

Renz Joshua Villanueva & Binod Comini Professore: Francesco Stranieri

Progetto in linguaggio ${\bf R}$

 ${\rm Milano} \\ {\rm ITS~Angelo~Rizzoli} \\ {\rm Artificial~Inteligence~\&~Machine~Learning} \\$

Contents

1	Dat	taset	3	
2	Correzione dei dati			
	2.1	valori NA	. 3	
	2.2	Colonne non necessarie	. 4	
	2.3	Rinominazione delle colonne	. 4	
	2.4	Codice	. 4	
3	Tipi di dati			
	3.1	Tipo di ogni attributo	. 5	
	3.2	Correzione dei tipi di dato	. 5	
	3.3	Codice	. 5	
	3.4	Livelli dei fattori	. 6	
4	Consistenza dei dati			
	4.1	max_hr	. 7	
	4.2	rest_bp diviso per sesso	. 9	
	4.3	Considerazioni finali	. 11	
5	Relazioni tra i dati			
	5.1	Regressione linare	. 12	
	5.2	age & rest_bp	. 12	
	5.3	age & max_hr	. 13	
6	Modelli di machine learning			
	6.1	Data frame per previsioni	. 15	
	6.2	Modelli di Machine Learning	. 16	
	6.3	K-Nearest Neighbors (KNN)	. 16	
	6.4	Perceptron Multistrato (MLP)	. 17	
	6.5	Random Forest (RF)	. 17	
7	Ex	tra	17	
8 Conclusioni Personali		17		

1 Dataset

Questo dataset contiene dati riguardanti il cuore. Per avere una prima visione delle variabili raccolte, abbiamo caricato il dataset "heart.csv" nella nuova variabile dataset con la funzione **read.csv**(*dataset*) e abbiamo, di conseguenza, analizzato la struttura dei dati raccolti.

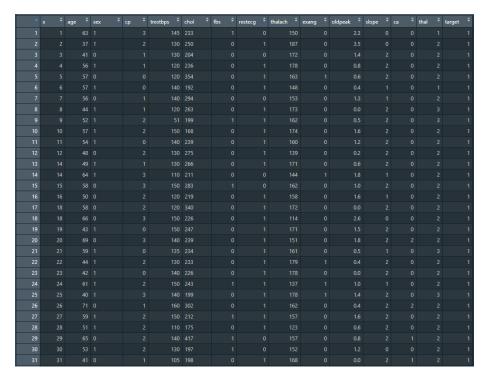


Figure 1: Screen dataset heart.csv

2 Correzione dei dati

2.1 valori NA

Osservando il dataset ci siamo accorti della presenza di valori NA dovuti, ad esempio, al data-entry manuale; quindi abbiamo creato un ciclo che controllasse la presenza di tali valori. Tutti i valori NA sono stati cancellati definitivamente dal dataset con la funzione *dataset* <- na.omit (*dataset*).

2.2 Colonne non necessarie

Abbiamo proseguito l'analisi del dataset controllando e rimuovendo le colonne non ritenute necessarie. Infatti, attraverso la funzione **view**(*dataset*), l'abbiamo visualizzato e, attraverso la funzione **subset()**, abbiamo rimosso le colonne superflue; in questo specifico caso la colonna X perché ritenuta inutile.

2.3 Rinominazione delle colonne

Terminato quest'ultimo passaggio, abbiamo rinominato le colonne in maniera appropriata, descrivendone, di ciascuna, il tipo di attributo. Abbiamo stampato il dataset con la funzione 'str(dataset)' e, con la funzione names(dataset)[names(dataset) == "vecchio"] <- "nuovo" sono state rinominate le colonne in modo appropriato. Come ultima azione abbiamo assegnato ad ogni attributo il suo tipo.

