Emmanuel Asante (933274210) ST 557 Final Project.

REPORT OF FINAL PROJECT.

BACKGROUND OF DATA

The data was obtained from a paper named "The Five Factor Model of personality and evaluation of drug consumption risk" published by Ferhman et al , in January, 2017. In this paper they tried to evaluate an individual's risk of drug consumption and misuse. They obtained data through conducting an online survey. The data set contained information on the consumption of 18 central nervous system psychoactive drugs. The database contains responses of 1885 respondents where 12 attributes were measured. Variables measured can be grouped into sets called Personality measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), as well as some demographic data such as level of education, age, gender, country of residence and ethnicity. All input attributes are originally categorical and are quantified. In addition, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers.

QUESTIONS ADDRESSED

1. Compare between the user classes of a specific drug Cannabis based on the attributes/variables using Hotelling's T2 and/or MANOVA. You can either use all the variables or a subset of variables like NEO-FFI-R.

2. Compare between the psychological variables for respondents between between different gender, country or age using Hotelling's T2 and/or MANOVA.

3. Perform PCA on variables to find important combinations of variables.

4. Perform Factor Analysis to find important factors from the variables.

6. Perform cluster analysis on the variables and compare the clusters with the drug consumption classes for different drugs.

7. Find classification rules for classifying between drug consumption classes based on the variables. Classification problem can either be seven class or two class problem.

FINDINGS OF ANALYSIS

1. **Hotellings Test Findings**

a. After conducting the Hotellings Test of equality of mean NEO-FFI-R scores between Users and Non-Users of Cannabis, I found that there is convincing evidence that the mean personality (NEO-FFI-R) scores of cannabis Users is different from the mean personality(NEO-FFI-R) scores of Non-users of cannabis. (p-value(<0.001), significance level=0.05)

b. There is convincing evidence that the mean personality (NEO-FFI-R) scores of heroin users is different from the mean personality(NEO-FFI-R) scores of Non-users of heroin. (p-value(<0.001), significance level=0.05)

MANOVA

a. There is convincing evidence that the mean personality (NEO-FFI-R) scores of cannabis Users is different among users of different Education backgrounds or levels (p-value<0.001, significance level=0.05).

b. There is no evidence that the mean personality (NEO-FFI-R) scores of Heroin Users is different among users of different Education backgrounds or levels (p-value=0.8775, significance level=0.05).

2. **MANOVA**

   a. There is convincing evidence that the mean psychological variables score of cannabis Users is different among different age groups of users (p-value<0.001, significance level=0.05).

   b. There is no evidence that the mean psychological scores of Heroin Users is different among users of different Age groups (p-value=0.1967, significance level=0.05).

3. **PRINCIPAL COMPONENTS ANALYSIS**

Since the variables are of different scales, I used the correlation matrix to in computing the principal components. I have decided to go with the first 3 principal components as a way of dimensionality reduction since, together they explain more than 50% of the total variation as indicated by the scree plot. It is however worthy of mention that the first 3 principal components explain significantly higher amount (about 50% of the total variation) of the total variation than each of the other principal components (i.e using the elbow rule).

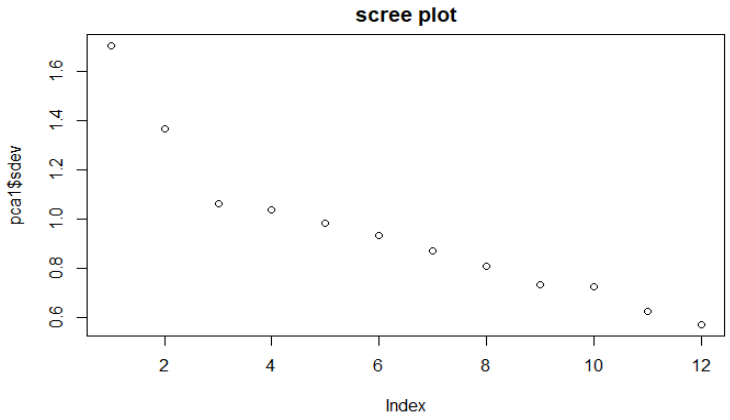
* From the loadings plot of principal component 1, it can be observed that the variables SS and Impulse contributed most in terms of weight.Thus individuals with a high component score will be those with high SS and Impulsive score as well as of younger age.

