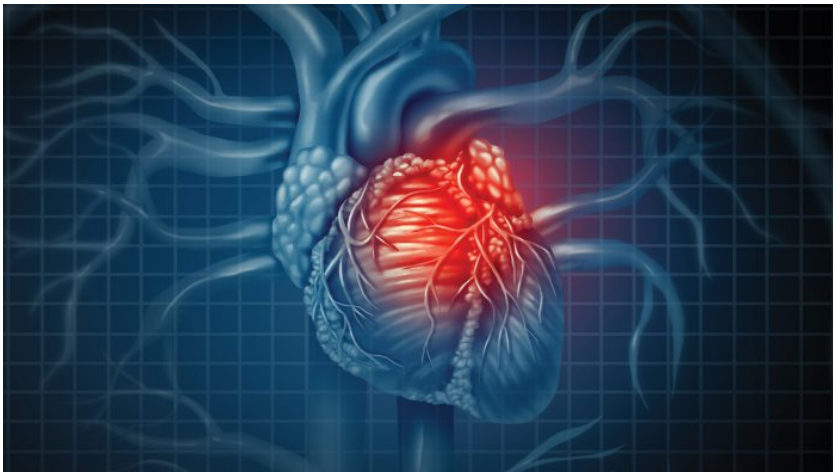
ITS Angelo Rizzoli
Corso ITS Machine Learning Specialist



Progetto in R: Heart Disease

Michele Segalini
Nicolò Manzoni

Corso 2020 - 2022

Indice

1 Introduzione

Questo progetto si focalizza sulle malattie cardiovascolari le quali possono colpire la struttura oppure le funzioni del cuore come ad esempio la disfunzione dell'arteria coronaria e quella vascolare.

A seguito del report dell'Organizzazione Mondiale della Salute, le malattie cardiovascolari rappresentano la prima causa di morte nel mondo mondo; secondo quanto riportato nel 2019 risultano decedute 8.9 MLN di persone a seguito di queste malattie (11 % di tutte le morti nel mondo).

Per studiare questa malattia, ci è stato fornito un dataset contenente dati reali forniti dai seguenti enti: Hungarian Institute of Cardiology (Budapest), University Hospital, Zurich (Switzerland), University Hospital (Basel, Switzerland), V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

2 Dataset

Il database principale contiene 76 attributi ma il nostro studio si baserà solamente su un dataset contenente 15 di questi attributi, fornito dal Cleveland database, per un totale di 303 osservazioni.

Nello svolgimento del progetto abbiamo utilizzato alcune librerie di r come "tidyverse" e "dbplyr" utilizzate nell'analisi dei dati, e la libreria "ggplot2" utilizzata per realizzare grafici piu accativanti durante la fase dell'Analisi Descrtittiva.

2.1 Analsi dei dati

Dopo aver caricato il dataset abbiamo subito proceduto con l'analisi del dataset stesso utilizzando il comando str(heart) per avere una panoramica del dataset.

Guardando il risultato del comando, notiamo la grande quantità di dati contenuta più precisamente 15 x 330, con ogni colonna rispettivamente relativa a una delle 7 categorie dei dati forniti e le righe invece rappresentano i pazienti dalla quale sono poi sono stati ricavati i dati.

Entriamo più nello specifico delle categorie presentate:

"x" -Nominale- nominale - Questa colonna rappresenta con il suo contenuto un id identificativo del paziente, il quale ha poi reso successivamente possibile il riconoscimento delle varie procedure eseguite al soggetto.

"Age" -Ordinale- Questa colonna contiene l'età dei pazienti

"Sex" -Nominale- Questa colonna contiene il sesso del paziente (1 = maschio, 0 = femmina)

"Cp" -Nominale- Questa colonna si sofferma sul tipo di dolore toracico del nostro paziente, gli studiosi che hanno creato il database lo hanno diviso in 4 tipologie (1. Tipico dolore dovuto da *Angina prectoris, 2. Dolore atipico dovuto da Angina Prectoris, 3. Dolore non dovuto all'Angina, 4. Asintomatico).

*L'Angina pectoris è una sindrome clinica caratterizzata da dolore e/o oppressione precordiale (porzione della parete toracica che ricopre la faccia anteriore del cuore) dovuta a ischemia miocardica transitoria (mancaza

di afluimento del sangue all'interno del cuore a seguito di sforzo o stress psicologico - spesso collegato all'infarto).

"Trestbps" -Ordinale- Questa colonna ci mostra il livello della pressione del sangue a riposo nel momento in cui i pazienti sono stati visitati in ospedale (ella può variare da un min di 0 mmHg a un massimo di 300 mmHg)

"Chol" -Ordinale- Questa colonna rappresenta la quantità di colesterolo totale presente nel *siero (in rapporto mg/dl)

*il Siero è un liquido di colore giallastro presente nel sangue; privato del fibrinogeno che rimane intrappolato nel coagulo sotto forma di fibrina, ha la stessa composizione del plasma sanguigno. Oltre a ciò contiene al suo interno il colesterolo.