2.4 Codice

```
# Codice con R tradizionale
dataset <- subset(dataset, select = - x)</pre>
names(dataset) [names(dataset) == "cp"] <- "chest_pain"</pre>
names(dataset)[names(dataset) == "trestbps"] <- "rest_bp"</pre>
names(dataset)[names(dataset) == "chol"] <- "cholesterol"</pre>
names(dataset)[names(dataset) == "thalach"] <- "max_hr"</pre>
names(dataset)[names(dataset) == "exang"] <- "exercise_angina"</pre>
names(dataset)[names(dataset) == "thal"] <- "thalassemia"</pre>
names(dataset)[names(dataset) == "target"] <- "heart_disease"</pre>
names(dataset)[names(dataset) == "ca"] <- "n_vessels"</pre>
names(dataset)[names(dataset) == "restecg"] <- "rest_ecg"</pre>
# Codice con Tidyverse
dataset <- dataset %>%
  select(-one_of("x")) %>%
 rename(
    chest_pain = cp,
   rest_bp = trestbps,
    cholesterol = chol,
    max_hr= thalach,
    exercise_angina = exang,
    thalessemia = thal,
   heart_disease = target,
   n_vessels = ca,
   rest_ecg = restecg
```

3 Tipi di dati

3.1 Tipo di ogni attributo

age int ORDINALE
sex chr NOMINALE
chest_pain int NOMINALE
rest_bp int DI RAPPORTO
cholesterol chr DI INTERVALLO
fbs int DI RAPPORTO
rest_ecg int NOMINALE
max_hr int DI INTERVALLO
exercise_angina int NOMINALE
oldpeak num ORDINALE
slope int NOMINALE
n_vessels int ORDINALE
thalassemia int NOMINALE
heart_disease int NOMINALE

3.2 Correzione dei tipi di dato

Successivamente, abbiamo eseguito un controllo per correggere la consistenza del tipo di dato per ogni variabile con la funzione

 $[colonna\ del\ dataset]$ <- as. $[tipo\ nel\ quale\ voglio\ cambiare\ i\ dati]$ $(colonna\ del\ dataset)$

così come segue:

- Abbiamo trasformato, nella colonna sex, i semplici valori "0" e "1" in "F" per femmina e in "M" per maschio e poi abbiamo cambiato il tipo di dato da int a factor (quindi diviso in più livelli) per gli attributi "F" e "M".
- Abbiamo cambiato il tipo di dato per la colonna chest pain da int a factor quindi diviso in più livelli (0 1 2 3).
- Per la colonna cholesterol abbiamo trasformato in primo luogo tutti i valori "undefined" nella mediana dei valori di tutta la mia colonna; in secondo luogo abbiamo trasformato il tipo di dato da char a integer.
- Abbiamo cambiato il tipo di dato per la colonna f
bs da int a factor quindi diviso in più livelli (1 0).
- Abbiamo cambiato il tipo di dato per la colonna rest_ecg da int a factor quindi diviso in piu' livelli (0 1 2).

3.3 Codice

[#] Codice con R tradizionale

```
levels(dataset$chest_pain) [levels(dataset$chest_pain) == 0 ] <-</pre>
    "asymptomatic"
levels(dataset$chest_pain)[levels(dataset$chest_pain)== 1 ] <-</pre>
    "nontypical_angina"
levels(dataset$chest_pain)[levels(dataset$chest_pain)== 2 ] <-</pre>
    "nonanginal_pain"
levels(dataset$chest_pain)[levels(dataset$chest_pain)== 3 ] <-</pre>
    "typical_angina"
levels(dataset$fbs)[levels(dataset$fbs)== 0 ] <- "False"</pre>
levels(dataset$fbs)[levels(dataset$fbs)== 1 ] <- "True"</pre>
levels(dataset$rest_ecg)[levels(dataset$rest_ecg)== 0 ] <-</pre>
    "Ventricular_hypertrophy"
levels(dataset$rest_ecg)[levels(dataset$rest_ecg)== 1 ] <- "Normal"</pre>
levels(dataset$rest_ecg)[levels(dataset$rest_ecg)== 2 ] <- "Anomaly"</pre>
levels(dataset$exercise_angina)[levels(dataset$exercise_angina)== 0 ] <-</pre>
levels(dataset$exercise_angina) [levels(dataset$exercise_angina) == 1 ] <-</pre>
--etc... con via con gli altri attributi
# Codice con Tidyverse
dataset <- dataset %>%
 mutate(
   age <- as.integer(age),</pre>
   sex = ifelse(sex == "1", "M", "F"),
   sex = as.factor(sex),
   chest_pain = as.factor(chest_pain),
   cholesterol =ifelse(cholesterol == "undefined",
        median(cholesterol), cholesterol),
   cholesterol = as.integer(cholesterol),
   fbs = as.factor(fbs),
   rest_ecg = as.factor(rest_ecg),
   exercise_angina = as.factor(exercise_angina),
   slope = as.factor(slope),
   thalassemia = as.factor(thalassemia),
   thalessemia = as.factor(thalessemia),
   heart_disease = as.factor(heart_disease)
 )
```

