presso la Clinic Foundation. Il valore è di tipo di rapporto.
- `chol` indica il livello di colesterolo sierico, misurato in mg/dl, presente dell'individuo osservato. Si ipotizza si tratti di colesterolo delle lipoproteine a bassa densità (**LDL**), il quale tende a depositarsi sulle superfici cardiovascolari, con la conseguente ostruzione del flusso sanguinio e il conseguente aumento della pressione sanguinia
- `fbs` indinca se il livello di glucosio presente nelle vene dell'osservato in condizioni di digiuno, registrato con il *Test orale di tolleranza al glucosio*[3]
- `restecg` indica i risultati elettrocardiografici dell'osservato a riposo. Il valore attribuito può essere: `0` se l'osservato mostra un'ipertrofia ventricolare sinistra probabile o definita secondo i criteri di Estes; `1` se l'sservato mostra risultati normali; `2` se l'osservato presenta un'anomalia dell'onda ST-T, con inversioni dell'onda T e/o elevazione o depressione ST maggiore di 0,05 mV
- `thalach` indica la frequenza cardiaca massima raggiunta dall'individuo osservato
- `exang` indica la presenza di angina indotta da esercizio fisico nell' osservato.Il valore attribuito può essere: `0` se l'angina presente nell'individuo osservato non è indotta dall'esercizio fisico; verb|1| se l'angina presente nell'individuo osservato è indotta dall'esercizio fisico;
- `oldpeak` indica la depressione del segmento ST, presente nell' ECG, indotta dall'esercizio rispetto al riposo
- `slope` indica l'alterazione del segmento ST[4], presente nell' ECG,indotto dall'eserczio fisico eseguito con prestazioni fisiche massime.Il valore attribuito può essere: `0` se è presente un sottoslivellamento del segmento ST;`1` se il segmento ST non presenta alcune alterazioni del livello;`2` se è presente un sopraslivellamento del segmento ST;
- `ca` indica il numero dei principali vasi sanuini colorati nell'individuo osservato sottoposto alla *fluoroscopia*.

[1] UCI Machine Learning Repository `http://archive.ics.uci.edu/ml/datasets/Heart+Disease`

[2] l'***Angina*** è, in sostanza, un dolore transitorio al torace o sensazione di pressione che si manifesta quando il muscolo cardiaco non riceve una sufficiente quantità di ossigeno. `https://www.msdmanuals.com/it-it/casa/disturbi-cardiaci-e-dei-vasi-sanguigni/coronaropatia/angina`

[3] Test orale di tolleranza al glucosio `https://it.wikipedia.org/wiki/Test_orale_di_tolleranza_al_glucosio`

[4] Valutazione del tratto ST su `https://it.my-ekg.com/come-leggere-ecg/tratto-st.html`

- `thal` indica i risultati dell' analisi della talassemia[5] sull'individuo osservato. Il valore attribuito può essere: `NA` il valore è eliminato dal precedente set di dati; `1` se è presente un difetto fisso, ovvero, il flusso sanguigno è assente in alcune parti del cuore; `2` se il flusso sanguigno è nella norma; `3` se è presente un difetto reversibile, ovveo si sono registrate nell'individuo ossevato flussi sanguigni anormali
- `target` indica la presenza di malattie cardiache nell'individuo, stabilite con analisi invasive sul corpo dell'individuo. Il valore attribuito può essere: `0` se l'individuo osservato presenta malattiache; `1` se l'individuo non presenta malattie cardiache

[5] Si parla di talassemia quando l'organismo sintetizza forme anomale di emoglobina nei globuli rossi

3 Analisi dei dati *Tecnicamente corretti e consistenti*

Prima di poter svolgere un analisi descrittiva dei dati è opprtuno controllare il dataset, verificandone la correttezza e la consistenza, attraverso processi di pulizia, correzione e trasformazione

3.1 Analisi dei dati tecnicamente corretti

Per iniziare l'analisi, si utilizza la funzione `glimpse()` della libreria *dplyr* per analizzare la struttura del dataset;

```
> heart %>% gimpse()
Rows: 303
Columns: 15
$ x        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 14, 15, 16, 18, 18, 19, 20, 21, 22, 23,
$ age      <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 50, 58, 66, 43, 69, 59
$ sex      <chr> "1", "1", "0", "1", "0", "1", "0", "1", "1", "1", "1", "0", "1", "1", "0", "0", "0
$ cp       <int> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0, 2, 0, 2, 3, 1, 2, 2
$ trestbps <int> 145, 130, 130, 120, 120, 140, 140, 120, 51, 150, 140, 130, 130, 110, 150, 120, 120
$ chol     <chr> "233", "250", "204", "236", "354", "192", "294", "263", "199", "168", "239", "275"
$ fbs      <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0
$ restecg  <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
$ thalach  <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 171, 144, 162, 158, 17
$ exang    <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0
$ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0.6, 1.8, 1.0, 1.6, 0.
$ slope    <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1, 2, 2, 1, 2, 2, 2, 2
$ ca       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 2, 0, 0
$ thal     <int> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 3, 2, 2, 2
$ target   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
```

Come si può notare dal blocco di codice di cui sopra, il dataset è strutturato in maniera errata:

- `sex` è visto chome un *character*
- `chol` è visto come un *character*
- `cp`,`fbs`,`restecg`,`exang`,`slope`,`ca`,`thal` e `target` sono visti come *integer*

Quindi è necessario trasformare queste variabili nel **tipo** corretto.

