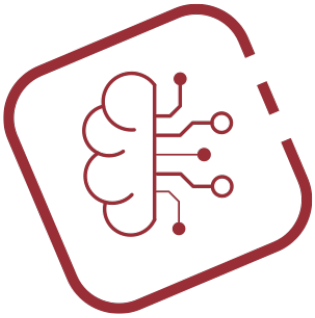
ITS Angelo Rizzoli
Corso ITS Machine Learning Specialist



Linguaggi di Programmazione
per il Machine Learning

Programmazione Base in R: Heart Disease

Marco Papa
Matteo Romanazzi

Corso 2020 - 2022

Indice

1 Introduzione e obiettivi

Lo scopo di questo report é quello di descrivere ciò cha abbiamo applicato sul nostro codice, mostrando tutto quello che abbiamo appreso durante il corso. Per far ciò abbiamo svolto un'analisi completa del dataset "heart.csv". Questo dataset riporta dei dati utili per svolgere previsioni sulla possibilità che una persona soffra di malattie cardiache o meno.

Per svolgere il lavoro è stato utilizzato il linguaggio r con l' IDE RStudio.

Il report é suddiviso in otto capitoli:
In questo (il primo) facciamo una piccola introduzione sui nostri obbiettivi. Nel secondo introduciamo un dataset, soffermandoci su che cos'é e che cosa contiene. Nel terzo è presente un' analisi descrittiva dei dati, con l' ausilio di grafici. Nel quarto svolgiamo una revisione dei dati, rendendoli tecnicamente corretti e consistenti, per far ciò analizziamo i tipi degli attributi (cambiandoli se necessario) e successivamente li classifichiamo in nominale, ordinale, di intervallo o di rapporto,infine facciamo pulizia dei valori "missing", rendendo così il nostro dataset tecnicamente corretto. Nel quinto capitolo visualiziamo graficamente i dati, prima e dopo le modifiche, utilizzando oltre i plot base di r, quelli della librerie ggplot2. Nel sesto capitolo analizziamo le relazioni tra le nostre variabili tramite la regressione lineare, visualizzando la retta di regressione, i residui e la dispersone in quantili. Nel settimo capitolo applichiamo un modello di machine learning misurandone l'accuratezza.

2 Descrizione del dataset

Il dataset in questione é "heart.csv", questo dataset contiene dei dati che sono utili per fare previsioni sulla possibilità che una persona soffra di malattie cardiache o meno, ed é composto dai seguenti attributi:

$x	Numero identificativo.
$age	Età.
$sex	Dolore al petto.
$cp	Pressione sanguigna a riposo.
$trestbps	Livello colesterolo.
$chol	Colesterolo.
$fbs	Zucchero nel sangue a digiuno.
$restecg	Risultati ecg a riposo: 0=normal 1= anormalità dell'onda ST-T (inversioni dell'onda T e /o elevazione o depressione dell'ST maggiore di 0,05 mV) 2=probabile o definita ipertrofia ventricolare sinistra secondo i criteri di Estes.
$thalach	Battito massimo registrato.
$exang	Angina provocata dall'esercizio (1 = si; 0 = no).
$oldpeak	Depressione ST causata dall'esercizio paragonata al riposo.
$slope	The slope of the peak exercise ST segment 1: upsloping 2: flat 3: downsloping
$ca	Numero di vasi principali (0-3) colorati con fluorosopia.
$thal	3 = normale; 6 = difetto risolto; 7 = difetto reversibile.
$target	int

Per ottenere maggiori informazioni sugli attributi e capire cosa vogliano dire abbiamo controllato su varie fonti e abbiamo riscontrato molte similitudini tra esse, di conseguenza siamo risaliti ai nomi interi e al loro significato.

