Report esame R

Perdicchia Stefano

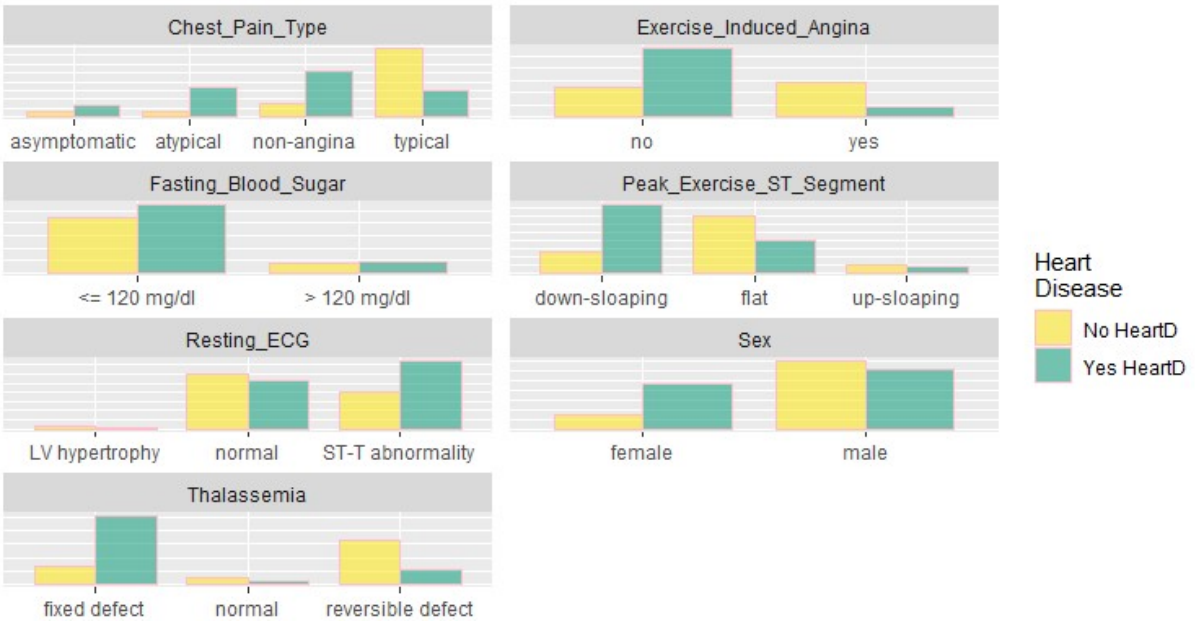
Capitolo 1: Introduzione e obiettivi

Come esame finale del modulo di R abbiamo preso in esame l'analisi dei dati del dataset "Heart.csv", questo dataset, dopo diverse ricerche abbiamo scoperto essere stato sviluppato e studiato grazie a diverse tecniche di machine learning dal: V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.. Come si può notare questo è solo un estratto del dataset originale che contiene 76 attributi, mentre nel database che andremo a studiare ne sono presenti solamente 15.
L'obbiettivo finale del nostro studio sarà di andare a comprendere meglio i dati, quindi dopo averli sistemati e dopo aver fatto le corrette analisi, andare ad implementare i tre algoritmi posti, quindi quello di forza bruta, del gradiente e infine un algoritmo di machine learning. Infine lo scopo principale è quello di comprendere meglio lo studio e i mezzi che R come linguaggio mette a disposizione, per poter elaborare dei dati complessi come quelli a cui andremo a fare riferimento.
Durante la stesura del codice sono state prese diverse decisioni sulla sistemazione dei dati, che sono tutte state motivare all'interno dei commenti e durante lo sviluppo stesso del codice, quindi per le mancate informazioni che non saranno riportate all'interno di questo report, si farà fede ai dati presenti ed esposti all'interno del file con estensione R.
Il primo scopo che ci siamo dati è stato quello di rendere i dati che sarebbero stati analizzati tecnicamente corretti e consistenti, che all'interno della traccia di esame erano segnalati come punti fondamentali dello sviluppo del linguaggio, anche se ne andremo a parlare in maniera più approfondita nei capitoli successivi.
Successivamente dopo aver sistemato e ricontrollato molteplici volte i dati abbiamo svolto delle indagini all'interno dei dati per comprendere come e su quali dati sviluppare una regressione lineare che avesse un buon valore di corrispondenza tra i dati stessi, alla fine dopo diversi tentativi siamo riusciti a trovare i dati di tipo numerico che corrispondevano alla nostra descrizione e abbiamo successivamente riutilizzato gli stessi dati applicandoli agli algoritmi, punto finale di questo esame.
Infatti per concludere abbiamo implementato i due algoritmi di orza bruta e del gradiente per ottenere dei risultati consistenti e infine abbiamo applicato due diversi algoritmi di machine learning per comprendere la correttezza dei dati esposti all'interno del codice.
Adesso entriamo nel dettaglio del lavoro che abbiamo svolto.

