**SUBMISSION PART 1**

**GROUPING BY CLUSTERING TECHNIQUE**

**# GETTING DATA**

data<-read.csv("sample\_data\_test.csv")

**# INPUT “NULL” VALUES THROUGH KNN TECHNIQUE**

install.packages("VIM")

library(VIM)

data$minor[which(data$minor=="NULL")]<-NA

final<-kNN(data,variable = "minor")

final$marital\_status[which(final$marital\_status=="NULL")]<-NA

final<-kNN(final,variable = "marital\_status")

final$sex[which(final$sex=="NULL")]<-NA

final<-kNN(final,variable = "sex")

final$educational\_status[which(final$educational\_status=="NULL")]<-NA

final<-kNN(final,variable = "educational\_status")

final$income[which(final$income=="NULL")]<-NA

final<-kNN(final,variable = "income")

final$date\_of\_birth[which(final$date\_of\_birth=="NULL")]<-NA

final<-kNN(final,variable = "date\_of\_birth")

final<-final[,-which(names(final)=="date\_of\_birth")]

# **GETTING SIGNIFICANT VARIABLES**

install.packages("Boruta")

library(Boruta)

set.seed(123)

boruta.train <- Boruta(dependents~., data = final, doTrace = 2)

final.boruta <- TentativeRoughFix(boruta.train)

boruta.df <- attStats(final.boruta)

print(boruta.df)

**#SEPERATING SIGNIFICANT VARIABLES ONLY :-**

DeleteNames<-rownames(boruta.df)[which(boruta.df$decision == "Rejected")]

listD<-0

lenA<-length(names(final))

lenB<-length(DeleteNames)

for(i in 1:lenA)

{

for(j in 1:lenB)

{

if(names(final)[i] == DeleteNames [j])

{ listD<-c(listD,i) }

}

}

final<-final[,-listD]

**# GROUPING THROUGH CLUSTER FORMATION**

install.packages("cluster")

library(cluster)

pamClus <- pam(final, 5)

**# SEPERATING CLUSTERS**

listA<-c(which(pamClus$clustering==1))

listB<-c(which(pamClus$clustering==2))

listC<-c(which(pamClus$clustering==3))

listD<-c(which(pamClus$clustering==4))

listE<-c(which(pamClus$clustering==5))

IDSET1<-final[listA,]

IDSET2<-final[listB,]

IDSET3<-final[listC,]

IDSET4<-final[listD,]

IDSET5<-final[listE,]

GROUP1 <- IDSET1$cid\_no

GROUP2 <- IDSET2$cid\_no

GROUP3 <- IDSET3$cid\_no

GROUP4 <- IDSET4$cid\_no

GROUP5 <- IDSET5$cid\_no

FINAL CLUSTERS FORMED ARE “IDSET” AND GROUPING IS DONE ON THE BASIS OF “ cid\_no”.

IF YOU WANT GROUPING BASED ON ANY VARIABLE JUST CHANGE THE LAST PART.