"Fbs" -Nominale- Questa colonna rappresenta una delle tipologie di test al glucosio, che viene utilizzato per appurare quanto il corpo del paziente sia capace di normalizzare il livello del glucosio nel sangue a seguito di alterazione (diabete). In questo preciso caso vediamo se il livello del paziente è maggiore o no di 120 mg/dl (1. Maggiore di 120mg/dl, 2. Minore di 120mg/dl)

"Restegc" -Nominale- In questa colonna possiamo vedere quale diagnosi il nostro paziente ha ottenuto a seguito di un elettrocardiogramma a riposo, ogni numero da 0 a 2 è rispettivamente una tipologia di diagnosi diversa. (0. Non son state trovate anomalie, 1. Anomalia riscontrata nel ST-Wave*: inversione della T-wave e/o ST Elevazione o Depressione ¿ 0,05 mV, 2. Ipertrofia probabile o definita secondo i criteri di Estes)

*Intervallo dell'elettrocardiogramma che si situa tra la fase di sistole (contrazione del cuore) e la fase di diastole (rilassamento del cuore).

"Thalach" -Ordinale- Massima freuenza cardiaca ottenuta dal paziente

"Exang" -Nominale- In questa colonna vediamo se il nostro paziente a seguito di esercizio fisico ha avuto fitte torachiche (sempre di tipologia Angina), se otteniamo 1 vuol dire che il paziente ha avuto questi malori invece 0 no.

"oldpeak" -Ordinale- Livello della depressione del ST Segment forzata tramite esercizio fisico

"slope" -Nominale- Pendenza a partire del J point* del ST segment rilevata nell'elettrocardiogramma (1. Crescente, 2. Piana, 3. Decrescente)

*J Point: Punto dell'ST Segment dalla quale parte la pendenza a seguito dell'attività fisica del paziente

"ca" -Nominale- In questa colonna verifichiamo se il nostro paziente è affetto da una mallattia relativa alla arteria cornaria, e la "flouroscopy" che è stata utilizzata è uno dei metodi non invasivi per farlo. I risultati vanno da 0 a 3 per verificare quanti dei condotti che legano il cuore a tutto il corpo si colorano a seguito di questo intervento.

"Thal" -Nominale- Questa colonna rappresenta i risultati ottenuti dai nostri pazienti a seguito del test dello sforzo che misura il flusso sanguigno al cuore che puo assumere tre tipologie di valori, normal, fixed defect e reversable defect (rispettivamente i valori 1, 2 e 3).

"Target" -Nominale- A seguito di tutti gli esami e test esguiti sui pazienti, la colonna target ha il ruolo di definire se tale paziente ha una possibilità maggiore o no del 50 perdcento di essere affetto da una malattia cardiaca. (1. Si, 0 No)

2.2 Trasformazione del tipo dei Dati

Abbiamo proceduto all'analisi dei dati; prima di tutto abbiamo dato un'occhiata ai tipi delle variabili (interi, stringhe, numeriche, ecc...) attraverso il comando "str(heart)"

```
'data.frame':   303 obs. of  15 variables:
$ x       : int  1 2 3 4 5 6 7 8 9 10 ...
$ age     : int  63 37 41 56 57 57 56 44 52 57 ...
$ sex     : chr  "1" "1" "0" "1" ...
$ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
$ trestbps: int  145 130 130 120 120 140 140 120 51 150 ...
$ chol    : chr  "233" "250" "204" "236" ...
$ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
$ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
$ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
$ exang   : int  0 0 0 0 1 0 0 0 0 0 ...
$ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
$ slope   : int  0 0 2 2 2 1 1 2 2 2 ...
$ ca      : int  0 0 0 0 0 0 0 0 0 0 ...
$ thal    : int  1 2 2 2 2 1 2 3 3 2 ...
$ target  : int  1 1 1 1 1 1 1 1 1 1 ...
```

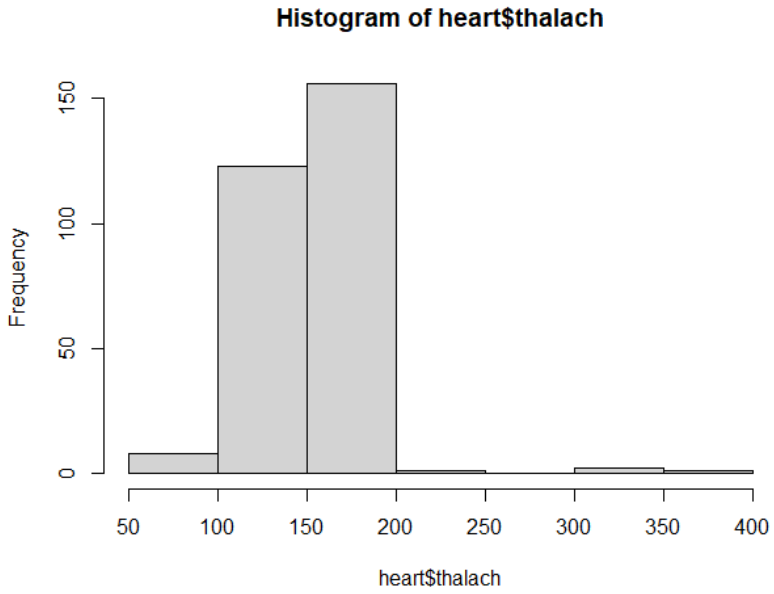
Successivamente abbiamo trasformato alcune colonne in "factor", poichè ritenevamo necessario che ci fossero delle suddivisioni in livelli all'interno di queste variabili : "sex", "cp", "chol", "fbs", "restecg", "exang", "slope", "ca", "thal" e "target". Dopo aver trasformato queste colonne in fattori, abbiamo rinominato i livelli di ognuno di essi utilizzando la sintassi "dplyr" che fa parte del pacchetto libreria "tidyverse" in modo che i dati fossero più comprensibili. Nella colonna "thal" siamo andati a rimuovere il livello 0, poichè la variabile "thal" può assumere soltanto valori interi da 1 a 3.

