**R code for MDR classification using machine learning**

**Step 1 : Data cleaning and building a Scoring system**

Library(ggplot2)

library(dplyr)

library(magrittr)

library(leaps)

library(corrplot)

library(readxl)

library(sqldf)

library(RWeka)

library(DT)

library(sqldf)

data1<-read.csv(“fulldata\_nti.csv”)

str(data1)

View(data1)

sum(is.na(data1))

**# Assign scores for resistance to Rifampicin and Isoniazid**

data1$class1 <- ifelse(data1$cdstInh == "Resistant",1,0)

data1$class2 <- ifelse(data1$cdstRif == "Resistant",1,0)

data1$classA <- ifelse(data1$ddstInh == "Resistant",1,0)

data1$classB <- ifelse(data1$ddstRif == "Resistant",1,0)

**#Other drug resistance**

data1$class3 <- ifelse(data1$cdstE == "Resistant",1,0)

data1$class4 <- ifelse(data1$cdstPyz == "Resistant",1,0)

data1$classC <- ifelse(data1$ddstE == "Resistant",1,0)

data1$classD <- ifelse(data1$ddstPyz == "Resistant",1,0)

data1$class5 <- ifelse(data1$cdstKan == "Resistant",1,0)

data1$class6 <- ifelse(data1$cdstAmi == "Resistant",1,0)

data1$classE <- ifelse(data1$ddstKan == "Resistant",1,0)

data1$classF <- ifelse(data1$ddstAmi == "Resistant",1,0)

data1$class7 <- ifelse(data1$cdstCap == "Resistant",1,0)

data1$class8 <- ifelse(data1$cdstOfl == "Resistant",1,0)

data1$classG <- ifelse(data1$ddstCap == "Resistant",1,0)

data1$classH <- ifelse(data1$ddstOfl == "Resistant",1,0)

data1$class9 <- ifelse(data1$cdstLev == "Resistant",1,0)

data1$class10 <- ifelse(data1$cdstMox == "Resistant",1,0)

data1$classI <- ifelse(data1$ddstLev == "Resistant",1,0)

data1$classJ <- ifelse(data1$ddstMox == "Resistant",1,0)

data1$class11 <- ifelse(data1$cdstPas == "Resistant",1,0)

data1$class12 <- ifelse(data1$cdstEth == "Resistant",1,0)

data1$classK <- ifelse(data1$ddstPas == "Resistant",1,0)

data1$classL <- ifelse(data1$ddstEth == "Resistant",1,0)

newtab2<-data1

require(dplyr)

class\_filt<-select(newtab2,class1,class2,class3,class4,

class5,class6,class7,class8,class9,class10,class11,

class12,classA,classB,classC,classD,classE,classF,

classG,classH,classI,classJ,classK,classL)

View(class\_filt)

sum(is.na(class\_filt))

class\_filt[rowSums(is.na(class\_filt))>23 , ] <- 5

sum(is.na(class\_filt))

View(class\_filt)

class\_filt[is.na(class\_filt)] <- 0

class\_filt$sumc<-class1+class2

class\_filt$sumd<-classA+classB

**#Final Score of MDR patients**

class\_filt$MDRSUM=class\_filt$sumc+class\_filt$sumd

class\_filt$sume<-class3+class4

class\_filt$sumf<-classC+classD

class\_filt$sumg<-class5+class6

class\_filt$sumh<-classE+classF

class\_filt$sumi<-class7+class8

class\_filt$sumj<-classG+classH

class\_filt$sumk<-class9+class10

class\_filt$suml<-classI+classJ

class\_filt$summ<-class11+class12

class\_filt$sumn<-classK+classL

**#Final score of Other drug resistant patients**

class\_filt$DRUG\_SUM<-class\_filt$sume+class\_filt$sumf+class\_filt$sumg+class\_filt$sumh+class\_filt$sumi+class\_filt$sumj+class\_filt$sumk+class\_filt$suml+class\_filt$summ+class\_filt$sumn

View(class\_filt)

newtab2$Resistance<-class\_filt$MDRSUM

newtab2$susceptible<-class\_filt$DRUG\_SUM

View(newtab2)

str(newtab2)

newtab2$MDR <- ifelse(newtab2$Resistance >=1,"yes","No")

newtab2$Drugrest <- ifelse(newtab2$susceptible >=1,"yes","No")

