install.packages("mice")

library(mice)

data <- iris

iris.mis <- subset(iris.mis, select = -c(Species))

iris.mis <- prodNA(iris, noNA = 0.02)

imputedD<-mice(iris.mis,m=5,maxit = 50,method= 'pmm',seed=500)

completeData <- complete(imputedD,2)

fit <- with(data = iris.mis, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))

combine <- pool(fit)

#for x=5 percent

iris.mis <- prodNA(iris, noNA = 0.05)

imputedD<-mice(iris.mis,m=5,maxit = 50,method= 'pmm',seed=500)

completeData <- complete(imputedD,2)

fit <- with(data = iris.mis, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))

combine <- pool(fit)

#for x=10 percent

iris.mis <- prodNA(iris, noNA = 0.1)

imputedD<-mice(iris.mis,m=5,maxit = 50,method= 'pmm',seed=500)

completeData <- complete(imputedD,2)

fit <- with(data = iris.mis, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))

combine <- pool(fit)

#for x=15 percent

iris.mis <- prodNA(iris, noNA = 0.15)

imputedD<-mice(iris.mis,m=5,maxit = 50,method= 'pmm',seed=500)

completeData <- complete(imputedD,2)

fit <- with(data = iris.mis, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))

combine <- pool(fit)

#for c=20 percent

iris.mis <- prodNA(iris, noNA = 0.2)

imputedD<-mice(iris.mis,m=5,maxit = 50,method= 'pmm',seed=500)

completeData <- complete(imputedD,2)

fit <- with(data = iris.mis, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))

combine <- pool(fit)

#for x=25 percent

iris.mis <- prodNA(iris, noNA = 0.25)

imputedD<-mice(iris.mis,m=5,maxit = 50,method= 'pmm',seed=500)

completeData <- complete(imputedD,2)

fit <- with(data = iris.mis, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))

combine <- pool(fit)

#B)Using second method of data imputation

#In this method we impute the data by replacing the missing data when mean

#When x = 2 percent

miris <- prodNA(iris, noNA = 0.02)

irismean <- impute(miris, fun = mean)

#When x = 5 percent

miris <- prodNA(iris, noNA = 0.05)

irismean <- impute(miris, fun = mean)

#When x = 10 percent

miris <- prodNA(iris, noNA = 0.1)

irismean <- impute(miris, fun = mean)

#When x = 15 percent

miris <- prodNA(iris, noNA = 0.15)

irismean <- impute(miris, fun = mean)

#When x = 20 percent

miris <- prodNA(iris, noNA = 0.2)

irismean <- impute(miris, fun = mean)

#When x = 25 percent

miris <- prodNA(iris, noNA = 0.25)

irismean <- impute(miris, fun = mean)

#C)Using third method of data imputation

#Using the miss forest package

#A non-parametric method that can handle categorical and continuous variables at the same time #via a very user-friendly interface in the missForest package.

#This iterative procedure fits a random forest model on the available data in order to predict the #missing values

#When percenatge of missing values is 2

set.seed(81)

iris2 <- subset(iris, select = -c(Species))

iris2<- prodNA(iris, noNA = 0.02)

iiris <-missForest(iris2)

#When percenatge of missing values is 5

set.seed(81)

iris2 <- subset(iris, select = -c(Species))

iris2<- prodNA(iris, noNA = 0.05)

iiris <-missForest(iris2)

#When percenatge of missing values is 10

set.seed(81)

iris2 <- subset(iris, select = -c(Species))

iris2<- prodNA(iris, noNA = 0.1)

iiris <-missForest(iris2)

#When percenatge of missing values is 15

set.seed(81)

iris2 <- subset(iris, select = -c(Species))

iris2<- prodNA(iris, noNA = 0.15)

iiris <-missForest(iris2)

#When percenatge of missing values is 20

set.seed(81)

iris2 <- subset(iris, select = -c(Species))

iris2<- prodNA(iris, noNA = 0.2)

iiris <-missForest(iris2)

#When percenatge of missing values is 25

set.seed(81)

iris2 <- subset(iris, select = -c(Species))

iris2<- prodNA(iris, noNA = 0.25)

iiris <-missForest(iris2)

#Performance of mice package

R2<-c(iris[8,1],iris[19,1],iris[24,1],iris[28,1],iris[29,1],iris[30,1],iris[33,1],iris[39,1],iris[49,1],iris[56,1],iris[57,1],iris[58,1],iris[71,1],iris[95,1],iris[119,1],iris[122,1],iris[135,1],iris[136,1],iris[139,1],iris[4,2],iris[10,2],iris[12,2],iris[17,2],iris[18,2],iris[36,2],iris[37,2],iris[45,2],iris[59,2],iris[80,2],iris[116,2],iris[117,2],iris[120,2],iris[149,2], iris[27,3],iris[29,3],iris[43,3],iris[50,3],iris[74,3],iris[82,3],iris[93,3],iris[106,3],iris[116,3],iris[144,3],iris[146,3]),iris[4,4],iris[10,4],iris[12,4],iris[17,4],iris[18,4],iris[36,4],iris[37,4],iris[45,4],iris[59,4],iris[80,4],iris[116,4],iris[117,4],iris[120,4],iris[149,4])