Fonti

1) https://www.kaggle.com/zhaoyingzhu/heartcsv/discussion/51213
2) https://archive.ics.uci.edu/ml/datasets/Heart+Disease
3) http://www.iaeng.org/publication/IMECS2010/IMECS2010_pp134-139.pdf

3 Analisi descrittiva

La prima cosa che abbiamo fatto é stata una rapida analisi strutturale del dataset. Abbiamo utilizzanto rStudio per visualizarne la struttura (attributi e tipi)e per fare un controllo a tappeto sull'intero dataset per controllare la presenza di valori mancanti o errati come:
na / unassigned / undefined e unspecified.
Con l'analisi strutturale abbiamo determinato la presenza di 15 attributi con i seguenti tipi.

$x	int	$trestbps	int	$thalach	int	$ca	int
$age	int	$chol	chr	$exang	int	$thal	int
$sex	chr	$fbs	int	$oldpeak	num	$target	int
$cp	int	$restecg	int	$slope	int		

Analizziamo i dati delle colonne con i grafici, questo serve per avere un'idea ancora più chiara su come sono strutturati i vari attributi.
nb:i grafici mostrano i valori dopo essere stati resi tecnicamente corretti e consistenti.

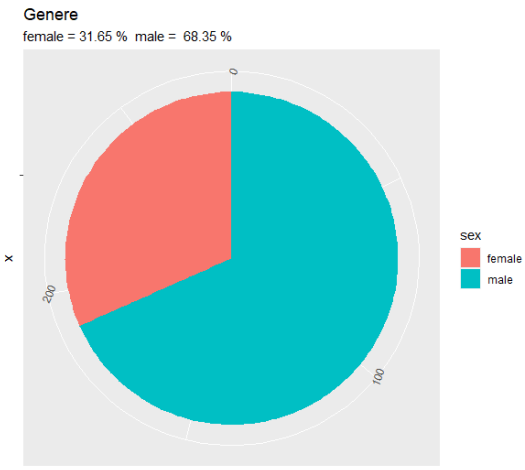


Figura 1: Iniziamo con l'attributo $sex, possiamo notare che il numero di pazienti maschi é decisamente più alto rispetto a quello delle donne. 31,65% - 68,35%

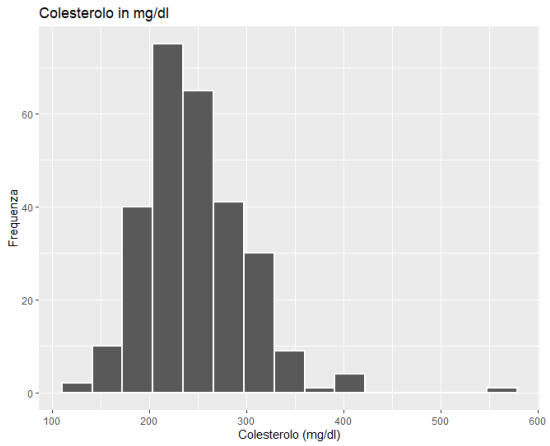


Figura 2: Possiamo notare che la maggior parte dei soggetti ha una concentrazion di colesterolo compresa tra poco più di 150mg/dl e 300mg/dl.

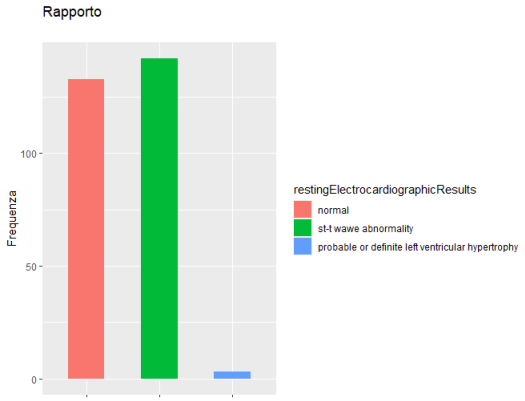


Figura 3: Mentre, da questo grafico risulta che solo una minima parte di pazienti soffre ipertrofia ventricolare sinistra.

