**Villanueva - Comini**

**Prova d’esame**

**Analisi descrittiva**

Per avere una prima visione delle variabili raccolte ho caricato il dataset “heart.csv” nella nuova variabile dataset con la funzione ‘read.csv’ e ho di conseguenza analizzato la struttura dei dati raccolti.

Leggendo il dataset mi sono accorto della presenza di valori NA dovuti ad esempio al data-entry manuale. Quindi ho creato un ciclo che controllasse la presenza dei valori. Tutti i valori NA sono stati cancellati definitivamente dal dataset con la funzione ‘dataset <- na.omit(dataset)’.

Ho continuato controllando e rimuovendo le colonne non ritenute necessarie, quindi ho visualizzato il dataset attraverso la funzione ‘view(dataset)’ e ho rimosso le colonne non necessarie attraverso la funzione ‘subset()’. Ho deciso di eliminare la colonna X perché in questo caso ritenuta inutile.

Terminato il passaggio dell’eliminazione delle colonne superflue, ho continuato rinominando le colonne in maniera appropriata e descrivendo il tipo di ogni attributo di ciascuna. Infatti ho stampato il mio dataset con la funzione ‘str(dataset)’, con la funzione ‘names(dataset)[names(dataset) == "vecchio"] <- "nuovo"’ ho rinominato le colonne in modo appropriato, come ultima azione ho assegnato ad ogni attributo il suo tipo.

names(dataset)[names(dataset) == "cp"] <- "chest\_pain"

names(dataset)[names(dataset) == "trestbps"] <- "rest\_bp"

names(dataset)[names(dataset) == "chol"] <- "cholesterol"

names(dataset)[names(dataset) == "thalach"] <- "max\_hr"

names(dataset)[names(dataset) == "exang"] <- "exercise\_angina"

names(dataset)[names(dataset) == "thal"] <- "thalassemia"

names(dataset)[names(dataset) == "target"] <- "heart\_disease"

names(dataset)[names(dataset) == "ca"] <- "n\_vessels"

names(dataset)[names(dataset) == "restecg"] <- "rest\_ecg"

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

Ho eseguito poi un controllo per correggere la consistenza del tipo di dato per ogni variabile, infatti, con la funzione ‘[colonna del dataset] <- as.[tipo nel quale voglio cambiare i dati](colonna del dataset)’, ho trasformato i dati in modo che siano tecnicamente corretti.

* Ho trasformato, nella colonna sex, i semplici valori "0" e "1" in "F" per femmina e in "M" per maschio e poi ho cambiato il tipo di dato da int a factor (quindi diviso in più livelli) per i miei attributi “F” e “M”.
* Ho 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 ho trasformato in primo luogo tutti i valori "undefined" nella mediana dei valori di tutta la mia colonna e in secondo luogo ho trasformato il tipo di dato da char a integer.
* Ho cambiato il tipo di dato per la colonna fbs da int a factor quindi diviso in più livelli (1 - 0).
* Ho cambiato il tipo di dato per la colonna rest\_ecg da int a factor quindi diviso in più livelli (0 - 1 - 2).
* Ho cambiato il tipo di dato per la colonna exercise\_angina da int a factor quindi diviso in più livelli (1 - 0).
* Ho cambiato il tipo di dato per la colonna slope da num a factor quindi diviso in più livelli (0 - 1 - 2).
* Ho cambiato il tipo di dato per la colonna thalassemia da int a factor quindi diviso in più livelli (0 - 1 - 2 - 3).
* Ho cambiato il tipo di dato per la colonna heart\_disease da int a factor quindi diviso in più livelli (1 - 0).

Per vedere se le modifiche fossero avvenute con successo ho stampato nuovamente il dataset.

Di conseguenza ho rinominato i livelli dei fattori per ogni colonna del dataset cosi 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"

Per far emergere gli outlier e le anomalie, come prima cosa ho, con la funzione hist(dataset$max\_hr), estratto un grafico a barre dei valori dei dati forniti in ingresso dal dataset. Di conseguenza ho 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.

Ho rappresentato con un altro grafico più espilcativo, il boxplot, gli outliers relativi alla variabile prezzo boxplot(datset$rest\_bp). Ho calcolato il 3° e il 1° quantile e li ho sostituiti nella formula IQR<-(Q3-Q1) per trovare lo scarto interquartile. Ho rilevato, di conseguenza, il range della differenza interquartile per creare il nuovo boxplot.

Ristampando a console il dataset, aggiornato e modificato, ho notato che è più consistente rispetto alla prima volta che l’ho visualizzato ma ci sono ancora delle modifiche da apportare. Quindi ho impostato, sull’attributo age, un controllo che non mi permettesse di inserire valori inferiori a 0 o superiori di 120 invece, sull’attributo rest\_bp, ho inserito un range che va da 70 a 150 per la pressione sanguigna a riposo. Alla fine di tutte queste modifiche ho ristampato il dataset e ho stabilito che non ci fossero più modifiche da apportare perché per me consistente.

Procedendo con l’analisi del dataset mi sono confrontato con la regressione lineare semplice e quindi ho dovuto mettere in relazione due variabili e vedere, tramite gli appositi grafici di regressione lineare, se ci fosse ho meno una correlazione.

Per prima cosa ho scelto i due attributi da mettere in relazione, nel mio caso age e rest\_pb, e ho stampato a video con la funzione “summary()” il loro contenuti suddiviso in quantili. Ho continuato con il disegnare la retta di regressione e quindi ho visto dal grafico che tra age e rest\_bp non c’è correlazione. Quindi ho dovuto cambiare attributi per poter continuare con il metodo della regressione. A questo punto ho scelto di mettere a confronto gli attributi age e max\_hr infatti, arrivati allo stesso punto di prima, ho visto che c’è una forte correlazione tra i due. E infine ho potuto rappresentare il tutto tramite una rappresentazione grafica. Continuando, ho potuto calcolare il coefficiente di correlazione lineare e il coefficiente di determinazione per i due attributi. Per ultimare questa fase di regressione lineare ho analizzato i residui e poi con un altro grafico ho potuto confrontare la distribuzione in quantili rispetto ad una distribuzione normale standard.

Come ultima consegna ci è stato chiesto di creare un data frame contente 10 osservazioni ovviamente non presenti nel dataset e quindi di effettuare delle previsioni. Quindi tramite python abbiamo creato uno script che ci permettesse di creare 10 osservazioni casuali per il nostro nuovo file “osservazioni.csv”. Grazie alla funzione “predict()” ho potuto predire l’intervallo di confidenza per il mio attributo. Infine, abbiamo applicato due modelli di machine learning per misurare l’accuratezza sul test set. È stato utilizzato il metodo k-Nearest Neighbors e il metodo Multi-Layer Perceptron e poi è stato creato un dotplot() del risultato modelli utilizzati.