R1<-c(CompleteData[8,1],CompleteData[19,1],CompleteData[24,1], CompleteData[28,1],CompleteData[29,1],CompleteData[30,1],CompleteData[33,1],CompleteData[39,1],CompleteData[49,1],CompleteData[56,1],CompleteData[57,1],CompleteData[58,1],CompleteData[71,1],CompleteData[95,1],CompleteData[119,1],CompleteData[122,1],CompleteData[135,1],CompleteData[136,1],CompleteData[139,1],CompleteData[4,2],CompleteData[10,2],CompleteData[12,2],CompleteData[17,2],CompleteData[18,2],CompleteData[36,2],CompleteData[37,2],CompleteData[45,2],CompleteData[59,2],CompleteData[80,2],CompleteData[116,2],CompleteData[117,2],CompleteData[120,2],CompleteData[149,2],CompleteData[27,3],CompleteData[29,3],CompleteData[43,3],CompleteData[50,3],CompleteData[74,3],CompleteData[82,3],CompleteData[93,3],CompleteData[106,3],CompleteData[116,3],CompleteData[144,3],CompleteData[146,3],CompleteData[4,4],CompleteData[10,4],CompleteData[12,4],CompleteData[17,4],CompleteData[18,4],CompleteData[36,4],CompleteData[37,4],CompleteData[45,4],CompleteData[59,4],CompleteData[80,4],CompleteData[116,4],CompleteData[117,4],CompleteData[120,4],CompleteData[149,4])

rmse(R1,R2)

#The output of root mean square package for mice is 0. That is, mice package predicted all the missing values precisely.

#Performance of miss forest package

is.mat4 <- rep(NA,150)

is.mat3 <- rep(NA,150)

is.mat2 <- rep(NA,150)

is.mat1 <- rep(NA,150)

for(i in 1:nrow(iris2)){

if(is.na(iris2[i,1]))is.mat2[i]=iris[i,1]}

for(i in 1:nrow(iris2)){

if(is.na(iris2[i,2]))is.mat2[i]=iris[i,2]}

for(i in 1:nrow(iris2)){

if(is.na(iris2[i,3]))is.mat2[i]=iris[i,3]}

for(i in 1:nrow(iris2)){

if(is.na(iris2[i,4]))is.mat2[i]=iris[i,4]}

na.exclude(is.mat1)

na.exclude(is.mat2)

na.exclude(is.mat3)

na.exclude(is.mat4)

finalm<-c(is.mat,is.mat2,is.mat3,is.mat4)

is.mat5 <- rep(NA,150)

is.mat6 <- rep(NA,150)

is.mat7 <- rep(NA,150)

is.mat8 <- rep(NA,150)

for(i in 1:nrow(iris2)){

if(is.na(iris2[i,1]))is.mat5[i]=iiris[[1]][i,1]}

for(i in 1:nrow(iris2)){

if(is.na(iris2[i,2]))is.mat6[i]=iiris [[1]][i,2]}

for(i in 1:nrow(iris2)){

if(is.na(iris2[i,3]))is.mat7[i]=iiris[[1]][i,3]}

for(i in 1:nrow(iris2)){

if(is.na(iris2[i,4]))is.mat8[i]=iiris[[1]][i,4]}

z1<-na.exclude(is.mat5)

z2<-na.exclude(is.mat6)

z3<-na.exclude(is.mat7)

z4<-na.exclude(is.mat8)

finalm2<-c(z1,z2,z3,z4)

rmse(finalm,finalm2)

#The output of rmse is 0 for miss forest package.

#Performance of mean method

q1<-mean(iris\$Sepal.Length)

miris <- subset(miris, select = -c(Species))

rmse(q1,q2)

0.9128709

#The result of rmse for mean method is 0.9128

#Therfore we observe that performane of mice and miss forest which uses multiple imputation #performs best. Both the

#multiple impuation have zero percent rmse error.

#Whereas mean method performs worst with 0.9128709 rmse error.

#Knn classifier to classify iris data set

#Random splitting of iris data as 70 percent train and 30 percent test datasets

ind <- sample(2, nrow(iris), replace=TRUE, prob=c(0.7, 0.3))

trainData <- iris[ind==1,]

testData <- iris[ind==2,]

#removing factorvariable from training and test datasets

trainData1 <- trainData[-5]

testData1 <- testData[-5]

iristrainlabels <- trainData\$Species

install.packages("class")

library(class)

iristestpred1<-knn(train=trainData1,test=testData1,cl=iristrainlabels,k=3,prob=TRUE)

#Evaluating the performance of knn

install.packages("gmodels")

library(gmodels)

CrossTable(x = iristestlabels, y = iristestpred1,prop.chisq=FALSE)

for(i in 1:nrow(iris)){

if(iris$Species[i]=="versicolor" && iris$Sepal.Length[i] > =6.5){iris\$Sepal.Length[i]='na'}}