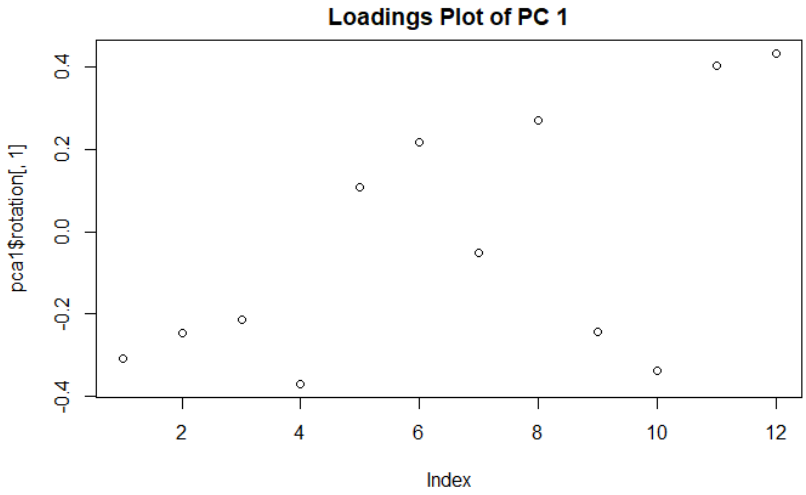
* The variables Oscore and Escore contributed most to the second principal component. Thus individuals with a high component score will be those with high NScore and low Oscore and Escore.

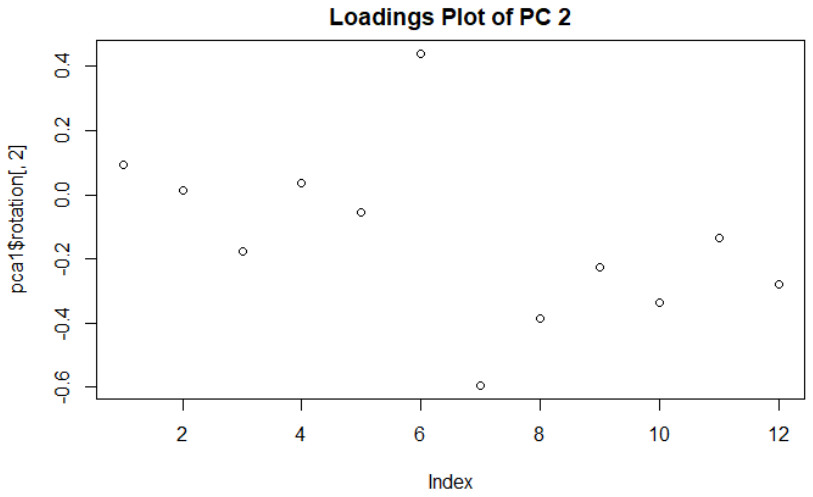
* The variables Education and Nscore contributed most to the Third principal component. Thus individuals with a high component score will be those with high scores in Education and NScore.