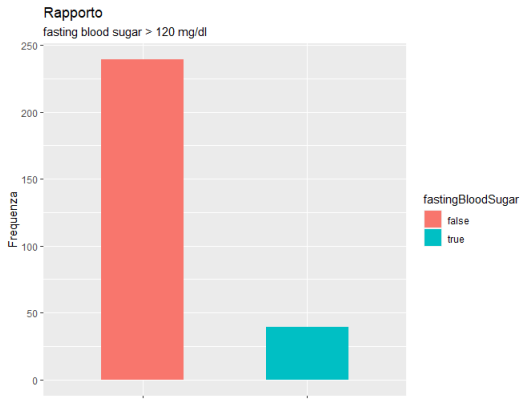


Figura 4: Questo grafico ci mostra che la maggior parte dei pazienti non presenta quantitativi eccessivi di zuccheri nel sangue.

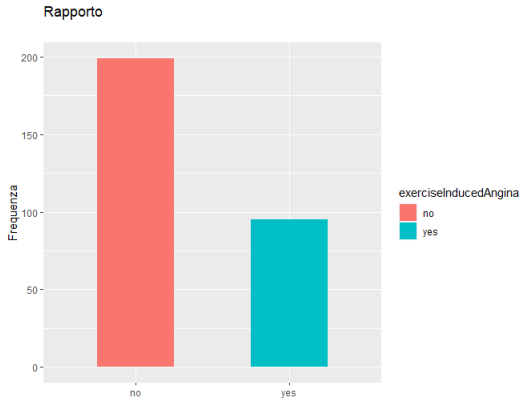


Figura 5: Anche quà possiamo notare una disparità tra i dati, si può notare che la maggior parte dei pazienti non sente dolore al petto durante l'esercizio

4 Analisi dei dati (tecnicamente corretti e consistenti)

Prima di classificare gli attributi abbiamo fatto un controllo sulla consistenza dei dati. Possiamo notare che visualizzando la struttura con str(dataset) molti attributi sono di tipo errato (evidenziati in rosso).

$x	int	$trestbps	int	$thalach	int	$ca	int
$age	int	$chol	chr	$exang	int	$thal	int
$sex	chr	$fbs	int	$oldpeak	num	$target	int
$cp	int	$restecg	int	$slope	int		

Dopo aver analizzato gli attributi siamo passati alla fase di classificazione, dove in base al tipo di operatore che ha senso applicare ai valori di quella colonna possiamo attribuire a un attributo i seguenti tipi:
- Categoriche: (Nominale, Ordinale).
- Numeriche: (Di intervallo, Di rapporto).

$x	NOMINALE	$chol	ORDINALE
$age	ORDINALE	$fbs	NOMINALE
$sex	NOMINALE	$restecg	NOMINALE
$cp	ORDINALE	$thalach	DI RAPPORTO
$trestbps	ORDINALE	$exang	NOMINALE
		$oldpeak	ORDINALE
		$slope	NOMINALE
		$ca	ORDINALE
		$thal	NOMINALE
		$target	NOMINALE

Basandoci sulla nostra classificazione e sui dati presenti nelle varie colonne abbiamo deciso di assegnare i seguenti tipi agli attributi(evidenziati in blu).

$x	int	$chol	integer	$oldpeak	num
$age	int	$fbs	factor	$slope	factor
$sex	factor	$restecg	factor	$ca	factor
$cp	factor	$thalach	int	$thal	factor
$trestbps	int	$exang	factor	$target	factor

Abbiamo deciso di assegnare i nomi dei livelli solamente agli attributi sex, fbs, restecge exang, questo perché basandoci su quello riportato dalla pagina del dataset i valori in alcuni casi non corrispondono, e per evitare di creare ulteriori ambiguità e per non alterare il risultato abbiamo deciso di lasciare i livelli senza nome ai seguenti attributi: cp (0,1,2,3 / 1,2,3,4), slope (0,1,2,3 / 1,2,3,4), thal (0,1,2 / 3,6,7).

Successivamente siamo passati alla fase di pulizia, qui rimuoviamo i dati NA, unasigned, undefined e unspecified, questi possono essere presenti sia in origine e sia dopo la conversione, infatti dopo aver cambiato il tipo degli attributi alcuni dati hanno assunto dei valori NA. (NA introdotte per coercizione).