3.4 Livelli dei fattori

Per vedere se le modifiche fossero avvenute con successo abbiamo stampato nuovamente il dataset, abbiamo rinominato i livelli dei fattori per ogni colonna, così da renderli più comprensibili.

```
Livelli per chest pain
0 = "asymptomatic"
1 = "nontypical_angina"
2 = "nonanginal_pain"
3 = "typical_angina"
Livelli per fbs
0 = "False"
1 = "True"
Livelli per rest_ecg
0 = "Ventricular_hypertrophy"
1 = "Normal"
2 = "Anomaly"
Livelli per exercise_angina
0 = "No"
1 = "Yes"
Livelli per slope
0 = "Descending"
1 = "Flat"
2 = "Ascending"
Livelli per thalessimia
0 = "non_existent"
1 = "defect_corrected"
2 = "normal_blood"
3 = "reversible_defect"
Livelli per heart disease
0 = "Yes"
1 = "No"
```

4 Consistenza dei dati

$4.1 \quad \text{max_hr}$

Per far emergere gli outlier e le anomalie della frequenza cardiaca più alta, come prima cosa abbiamo estratto con la funzione

```
hist(dataset$max_hr)
```

un grafico a barre dei valori dei dati forniti in ingresso dal dataset. In ggplot la funzione diventa

```
dataset %>%
  ggplot(aes(x, fill = f)) +
  geom_histogram()
```

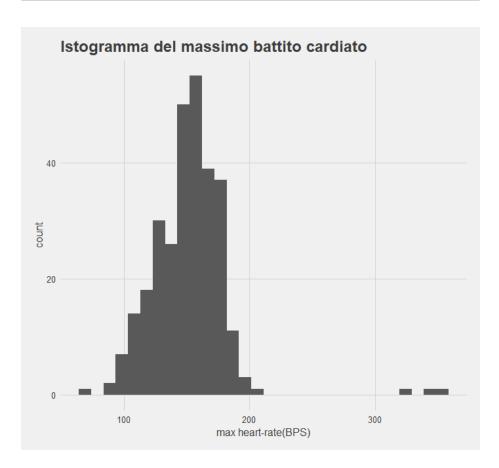


Figure 2: GRAFICO HIST max_hr con ggplot

Abbiamo deciso che il numero maggiore di battiti cardiaci non sia superiore a 222 e che il numero minore di battiti cardiaci sia il valore medio della variabile.

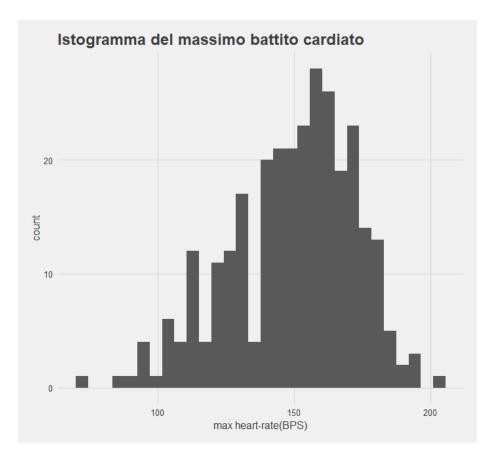
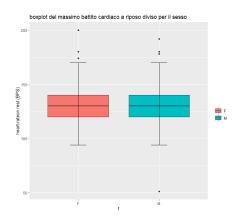


Figure 3: GRAFICO HIST max_hr con ggplot

4.2 rest_bp diviso per sesso

Per far emergere gli outlier e le anomalie della pressione sanguigna a riposo della persona, oltre ad un istogramma, abbiamo utilizzato un altro grafico più esplicativo: il boxplot [Figure 4 e 5].



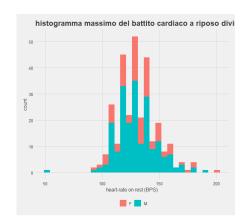
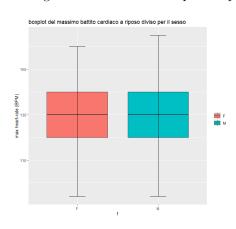


Figure 4: GRAFICO BOXPLOT max.hr con ggplot

Figure 5: GRAFICO HIST max_hr con ggplot

Abbiamo rappresentato gli outliers relativi alla variabile rest_bp', abbiamo calcolato il terzo e il primo quantile e li abbiamo sostituiti nella formula $\mathbf{IQR} < -(\mathbf{Q3} - \mathbf{Q1})$ per trovare lo scarto interquartile ed infine abbiamo rilevato il range della differenza interquartile per creare il nuovo boxplot.