Capitolo 2: Descrizione del Dataset

Dopo aver caricato il dataset all'interno dell'ambiente di lavoro, quindi abbiamo controllato che la Directory fosse impostata correttamente e che R fosse perfettamente aggiornato, abbiamo richiamato diverse funzioni di calcolo su R per andare a controllare e per svolgere una prima analisi sul dataset che ci era stato assegnato.
Questa analisi è stata suddivisa in diverse fasi per una comodità pratica in quanto i dati presenti all'interno del dataset erano più di 4000, quindi abbiamo deciso di fare un'analisi inizialmente su quelle che erano le colonne e i significati che potevano assumere i record presenti all'interno di ogni colonna.
Abbiamo quindi analizzato le 15 colonne che erano presenti all'interno del dataset "Heart", abbiamo subito notato che la prima colonna, che era denominata "x" possedeva dei dati pressoché insignificanti per lo scopo che ci eravamo prefissati, infatti avere all'interno del dataset una colonna che semplicemente mi indica il numero dal 1 al 303 (con anche degli errori all'interno) era pressoché inutile, dato che abbiamo ipotizzato non fosse l'id del paziente, visto che era un banale elenco.
Quindi abbiamo deciso di rimuovere dall'intero dataset questa colonna, ed è stata l'unica scelta che abbiamo preso sulla rimozione delle colonne che abbiamo ritenuto inutili al fine dello studio.
Successivamente abbiamo notato che i nomi che erano stati attribuiti alle colonne erano poco chiari e non sempre facilmente comprensibili, quindi abbiamo fatto delle ricerche online per andare a comprendere che cosa indicavano. Dopo diverse ricerche abbiamo deciso di sistemare i nomi delle 14 colonne rimanenti così da poter procedere in maniera più facile allo studio del dataset.

Quindi eravamo rimasti con un dataset di 303 righe e di 14 colonne. In secondo luogo dando uno sguardo al nostro dataset ci siamo resi conto che molti dati richiedevano una sistemata, quindi abbiamo attuato delle tecniche di data cleaning per poter rendere più facile da leggere e ottimizzato al meglio il nostro dataset. Infatti i dati che stavamo osservando, soprattutto nella colonna nominata Sex e in quella che era relativa al Serum_Cholesterol possedevano dei dati che risultavano come unspecified/undefined e che quindi rendevano i dati relativi a quelle due colonne non corretti in modo tecnico, quindi abbiamo sistemato i dati rimuovendoli per rimediare ai danni che aveva subito il nostro dataset. Per ultimo abbiamo sistemato i dati che erano nulli, quindi quelli che su R vengono considerati come NA, abbiamo omesso ogni dato presente così da poter lavorare solamente su dati corretti tecnicamente e consistenti.
Infine abbiamo studiato quelli che erano i significati dei fattori nelle diverse colonne, le colonne rimanenti si dividevano in 7 colonne di dati "numerici", ossia che davano un valore numerico consistente in una unità di misura prestabilita, mentre le rimanenti erano colonne che davano come esiti o valori booleani o ad ogni numero andava attribuito un significato, Questi significati abbiamo deciso di svilupparli in due grafici per poter vedere con occhio a che cosa ci riferivamo, prendendo in considerazione la colonna Diagnosis_Heart_Disease, che aveva come valori booleani 0,1 e che andavano a definire se una persona non Soffre o se soffre di problemi cardiaci.

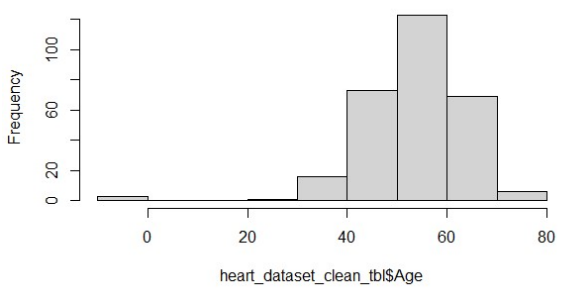


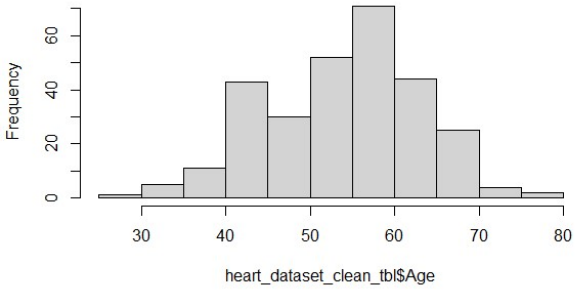
Dopo di che abbiamo continuato il nostro lavoro sviluppando la parte di Analisi dei dati da tecnicamente corretti e consistenti.