- `sex` da *character* a *integer*
- `chol` da *character* a *integer*
- `cp`,`fbs`,`restecg`,`exang`,`slope`,`ca`,`thal` e `target` da *integer* a fattori

Essendo le ultime variabili sopracitate codificate con dei numeri, è più appropriato scriverne direttamente i valori, per una più facile lettura del dataset. Ad esempio, come definito nel paragrafo 2.1, per la variabile cp (tipo di dolore al petto) il valore:

- 0 indica un paziente asintomatico
- 1 indica un paziente che presenta un dolore anginale atipico
- 2 indica un paziente che non presenta un dolore di tipo anginale
- 3 indica un paziente che presenta un dolore anginale tipico

Quindi nel dataset al posto dei valori (0-3) si troveranno i valori: `"asintomatico","angina atipica","non anginale","angina t`
Dopo aver trasformato le variabili nel tipo corretto, verifichiamo di nuovo la struttura per vedere se le nostre modifiche sono state eseguite. Ora che le variabili sono state trasformate, apriamo il nostro dataset Heart (view): Notiamo che

alcune righe hanno un numero di vasi sanguigni (ca) superiore a 3, i quali si riferiscono alle arterie, alle vene e ai capillari. Si trasformano in NA queste righe e in seguito eliminate poichè è impossibile avere 4 vasi sanguigni:

```
heart$ca[heart$ca == 4] <- NA
```

age	sex	cp	trestbps	chol	restecg	thalach	exang	oldpeak	slope	ca	thal
52	maschio	Non aginale	138	223	normale	169	no	0	In salita	4	2
58	maschio	Agina atipica	125	220	normale	144	no	0.4	piatto	4	3
38	maschio	Non aginale	138	175	normale	173	no	0	In salita	4	27
43	maschio	asintomatico	132	247	ipertrofia	143	si	0.1	piatto	4	3

Stessa cosa per la variabile thal, infatti il valore 0 indica il valore NA:

```
heart$thal[heart$thal == 0] <- NA
```

age	sex	cp	trestbps	chol	restecg	thalach	exang	oldpeak	slope	ca	thal
53	femmina	Non aginale	128	216	ipertrofia	115	no	0	In salita	0	0
52	maschio	asintomatico	128	204	normale	56	si	1.0	piatto	0	0

come notiamo, oltre alla presenza degli `NA`, ci sono valori `'unspecified'` e `'undefined'`, rispettivamente nella colonna

di `sex` e `chol`. Per risolvere, dunque trasformiamo questi valori in `NA` in questo modo:

```
heart$sex[heart$sex == 'unspecified' ] <- NA
heart$chol[heart$chol == "undefined"] <- NA
```

procediamo quindi con la pulizia del nostro dataset mediante l'eliminazione delle righe in cui sono presenti gli `NA`.

3.2 Rinomina ed eliminazione degli attributi

Il passo successivo è quello di rinominare le variabili in maniera appropriata e piÃ¹ significativa:

- `cp` è rinominato in `chest_pain_type`
- `trestbps` è rinominato in `rest_blood_pressure`
- `chol` è rinominato in `cholesterol`
- `restecg` è rinominato in `rest_electrocardio_result`
- `thalach` è rinominato in `maximum_heart_rate`
- `exang` è rinominato in `exercise_angina`
- `ca` è rinominato in `n_vessels`

Infine eliminiamo la colonna x in quanto ci fornisce un dato irrilevante per lâ€™analisi descrittiva del set, ovvero l'identificativo del paziente.

3.3 Dati consistenti

Ora che i nostri dati sono tecnicamente corretti dobbiamo renderli consistenti, ovvero verificare se ci sono degli errori o dei problemi a livello di contenuti:

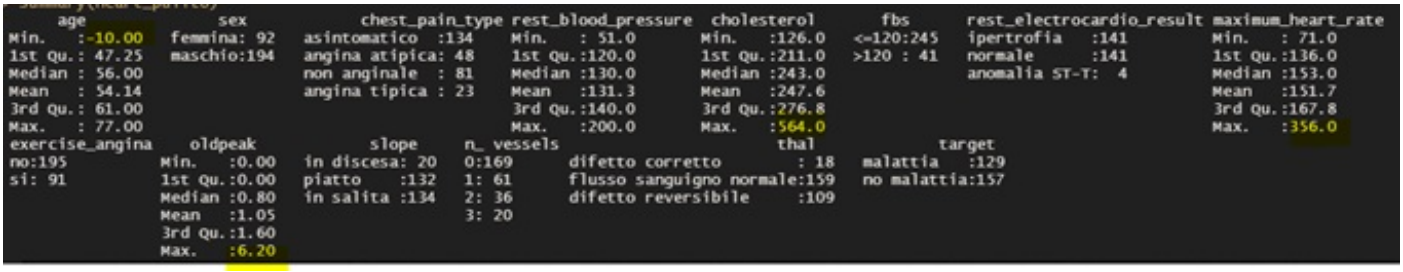


Figure 3.1: Restituito della funzione **summary()** con argomento il dataset

Come si osserva dal summary il valore minimo della variabile age è -10, un risultato che nella realta non può esistere. Un'altra variabile inconsistente è il colesterolo in quanto arriva ai 564 , valore che non verrebbe mai raggiunta nella realtà, infatti già per livelli di colesterolo maggiori di 240 risultano eccessivi. Stesso ragionamento per la variabile

`maximum_heart_rate`, una frequenza cardiaca pari a 356 risulta impossibile nella vita reale. Cosi anche per la `oldpeak` e `rest_blood_pressure`. Applichiamo quindi dei filtri per rimuovere questi dati incoerenti:

```
heart <- heart[heart$age > 0,]]
heart $maximum_heart_rate[heart $maximum_heart_rate > 222 ] <- mean(heart $maximum_heart_rate)
```

