**1. Perform the below given activities:**

**a. Take a sample data set of your choice**

Ans:

heart <- read.csv("C:/Users/hemakumar/Downloads/heart.csv")

head(heart)

heart$target<-as.factor(heart$target)

heart$sex<-as.factor(heart$sex)

heart$fbs<-as.factor(heart$fbs)

heart$exang<-as.factor(heart$exang)

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

data <- heart[ind==1,]

newdata <- heart[ind==2,]

head(data)

str(data)

xtabs(~target+sex+fbs+exang,data=heart)

b. Apply random forest, logistic regression using Spark R

Ans:

#logistic regression

my\_modle<-glm(target~.,data=data,family = "binomial")

summary(my\_modle)

#or

control <- trainControl(method = 'repeatedcv',

number = 10,

repeats = 3)

seed <-7

metric <- 'Accuracy'

set.seed(seed)

lg\_reg <- train(target~.,

data = data,

method = 'glm',

metric = metric,

trControl = control,

maxit=100)

#random forest

control <- trainControl(method = 'repeatedcv',

number = 5,

repeats = 3)

library(randomForest)

set.seed(7)

mtry <- sqrt(ncol(data))

rf\_random <- train(target~.,

data = data,

method = 'rf',

metric = 'Accuracy',

tuneLength = 7,

trControl = control)

print(rf\_random)

plot(rf\_random)

**c. Predict for new dataset**

Ans:

#logistic regression

#predictions

p1<-predict(my\_modle,data,type = "response")

head(p1)

#accuracy

pred1<-ifelse(p1>0.5,1,0)

tab1<-table(pred1,data$target)

accuracy<-sum(diag(tab1))/sum(tab1)

#prediction for new data

p2<-predict(my\_modle,newdata,type = "response")

head(p2)

#accuracy

pred2<-ifelse(p2>0.5,1,0)

tab2<-table(pred2,newdata$target)

accuracy<-sum(diag(tab2))/sum(tab2)

#cross validation

#predictions

p1<-predict(lg\_reg ,data)

head(p1)

#accuracy

tab1<-table(p1,data$target)

accuracy<-sum(diag(tab1))/sum(tab1)

#prediction for new data

p2<-predict(lg\_reg,newdata)

head(p2)

#accuracy

tab2<-table(p2,newdata$target)

accuracy<-sum(diag(tab2))/sum(tab2)

#random forest

predictions<- predict(rf\_random,data)

#accuracy

tab1<-table(predictions,data$target)

accuracy<-sum(diag(tab1))/sum(tab1)

#or

confusionMatrix<- confusionMatrix(predictions,data$target)

confusionMatrix

#for new data

predictions1<- predict(rf\_random,newdata)

#accuracy

tab2<-table(predictions1,newdata$target)

accuracy<-sum(diag(tab2))/sum(tab2)