Dopo aver trasformato le varibili nel tipo corretto, facendo una verifica più approfondita nelle colonne, abbiamo notato la presenza di alcuni valori "unspecified, undefined" tramite l'utilizzo del comando "levels(...)" nelle rispettive variabili "sex" e "chol". Abbiamo trasformato tali valori in NA per rendere più semplice la rimozione di tutti i valori NA all'interno del dataset attraverso il comando (na.omit). In aggiunta abbiamo deciso di rimovere la colonna "x", ritenuta inutile poichè non sarebbe stata di rilievo per le analisi del dataset.

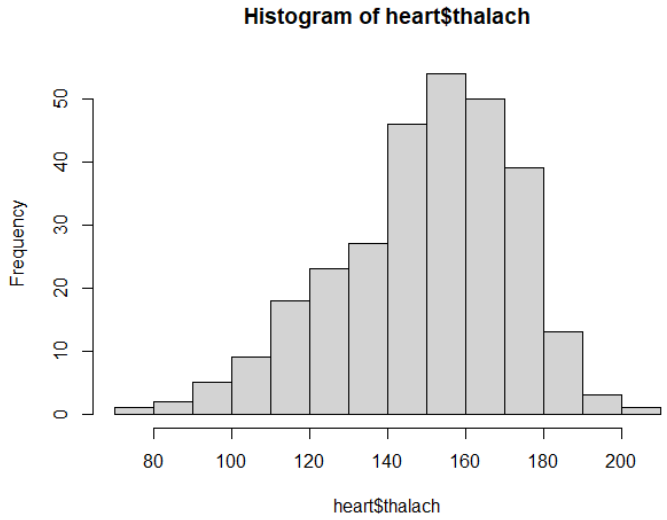
2.3 Consistenza dei Dati

Inizialmente abbiamo verificato la consistenza dei dati tramite il comando summary, notando che in alcune colonne erano presenti dati non consistenti rispetto agli standard.

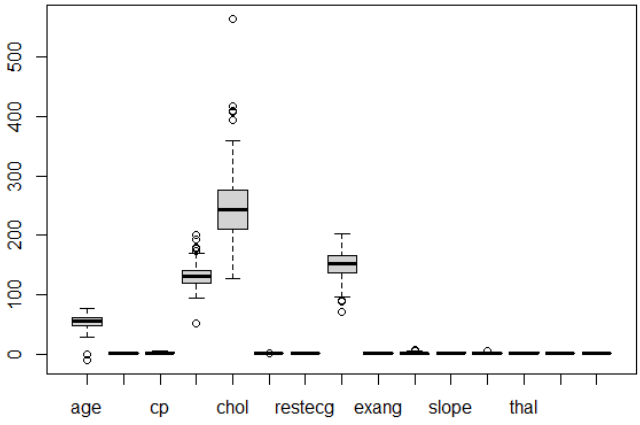
Successivamente abbiamo denotato una inconsistenza presente nella colonna "thalach", visto che vi erano presenti valori superiori a 222bpm e per natura la frequenza cardiaca non può essere così alta.



Per vedere i dati erronei più in dettaglio abbiamo deciso di fare un instogramma rappresentate la colonna. Dopo aver ragionato tra di noi per trovare una soluzione a questo problema, abbiamo deciso di sostituire tali dati incorretti con la media di tutti i valori presenti nella colonna (eccetto quelli maggiori di 222).

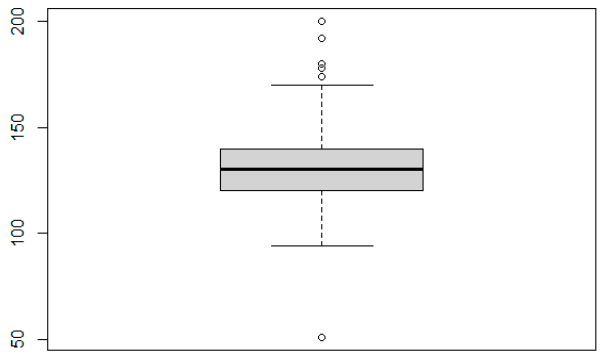


Successivamente per verificare sempre la consistenza abbiamo visualizzato una panoramica generica degli outliers di tutto il dataset:



Visto che il processo di rimozione degli outliars è il medesimo per tutte le colonne, mostremo semplicemente la rimozione di tali outliers prendendo la colonna "trestbps" come esempio.

