```
######################### Esame 28/01/2019 ################################
#
###########################################################################

####### 1 #######

### Importo le librerie di base
library("ggplot2") # plotting
library(MASS) # lda, qda
library(class) # knn
library(boot)
library(leaps)
library(outliers)

### Caricamento del dataset
mydforig <- read.csv("cesario.csv", sep =",")


mydforig <- read.csv("D:/spigama/Documenti/Dropbox/Marco-Mattia/Studio Universitario/04 - Metodi Statistici Data
Science/03 - Esercitazione/00 ESAME Gennaio/cesario.csv", sep =",")

#mydforig <- read.csv("C:/Users/Marco/Dropbox/Marco-Mattia/Studio Universitario/04 - Metodi Statistici Data
Science/03 - Esercitazione/00 ESAME Gennaio/cesario.csv", sep =",")

mydforig <- read.csv("/Users/mattiaspiga/Dropbox/Marco-Mattia/Studio Universitario/04 - Metodi Statistici Data
Science/03 - Esercitazione/00 ESAME Gennaio/cesario.csv", sep =",")

mydforig <- read.csv("/Users/mattiaspiga/Dropbox/Marco-Mattia/Studio Universitario/04 - Metodi Statistici Data
Science/03 - Esercitazione/00 - ESAME FINALE/01 - Cesarian 2Cat/cesario.csv",sep=",")

# 1.1 ANALISI ESPLORATIVA
# Visualizzo il data set e faccio una ANALISI ESPLORATIVA con le funzioni base
# noto che il data set ha una dimensione ridotta
# ci sono 80 osservazioni e 6 predittori + una variabile di risposta (Caesarian)
# Le osservazioni riguardano 80 parti di donne, in cui si analizzano alcuni dati sanitari fondamentali
# e probabilmente correlabili alla fine di valutare la necessit? di un taglio cesario
# Osservo che la variabile Caesarian

# Numero dei casi e delle variabili
dim(mydforig)

names(mydforig)

# Visualizza il dataset
View(mydforig)

# Le prime osservazioni del Dataset
head(mydforig)

# Le ultime osservazioni del Dataset
tail(mydforig)

# Struttura e classi del Dataset
str(mydforig)

# Osservazione generale
summary(mydforig)


# 1.2 DATA CLEANING e DATA TRANSFORMATION

# faccio una copia di mydf Origine al fine di preservarlo dalle successive modifiche
# il data set ha anche dimensioni ridotte pertanto un duplicato non ha un particolare onere di memorizzazione
mydf = mydforig


### La summary non ha segnalato  NA
### Analisi ulteriore presenza di variabili NA (not assigned)
# non ci sono na nel data set
sum(is.na(mydf))


# Rivedo alcune variabili per ridefinire in modo qualitativo

# Rimuovo la variabile id, essendo un progressivo potrebbe distorcere le valutazioni
# mydf = mydf[, -which(colnames(mydf)=="id")]

#1 Lascio Age, osservo una mediana a 27 anni
# non osservo valori anomali
quantile(mydf$Age)

#2. Delivery number (motivo del parto)
# il termine Programmato si riferisce ad un cesario programmato?
mydf$Delivery_number = factor(mydf$Delivery_number)
levels(mydf$Delivery_number) = c("Minaccia Immediata", "Compromessa, senza pericolo", "Nessuna condizione, ma
```

```
prematuro", "Programmato")


# 3. Delivery time (parto): {0 = Nato a termine, 1 = Nato prematuro, 2 = Nato post-termine}
mydf$Delivery_time = factor(mydf$Delivery_time)
levels(mydf$Delivery_time) = c("Nato a termine", "Nato prematuro", "Nato post-termine")


# 4. Blood Pressure (pressione del sangue): {0 = Bassa, 1 = Normale, 2 = Alta}
mydf$Blood_Pressure = factor(mydf$Blood_Pressure)
levels(mydf$Blood_Pressure) = c("Bassa", "Normale", "Alta")


# 5. Heart Problem (problemi al cuore): {0 = No, 1 = S?}
mydf$Heart_Problem = factor(mydf$Heart_Problem)
levels(mydf$Heart_Problem) = c("No", "Si")

# 6. Caesarian (parto con taglio cesario): {0 = No, 1 = S?}
mydf$Caesarian = factor(mydf$Caesarian)
levels(mydf$Caesarian) = c("No", "Si")




### 1.3 RICERCA DI OUTLIERS O HIGH LEVERAGE POINTS
# Ci possono essere dei valori che non seguono l'andamento generale del data set
# pertanto sono da analizzare con estrema attenzione e valutare caso per caso il da farsi
# la ricerca e' supportata da grafici

summary(mydf)

# Analizzo Age, reputo altamente significativa la distribuzione di frequenza dell'istogramma
# sono dati ospedaliari, sono stati presi per una ricerca? Sembrano distribuiti
hist(mydf$Age)



# calcoliamo gli outlier
# non trovo significativite motivazioni per eliminare id 40
mydf[outlier( mydf$Age, opposite = FALSE, logical = FALSE),]

# non noto outliers, non eseguo rimozioni
boxplot(mydf$Age)


### 1.4 ANALISI GRAFICA DELLE RELAZIONI TRA LE VARIABILI

# correlazioni
# non posso esseguire l'analisi delle correlazioni con corplot
# perch? l'unica variabile numerica ? age, le altre sono state impostate
# come qualitative

# Analiziamo con pairs il rapporco con tutte le variabili a coppie

pairs(mydf, main="Matrice degli scatterplot", col="blue")