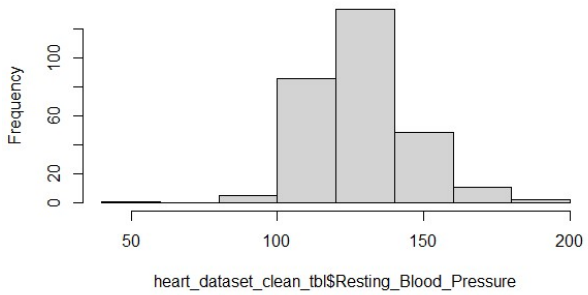
Capitolo 3: Analisi dei dati (tecnicamente Corretti e Consistenti)

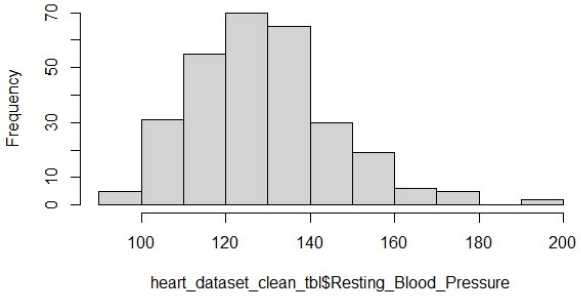
Dopo le prime analisi che abbiamo svolto sul dataset e i concetti teorici sopra riportati che ci eravamo prefissati abbiamo attuato le procedure, sotto forma di funzioni per la correzione dei dati. Abbiamo lavorato su diversi fronti, inizialmente abbiamo deciso di trasformare i dati in tecnicamente corretti, quindi abbiamo sistemato le due colonne (Sex e Serum_Cholesterol), cosi da andare a modificare da carattere a numerici, e in questo caso siamo riusciti a sistemare in maniera veloce i dati che risultavano tecnicamente non corretti, coi da lavorare invece su un dataset che possedeva i dati sotto formati corretti, successivamente dopo aver lavorato sui livelli e dopo aver dichiarato (come nel grafico precedente) a che cosa corrispondevano i dati su cui avremmo lavorato, abbiamo sistemato i dati cosi da renderli consistenti, infatti avevamo subito notato che certi dati non erano compatibili con le realtà effettive, per esempio erano presenti delle età minori di 0, cosa che effettivamente non può essere, poi abbiamo sistemato i dati relativi alla pressione sanguigna, dato che erano presenti dei dati che erano troppo bassi e incompatibili, allora li abbiamo sostituiti per scelta, per non perdere altri dati, con il valore medio dei dati già consistenti, successivamente abbiamo analizzato il caso dei dati relativi al colesterolo, caso analogo al precedente erano presenti dei valori troppo alti per una persona e come nel caso precedente abbiamo sistemato i dati

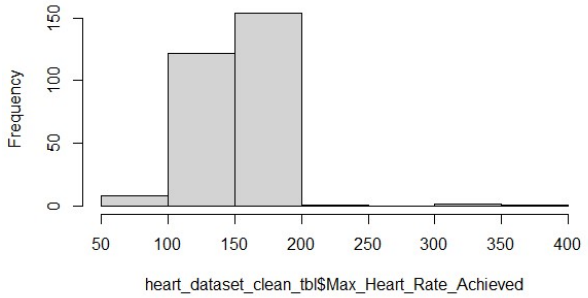
cosi da renderli consistenti, per finire abbiamo lavorato sui battimi massimi di una persona a riposo, come da consegna avremmo dovuto sistemare i dati che superavano i 222, e sostituirli alla media dei dati, e abbiamo sviluppato quindi un grafico corretto con i dati consistenti.
Qui sotto inseriremo i grafici con una visione di prima e dopo dei dati che abbiamo analizzato e sistemato per mostrare come sono cambiati durante la nostra analisi.