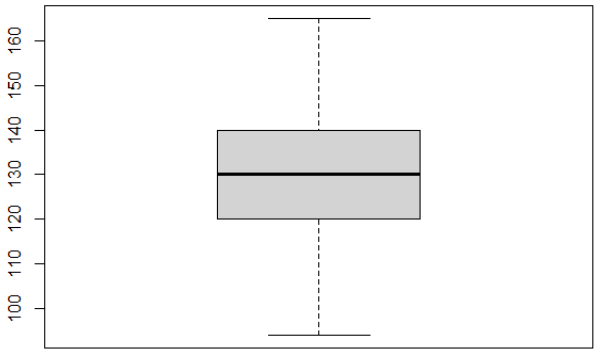
Per questa dimostrazione partiamo con la visualizzazione del boxplot della colonna presa in esame:



Abbiamo calcolato il valore del terzo e del primo quartile, con lo scopo di calcolare successivamente la differenza interquartile (Q3- Q1) e il min e max valore della colonna.

Dopo aver calcolato tali valori, abbiamo rimosso gli outliers attraverso un filtro che includeva il val min e max calcolati precedentemente.

Infine visualizziamo il boxplot a seguito della rimozione dei valori.

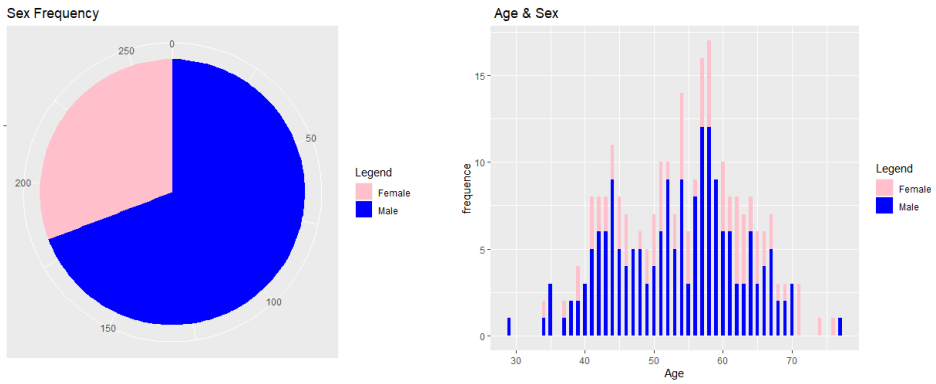


Questa operazione è stata effettuata anche sulle variabili "age", "chol" e "oldpeak".

3 Analisi Descrittiva Approfondita

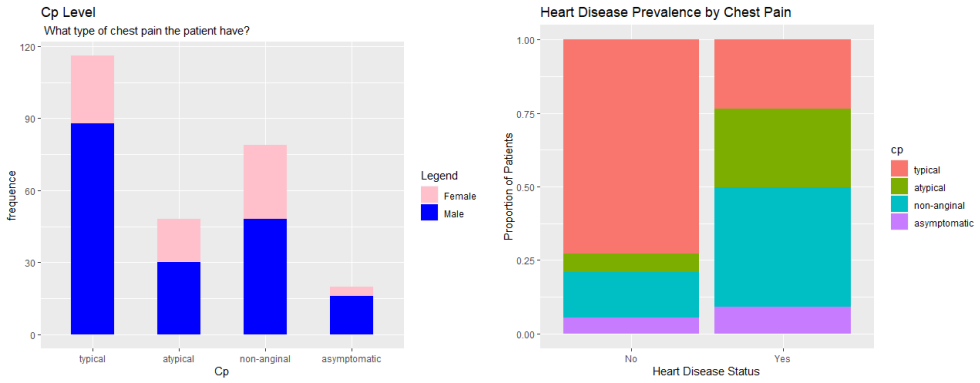
Grazie alla libreria *ggplot2* siamo riusciti a effettuare analisi più approfondite dei dati delle singole colonne e anche in relazione tra di loro.

Analizziamo la variabile *sex* in relazione ad *age*:



Dal grafico a torta si può notare molto semplicemente che la quantità di soggetti maschi sottoposta alla analisi è prettamente maggiore della quantità di soggetti femminili. Sull'istogramma di sinistra si può denotare come il rapporto *età-sesso* sia più omogeneo per gli uomini che per le donne, le quali risultano divise a scaglioni. Se si vuole prendere in consideraione solo il fattore età dell'istogramma si può vedere che la moda è di circa sessanta anni

Analizziamo la variabile *chest pain* (dolore toracico, cp) in relazione con *sex* nel grafico sulla sinistra e *heart-disease* nel grafico sulla destra:

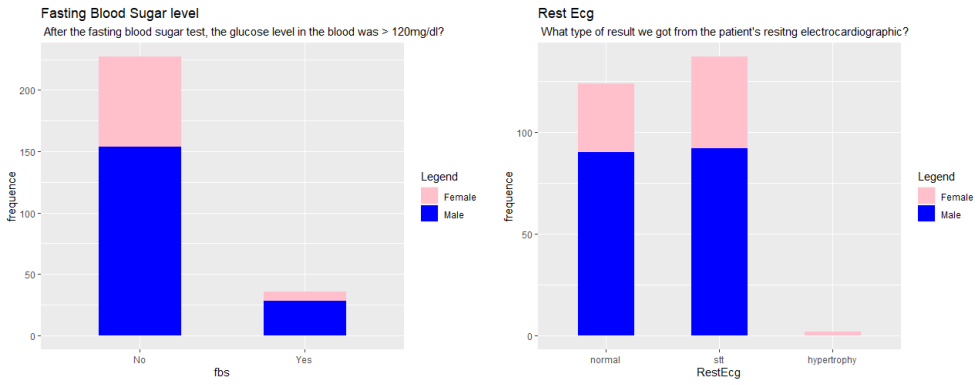


Dal primo grafico si può notare che tra gli uomini è molto più comune essere affetti da *typical angina cp* invece per le donne non c'è una grande differenza tra i soggetti affetti da *typical angyna cp* o *atypical angyna cp*.

La presenza di *asymptomatic cp* a differenza di tutte le altre presenta una frquenza molto bassa da parte di tutte e due i sessi (in particolare quasi nulla da parte del genere femminile).

Invece nel secondo grafico possiamo notare come i soggetti non affetti da malattie cardiache tendono ad avere una tipologia di *cp* del tipo *typical angyna*, invece coloro affetti da patologie cardiache sono propensi ad avere una tipologia di cp *non-angynal*.

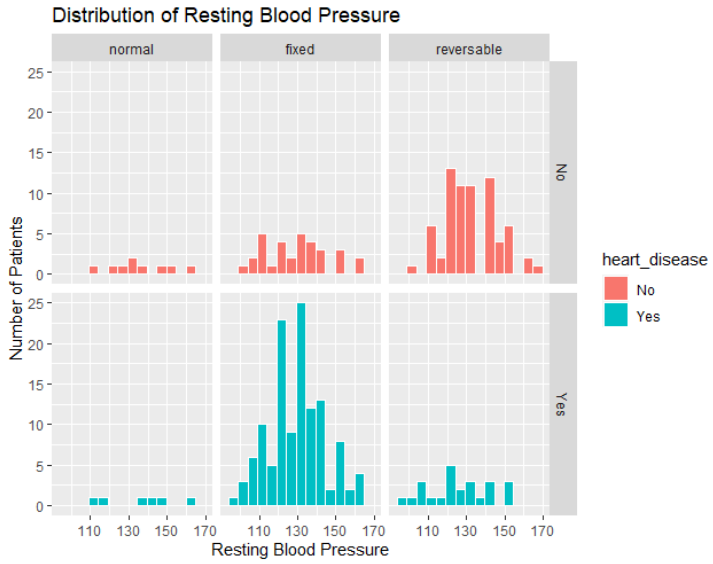
Analizziamo il rapporto tra *sex* e le variabili *fbs* e *restecg*:



Dal primo grafico si può semplicemente constatare come sia poco frequente il risultato positivo di questo test, e in particolare da parte del genere femminile.

Nel secondo grafico si può denotare fin da subito come il risultato *hypertrophy* dall'elettrocardiogramma dei soggetti sottoposti sia praticamente quasi nullo eccetto per qualche paziente di sesso femminile. Per il resto dei risultati invece si ha un equilibrio tra i soggetti maschili che hanno avuto *normal* e *abnormal st segment* come risultati.

Analizziamo la variabile *trestbps* in relazione a *heart-disease* e *thal*:



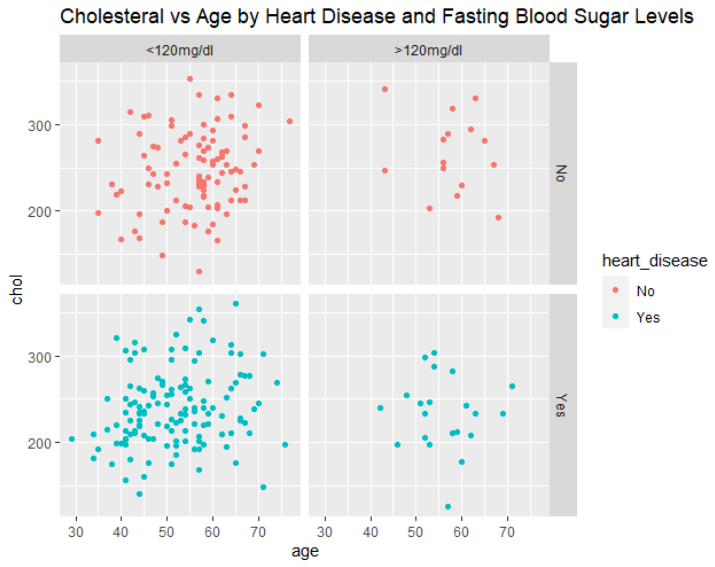
In questo grafico possiamo osservare che coloro affetti da malattie cardiache vengono principalmente diagnosticati con la tipologia identificata come *fixed defect* dal test dello sforzo. Oltre a ciò i soggetti identificati in questo riquadro presentano una moda di *trestbps* intorno ai *130mmHg*.