#GRAFICO7
# Boxplot Et? (continua) - Cesario (categoria)
# Essendo Caesarian una variabile categoria, esploro la relazione mediante un Boxplot
# si osservano cosÃ¬ le relazioni con Age.
boxplot(Age~Caesarian, data = mydf,
        horizontal=TRUE,
        xlab="Eta' della donna",
        col=c("thistle","wheat"),
        main="Cesario in riferimento all'Eta' della donna")


#GRAFICO8
# Grafico a Punti Motivo del parto (categorica) - Et? (continua), colore ceario
# grafico per trovare una correlazione nei dati
ggplot2::ggplot(mydf, aes(x = Delivery_time, y = Age, colour = Caesarian, shape = Heart_Problem)) +
  geom_point() +
  labs(x = "Motivo del Parto", y = "Eta'") +
  ggtitle("Motivo del parto in funzione dell'eta'")


#GRAFICO9
# Grafico a punti Problemi al cuore (categorica) - Eta' (continua), colore cesario
# grafico per trovare una correlazione nei dati
# Trovo una certa significativit? tra problemi al cuore e fare un cesario
ggplot2::ggplot(mydf, aes(x = Heart_Problem, y = Age, colour = Caesarian)) +
  geom_point() +
  labs(x = "Problemi al cuore", y = "Eta'") +
  ggtitle("Motivo del parto in funzione dell'eta'")
```

```
#GRAFICO10
# Grafico di Densit? per eta'
# l'eta' sembra distribuita' uniformemente
ggplot2::ggplot(data = mydf) +
  geom_density(  mapping= aes(x=Age), alpha=0.3, fill="Red") + xlab("Age") + ylab("Densit?") + ggtitle("Grafico di
densit? per l'et? delle puerpere")



#GRAFICO11
# Torta suddivisione fasce di eta'- Age (continua raggruppata in quantili)
# eseguo i breaks per i quantili, ne analizzo la distribuzione
group = cut(mydf$Age, breaks = quantile(mydf$Age))
somma = as.data.frame(group)
somma[,"conta"] = 1
suddivisione = as.data.frame(aggregate(conta ~ group, data = somma, sum))
ggplot(suddivisione, aes(x="", y=suddivisione$conta, fill=suddivisione$group)) + geom_bar(width = 1, alpha = 0.7,
stat = "identity") + coord_polar("y", start=0) + ggtitle("Suddivisione delle donne per gruppi di et?")



#GRAFICO12
# L'eta' media delle donne che hanno avuto un cesareo ? superiore
#
ggplot2::ggplot(data = mydf ) +
  stat_summary( mapping = aes(x = Caesarian, y = Age, fill = Caesarian), fun.y = mean, geom = "bar",
                colour = "black") +
  xlab("Ceario") + ylab("Media delle medie delle donne") +
  ggtitle("Eta' media delle donne che hanno avuto un cesario")


#GRAFICO13
#oltre il 50% delle donne hanno censite avevano un delivery number con "minaccia immediata"
# vediamo se c'? una correlazione con il cesario
# Si pu? notare che pi? aumenta la minaccia per la donna/feto e pi? aumenta la probabilit? di un cesario
ggplot( data = mydf)  + geom_bar( mapping = aes( x = Delivery_number, fill = Caesarian))  +
  xlab("Motivo del parto") + ylab("Numero di parti suddivisi per cesario si/no") +
  ggtitle("Motivo del parto e relazione sul cesario")