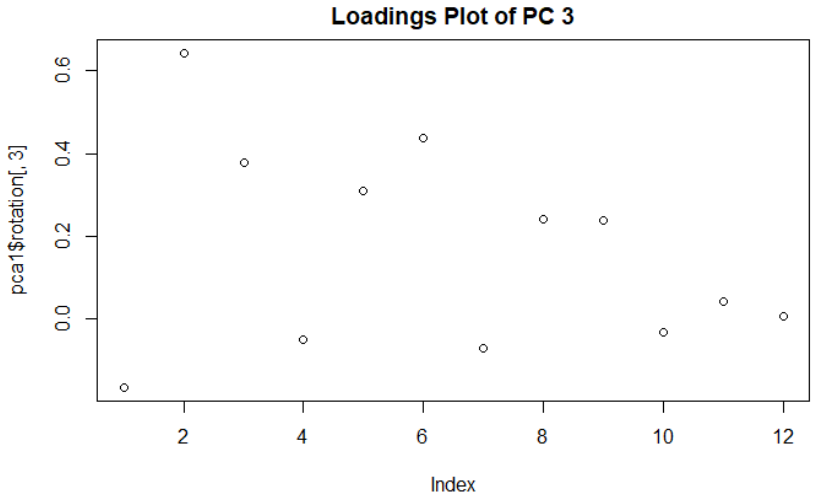
It is worth noting that the original variables are not strongly correlated except for the correlation between the SS and Impulsive score variables. PCA performs better when the original variables are highly correlated.

Below are the Scree plot and Loadings plot of Principal Components 1,2, and 3.