La prima è utilizzata per eliminare i valori di age minori di 0, la seconda per sostituire i valori di `maximum_heart_rate` maggiori di 222 con la sua media. Per cancellare le altre variabili inconsistenti invece utilizziamo la 1.5xIQR Rule

4 | Analisi descrittiva

Dopo aver reso il dataset tecnicamente corretto e consistente, bisogna fare una **analisi descrittiva** dei dati raccolti in modo tale da:

- avere una prima visione delle variabili raccolte
- controllare la presenza di errori, dovuti ad esempio al data-entry manuale
- valutare qualitativamente ipotesi e assunti
- determinare qualitativamente le relazioni tra le variabili

4.1 Grafici

Per prima cosa andiamo a vedere come le nostre variabili (qualitative e quantitative) sono distribuite, per avere una prima visione generale del dataset *heart*quindi utilizziamo i seguenti grafici: Il nostro dataset presenta molti pazienti compresi tra i 45 e 65 anni di eta circa **(figura 1.a**) e una prevalenza di pazienti di sesso maschile rispetto al sesso femminile (**figura 1.b**).

La maggior parte degli individui presenta un livello di colesterolo compresa tra i 200 e 250 (**figura 1.e**), mentre per la pressione sanguigna tra i 120 e 140 mm Hg (**figura 1.d**).

Altro da notare in questo dataset e che ci sono pochi pazienti che presentano un'anomalia dell'onda ST-T nei risultati elettrocardiografici (**figura 2.a**), cio fa intendere che e molto rara.

Infine, in questo set di dati, si puo notare che la distribuzione di chi soffre di una malattia cardiovascolare e leggermente minore da chi non ne presenta (**figura 2.h**) Ora determiniamo qualitativamente le relazioni tra la variabile *sesso* con le altre variabili:

4.1.1 target

Mettendo in relazione il sesso dei pazienti e la presenza di una malattia cardiovascolare, possiamo osservare che i malati sono principalmente maschi. Possiamo dunque ipotizzare che gli uomini hanno piu probabilita di avere una malattia cardiovascolare rispetto le donne.

4.1.2 Rest blood pressure

Guardando il boxplot affianco, possiamo affermare che non c'e una differenza di pressione sanguigna (a riposo) tra un paziente di sesso maschile e una femminile, infatti tutte e due stanno tra le 120 e le 140 mm Hg.

4.1.3 oldpeak

Relazionando la variabile sesso con la depressione ST, invece, si osserva che per i maschi si ha una depressione maggiore rispetto le femmine. Determiniamo qualitativamente le relazioni tra la variabile *target* (presenza d malattia cardiovascolare) con le altre variabili:

4.1.4 age

L'intervallo di pazienti con un disagio cardiovascolare, come si osserva dal boxplot, si aggira tra i 50 e 60 anni e sono piu anziane dei pazienti che non presentano disagi al cuore.

Ipotizziamo quindi che i pazienti con malattie cardiovascolari sono anziane; Difficilmente una persona giovane ha problemi cardiovascolari.

4.1.5 chest pain type

Mettendo in relazione la variabile target con i tipi di dolore al petto, notiamo molti individui asintomatici, ma che riportano malattie al cuore. Al contrario, chi presenta dolori di tipo anginale (tipico-atipico) e non anginale sono tipicamente pazienti che non presentano malattia cardiache.

4.1.6 cholesterol

Si puo confermare che i pazienti con disagi cardiaci hanno livelli di colesterolo piu alti rispetto ai livelli di un paziente sano, come si puo notare nella figura accanto.

Da notare anche la presenza di*outliers//anomalie*, ovvero la presenza di individui che non presentano malattie al cuore, ma con livelli di colesterolo molto elevate.

4.1.7 fbs

Lo zucchero nel sangue a digiuno maggiore di 120, per pazienti che riscontrano una malattia cardiovascolare sono minimi in confronto a coloro che hanno dei livelli di zucchero inferiore a 120. Si puo dire che coloro che non presentano una malattia al cuore, ma che hanno un livello di zucchero elevato nel sangue soffrono di diabete.

4.1.8 Maximum heart rate

I pazienti senza un disagio cardiaco hanno una frequenza cardiaca massima significativamente piu alta rispetto ai pazienti malati.

Da notare anche qui la presenza di *otlier/anomalie* nel boxplot dei non malati, mentre per l'otlier nel boxplot dei malati sicuramente si tratta di un errore dovuto al data-entry, in quanto la frequenza cardiaca di una persona a riposo si trova tra i 60 e 100 battiti al minuto.

4.1.9 oldpeak

Osserviamo che gli individui che presentano una malattia hanno una depressione ST piu alta dei pazienti sani. Anche in

questo boxplot si hanno delle *anomalie/outliers*

4.1.10 Altre variabili messe in relazione

age-maximum heart rate

Mettendo in relazione la variabile eta con la variabile della frequenza cardiaca massima possiamo affermare che piu un paziente invecchia, la sua frequenza cardiaca massima diminuira.

age-cholesterol

Si osserva dallo scatter plot che esiste una correlazione, seppur debole, tra l'eta e il colesterolo dei pazienti: piu l'eta avanza, maggiore saranno i livelli di colesterolo.