Nella fase di pulizia dei dati utiliziamo la funzione sum(is.na(dataset)) che ritorna in numero di valori NA presenti nel dataset,e se sono presenti dai valori NA li rimuoviamo con la funzione na.omit(dataset) che elimina la riga nel quale é presente il valore NA.

Successivamente con un editor esterno (excel) controlliamo tutto il dataset per trovare i dati "missing" (Unasigned, Undefined e Unspecified), e se sono presenti li convertiamo in NA dataset$sex[dataset$sex == "unspecified"] ¡- NA e poi li rimuoviamo con la funzione na.omit(dataset).

Dopo aver ripulito tutti i dati e averli resi tecnicamente corretti abbiamo fatto un'ulteriore controllo sul "senso" dei dati. È risultato che nella colonna "age" sono presenti dei valori minori di 0, visto che una persona non può avere un'età inferiore a 0 abbiamo deciso di rimuoverli.
dataset (assegnazione) subset(dataset, age (maggiore di) 0).

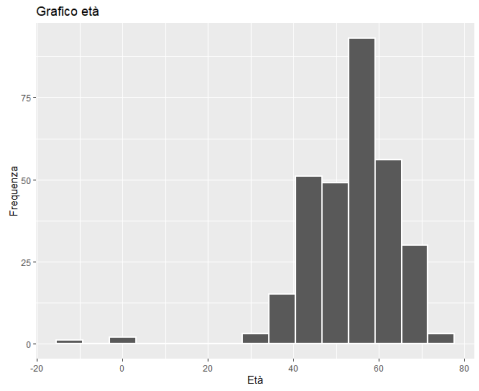
Infine abbiamo rinominato le colonne per renderle più comprensibili (rimuovendo le abbreviazioni).

$x	id
$cp	chestpain
$trestbps	restingBloodPressure
$chol	cholesterol
$fbps	fastingBloodSugar
$restecg	restingElectrocardigraphicResult
$thalach	maxHeartRate
$exang	exerciseInducedAngina
$ca	nMainVesselsStainedfluorosopy
$target	vaselNarrowing

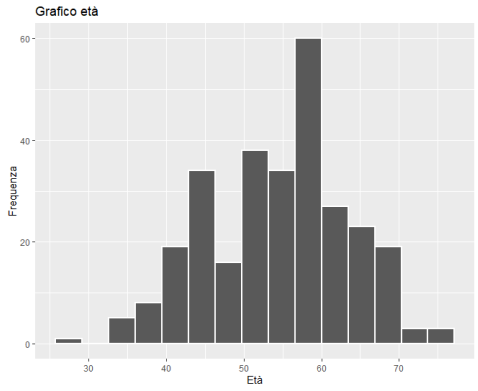
Abbiamo deciso di rimuovere la colonna id, questo perché il dataset contiene dati utili per fare predizioni e avere un attributo id per identificare le colonne non é necessario.

5 Visualizzazione dati Prima e Dopo

Paragone dopo aver rimosso i dati dell'età minori di 0.