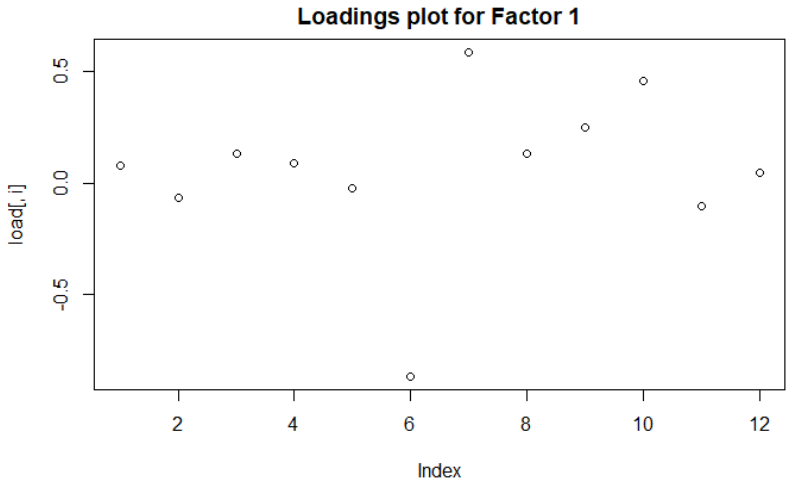


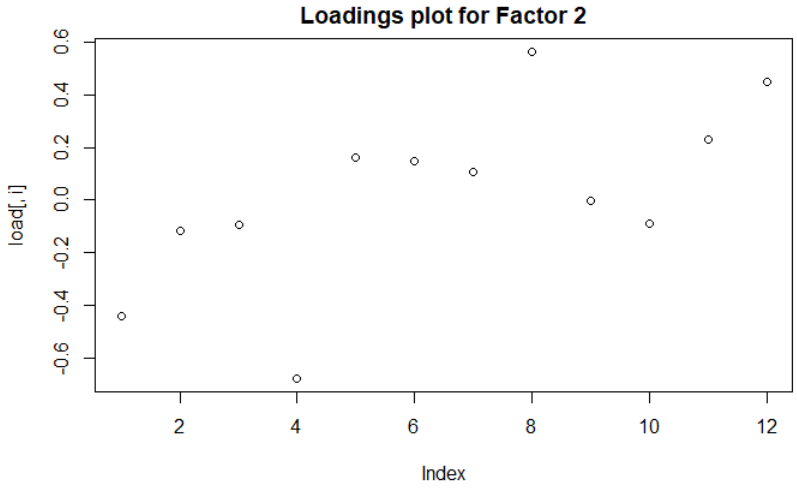
4. FACTOR ANALYSIS

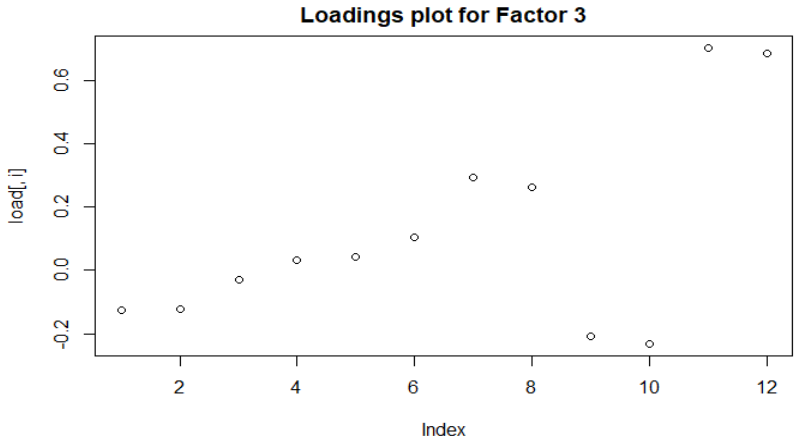
Looking at the p-values, it was best to do factor analysis with 4 factors since it gave the p-value=0.1166 which led to a conclusion to use 4 factors with the ML test.

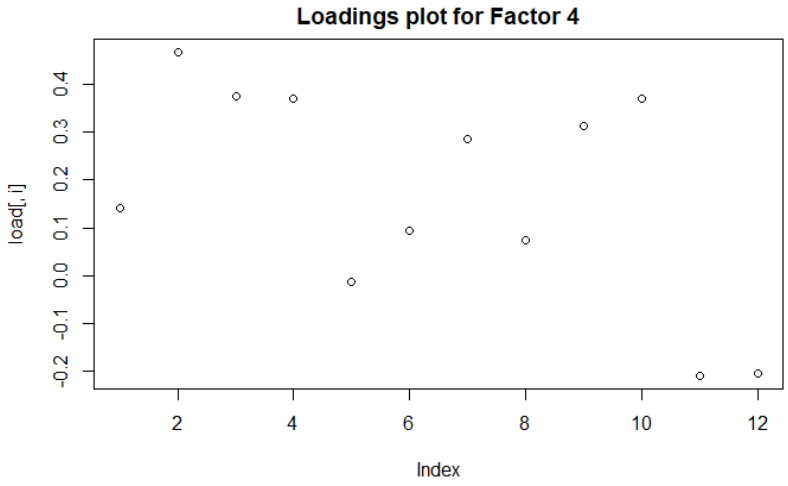
The Nscore and Escore are highly influenced or driven by the factor 1. We may name factor 1 the "Nscore-Escore Factor".The variables Country and Oscore are highly influenced by Factor 2. We may as well call it the "Country-Oscore factor". The factor 3 is common to Impulsive and SS scores. It can be called the "Risk Factor" according to my understanding of the meanings of the variables. Factor 4 seem to influence all variables evenly but especially Gender, Education and Country.

Below are Loadings Plots


Loadings plot for Factor 1
load[, i]
0.5
0.0
-0.5
2
4
6
8
10
12
Index


Loadings plot for Factor 2
load[, i]
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
2
4
6
8
10
12
Index


Loadings plot for Factor 3
load[, i]
0.6
0.4
0.2
0.0
-0.2
2
4
6
8
10
12
Index



5. CLUSTER ANALYSIS.

I used both Hierarchical Clustering and the K-medoids Method to cluster the data. It is observed that there is a high level of misclassification using the hierarchical clustering with complete linkage.

Results for Hierarchical clustering using information from users of cannabis and heroin are as follow respectively:

```
Non-User     User
    1673      212
          cl
                1    2
  Non-User    607 1066
  User        122   90
```

```
Non-User     User
     620     1265
          cl
                1   2
  Non-User    157 463
  User        572 693
```

For the K-Medoids algorithm for two clusters, there were more misclassification errors for some drugs but performed worse with other drugs. It was a better clustering rule for the data as shown below in table 6 and 11:

```
     1   2
1   19 952
2   16 898
```

```
     1   2
1 888  83
2 647 267
```

In fact both methods performed poorly on the data. The clustering algorithm gave a fairly good fit for a few drugs like the legalh, Ecstacy and Amphetamine. For these drugs there were fewer

misclassifications relative to the other drugs. Misclassification errors were very high in Alcohol, chocolate, heroine and crack.