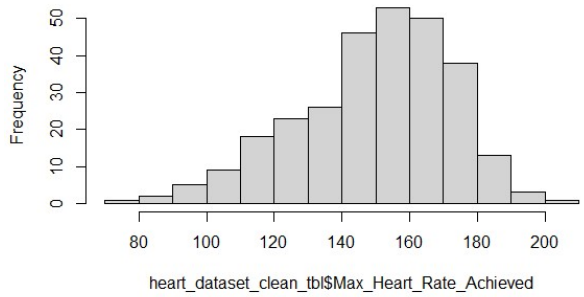


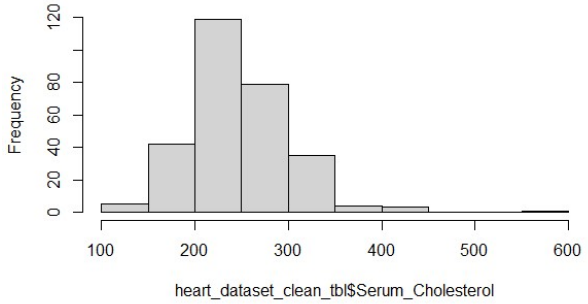


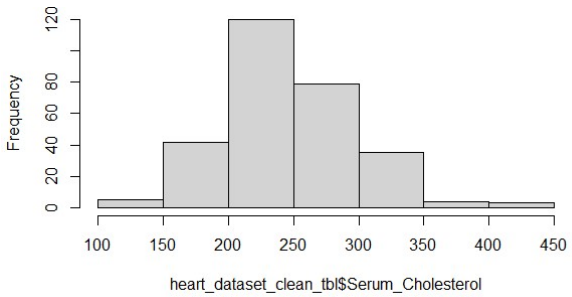












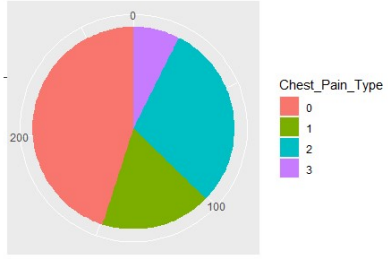
Possiamo notare come i grafici dei dati che abbiamo sistemato e reso tecnicamente corretti siano esteticamente e graficamente migliori dei precedenti, quindi ora avevamo il problema dei dati che venivano considerati come degli outlier, e per facilitare le cose e per poter lavorare su dei dati che principalmente possiamo notare a livello visivo abbiamo deciso di sviluppare un grafico con le funzioni di R che ci mostrasse tutti i dati che avevano degli outlier che non consideravano la IQR rule, sempre in relazione ai dati dei pazienti se erano malati o meno per poter anche comprendere le differenze tra i due.

Vedendo questo grafico possiamo notare come siano presenti diversi dati che vanno sistemati, vista anche la richiesta del docente abbiamo sistemato i dati rispettando la regola IQR e sviluppata anche questa parte di codice dove sono presenti i grafici, abbiamo salvato tutti i nuovi dati all'interno del dataset Heart_Disease, che sarà il nostro riferimento da questo punto in poi dello studio, visto che grazie agli operatori di R siamo riusciti a ricostruire un dataset con tutti i dati all'interno che abbiano un senso logico, consistenti e tecnicamente corretti.

Capitolo 4: Analisi Descrittiva

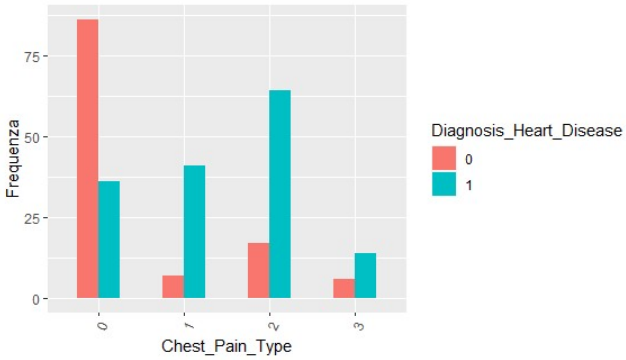
Dopo aver fatto una analisi sui dati che abbiamo sistemato e resi corretti, consistenti e reali abbiamo sviluppato, prima di procedere con le operazioni finali del codice, a sviluppare una analisi descrittiva approfondita con i dati che abbiamo ricevuto.
In primo luogo abbiamo osservato che il numero delle righe risultava uguale al numero delle osservazioni su cui stavamo lavorando, questo ci fa intendere che abbiamo sistemato correttamente i dati presi in esame, e che il numero di colonne risultava identico al numero delle variabili su cui potevamo lavorare, successivamente abbiamo notato come tra tutti i dati presenti all'interno del dataset quelli che avevano maggiore impatto sono: Chest_Pain_Type, Num_Majora_Vassel_Flouro, Thalassemia, Max_Heart_Rate_Achived. Per controllare questi dati possiamo notare che all'interno dei vari grafici riusciamo a dare diverse spiegazioni su quello che ci rivelano i dati:

- Chest_Pain_Type:
    - Circa la metà di tutti i pazienti presenta una angina ma in forma asintomatica
    - Circa un quarto dei pazienti ha dolori che differiscono dall'angina
    - Poco meno del 25% presenta una angina di classe tipica
    - Una piccola parte invece presenta una angina classica



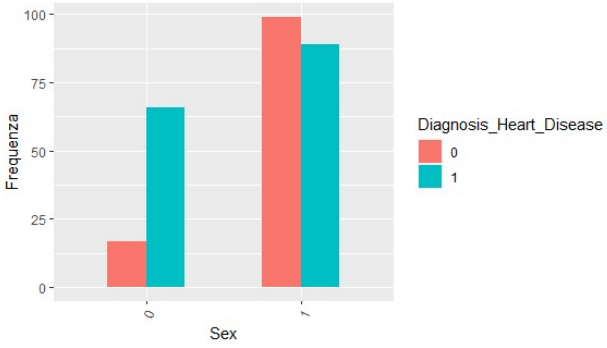
Successivamente siamo andati a sistemare questi dati all'interno di un grafico che ci permetteva di vedere i dati sopra riportati in relazione alla diagnosi finale del paziente, per vedere cosa succedeva e infatti abbiamo avuti diversi risultati che non ci aspettavamo:

- Chest_Pain_Type$Diagnosis_Heart_Disease
    - Il pazienti che abbiamo preso in considerazione che nel caso di prima presentano una angina senza sintomi hanno meno/soffrono meno di malattie cardiache.
    - I pazienti che presentano una angina atipica invece sono quasi tutti con problemi cardiaci, quindi hanno un rischio più alto rispetto a tutti gli altri.
    - Invece i pazienti che hanno un dolore non anginoso notiamo che come nel caso precedente sono principalmente tutti con problemi cardiaci
    - Infine i pazienti con un angina ti classe tipica ci sono per metà circa pazienti sani e pazienti malati.