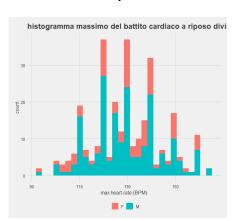


Figure 6: GRAFICO BOXPLOT max_hr con ggplot

Figure 7: GRAFICO HIST max_hr con ggplot

4.3 Considerazioni finali

Ristampando a console il dataset, aggiornato e modificato, abbiamo notato che è più consistente rispetto alla prima volta che l'avevamo visualizzato; tuttavia ci sono ancora delle modifiche da apportare. Abbiamo quindi impostato sull'attributo age un controllo che non permettesse di inserire valori inferiori a 0 o superiori di 120; mentre sull'attributo rest_bp, abbiamo inserito un range che va da 70 a 150 per la pressione sanguigna a riposo. Al termine di queste modifiche abbiamo ristampato il dataset abbiamo stabilito che non ci fossero più modifiche da apportare poiché, per noi ritenuto consistente.

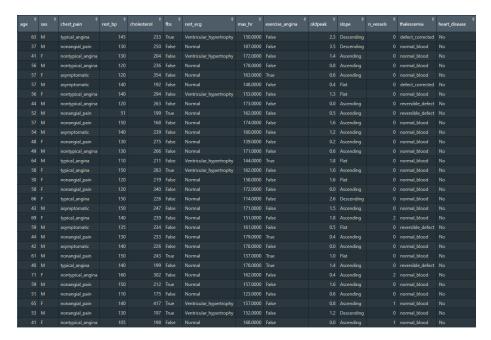


Figure 8: Screen dataset heart.csv aggiornato

5 Relazioni tra i dati

5.1 Regressione linare

Procedendo con l'analisi del dataset ci siamo confrontati con la regressione lineare semplice e quindi abbiamo messo in relazione due variabili per vedere, tramite gli appositi grafici di regressione lineare, se ci fosse o meno una correlazione.

5.2 age & rest_bp

Per prima cosa abbiamo scelto i due attributi da mettere in relazione, nel nostro caso l'età(age) e la pressione sanguigna a riposo della persona (rest_pb), e abbiamo stampato a video con la funzione **summary**(*dataset*) il loro contenuti suddivisi in quantili. Abbiamo continuato disegnando la retta di regressione e abbiamo notato, dal grafico, che tra age e rest_bp non c'era un apparente correlazione.

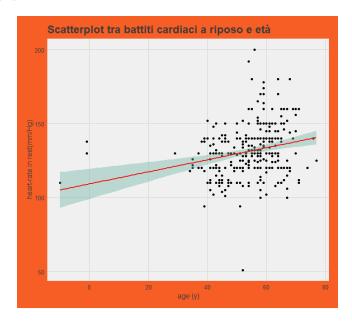


Figure 9: Relazione tra age e rest_bp

5.3 age & max_hr

Per poter continuare con il metodo della regressione abbiamo dovuto, perciò, cambiare attributi. Abbiamo scelto di mettere a confronto gli attributi age e max_hr; infatti, arrivati allo stesso punto di prima, abbiamo visto che c'era una forte correlazione tra i due. Infine, abbiamo potuto rappresentare il tutto tramite una rappresentazione grafica.

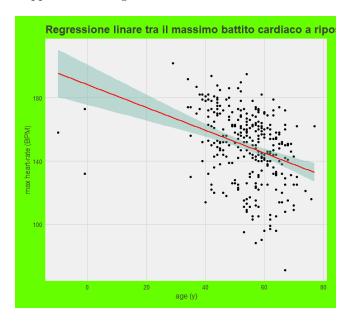


Figure 10: Relazione tra age e max_hr

Continuando la procedura, abbiamo calcolato il coefficiente di correlazione lineare e il coefficiente di determinazione per i due attributi (R^2) . Per ultimare la fase di regressione lineare abbiamo analizzato i residui con apposito grafico e, con un altro grafico, abbiamo potuto confrontare la distribuzione in quantili rispetto ad una distribuzione normale standard. E possiamo vedere nella [Figure 12] che i valori sono equidistribuiti intorno alla retta.