Per i pazienti ai quali non è stata rilevata alcuna malattia cardiaca si è notata un'alta frequenza di *reversable* come risultato dal test dello sforzo. In

questo caso la trestbps (pressione sanguigna a riposo) registrata è in gran parte compresa tra *120mmHg* e *150mmHg*.

Infine dal grafico si può denotare come coloro affetti e non affetti da una malattia cardiaca alla quale non è stata identificata alcuna malformazione attraverso il test dello sforzo rappresentano solo una piccola parte dei pazienti presi in considerazione.

Analizzimao la variabile *chol* in relazione a *heart-disease*, *age*:

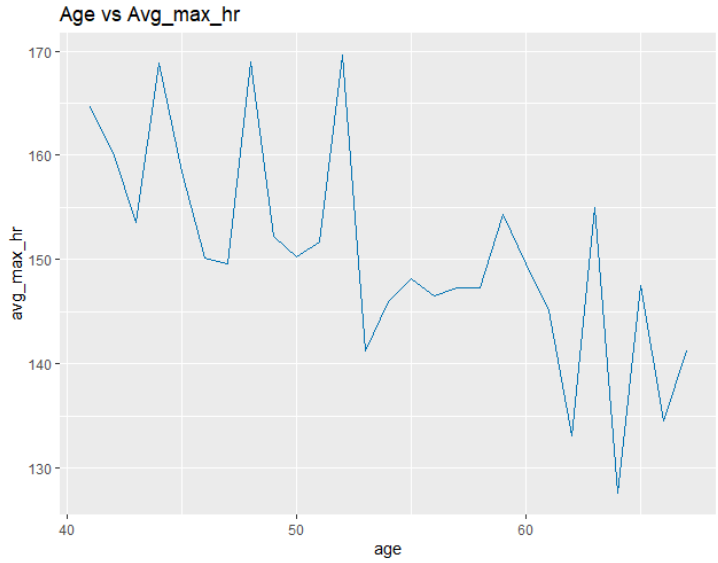


Da primo impatto notiamo come ci siano poche persone all'interno di questo studio che abbiano ottenuto un risultato *>120mg/dl* dal test del glucosio.

Dal grafico poi denotiamo come coloro non affetti da malattie cardiache che hanno ottenuto come risutato dal test al glucosio *<120mg/dl* sono maggiormente distribuiti tra i *55* e i *65* anni e tendono ad avere un valore di

chol tra i *200mg/dl* e *300mg/dl*, invece coloro che hanno avuto il medesimo risultato dal *fbs*, ma alla quale è stata diagnosticata una malattia cardiaca sono principalmente disposti tra i *40* e i *55* anni e sono orientati ad avere un *chol* tra *200mg/dl* e *250mg/dl*.

Analizzimao la variabile *thalach* in relazione a *age*:

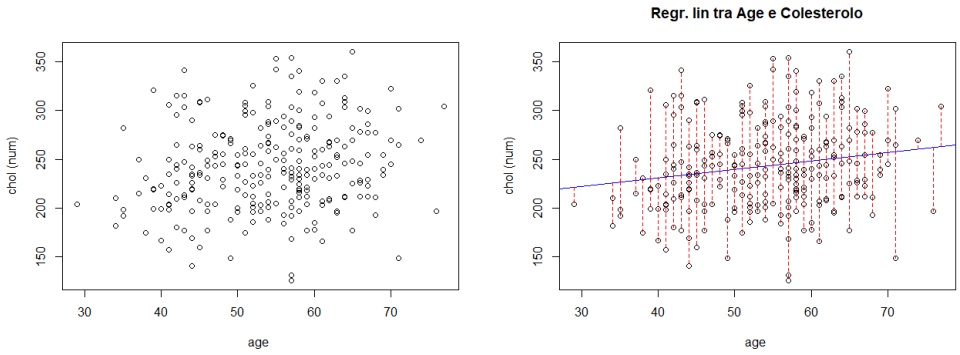


In questo grafico sono state prese in considerazione soltanto le età con almeno una frequenza di 5 pazienti per ogni età (per questo nel grafico precedente è possibile vedere pazienti con età minore di 40 o maggiore di 70).

Come si può notare dall'immagine, per i pazienti che hanno una età maggiore di 53 anni, la pressione cala in modo determinante senza mai raggiungere valori medi maggiori di 155bpm. Per i pazienti con età inferiore ai 53 anni, si verifica esattamente l'opposto: la pressione media risulta sempre maggiore di 150bpm, con picchi che arrivano circa a 170bpm.