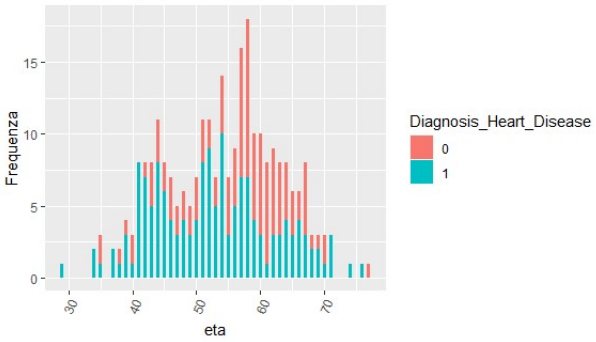
Dopo aver sistemato questi dati abbiamo deciso di procedere con i paragoni per poter vedere in base al sesso del paziente, quindi se tra uomini e donne le malattie cardiache chi colpivano di più, e siamo riusciti ad ottenere i grafici per poterli analizzare e i dati che abbiamo notato sono i seguenti:

- Sex$Diagnosis_Heart_Disease
    - Abbiamo constatato che dei pazienti di sesso maschile, per la metta sono o hanno avuto problemi a livello cardiaco, mentre l'altra metà no
    - Al contrario notiamo che una piccola parte delle donne sono affette da problemi cardiaci, mentre una fetta maggiore risulta sana e senza problemi di questo tipo.



Andando avanti ci siamo posti la domanda se influissero i dati relativi all'età con quelli che riguardano la diagnosi di una malattia cardiaca, quindi se una persona più anziana avesse più affinità con questo tipo di dolori, ne abbiamo conseguito, sviluppando i grafici, che:

- Age$Diagnosis_Heart_Disease
    - All'interno della fascia di età che viene compresa dai 40 ai 60 anni circa la probabilità di avere una malattia cardiaca è maggiore che di avere una malattia dello stesso tipo in una fascia minore e/o maggiore (riprende a livello di grafico la teoria del 10-80-10)



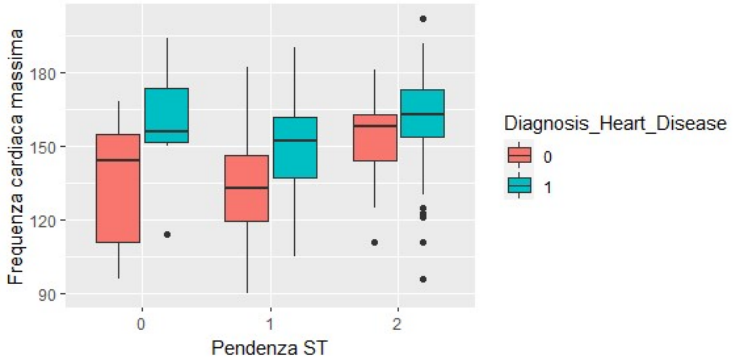
Proseguendo con i dati che avevamo in mano, abbiamo deciso di approfondire lo studio che riguarda i dati tra età e pressione, visto che a livello teorico una persona anziana dovrebbe avere una pressione sanguigna maggiore di un giovane, specialmente a riposo, avendo per conoscenze personali questa idea abbiamo deciso di controllare che anche i dati ci fornissero queste informazioni, e infatti abbiamo conseguito che:

- Age$Resting_Blood_Pressure x Sex
    - Basandoci sui dati che avevamo raccolto possiamo notare come i più giovani a riposso hanno una pressione sanguigna minore di una persona più anziana, anche se alcuni giovani possiedono una pressione elevata, e qualche caso di persona anziana che invece aveva dei valori inferiori alla media. Questo ci ha fatto riflettere che non possiamo dare nulla per scontato, ogni paziente è a se.
    - Avendo incrociato i dati con il sesso dei diversi pazienti abbiamo potuto notare come ne uomini e donne hanno grandi variazioni, quindi il sesso non influisce sulla pressione sanguigna a riposo dei pazienti, anzi con stupore abbiamo notato che non ci sono differenze di casi.