6. CLASSIFICATION.

I used the two-class random forest algorithm to classify respondents into users of cannabis and non-users of cannabis. I also classified the data into consumers of alcohol and non-consumers. There are a few misclassifications. 47 non-users have been wrongly classified as Cannabis Users and 21 users of cannabis have been classified as Non-users.

```
Non-User     User
     148      237
          prediction
           Non-User User
  Non-User      101   47
  User           21  216
```

All 11 non-users in the test data have been incorrectly classified as users of alcohol. All users of alcohol are correctly classified as users of alcohol.

```
Non-User     User
      11      374
          prediction
           Non-User User
  Non-User        0   11
  User            0  374
```

OBSTACLES AND SUGGESTED SOLUTIONS:

I could not use some methods such as the canonical correlation analysis on my data mainly because all the variables were categorical even though they had been converted to numerical data. Also, I realized that the PCA has reduced the dimensionality of the data too drastically since the data originally does not have strong correlations between variables. I believe that to solve these problems, more quantitative data could be collected in a way as to achieve the same aim of measuring psychological and personality variables. Also, variables that may be highly correlated should be included in the data for the PCA to give a better performance.

FUTURE WORK

Future studies could explore even more the relationships between the variables. Also, there could be post hoc test of means carried out after the MANOVA tests. Pairwise equality of means tests could be done on the variables to know whether the differences in means are positive or negative depending on the outcome of the MANOVA Tests. Future studies could also construct confidence intervals for the mean vectors.

APPENDIX

---

title: "ST 557 Final Project"

author: "Emmanuel Asante (933274210)"

date: "November 28, 2017"

output: pdf_document

```{r,include=F,echo=F,warning=F,message=F}

library(mvtnorm)

library(cluster)

library(ICSNP)

library(plotrix)

library(graphics)

library(tidyverse)

library(purrr)

library(dplyr)

library(mclust)

library(e1071)

library(randomForest)

```

```{r, include=FALSE}

setwd("C:/Users/Nana Kwame/Desktop/Multivariate Final Project")

drug <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-databases/00373/drug_consumption.data",header = F)

```

```{r}

colnames(drug) <-
c("ID","Age","Gender","Education","Country","Ethnicity","Nscore","Escore","Oscore","Ascore","Csc
ore","Impulsive","SS","Alcohol","Amphet","Amyl","Benzos","Caff","Cannabis","Choc","Coke","Crack
","Ecstasy","Heroine","Ketamine","Legalh","LSD","Meth","Mushrooms","Nicotine","Semer","VSA")

```

#1

##Hotellings For Cannabis

Hotelling's T2 test for the means of NEO-FFI-R scores of Users of cannabis and Non-users of Cannabis.

$H_0:\mu_{Non-users}=\mu_{Users}$ vs $H_a:\mu_{Non-users}\neq\mu_{Users}$

```{r}

group1 <- drug[drug$Cannabis=="CL0"|drug$Cannabis=="CL1",c(7:11)]

group2 <-
drug[drug$Cannabis=="CL2"|drug$Cannabis=="CL3"|drug$Cannabis=="CL4"|drug$Cannabis=="CL5"
|drug$Cannabis=="CL6",c(7:11)]

HotellingsT2(group1, group2)

```

##Hotellings For Heroine

Hotelling's T2 test for the means of NEO-FFI-R scores of Users of heroine and Non-users of heroine.