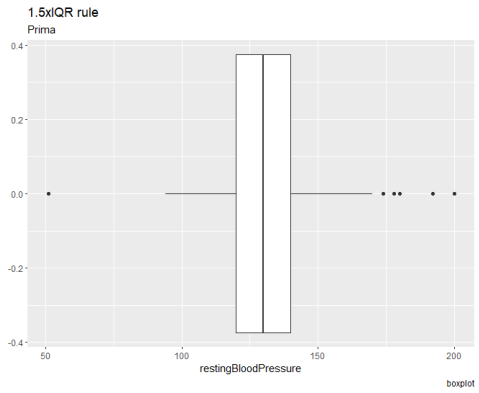


(a) Prima

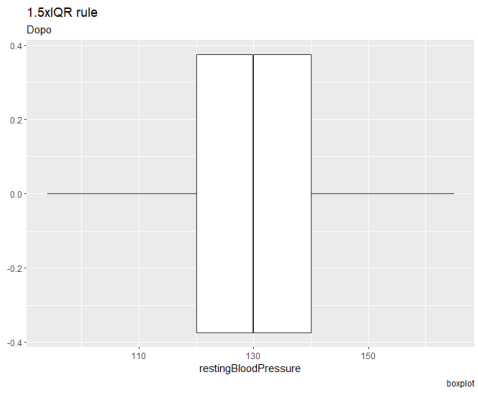


(b) Dopo

Assumere come outlier, i valori relativi alla pressione sanguigna a riposo che non rispettano la 1.5xIQR Rule.

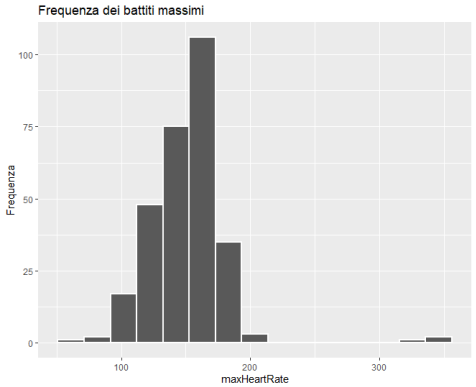


(a) Prima

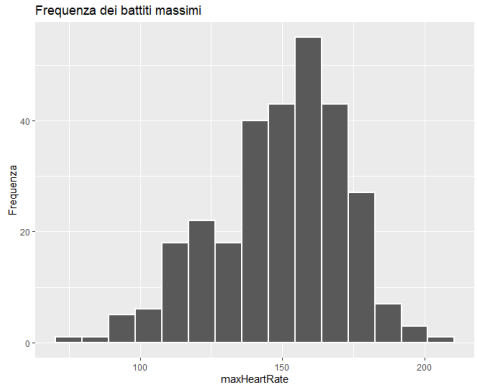


(b) Dopo

Oltre a quello abbiamo assunto che tutti i dati nella colonna maxHeartRate con un valore superiore a 222 debbano essere impostate con il valore medio della variabile.
dataset$mhr[dataset$mhr ¿ 222] ¡- median(dataset$mhr)



(a) Prima



(b) Dopo

6 Regressione lineare

Attraverso la regressione lineare, abbiamo deciso di analizzare la relazioni tra le variabili del nostro dataset, age e Cholesterol. Abbiamo richiamato la funzione summary(la quale ci permette di ottenere i valori dell'intercetta, del coefficiente angolare e di R^ 2) sulla variabile dataset_reg contenente il valore ritornato dalla funzione lm().
Per renderci conto se ci sia una correlazione lineare tra le due variabili abbiamo:

- Visualizzato la retta di regressione, tramite un grafico a dispersione, rappresentando nel grafico le variabili age e Cholesterol.
(figura 6)

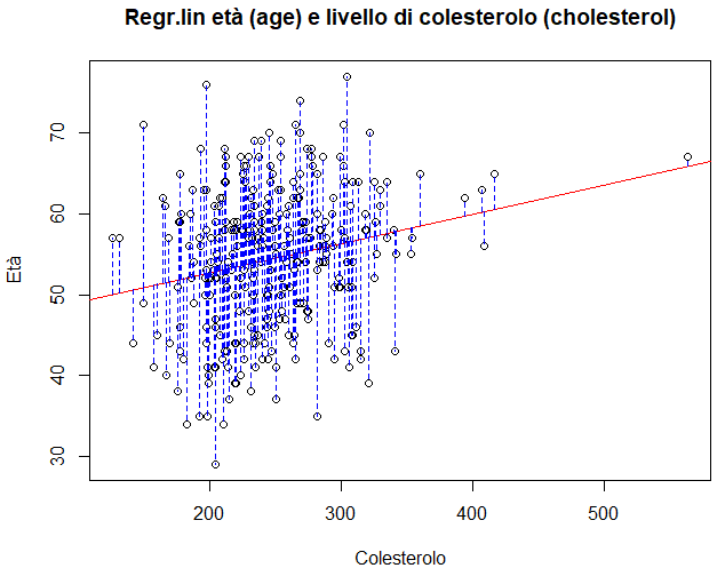


Figura 6

- Analizzato i residui tramite un altro grafico di dispersione (scatterplot), utilizzando come parametri per la funzione di creazione del grafico le variabili "fitted " e "residuals". (figura 7)