5 | Regressione lineare

Dopo l'analisi descrittiva, andiamo ad analizzare la relazione tra due variabili del dataset . Le variabili utilizzate per questa analisi sono age (eta) e rest_blood_pressure (pressione sanguigna del paziente a riposo). Andiamo a fare uno scatterplot:

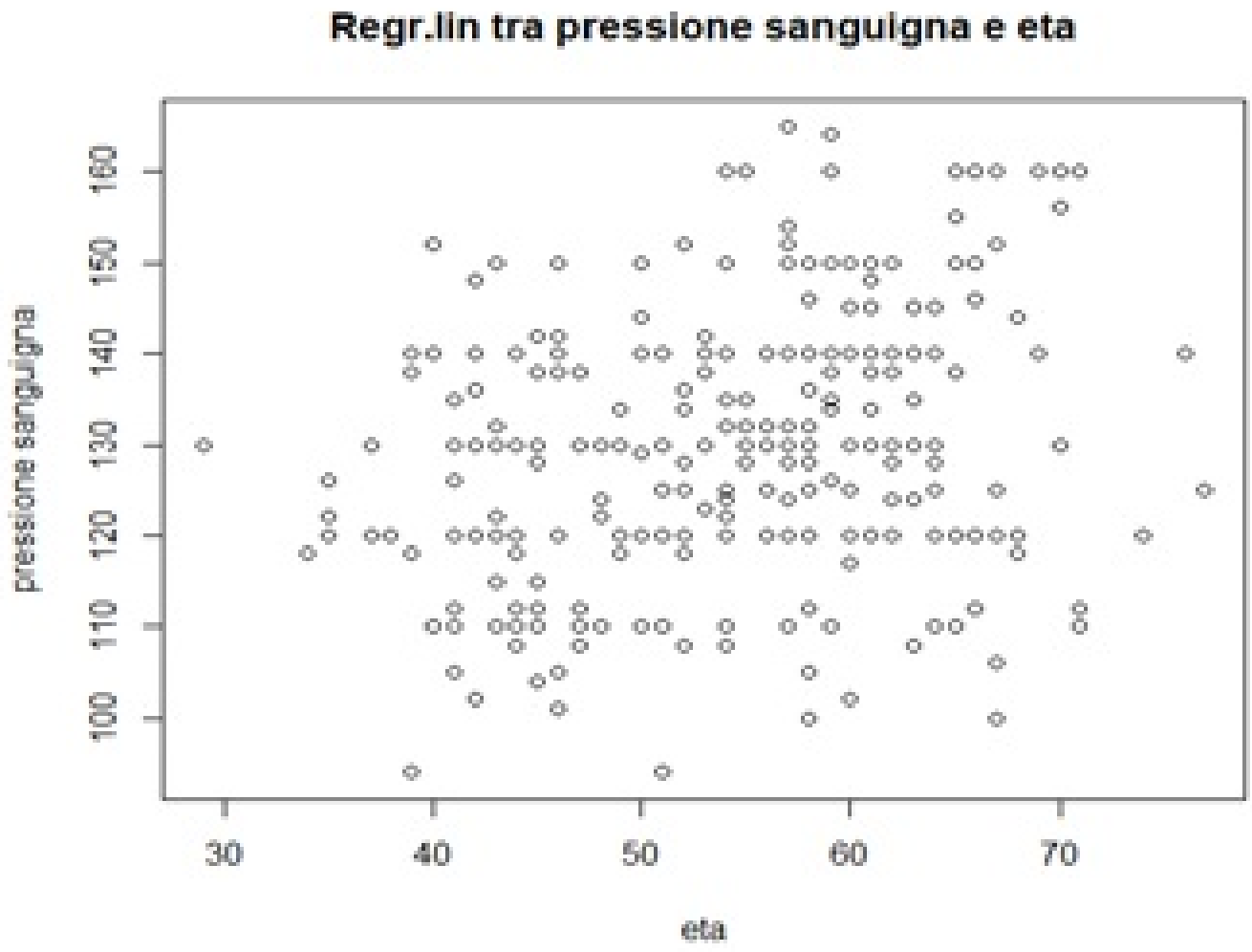


Figure 5.1: Scatterplot della regressione lineare tra pressione sanguinia e età

Secondo il grafico possiamo dire che e plausibile che la relazione sia **lineare**, notiamo che l'andamento dei valori tende a salire, cioe al crescere di eta cresce anche la pressione sanguigna (correlazione positiva). Attraverso la seguente linea di codice:

```
reg <- lm(blood_pressure ~ eta)
```