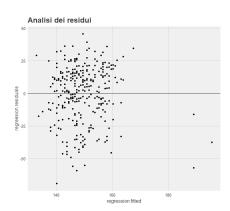


Figure 11: Grafico dei residui

Figure 12: Distribuzione in quantili confrontabile con quella di una normale

6 Modelli di machine learning

6.1 Data frame per previsioni

Come ultima consegna ci è stato chiesto di creare un data frame contente 10 osservazioni, non presenti nel dataset e quindi di effettuare delle previsioni.

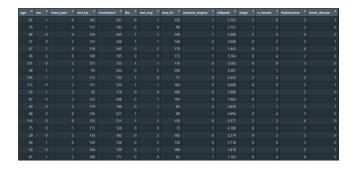


Figure 13: Dataset osservazioni.csv

Tramite python abbiamo creato uno script che ci permettesse di creare 20 osservazioni casuali per il nostro nuovo file "osservazioni.csv"

```
1 Security residue as: n

of (0)

sqc_min = 0
sqc_min
```

Figure 14: Script python

e grazie alla funzione $\mathbf{predict}(dataset)$ abbiamo potuto predire l'intervallo di confidenza per ogni mio attributo

6.2 Modelli di Machine Learning

Infine, abbiamo applicato tre modelli di machine learning (k-Nearest Neighbors, Multi-Layer Perceptron, Random Forest) per misurare l'accuratezza sul test set e, di conseguenza, è stato creato un **dotplot()** del risultato modelli utilizzati.

Risultati dei modelli di machine learning

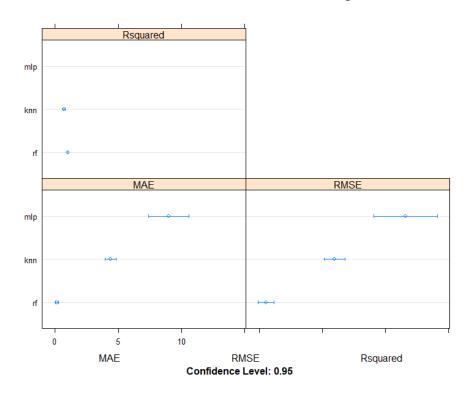


Figure 15: Grafico dei modelli di Machine Learning

6.3 K-Nearest Neighbors (KNN)

Il modello K-Nearest Neighbors (KNN) è uno degli altri algoritmi diffusi nel machine learning. Può essere utilizzato sia per problemi di classificazione che di regressione, anche se è più utilizzato nei primi. La forza di quest'algoritmo è che permette di memorizzare tutte le istanze disponibili e di classificarle valutando la distanza rispetto ai suoi vicini. L'istanza verrà assegnata alla classe che include il data point più vicino all'istanza stessa.

6.4 Perceptron Multistrato (MLP)

Il modello perceptron multistrato (MLP) è una rete neurale artificiale feedforward che genera un insieme di output da un insieme di input. Un MLP è caratterizzato da diversi livelli di nodi di input collegati come un grafico diretto tra i livelli di input e output. MLP utilizza la backpropogation per addestrare la rete. MLP è un metodo di apprendimento profondo.

6.5 Random Forest (RF)

Il modello Random Forest (RF)è un metodo versatile di machine learning, capace di affrontare sia compiti di classificazione che di regressione. Con le foreste casuali è anche possibile applicare metodi per la riduzione della dimensionalità, gestire dati mancanti, valori degli outlier ed altri passaggi essenziali di esplorazione dei dati, producendo buoni risultati.

7 Extra

Per facilitare la stesura del codice abbiamo utilizzato le librerie:

library(tidyverse)

library(caret)

library(ggthemes)

Per avere una condivisione più efficiente di tutti i file è stato usato GitHub.

Link di GitHub: https://github.com/r-vil/heart_in_r

8 Conclusioni Personali

Grazie a questo progetto, abbiamo potuto lavorare in squadra come se fossimo realmente dei programmatori in un ambito lavorativo.

Infatti, ci siamo scambiati le risorse con il software Github e abbiamo imparato a stilare una relazione descrittiva del nostro progetto con un ulteriore software, quale LaTex.

Ringraziamo tutte e due i docenti per averci fatto imparare un nuovo linguaggio di programmazione e speriamo di ritrovarci in un'altra unità formativa.

Grazie, Villanueva & Comini