#GRAFICO14
# Come al cresce dell'eta' della donna aumentino i casi sdi problemi cardiaci e di cesario?
ggplot(mydf, aes(x = reorder(id, Age), y = Age, colour = Caesarian, shape = Heart_Problem)) +
  geom_point() +
  labs(x = "Id", y = "Eta'") +
  ggtitle("Eta' delle donne cesario e problemi cardiaci")

# Se si ha la pressione normale il rischio di avere problemi al cuore si riduce di un terzo
table(mydf$Blood_Pressure, mydf$Heart_Problem)




####### 2 #######
#
# Sviluppare adeguati  modelli previsionali, utilizzando come variabile di risposta Caesarian. Commentare
# i modelli che si e' sviluppato, motivando quello ritenuto migliore (considerando sia l'aspetto predittivo,
# sia quello inferenziale)
#


# Si tratta anzitutto di un problema di classificazione. Infatti Caesarian e' una variabile qualitativa
# che puo' assumere due valori, vero o falso. Data la condizione binaria, il primo modello testabile e'
# una regressione logistica

# dai grafici ho potuto notare l'importanza della variabile
# delivery number (motivo del parto), oltre il 50% delle donne con minaccia del parto hanno subito un cesario
# hear problem, le donne con problemi al cuore hanno piu' probabilita' di subiro un cesario


# Rimuovo una variabile ID in quanto non utile all'elaborazione
mydf = mydf[, -which(colnames(mydf)=="id")]


# per realizzare il modello distinguo tra training set e test set
# con il data set preparo il modello
# con il test set ne valuto l'accuratezza
# il numero delle osservazioni ? abbastanza contenuto, potrebbe essere un problema applicando la logistica
set.seed(1)
nperc = 0.8
index_train = sample(1:nrow(mydf), nperc * nrow(mydf), replace = F)
```

```
### 2.1 Modello Regressione Logistica


# costruisco il modello con tutte i predittori presenti del data set tranne id
# ho rimosso il Delivery_number perch? da dei valori anomali sulla deviazione standard
# il quantile dei residui sono abbastanza centrati sullo zero, senza code anomale
mod_logistica <- glm(Caesarian ~  Delivery_time +  Heart_Problem, data = mydf, family = "binomial",  subset =
index_train)



summary(mod_logistica)

# la linea dei residui vs fitted resta centrata sullo zero senza deviazioni
# anche la scale location sembra confermare il modello
plot(mod_logistica)




# eseguo la verifica di accuratezza con il test set
pred_logistica = predict(mod_logistica, mydf[-index_train,], type="response" )


pred = rep("No", length(pred_logistica) )
# soglia di probabilita'
soglia = 0.5
pred[pred_logistica > soglia] = "Si"


# matrice di confusione
table(pred, mydf[-index_train, which(colnames(mydf)=="Caesarian")])


# il sistema riesce ad individuare il 68% dei casi
mean(pred == mydf[-index_train, which(colnames(mydf)=="Caesarian")])

# il tasso di errore a livello globale ? pari al 31%
1 - mean(pred == mydf[-index_train, which(colnames(mydf)=="Caesarian")])


# PROVO ALTRO CASO

mod_logistica2 <- glm(Caesarian ~  Blood_Pressure +  Heart_Problem, data = mydf, family = "binomial",  subset =
index_train)



### Proviamo un modello con tutte le variabili

mod_logisticaALL <- glm(Caesarian~., data = mydf, family = "binomial",  subset = index_train)

# eseguo la verifica di accuratezza con il test set
pred_logisticaALL = predict(mod_logisticaALL, mydf[-index_train,], type="response" )
predALL = rep("No", length(pred_logisticaALL) )
# soglia di probabilit?
sogliaALL = 0.5
predALL[pred_logisticaALL > sogliaALL] = "Si"


# matrice di confusione
table(predALL, mydf[-index_train, which(colnames(mydf)=="Caesarian")])


# il sistema riesce ad individuare il 50% dei casi
mean(predALL == mydf[-index_train, which(colnames(mydf)=="Caesarian")])


####### 3.2 #######
# Analisi con subset selection per confermare quanto analizzato sopra


subsets = regsubsets(Caesarian~., mydf)

subsets.summary = summary(subsets)

reg.summary = summary(subsets.summary, matrix.logical=TRUE)
## Forced in e Forsed out, forza a far entrare o a non far entrare nel modello variabilei

plot(subsets)
subsets.summary$rsq


#Disegnamo
```

```
#quale Ã¨ il modello che fa regisgtrare R^2 aggiustato piu alto?
which.max(subsets.summary$adjr2)

#Modello 11, raficamente

#scale= "Cp", "adjr2", "r2" or "bic"
plot (subsets, scale="adjr2")