andiamo a disegnare la retta di regressione

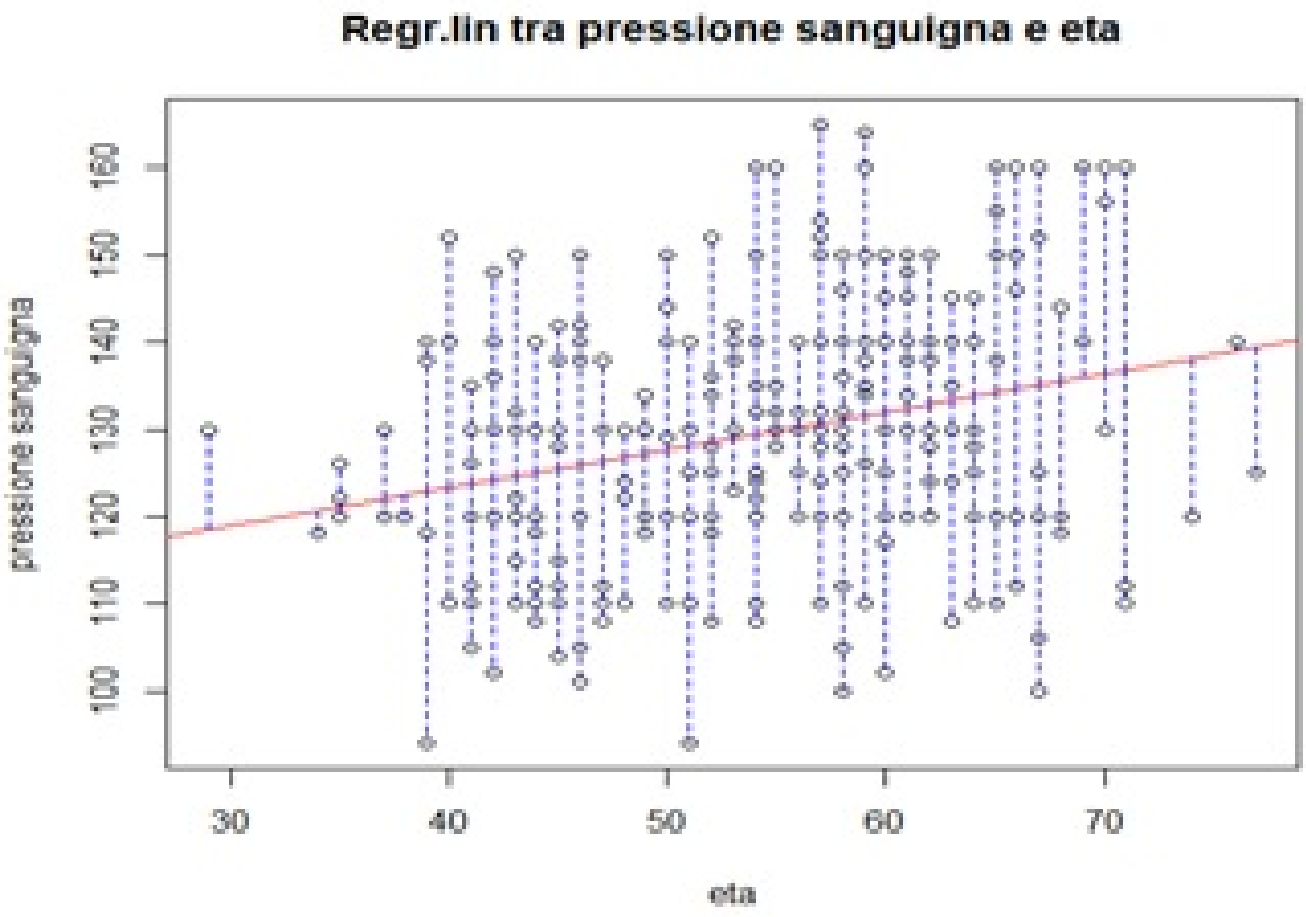


Figure 5.2: Scatterplot della regressione lineare tra pressione sanguinia e età con indicazione dello scostamento tra la retta di regressione e le osservazioni

Attraverso la funzione `summary()` troviamo le stime dei coefficienti di regressione `b0`, cioe il punto d'intersezione della retta con l'asse delle ordinate e `b1`, coefficiente angolare della retta :

```
b0 <- 105.80482
b1 <- 0.43744
```

quindi la nostra retta di regressione e data da: `rest_blood_pressure` = 105.80482 + 0.43744 * `age` sempre nella summary e presente anche il coefficiente di determinazione R^2 il quale assume il valore 0.07266 e possiamo dunque dire che il modello di regressione lineare e discreto. Riguardo il tipo di relazione utilizziamo il coefficiente di correlazione lineare r , che assume il valore 0.269552, possiamo dire che la relazione lineare e positiva (al crescere di eta cresce la pressione sanguigna), ma debole.

5.1 Analisi dei residui

L'analisi dei residui conferma che questi si distribuiscono uniformemente attorno all'asse. Si puo quindi confermare l'ipotesi di distribuzione casuale, infatti i valori sono equidistribuiti intorno alla retta e sono distribuiti sia sopra che sotto di essa con media nulla; In oltre sono incorrelati tra di loro.