Successivamente abbiamo deciso di fare una analisi maggiormente approfondita per poter vedere se ci fosse una correlazione di dati tra la malattia effettiva di un paziente, quindi se aveva avuto un problema cardiaco e abbiamo deciso di andare ad indagare se aveva influito sul battito cardiaco a riposo, abbiamo inoltre indagato e abbiamo scoperto che il dato che fa riferimento alla pendenza ST potrebbe avere dei legami con i due valori precedentemente analizzati, quindi per comodità siamo andati ad analizzare un box plot, per andare a capire se nei tre dati che erano presenti cerano grandi differenze, questo studio ci ha condotto a diverse soluzioni, infatti:

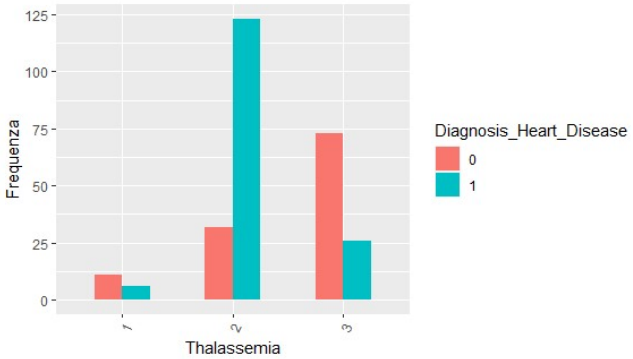
- Diagnosis_Heart_Rate$Max_Heart_Rate_Achived x Peak_Exercise_ST_Achieved
    - I pazienti che hanno avuto problemi cardiaci generalmente risultano gli stessi pazienti che hanno una frequenza cardiaca elevata, o comunque più alta rispetto ai pazienti che non hanno sofferto di malattie di questo tipo.
    - I pazienti che non hanno avuto problemi cardiaci, hanno una dispersione di dati maggiormente elevata sul frammento ST, rispetto a chi ha avuto problemi cardiaci.
    - Possiamo però notare come la frequenza cardiaca rilevata nei pazienti che hanno il dato della pendenza ST che volta verso l'alto, quindi con un dato in salita, ha una buona distribuzione di dati.

- Infine nei dati che analizziamo con la pendenza del segmento ST piatta, abbiamo notato che i pazienti che sono affetti da malattie cardiache possiedono diversi outlier, quindi ci fa capire che ci sono diversi pazienti con una frequenza cardiaca bassa rispetto alla media, e un valore che invece è estremamente alto, possiamo dire che i pazienti affetti da malattie cardiache e che hanno un andamento lineare del frammento ST hanno o un battito elevato, o un battito cardiaco che può diventare bradicardico.



Successivamente per fare una analisi approfondita dei dati abbiamo deciso di analizzare i dati che riguardano il dato sulla talassemia, questo dato tra i membri del gruppo di ricerca è stato molto discusso, perché non riuscivamo inizialmente a comprendere che cosa andassero ad indicare i dati che stavamo analizzando. Ci siamo documentati e abbiamo diviso i dati, all'interno di un grafico a torta abbiamo guardato quanti ne fossero presenti, e abbiamo concluso che ci sono tre forme di Talassemia, i dati erano "ben distribuiti", visto che circa la metà dei pazienti che avevamo in esame riportavano un livello normale, un'altra fetta di pazienti riportavano dei dati buoni, visto che il difetto presente sulla talassemia risulta reversibile e quindi curabile. Mentre una piccola parte dei pazienti presentava un difetto non modificabile, probabilmente in uno stato avanzato. Però osservando l'insieme sono dati buoni visto che sono molto pochi i pazienti con quest'ultimo tipo di Talassemia.
Quindi abbiamo indagato se, questi dati influissero particolarmente con la possibilità di aver avuto una malattia cardiaca, e dopo aver graficato i dati abbiamo ottenuto che:

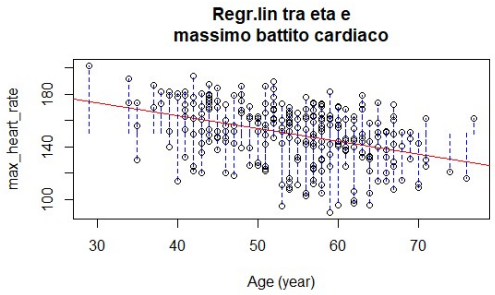
- Thallium$Diagnosis_Heart_Disease
  - Possiamo notare come i pazienti che hanno un dato sistemabile sul parametro talassemia, solamente un terso di quelli che sono i pazienti ha avuto problemi a livello cardiaco
  - Mentre per i pazienti che hanno un livello normalizzato all'interno del corpo di questo campo notiamo come solo una minima parte dei pazienti sia affetto da malattie di tipo cardiaco
  - Infine i pazienti che hanno un difetto talassemico non reversibile dimostrano che hanno solo una piccola percentuale di problemi cardiaci, quindi i dati mostrano che solo un 20% dei pazienti presenta problemi a livello cardiaco.
  - In sostanza per concludere le osservazioni notiamo che i pazienti ricoverati che hanno avuto un risultato normale riguardante il valore della talassemia hanno una maggiore compatibilità con i problemi cardiaci, non siamo riusciti a darci una spiegazione.



Con questi dati concludiamo la nostra analisi descrittiva, abbiamo quindi approfondito i campi che erano presenti all'interno del dataset per poter essere più precisi possibili nei dati che saremmo poi andati a sviluppare con gli algoritmi presenti per l'esame.
Per diversi punti non siamo riusciti a motivarci perché cera più probabilità o meno di avere problemi a livello cardiaco, e siamo soprattutto rimasti con un filo di dubbio, anche perché erano dati da analizzare di cui non avevamo mai sentito parlare, e anche con la documentazione online non siamo riusciti a darci ottime risposte.
Restiamo comunque soddisfatti dei dati che siamo riusciti ad osservare e analizzare, visto che in molti casi le soluzioni sono diverse da quello che pensavamo e che quindi siamo riusciti ad approfondire e studiare meglio. Adesso procediamo con le applicazioni dei vari algoritmi.