$H_0:\mu_{Non-users}=\mu_{Users}$ vs $H_a:\mu_{Non-users}\neq\mu_{Users}$

```{r}
group3 <- drug[drug$Heroine=="CL0"|drug$Heroine=="CL1",c(7:11)]

group4 <-
drug[drug$Heroine=="CL2"|drug$Heroine=="CL3"|drug$Heroine=="CL4"|drug$Heroine=="CL5"|dr
ug$Heroine=="CL6",c(7:11)]

HotellingsT2(group3, group4)
```

##MANOVA of NEO-FFI-R scores for Users of Cannabis of different Education Levels

H0: Means are the same for all Education levels

HA: At least one mean differ from the others.

```{r}
data_manova <-
drug[drug$Cannabis=="CL2"|drug$Cannabis=="CL3"|drug$Cannabis=="CL4"|drug$Cannabis=="CL5"
|drug$Cannabis=="CL6",c(7:11,4)]

mn <- manova(as.matrix(data_manova[,1:5])~Education,data=data_manova)

summary(mn)
```

##MANOVA of NEO-FFI-R scores for Users of Heroine of different Education Levels

H0: Means are the same for all Education levels

HA: At least one mean differ from the others.

```{r}
```

```
dat_manova <-
drug[drug$Heroine=="CL2"|drug$Heroine=="CL3"|drug$Heroine=="CL4"|drug$Heroine=="CL5"|dr
ug$Heroine=="CL6",c(7:11,4)]

mn1 <- manova(as.matrix(dat_manova[,1:5])~Education,data=dat_manova)

summary(mn1)
```

#2

##MANOVA of Psychological Variables among Age groups for Cannabis Users

H0: Means are the same for all Age groups

HA: At least one mean differ from the others.

```{r}
data_man <-
drug[drug$Cannabis=="CL2"|drug$Cannabis=="CL3"|drug$Cannabis=="CL4"|drug$Cannabis=="CL5"
|drug$Cannabis=="CL6",c(7:13,2)]

mn2 <- manova(as.matrix(data_man[,1:7])~Age,data=data_man)

summary(mn2)
```

##MANOVA of Psychological scores for Users of Heroine of different Age groups

.

```{r}
dat_m <-
drug[drug$Heroine=="CL2"|drug$Heroine=="CL3"|drug$Heroine=="CL4"|drug$Heroine=="CL5"|dr
ug$Heroine=="CL6",c(7:13,2)]

mn2 <- manova(as.matrix(dat_m[,1:7])~Age,data=dat_m)

summary(mn2)
```

```

#3

##PCA

```{r}
drug_pca <- drug[,2:13]

cor(drug_pca)

pca1 <- prcomp(drug_pca, scale. = T, retx=T)
plot(pca1)
#biplot(pca1)
summary(pca1)
plot(pca1$sdev,main = "scree plot")
plot(cumsum(pca1$sdev)/sum(pca1$sdev),main = "Cumulative Scree Plot")

pca1$rotation
plot(pca1$rotation[,1],main = "Loadings Plot of PC 1")
plot(pca1$rotation[,2],main = "Loadings Plot of PC 2")
plot(pca1$rotation[,3],main = "Loadings Plot of PC 3")
# biplot(pca1)
# abline(h=0)
```

#4

##Factor Analysis
```

```{r,echo=FALSE,message=FALSE,warning=FALSE}

library(psych)

x <- scale(drug_pca)

cor.x <- cor(x[,2:9])

p <- dim(x)[2]-8

for (i in 1:p){

  p.vals <- fa(cor.x, nfactors=i, rotate="varimax", scores="regression",n.obs = dim(x)[1], fm="ml")$PVAL

  print(p.vals)

  }

fa.ml <- factanal(x,factors=4, rotation="varimax",fm="ml")

fa.ml

load <- fa.ml$loadings[,1:4]

for(i in 1:4){ plot(load[,i],main = paste("Loadings plot for Factor",i))}

#plot(load,type="n")

#text(load,labels=names(x),cex=.7)

```

#6.