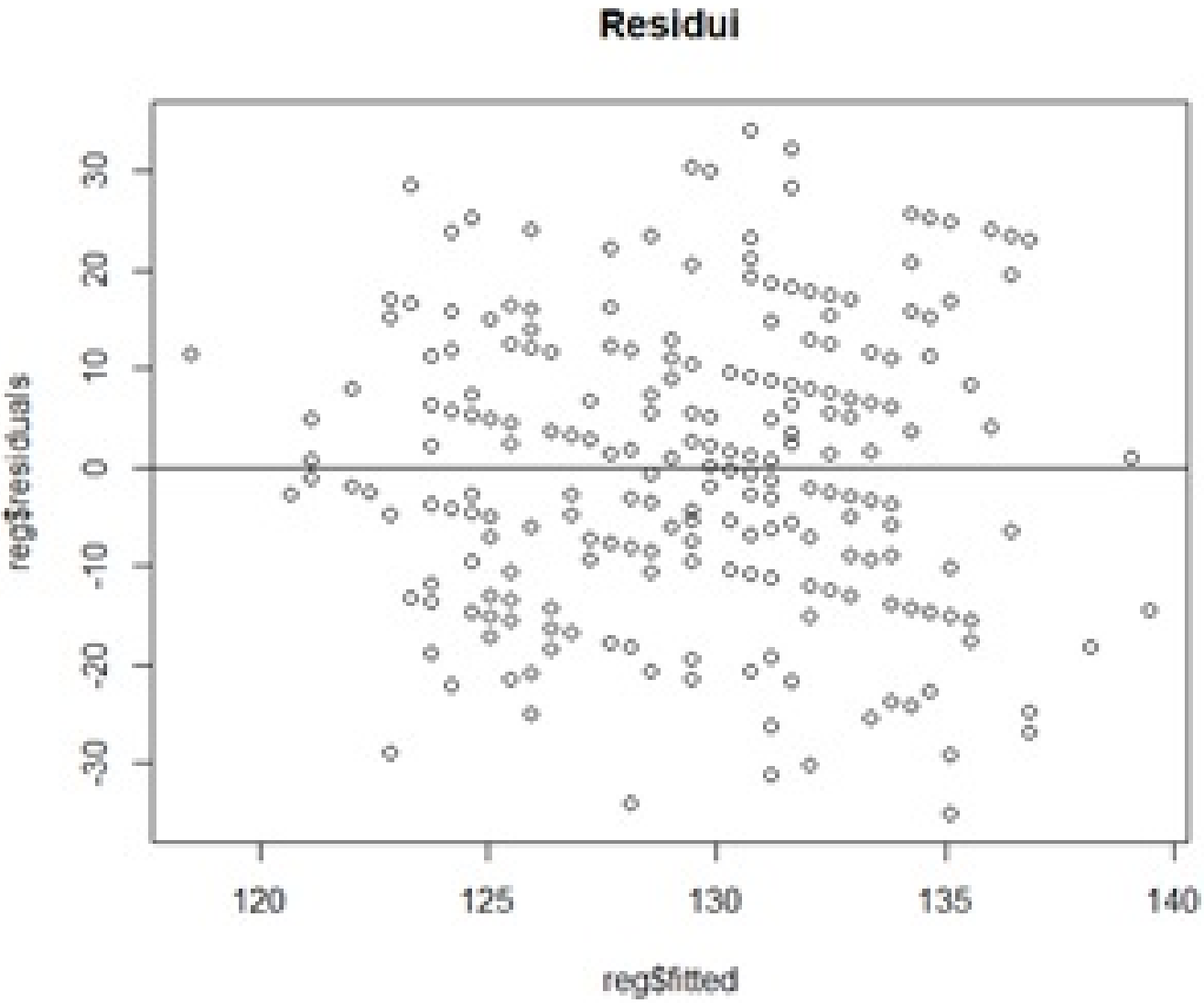


Figure 5.3: Q-Qplot

6 | Analisi della distribuzione in quantili

La distribuzione in quantili dei residui e confrontabile con quella di una normale.

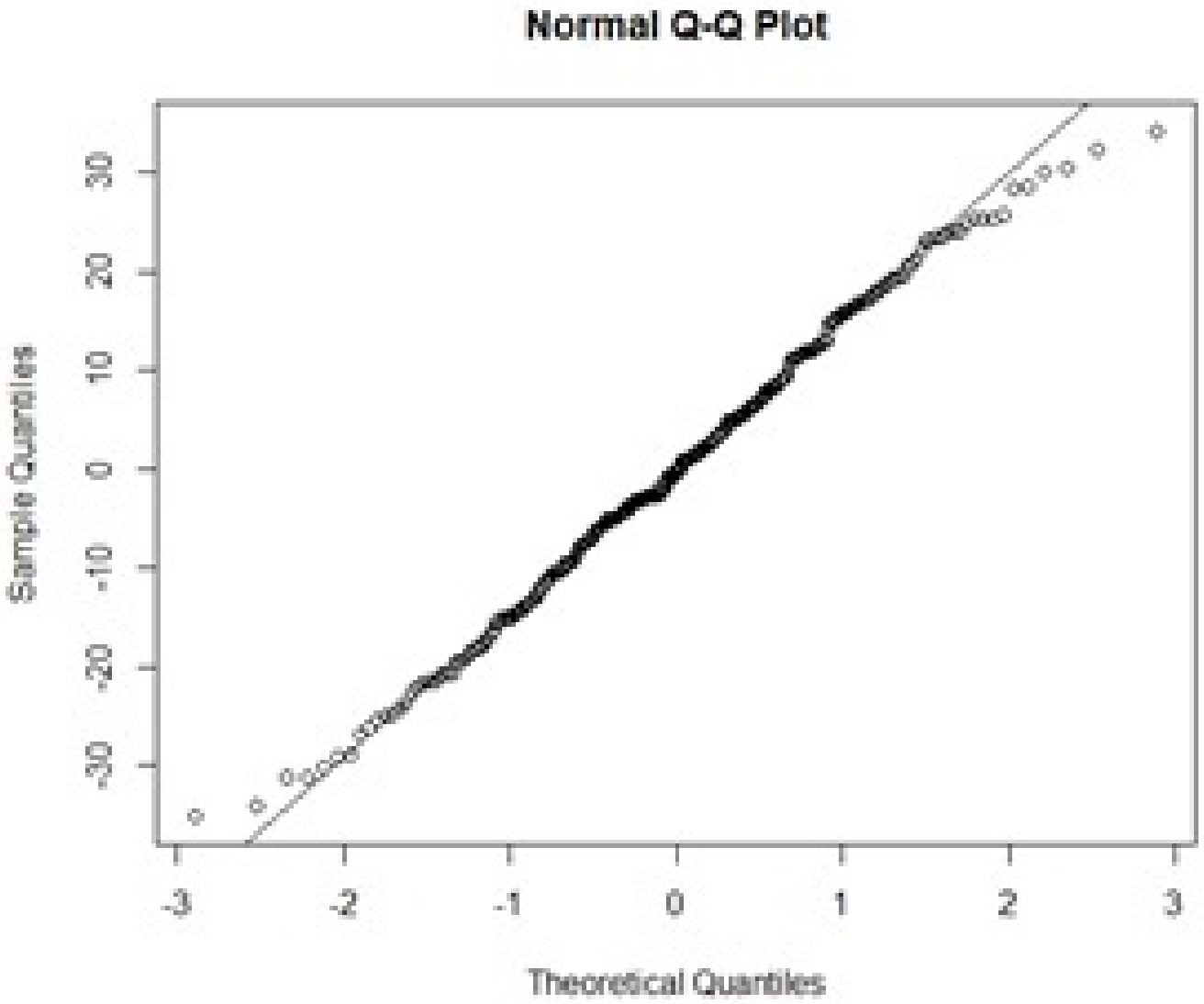


Figure 6.1: Q-Qplot

6.1 Previsioni

Dato che la retta di regressione ci permette di fare una stima del valore di y conoscendo il valore di x creiamo una dataframe contenente 10 osservazioni, cioe i valori dei predittori. Nel nostro caso proviamo a predire la pressione sanguigna a riposo