Capitolo 5: Regressione Lineare

Dopo aver sviluppato la nostra analisi descrittiva dettagliata (i dati presenti sono stati analizzati da noi senza effettivamente essere completamente competenti in materia quindi potrebbero contenere degli errori di lettura), procediamo con i punti sviluppando la regressione lineare.
Lo scopo era di cercare due dati da rapportare tra di loro per poi utilizzare le due variabili scelte sia per la regressione lineare che per gli studi successivi.
Abbiamo lavorato diverso tempo nello studio della regressione lineare, dato che abbiamo analizzato diversi capire per andare a trovare i migliori che potevano essere sviluppati.
Quindi dopo diverse analisi abbiamo scelto come dati per sviluppare la regressione lineare, la relazione tra le variabili "Age" e "Max_Heart_Rate_Achived". Successivamente abbiamo chiesto al nostro calcolatore R di riassumere i valori di queste due variabili cosi da poter avere una visualizzazione iniziale dei dati e per poterne capire la dispersione.
Successivamente abbiamo applicato un calcolo di regressione lineare predefinito in R, cosi da poter ricevere il grafico e i dati che interessano allo studio.

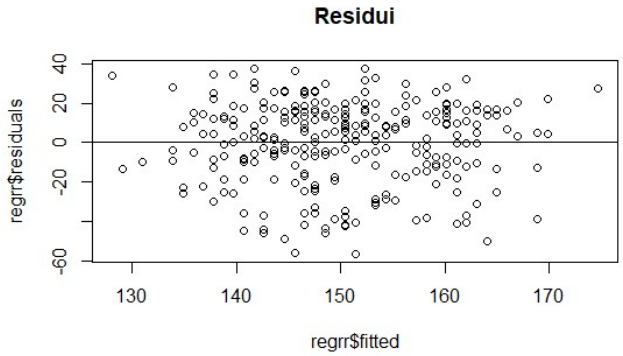


Come possiamo vedere i dati sono tanti e quindi il grafico non è dei migliori, eppure sono i migliori dati con cui sviluppare e studiare la regressione lineare all'interno del dataset proposto, anche se questa è una mia opinione. Infine abbiamo controllato i valori dell'intercetta e dello "scope", cosi da fare una analisi migliore:

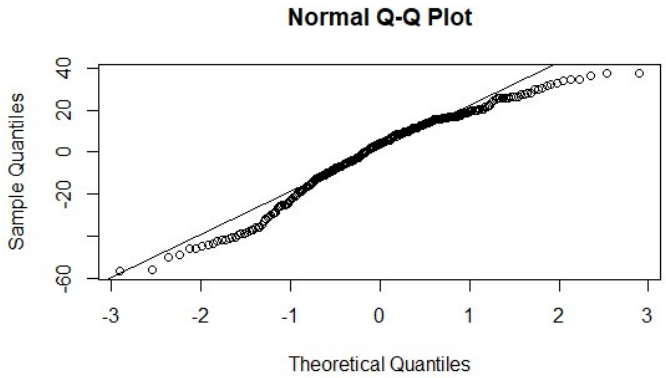
- Coefficiente angolare della retta proposta: 0.9717
- Intersezione con l'asse delle ordinate: 26.76

La relazione tra i dati presenti però, notando anche l'andamento della retta di regressione possiamo ipotizzare essere negativo, e non alto come valore, dato che siamo intorno al 40% di accuratezza.

Abbiamo inoltre studiato i residui della regressione, quindi mediante le funzioni di R siamo riusciti a tracciare un grafico che rappresentasse questi valori, che potremo notare essere non omogeneamente distribuiti all'interno del grafico:

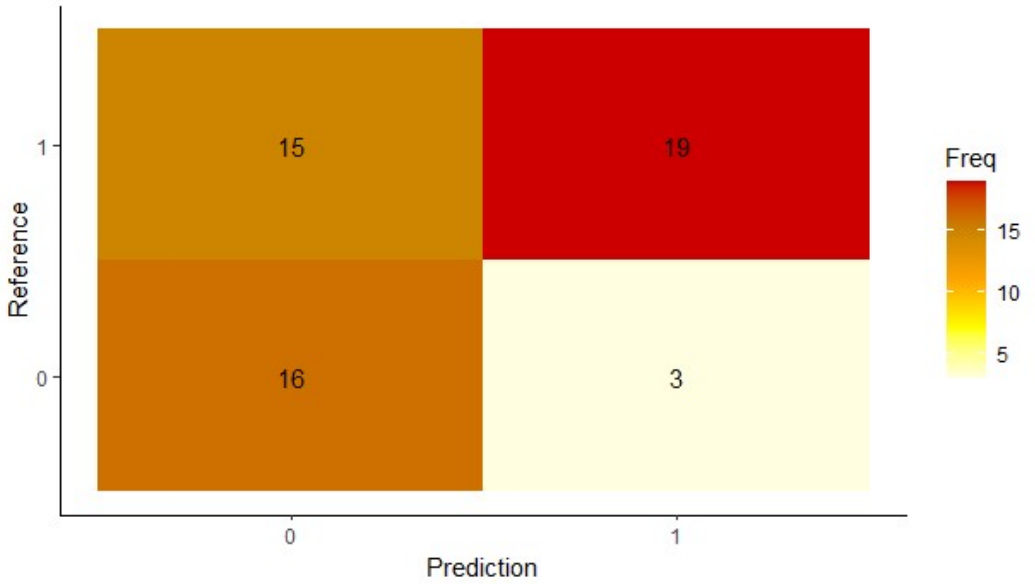


Infine andiamo a sviluppare il grafico con la distribuzione in quantili e lo sviluppo della retta che vada ad intercettare questi dati. Infatti il grafico serve per andare a mettere a confronto i quantili delle due distribuzioni prese in esame. Se la distribuzione della variabile ha la forma di una S, la distribuzione è asimmetrica; Se invece è al di sotto o al di sopra della linea, la distribuzione è platicurtica (la densità è maggiore sulle code), o leptocurtica (la densità è maggiore del previsto intorno ai valori centrali). Come nel nostro caso.



Capitolo 6: Machine Learning

Dopo aver analizzato approfonditamente i dati, e dopo aver avuto i risultati dei precedenti due algoritmi, quindi quello del gradiente e quello di forza bruta abbiamo provato ad applicare l'algoritmo predefinito di machine learning, fornito in sede di esame dal docente. Questo file era da implementare ai nostri dati per poter vedere se riuscivamo ad applicare e prevedere i dati, per prima cosa abbiamo dovuto installare e applicare la libreria "lattice" che ci permette di applicare questo algoritmo, e siamo andati avanti con lo studio.
Purtroppo mediante questo algoritmo non siamo riusciti ad ottenere dei dati che fossimo in grado di identificare e di analizzare propriamente, quindi abbiamo applicato un altro algoritmo di machine learning per poter portare dei risultati, questo algoritmo invece risulta più ottimizzato e più veloce.
Siamo riusciti ad ottenere dei risultati andando ad indagare sulla variabile Max_Heart_Rate_Achived, inizialmente abbiamo impostato una divisione randomica de nostro campo di analisi, impostando il nostro seed a 5000, cosi da avere una divisione del dataset sulla nostra variabile scelta, quindi sul battito cardiaco, noi abbiamo scelto di dividerlo in percentuale 80/20, questo perché tendenzialmente le scelte sono tra il 65 e l'85.
Dopo aver impostato la nostra percentuale scelta abbiamo dovuto "allenare" il nostro algoritmo, mentre la restante parte la abbiamo utilizzata per testare i dati.
Successivamente abbiamo graficato il tutto per ottenere dei risultati, sia per poterli vedere sotto forma di grafico, sia per riuscire a commentarli.



All’interno dei risultati ottenuti sappiamo che la nostra accuratezza, rispetto all’algoritmo che abbiamo utilizzato si aggira tra il 71% e il 59%, quindi effettivamente sono degli ottimi risultati, per un modello da allenare come questo. Molto probabilmente abbiamo ipotizzato che ottenendo altri dati sullo stesso dataset potevamo migliorare questo dato.
In particolare quella del nostro dataset e dell’algoritmo che abbiamo allenato è del 66,04%.
All’interno del grafico possiamo notare come, la matrice di confusione ci indichi nella diagonale dal basso all’alto i dati corretti, che ovviamente sono più densi e quindi corretti, mente nella diagonale opposta dall’alto al basso gli errori, notiamo come siano presenti molti errori nel primo quadrante preso in esame, purtroppo non siamo riusciti a sistemare i dati in maniera migliore di questa.
Restiamo comunque soddisfatti dei risultati ottenuti durante il nostro studio e la applicazione di questo algoritmo.

Capitolo 7: Conclusioni

Abbiamo dunque concluso i nostri studi con questi algoritmi, che ci sono serviti per poter applicare nuove funzioni all’interno dell’ambiente R, che abbiamo utilizzato per tre mesi come strumento di analisi dei dati.
In particolare sviluppando una analisi approfondita che riguardava la diagnosi di problemi cardiaci siamo riusciti ad analizzare diversi dati interessati, molti dei quali non conoscevamo e siamo andati a studiarli personalmente. I dati che abbiamo studiato e allenato ci hanno anche portato a dei risultati più che soddisfacenti e quindi rispetto al lavoro svolto posso dire che siamo rimasti molto soddisfatti rispetto alla prima volta.
Inoltre siamo contenti che gli strumenti che abbiamo utilizzati siano risultati completi e corretti, per di più ce stato un maggiore studio dietro agli strumenti che fornisce R come linguaggio di analisi dei dati.