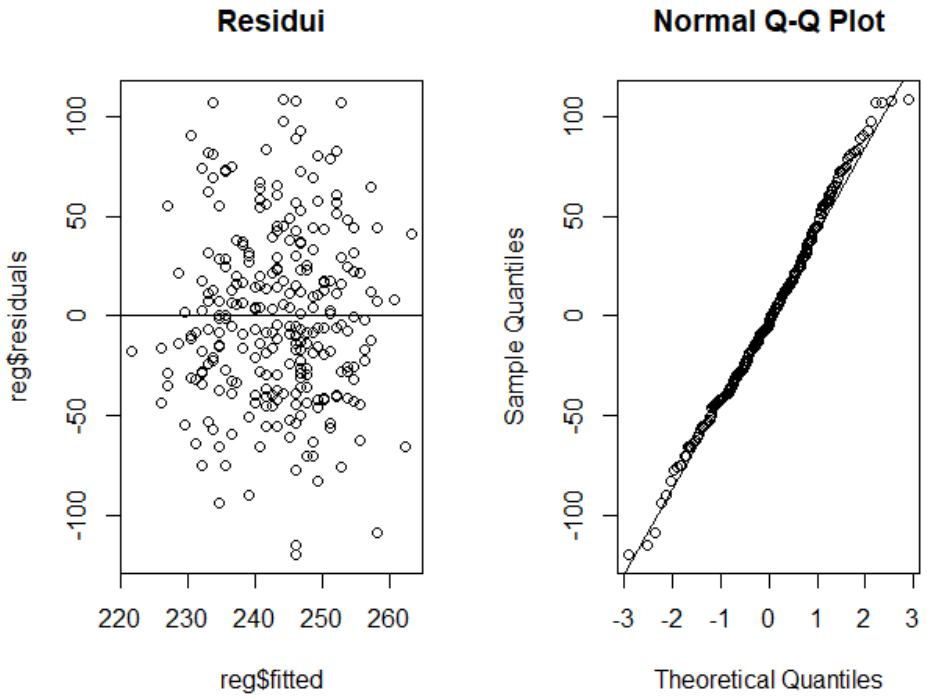
4 Regressione Lineare

In questa fase del progetto abbiamo analizzato la relazione tra due variabili del dataset. Nel nostro caso abbiamo messo in relazione la variabile *age* con *chol*.



Abbiamo incominciato con il grafico di dispersione, chiamato anche "scatterplot". Successivamente abbiamo disegnato una retta per capire il tipo di correlazione tra le 2 variabili prese in considerazione. A prima vista non sembra esserci una correlazione molto forte. Tramite il comando summary siamo andati a vedere il valore R-squared il quale ci fa capire se c'è una buona correlazione nel caso il valore si avvicini ad 1, mentre nel caso il valore sia tendente allo 0, significa che i dati non sono completamente correlati con conseguente alta varianza dei residui. Nel nostro caso il valore di R-squared risulta essere (0.031), da questo deduciamo che ci sia una alta varianza dei redisui.

Siamo passati poi alla analisi dei residui.



Dallo scatterplot di sinistra non si nota alcun pattern particolare, quindi sembrerebbero disposti in maniera del tutto casuale. Anche se non si denota una particolare correlazione grazie all'annalisi dei residui possiamo affermare che i dati sono qualitativamente accettabili.

Il grafico di destra conferma l'ipotesi di distribuzione casuale dei residui, poiché i valori sono equidistribuiti intorno alla retta.

4.1 Effettuare Previsioni

In questa fase del progetto abbiamo costruito un nuovo dataframe nel quale abbiamo inserito 10 osservazioni che non fossero presenti nel dataset precedentemente utilizzato per le analisi. Abbiamo scelto di inserire 10 valori di *age* per poter prevedere quale fosse la quantità di colesterolo (chol) totale presente nel siero di ogni paziente.

Per visualizzare gli intervalli di confidenza al 95% intorno alla media delle previsioni viene specificata l'opzione intervallo = "confidenza":

```
> heart_prevision <- data.frame("age" = c(47, 36, 56, 31, 29, 40, 42, 50, 35, 58))
> predict(reg, heart_prevision, interval = "confidence")
       fit      lwr      upr
1  237.1899 230.4566 243.9232
2  227.6667 215.8580 239.4753
3  244.9817 239.4947 250.4686
4  223.3379 208.8579 237.8179
5  221.6064 206.0319 237.1809
6  231.1297 221.3514 240.9079
7  232.8612 224.0397 241.6826
8  239.7872 233.9490 245.6253
9  226.8009 214.4674 239.1345
10 246.7132 240.8590 252.5673
>
```

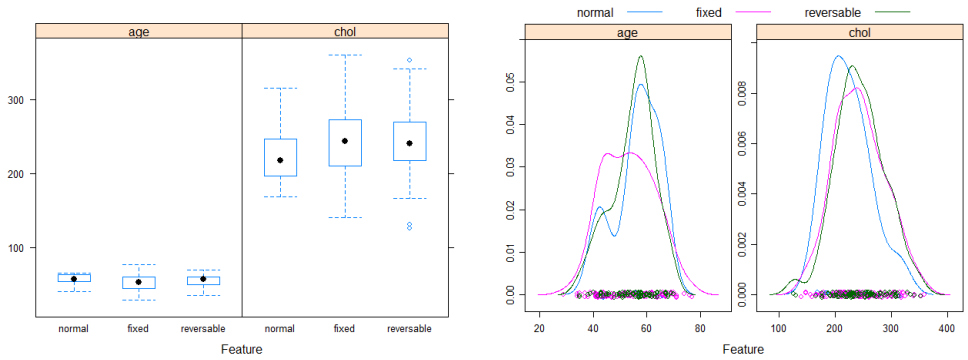
Come si può vedere dallo screenshot sopra, l'output della funzione "predict" ritorna 3 tipi di valori: "fit", "lwr", "upr". Prendiamo ad esempio la prima riga; il valore "fit" (237.1899) rappresenta il livello di colesterolo previsto per il paziente di età 47, le colonne *lower* (lwr) e *upper* (upr) identificano i limiti di confidenza inferiore e superiore per i valori attesi. Ad esempio, l'intervallo di confidenza del 95% associato ad un paziente di età 47 è (230,4566, 243,9232). Ciò significa che un paziente di 47 anni ha in media un livello di colesterolo compreso tra *230,4566* e *243,9232 mg/dl*.