## Se Ã¨ presente il quadrato nero, vuol dire che la variabile Ã¨ nel modello, scale di grigio Ã¨ la forza.
plot(subsets.summary$adjr2)
#Restituisce la posizione fisica
which.max(subsets.summary$adjr2 )

#Vediamo i coefficienti
coef(subsets, 4)

#Le entrate al di fuori delle variabili, da usare per elaboarazioni
subsets.summary$which

#Versione grafica del which
subsets.summary$outmat

plot(subsets, scale="adjr2")


# TODO ESEGUO IL TEST CON ANOVA

anova(mod_logistica, mod_logistica2, mod_logisticaALL, test = "LRT")



### LDA

# provo a verificare con la linear discriminant analysis
# i risultati dovrebbero essere migliori
# lda funziona meglio della logistica se si hanno pochi dati, ma si deve ipotizzare una distribuzione normale

mod_lda <- lda(Caesarian ~ Heart_Problem, data = mydf, subset = index_train )

# FARE LA Receiver Operating Characteristics (ROC)

summary(mod_lda)
plot(mod_lda)

# faccio la predizione sul test set e valuto l'accuratezza
pred_lda = predict(mod_lda, mydf[-index_train,] )

# i risultati che ottengo si avvicinano molto a quelli ottenuti con la logistica
table(pred_lda$class, mydf[-index_train,  which(colnames(mydf)=="Caesarian")])

mean(pred_lda$class == mydf[-index_train, which(colnames(mydf)=="Caesarian")])
(1 - mean(pred_lda$class ==  mydf[-index_train,  which(colnames(mydf)=="Caesarian")] ))*100 #tasso di err. clas.


#GRAFICO15
### Plot ROC - Almassimo sino a 2 classificatori
library(ROCR)
# choose the posterior probability column carefully, it may be
# lda.pred$posterior[,1] or lda.pred$posterior[,1], depending on your factor levels
pred <- prediction(pred_lda$posterior[,2], mydf[-index_train,]$Caesarian)
perf <- performance(pred,"tpr","fpr")
plot(perf,colorize=TRUE)



# i due precedenti modelli non si sono distinti significativamente
# pertanto utilizziamo qda

mod_qda <- qda(Caesarian ~ Delivery_time +  Heart_Problem  , data = mydf, subset = index_train )

summary(mod_qda)


# faccio la predizione sul test set e valuto l'accuratezza
pred_qda = predict(mod_qda, mydf[-index_train,] )

# i risultati che ottengo si avvicinano molto a quelli ottenuti con la logistica
table(pred_qda$class, mydf$Caesarian[-index_train])

mean(pred_qda$class == mydf$Caesarian[-index_train])
(1 - mean(pred_qda$class ==  mydf$Caesarian[-index_train] ))*100 #tasso di err. clas.
```

```
####### 2 - DA RIVEDERE


# i due precedenti modelli non si sono distinti significativamente
# verifichiamo con KNN
# il modello e' sviluppato solo con le due variabili e facciamo il cbind

train_knn = cbind(mydf$Delivery_time[index_train], mydf$Heart_Problem[index_train])
test_knn = cbind(mydf$Delivery_time[-index_train], mydf$Heart_Problem[-index_train])

knn_pred = knn(train_knn, test_knn, mydf$Caesarian[index_train], k = 3)

# Verifica accuratezza del modello
mean(knn_pred == mydf$Caesarian[-index_train])

# Matrice di confusione
table(knn_pred, mydf$Caesarian[-index_train])





####### 3 #######

# 3.1 Kfold CV

### TODO interpretazione CV
set.seed(123)

# kfold = nrow(mydf)
kfold = 10


# dall'help ricavo che le risposte binarie vanno gestite personalizzanzdo
# la funzione di costo:
# leave-one-out and 11-fold cross-validation prediction error for
# the nodal data set.  Since the response is a binary variable an
# appropriate cost function is
cost <- function(r, pi) mean(abs(r-pi)> 0.5)

mod_logistica_CV <- glm(Caesarian ~  Delivery_time +  Heart_Problem, data = mydf, family = "binomial" )
CV1 = cv.glm(mydf,mod_logistica_CV, cost, K = kfold)

mod_logistica_CV_1 <- glm(Caesarian ~., data = mydf, family = "binomial")
CV2 = cv.glm(mydf,mod_logistica_CV_1,cost, K = kfold)


#delta  spiegazione
# A vector of length two. The first component is the raw cross-validation estimate of prediction error.
# The second component is the adjusted cross-validation estimate. The adjustment is designed to compensate
# for the bias introduced by not using leave-one-out cross-validation.

# schelgo il secondo modello perche' presenta un minore tasso di errata classificazione
CV1$delta
CV2$delta


cost <- function(r, pi) mean(abs(r-pi)> 0.5)

LOOCV1 = cv.glm(mydf, mod_logistica_CV, cost)
LOOCV2 = cv.glm(mydf, mod_logistica_CV_1, cost)


LOOCV1$delta[1]
LOOCV2$delta[1]




#LOOCV MANUALE
count=0
for(i in 1:nrow(mydf)){
      fit = glm(formula = Caesarian ~ Delivery_time +  Heart_Problem, family = binomial, data = mydf[-i,])
      prob = predict(fit, newdata = mydf[i,], type = "response")
      pred = "No"
      if(prob > 0.5){
        pred="Si"
      }
      if(pred != mydf$Caesarian[i]){
        count = count + 1
      }
}
count/nrow(mydf)
```