##Cluster Analysis

```{r,include=FALSE}
```

```
regroup<-mutate(drug[,2:32], Alcohol2=ifelse(Alcohol=="CL0"|Alcohol=="CL1",1,2),
                Amphet2=ifelse(Amphet=="CL0"|Amphet=="CL1",1,2),
                Amyl2=ifelse(Amyl=="CL0"|Amyl=="CL1",1,2),
                Benzos2=ifelse(Benzos=="CL0"|Benzos=="CL1",1,2),
                Caff2=ifelse(Caff=="CL0"|Caff=="CL1",1,2),
                Cannabis2=ifelse(Cannabis=="CL0"|Cannabis=="CL1",1,2),
                Choc2=ifelse(Choc=="CL0"|Choc=="CL1",1,2),
                Coke2=ifelse(Coke=="CL0"|Coke=="CL1",1,2),
                Crack2=ifelse(Crack=="CL0"|Crack=="CL1",1,2),
                Ecstasy2=ifelse(Ecstasy=="CL0"|Ecstasy=="CL1",1,2),
                Heroine2=ifelse(Heroine=="CL0"|Heroine=="CL1",1,2),
                Ketamine2=ifelse(Ketamine=="CL0"|Ketamine=="CL1",1,2),
                Legalh2=ifelse(Legalh=="CL0"|Legalh=="CL1",1,2),
                LSD2=ifelse(LSD=="CL0"|LSD=="CL1",1,2),
                Meth2=ifelse(Meth=="CL0"|Meth=="CL1",1,2),
                Mushrooms2=ifelse(Mushrooms=="CL0"|Mushrooms=="CL1",1,2),
                Nicotine2=ifelse(Nicotine=="CL0"|Nicotine=="CL1",1,2),
                Semer2=ifelse(Semer=="CL0"|Semer=="CL1",1,2),
                VSA2=ifelse(VSA=="CL0"|VSA=="CL1",1,2))
```

##K-medoids Clustering

```{r}
x <- scale(drug[,2:13])
#str(drug)
```

```
n <- dim(x)[1]
w <- as.matrix(dist(x))  # Generating the distance matrix



## Cluster silhouette
avg.sil <- rep(0, 10)
for(m in 2:9) {
    cl <- pam(x, k=m)  # k-medoids clustering
    sil <- silhouette(cl$clustering, dist(x))
    avg.sil[m] <- mean(sil[,3])          ##Finding average silhouette for the clustering
}
avg.sil

##perform Partitioning Around Medoids
cl<-pam(x,k=2)
cl.table<-table(cl$clustering)
# cl.table<-t(as.matrix(cl.table))
# colnames(cl.table)<-c("cluster1","cluster2")
cl.table

table(cl$clustering,regroup$Alcohol2)

table(cl$clustering,regroup$Amphet2)

table(cl$clustering,regroup$Amyl2)

table(cl$clustering,regroup$Benzos2)
```

```
table(cl$clustering,regroup$Cannabis2)

table(cl$clustering,regroup$Choc2)

table(cl$clustering,regroup$Coke2)

table(cl$clustering,regroup$Crack2)

table(cl$clustering,regroup$Ecstasy2)

table(cl$clustering,regroup$Heroine2)

table(cl$clustering,regroup$Ketamine2)

table(cl$clustering,regroup$Legalh2)  ##best clustering fit

table(cl$clustering,regroup$Meth2)

table(cl$clustering,regroup$Mushrooms2)

table(cl$clustering,regroup$Nicotine2)

table(cl$clustering,regroup$Semer2)

table(cl$clustering,regroup$VSA2)
```

Heirarchical Clustering for Cannabis

```{r}

cannabis_data <- drug[,c(2:13,19)]

cannabis_data[,14] <- NA

cannabis_data$V14[which(cannabis_data$Cannabis=="CL0"|cannabis_data$Cannabis=="CL1")]<-"Non-User"

cannabis_data$V14[which(cannabis_data$Cannabis=="CL2"|cannabis_data$Cannabis=="CL3"|cannabis_data$Cannabis=="CL4"|cannabis_data$Cannabis=="CL5"|cannabis_data$Cannabis=="CL6"|cannabis_data$Cannabis=="CL7")]<-"User"

cannabis_data <- cannabis_data[,-13]

names(cannabis_data)[names(cannabis_data) == "V14"] <- "Cannabis"

cannabis_data$Cannabis <- as.factor(cannabis_data$Cannabis)


x <- cannabis_data[,1:12]

hc <- hclust(dist(x), method="complete")

cl <- cutree(hc, k=2)

table(cannabis_data$Cannabis)

table(cannabis_data$Cannabis,cl)
```