5 Modello Machine Learning

Come ultimo step dell'analisi del dataset Heart, abbiamo predisposto una parte del dataframe stesso sul quale eseguire un algoritmo di Machine Learning. Per la creazione del nuovo dataframe abbiamo tenuto in considerazione soltanto alcune delle colonne presenti nel dataframe di partenza, nello specifico : "age", "chol", "thal", "heart-disease". Tramite il comando str abbiamo visualizzato il dataset appena creato.

```
> str(heart_ml)
'data.frame':   263 obs. of  4 variables:
 $ age          : int  63 37 41 56 57 57 56 44 57 54 ...
 $ chol         : num  233 250 204 236 354 192 294 263 168 239 ...
 $ thal         : Factor w/ 3 levels "normal","fixed",..: 1 2 2 2 2 1 2 3 2 2 ...
 $ heart_disease: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
>
```

Successivamente abbiamo effettuato alcune operazioni per analizzare la struttura dei dati, tramite dei boxplot abbiamo verificato la distribuzione delle varie classi e dei valori per ogni classe. Qui di seguito i grafici realizzati per questo scopo:

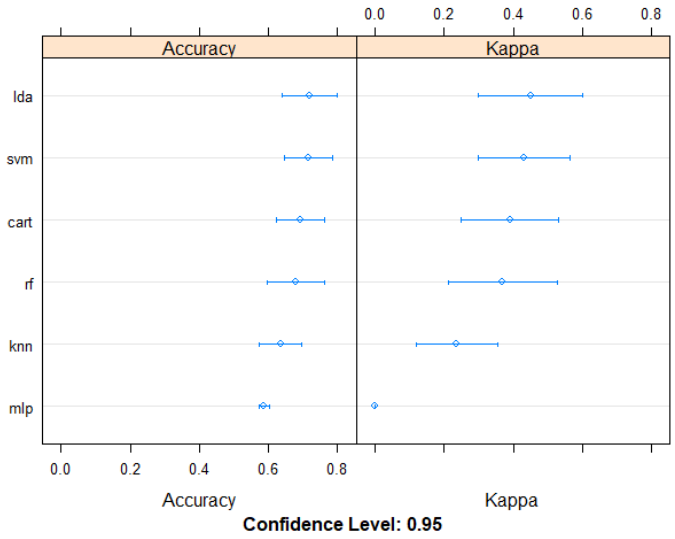


Per poter elaborare un algoritmo che effettui predizioni, abbiamo bisogno di suddividere il nostro dataframe creato in precedenza: per questo scopo si è deciso di creare una matrice che contenga l'80% delle osservazioni (questa tabella verrà poi utilizzata come training set), mentre il rimanente 20% dei dati verrà utilizzato come test set.

Per trovare l'algoritmo più efficiente per il nostro dataset andremo a testare 6 modelli diversi con lo scopo di stimarne la loro accuratezza. Per questo fine utilizzeremo algoritmi lineari come l'LDA (linear discriminant

analysis), algorimi non lineari come il CART (classification and regression trees) e un algoritmo avanzato come il RF (random forest).

Dopo aver fatto elaborare gli algoritmi dal pc, visualizziamo i risultati ottenuti da ognuno di esso tramite il comando *results*, per avere una visione più chiara abbiamo deciso di inserirli all'interno di un dotplot:



Il Grafico si divide in due sezioni, una rispetto al parametro Accuracy e una rispetto al parametro Kappa.

Accuracy paragona il dataset iniziale (creato per testare gli algoritmi) e quello di test per vedere se le diagnosi ottenute nelle previsioni rispecchiano quelle presenti nel dataframe di partenza.

Kappa invece paragona i singoli dati previsti nel dataset di testing e i loro rispettivi nel dataset base per vedere quanto si è avvicinato l'algoritmo ai dati di partenza.

6 Conclusioni

Dopo aver controllato ogni previsione creata dal nostro algoritmo di Machine Learning, abbiamo concluso che "lda" è risultato essere l'algoritmo con il grado di accuratezza più alto rispetto agli altri (Accuracy $\sim 0.7, Kappa \sim 0.45$).

Anche se i risultati raggiunti non hanno avuto un livello di accuratezza alto, siamo rimasti soddisfatti per gli obbiettivi raggiunti nella realizzazione del nostro primo progetto di machine learning su R.