dei pazienti in base all'eta:

eta	**Pressione sanguigna(previsioni)**
30	118.9281
40	123.3026
50	127.6770
60	132.0514
80	140.8003
70	136.4259
55	129.8642
45	125.4898
56	130.3017
77	139.4880

7 | Machine Learning

7.1 Preludio al capitolo

Con il seguente capitolo si descrive l'analisi del modello predittivo di Machine Learning migliore tra un set di 7 algoritmi. Seguono degli abstract dei modelli ML e grafici delle predizioni calcolate dall'elaborazione dei modelli ML interessati

7.2 LDA Linear Discriminant Analysis

Di fronte all'analisi di più di due classi di classificazione, l'algoritmo più adatto è l'Analisi Lineare del Discriminante. L'algoritmo regola la distribuzione dei predittori X separatamente in ogni classe e utilizza il teorema di Bayes per ottenere delle stime per ogni probabilità delle categorie dato il valore predittore X stimato come:

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + log(\hat{\pi}_k) \tag{1}$$

- $\hat{\delta}_k(x)$ indica il valore stimato del discriminante che ricade nella k-esima classe assieme alla variabile di risposta,sulla base del valore del predittore x
- $\hat{\mu}_k$ indica la media di tutte le osservazioni di training a partire dalla k-esima classe
- $\hat{\sigma}^2$ indica la media di tutte le osservazioni di training a partire dalla k-esima classe
- $\hat{\pi}_k$ indica la probabilità precedente tale che un osservazione appartenga alla k-esima classe

7.3 CART Classification And Regression Trees

E' l'algoritmo ML predittivo (non-lineare) più usato, sia per la classificazione e sia per la regressione. Tale algoritmo esegue sui dati l' approccio del recursive partitioning, ovvero partiziona in ripetizione i dati in molteplici sottospazi, fintanto che ogni sottospazio è il più omogeneo possibile. Al termine, l'algoritmo restituisce un set di regole (visualizzate in un albero binario) per predire la variabile risultante, che può essere una variabile continua (per gli alberi di regressione), o una variabile categorica (per gli alberi di classificazione). Le regole sono definite dalla ripetuta frammentazione delle variabili di predizione, iniziando dalla variabile che è la più associata con la variabile di risposta, e finendo quando le variabili coincidono con alcuni vincoli di termine. L'albero si compone di nodi decisionali (nodi "radice"), nodi interni (nodi "ramificazioni") e nodi secondari (nodi "foglia"). Durante l'esecuzione dell'algoritmo, l'albero si espande fintanto che:

- Tutti i nodi secondari hanno una sola classe
- Esiste un numero degli elementi del campione che non può essere assegnato ad ogni nodo secondario
- Il numero degli elementi osservati nel nodo secondario ha raggiunto ii numero minimo pre-specificato

7.4 SVM Support Vector Machine

L'algoritmo cerca un iperpiano (ovvero un sottospazio lineare di n-1 esima dimensione) in una n-esima dimensione che classifica con distinzione le osservazioni del set di training. L'operazione di classificazione avviene tramite la separazione di due classi del set di training, al fine di trovare il margine superiore e massimizzare la distanza per classificare future osservazioni con più confidenza.

Gli iperpiani (nel caso in cui si tratti di uno spazio vettoriale $\mathbb{R}^3$ sono piani tridimensionali, altrimenti nel caso in cui si tratti di uno spazio vettoriale $\mathbb{R}^2$ sono linee) possono essere identificati come confini decisionali, nel quale, le osservazioni che ricadono ai lati dell' iperpiano sono attribuiti a classi differenti.

Infine, determina i punti più vicini nell'iperspazio, modificandone la posizione e l'orientamento. Tali punti sono identificati come vettori di supporto (support vectors), e la modifica di tali vettori comporta una corrispettiva modifica della posizione e dell'orientamento dell' iperpiano

7.5 kNN k-Nearest Neighbours

L'algoritmo k-Nearest Neighbors (vicini k-simili) è un modello non supervisionato e di classificazione non parametrico, nel quale sono presenti (anziché i parametri modello scoperti durante la fase di training), parametri di calibrazione (tuning parameters) i quali determinano l'esecuzione della fase di training. Tale algoritmo si presta particolarmente in analisi predittive sia di classificazione e sia di regressione,assegnando un etichetta di classe e valutando la distanza di una determinata osservazione a una simile osservazione trovata nel set analizzato. La distanza è calcolata sulla base di molteplici metodi di calcolo, tra cui il più utilizzato il metodo della distanza euclidea. La distanza euclidea è calcolata tramite la seguente formula:

$$\sqrt{\sum_{i=1}^{n}(x_i + y_i)^2} \tag{7.1}$$

Altri metodi comprendono la distanza di Hamming (distanza tra i vettori), la distanza di Manhattan (distanza tra i vettori calcolata sommando la differenza di tali vettori), e la distanza di Minkowski, la quale, attraverso una costante p, è definita come:

$$\left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p} \tag{7.2}$$

7.6 RF Random Forest

L'algorintmo *Random Forest* è un modello supervisionato il quale presenta un insieme di alberi decisionali che operano sulla base dell'apprendimento Ensemble. Nella prima fase, l'algoritmo:

- Seleziona casualmente gli elementi `K` dal totale di `m` elementi dove k $<$ m
- Tra gli elementi `K` calcola il nodo `d` utilizzando il miglior punto di divisione
- Ripetere i passi da `a` a `c` fintantoche raggiunge il numero `l` di nodi
- Genera la foresta ripetendo i passi da `a` a `d` per un numero `n` di volte per creare un numero `n` di alberi

Nella seconda e ultima fase:

- Prende gli elementi del test e usa le regole di ogni albero decisionale creato casualmente, al fine di per prevedere il risultato. Memorizza, quindi, il risultato previsto (target)
- Calcola i voti per ogni obiettivo previsto
- Considera l'obiettivo predetto con i voti più alti come la previsione finale dell'algoritmo *Random Forest*

7.7 MLP Multi-Layer Perceptron

L'algoritmo si presenta come una rete di percettroni. Un percettrone si compone di una funzione di attivazione `f(x)`, il quale può essere una funzione sigmoidale , oppure una funzione *RELU.*Come il nome suggerisce, i percettroni sono collegati su più livelli

7.8 Sommario dell'accuratezza dei modelli

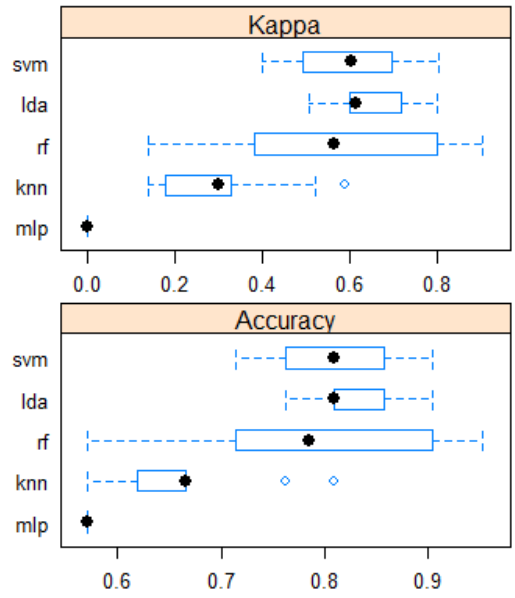


Figure 7.1: Boxplot con estremi, per la comparazione dell'accuratezza e del coefficiente **k** di Cohen

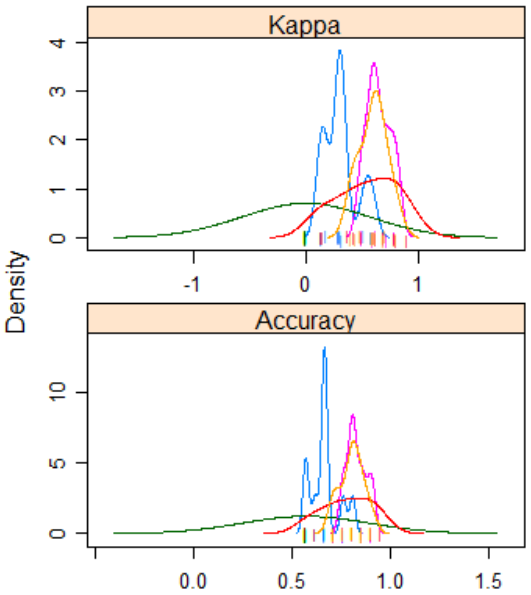


Figure 7.2: Plot di densità di comparazione dei modelli

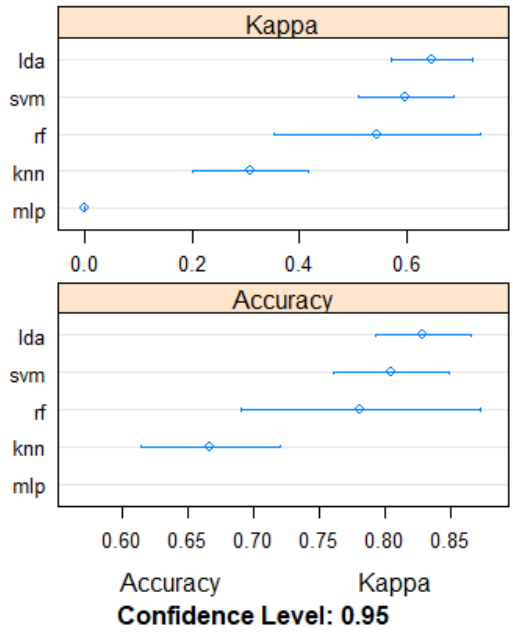


Figure 7.3: Dotplot di comparazione dell'accuratezza e del coefficiente **k** di Cohen

8 Conclusioni

Per identificare un paziente con una malattia cardiovascolare da uno sano abbiamo visto che si prendono in considerazioni vari fattori, come ad esempio la frequenza cardiaca massima, il colesterolo, l'età ecc.. Ad esempio un paziente con un alto colesterolo , che ha 55 anni, e la frequenza cardiaca bassa ,probabilmente è malato Dato che per un essere umano individuare se un paziente presenta un malattia risulta una procedura lenta e soggetta ad errori è possibile utilizzare degli algoritmi di machine learning per rendere la procedure piu efficace e meno soggetta ad errori. In conclusione possiamo dire che l'algoritmo piu adatto all' individuare i soggetti malati è lda.