**#Removing the extra classes created for scoring table**

newtab2<-within(newtab2, rm(class1))

newtab2<-within(newtab2, rm(class2))

newtab2<-within(newtab2, rm(class3))

newtab2<-within(newtab2, rm(class4))

newtab2<-within(newtab2, rm(class5))

newtab2<-within(newtab2, rm(class6))

newtab2<-within(newtab2, rm(class7))

newtab2<-within(newtab2, rm(class8))

newtab2<-within(newtab2, rm(class9))

newtab2<-within(newtab2, rm(class10))

newtab2<-within(newtab2, rm(class11))

newtab2<-within(newtab2, rm(class12))

newtab2<-within(newtab2, rm(classA))

newtab2<-within(newtab2, rm(classB))

newtab2<-within(newtab2, rm(classC))

newtab2<-within(newtab2, rm(classD))

newtab2<-within(newtab2, rm(classE))

newtab2<-within(newtab2, rm(classF))

newtab2<-within(newtab2, rm(classG))

newtab2<-within(newtab2, rm(classH))

newtab2<-within(newtab2, rm(classI))

newtab2<-within(newtab2, rm(classJ))

newtab2<-within(newtab2, rm(classK))

newtab2<-within(newtab2, rm(classL))

**#Renaming the columns and converting them to factor variable**

colnames(newtab2)[colnames(newtab2) == 'X.18'] <- 'Smokestatus'

colnames(newtab2)[colnames(newtab2) == 'X.19'] <- 'Drinkstatus'

colnames(newtab2)[colnames(newtab2) == 'X.20'] <- 'diabetic'

colnames(newtab2)[colnames(newtab2) == 'X.21'] <- 'Hypertension'

colnames(newtab2)[colnames(newtab2) == 'X.23'] <- 'HIVstatus'

colnames(newtab2)[colnames(newtab2) == 'X.37'] <- 'TBcontact'

colnames(newtab2)[colnames(newtab2) == 'X.41'] <- 'Finstatus'

colnames(newtab2)[colnames(newtab2) == 'X.3'] <- 'Citystatus'

colnames(newtab2)[colnames(newtab2) == 'X.7'] <- 'Gender'

colnames(newtab2)[colnames(newtab2) == 'X.11'] <- 'Marital'

colnames(newtab2)[colnames(newtab2) == 'X.12'] <- 'Education'

colnames(newtab2)[colnames(newtab2) == 'X.13'] <- 'Jobtype'

newtab2$MDR<-as.factor(newtab2$MDR)

newtab2$Drugrest<-as.factor(newtab2$Drugrest)

newtab2$Smokestatus<-as.factor(newtab2$Smokestatus)

newtab2$Drinkstatus<-as.factor(newtab2$Drinkstatus)

newtab2$diabetic<-as.factor(newtab2$diabetic)

newtab2$Hypertension<-as.factor(newtab2$Hypertension)

newtab2$HIVstatus<-as.factor(newtab2$HIVstatus)

newtab2$TBcontact<-as.factor(newtab2$TBcontact)

newtab2$Finstatus<-as.factor(newtab2$Finstatus)

newtab2$Citystatus<-as.factor(newtab2$Citystatus)

newtab2$Gender<-as.factor(newtab2$Gender)

newtab2$Marital<-as.factor(newtab2$Marital)

newtab2$Education<-as.factor(newtab2$Education)

newtab2$Jobtype<-as.factor(newtab2$Jobtype)

Fin\_data\_clean<-select(newtab2,statecode,Citystatus,age,Gender,Marital,Education,Jobtype,Smokestatus,Drinkstatus,diabetic,Hypertension,HIVstatus,TBcontact,Finstatus,Resistance,MDR,Drugrest)

str(Fin\_data\_clean)

sum(is.na(Fin\_data\_clean))

clean\_data <- Fin\_data\_clean[complete.cases(Fin\_data\_clean), ]

newtab4=clean\_data

**Step 2 : PREPARE THE DATA FOR CLUSTERING**

**#Removing MDR and drug resistance variables**

newtab4<-within(newtab4, rm(Resistance))

newtab4<-within(newtab4, rm(MDR))

**# Checking the MDR counts against each category**

table(newtab4$Marital,newtab4$MDR)

table(newtab4$Citystatus,newtab4$MDR)

table(newtab4$Education,newtab4$MDR)

table(newtab4$Gender,newtab4$MDR)

table(newtab4$Drugrest,newtab4$MDR)

table(newtab4$Smokestatus,newtab4$MDR)

table(newtab4$Drinkstatus,newtab4$MDR)

table(newtab4$Jobtype,newtab4$MDR)

table(newtab4$Finstatus)

newtab4\_prev<- sqldf("Select \* from newtab4

where Finstatus ='Previously Treated'")

newtab4\_prev\_cases=newtab4\_prev

newtab4\_prev\_cases<-within(newtab4\_prev\_cases, rm(MDR))

str(newtab4\_prev)

newtab4\_fresh\_cases <- sqldf("Select \* from newtab4

where Finstatus ='New'")

cl1 <- SimpleKMeans(newtab4\_prev\_cases, Weka\_control(N = 4))

attributes(cl1)

newtab4\_prev\_cases$Cluster <- cl1$class\_ids

newtab4\_prev\_cases$MDR<-newtab4\_prev$MDR

attach(newtab4\_prev\_cases)

table(newtab4\_prev\_cases$Cluster,newtab4\_prev\_cases$MDR)