Heirarchical Clustering for Heroine


```{r}

heroine <- drug[,c(2:13,24)]

heroine[,14] <- NA

heroine$V14[which(heroine$Heroine=="CL0"|heroine$Heroine=="CL1")]<-"Non-User"

heroine$V14[which(heroine$Heroine=="CL2"|heroine$Heroine=="CL3"|heroine$Heroine=="CL4"|heroine$Heroine=="CL5"|heroine$Heroine=="CL6"|heroine$Heroine=="CL7")]<-"User"

heroine <- heroine[,-13]
```

```
names(heroine)[names(heroine) == "V14"] <- "Heroine"

heroine$Heroine <- as.factor(heroine$Heroine)


x <- heroine[,1:12]

hc <- hclust(dist(x), method="complete")

cl <- cutree(hc, k=2)

table(heroine$Heroine)

table(heroine$Heroine,cl)
```


#7

##Classification

###Classification for Cannabis


```{r ,message=FALSE,warning=FALSE}

cannabis_data <- drug[,c(2:13,19)]

cannabis_data[,14] <- NA

cannabis_data$V14[which(cannabis_data$Cannabis=="CL0"|cannabis_data$Cannabis=="CL1")]<-
"Non-User"

cannabis_data$V14[which(cannabis_data$Cannabis=="CL2"|cannabis_data$Cannabis=="CL3"|cann
abis_data$Cannabis=="CL4"|cannabis_data$Cannabis=="CL5"|cannabis_data$Cannabis=="CL6"|can
nabis_data$Cannabis=="CL7")]<-"User"

cannabis_data <- cannabis_data[,-13]

names(cannabis_data)[names(cannabis_data) == "V14"] <- "Cannabis"

cannabis_data$Cannabis <- as.factor(cannabis_data$Cannabis)


train_cannabis <- cannabis_data[1:1500,]

test_cannabis <- cannabis_data[1501:1885,]
```

```
##random forest

set.seed(9999)

mod.cannabis<-randomForest(Cannabis ~., data = train_cannabis, ntree = 1000,importance=T)


prediction<-predict(mod.cannabis,test_cannabis)


table(test_cannabis$Cannabis)

table(test_cannabis$Cannabis,prediction)
```


###Classification for Alcohol


```{r ,message=FALSE,warning=FALSE}

alcohol_data <- drug[,c(2:13,14)]

alcohol_data[,14] <- NA

alcohol_data$V14[which(alcohol_data$Alcohol=="CL0"|alcohol_data$Alcohol=="CL1")]<-"Non-User"

alcohol_data$V14[which(alcohol_data$Alcohol=="CL2"|alcohol_data$Alcohol=="CL3"|alcohol_data$Alcohol=="CL4"|alcohol_data$Alcohol=="CL5"|alcohol_data$Alcohol=="CL6"|alcohol_data$Alcohol=="CL7")]<-"User"

alcohol_data <- alcohol_data[,-13]

names(alcohol_data)[names(alcohol_data) == "V14"] <- "Alcohol"

alcohol_data$Alcohol <- as.factor(alcohol_data$Alcohol)


train_alcohol <- alcohol_data[1:1500,]

test_alcohol <- alcohol_data[1501:1885,]


##random forest

set.seed(9999)

mod.alcohol<-randomForest(Alcohol ~., data = train_alcohol, ntree = 1000,importance=T)
```

```
prediction<-predict(mod.alcohol,test_alcohol)


table(test_alcohol$Alcohol)

table(test_alcohol$Alcohol,prediction)
```