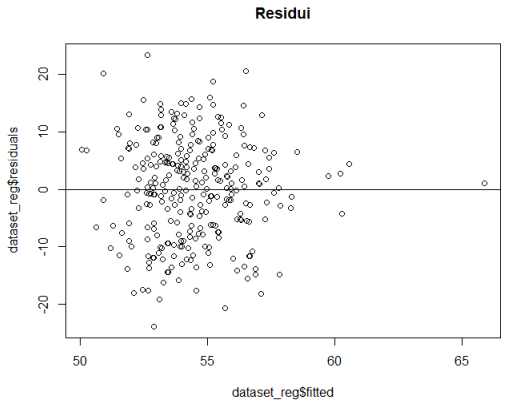


Figura 7

- Analizzato la distribuzione in quantili tramite grafico di dispersione(con le funzioni qqnorm e qqline). (figura 8)

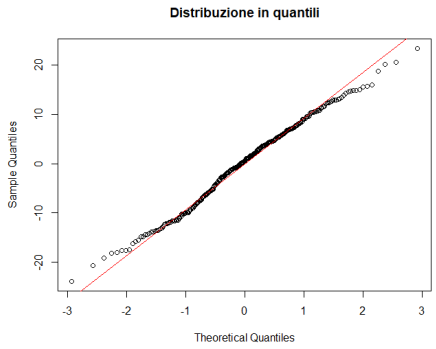


Figura 8

Da questo ultimo grafico abbiamo osservato che gli elementi non sono distribuiti in modo equilibrato lungo la retta,non è quindi confermata l'ipotesi di distribuzione casuale dei residui.

Abbiamo infine calcolato il valore di r (coefficiente di correlazione) per capire quanto sia forte la correlazione lineare tra le due variabili. Il risultato della variabile ottenuto è 0.2028307 Essendo, come valore, molto lontano da 1 e vicino a 0, concludiamo che tra le due varibili(age e Cholesterol) non ci sia una forte correlazione. Una volta ottenuto r abbiamo richiamato ancora una volta la funzione summary per visualizzare $R\hat{2}$(coefficiente di determinazione) per stabilire la bontà di adattamento del modello di regressione(come osservato anche dai grafici). Il valore ottenuto è 0.04114028,giungiamo quindi alla conclusione che il modello di regressione lineare non possa spiegare la variabiltà delle osservazioni.

7 Machine Learning

Infine, abbiamo voluto misurare il livello di accuratezza del modello di machine learning non lineare KNN. Il livello di accuratezza ottenuto è di 56%(visualizzato con la funzione di r confusionMatrix()).

Breve descrizione del modello:
KNN sta per K-Nearest Neighbors.
Si tratta di un algoritmo utilizzato per la classificazione di oggetti basandosi sulle caratteristiche degli oggetti vicini a quello considerato,servendosi del parametro k. Il modello in questione può essere utilizzato per la classificazione o la regressione.
Da questo dipende l' output:
Per la classificazione , l'output è l'appartenenza a una classe. Ad esempio,se k=1, l'oggetto viene assegnato alla classe del vicino più prossimo.
Per la regressione l' output è il valore della proprietà dell' oggetto. Questo valore è la media dei valori di k più vicini.
L' input invece è, in entrambi i casi, costituito dai k esempi di addestramento più vicini.

8 Conclusioni

Con questo progetto abbiamo applicato tutte le tecniche imparate durante il corso, partendo dalle semplici guide di stile, all'analisi dei dati e la regressione lineare fino all'applicazione dei modelli di machine learning.