View(newtab4)

require(plyr)

**STEP 3 : Estimating the distribution of MDR vs NON-MDR vs X-DR cases**

newclust\_MDR <- sqldf("Select \* from newtab4\_prev\_cases

where cluster = 0 ")

only\_MDR <- sqldf("Select \* from newtab4\_prev\_cases

where MDR=='yes' AND Drugrest=='No' ")

only\_Drug <- sqldf("Select \* from newtab4\_prev\_cases

where MDR=='No' AND Drugrest=='yes' ")

MDR\_AND\_DRUG<-sqldf("Select \* from newtab4\_prev\_cases

where MDR=='yes' AND Drugrest=='yes' ")

NO\_MDR\_AN\_DRUG<-sqldf("Select \* from newtab4\_prev\_cases

where MDR=='No' AND Drugrest=='No' ")

**Building Naïve Bayes Conditional probability table**

require(e1071)

newclust\_MDR<-within(newclust\_MDR, rm(Cluster))

statecode=as.factor(newclus=t\_MDR$statecode)

bayes\_0<-naiveBayes(MDR~.,data=newclust\_MDR)

bayes\_0$apriori

bayes\_0$tables

count(newclust\_MDR,"TBcontact")

count(newclust\_MDR,"Finstatus")

count(newclust\_MDR,"HIVstatus")

count(newclust\_MDR,"Hypertension")

count(newclust\_MDR,"diabetic")

count(newclust\_MDR,"Smokestatus")

count(newclust\_MDR,"Drinkstatus")

**# CALCULATING ODDS RATIO**

logitmod<-glm(MDR~.,family=binomial(),data=newclust\_MDR)

contrasts(newclust\_MDR$MDR)

contrasts(newclust\_MDR$diabetic)

contrasts(newclust\_MDR$cluster)

contrasts(newclust\_MDR$HIVstatus)

contrasts(newclust\_MDR$Smokestatus)

contrasts(newclust\_MDR$Drinkstatus)

summary(logitmod)

attributes(logitmod)

expb<-exp(coef(logitmod))

expb

exp(cbind(coef(logitmod), confint(logitmod)))

pred.probs<-predict(logitmod,type="response")

pred.probs<-as.data.frame(pred.probs)

View(pred.probs)

tail(pred.probs)

str(pred.probs)

pred<-ifelse(pred.probs < 0.4 ,"no","yes")

pred<-as.factor(pred)

newclust\_MDR$predicted=pred

str(pred)

head(pred)

tail(pred)

table(newclust\_MDR$MDR,pred)

accuracy<-mean(pred==newclust\_MDR$MDR)

accuracy

misc1<-mean(pred!=newclust\_MDR$MDR)

misc1

**BUILDING PREDICTION MODELS**

newtab6=newclust\_MDR

str(newtab6)

newtab6<-within(newclust\_MDR,rm(Cluster))

require(randomForest)

require(caret)

n<-sample(341,200)

#head(newtab6[,-10])

train\_set<-newtab6[n,]

test\_set<-newtab6[-n,]

head(train\_set)

head(test\_set)

rf<-randomForest(MDR~.,data=train\_set,ntree=500)

rf

varImpPlot(rf)

importance(rf)

pred1<-predict(rf,type="prob")

pred0<-predict(rf,train\_set)

pred2<-predict(rf,test\_set)

confusionMatrix(pred0,train\_set$MDR,positive="yes")

confusionMatrix(pred2,test\_set$MDR,positive="yes")

library("e1071")

svm\_model <- svm(MDR ~ ., data=train\_set)

svm\_pred<-predict(svm\_model,train\_set)

confusionMatrix(svm\_pred,train\_set$MDR,positive="yes")

attributes(pred)

table(train\_set$MDR,pred1)

head(pred1)

library(ROCR)

perf<-prediction(pred1[,2],train\_set$MDR)

auc<-performance(perf,"auc")

auc

pred3 = performance(perf, "tpr","fpr")

plot(pred3,main="ROC Curve for Random Forest",col=2,lwd=2)

abline(a=0,b=1,lwd=2,lty=2,col="gray")