Homework3

Julissa Duenas

2/24/2021

```
drug_use <- read_csv('drug.csv',
col_names = c('ID','Age','Gender','Education','Country','Ethnicity','Nscore',
              'Escore','Oscore','Ascore','Cscore','Impulsive','SS','Alcohol',
              'Amphet','Amyl','Benzos','Caff','Cannabis','Choc','Coke','Crack',
              'Ecstasy','Heroin','Ketamine','Legalh','LSD','Meth','Mushrooms',
              'Nicotine','Semer','VSA'))
```

Question 1

```
drug_use <- drug_use %>% mutate_at(as.ordered, .vars=vars(Alcohol:VSA))
drug_use <- drug_use %>%mutate(Gender = factor(Gender,
                                               labels=c("Male", "Female"))) %>%
  mutate(Ethnicity = factor(Ethnicity, labels=c("Black", "Asian", "White","Mixed:White/Black",
                                                "Other","Mixed:White/Asian","Mixed:Black/Asian")))%>%
  mutate(Country = factor(Country,labels=c("Australia", "Canada", "New Zealand","Other", "Ireland",
                                           "UK", "USA")))
```

(a)

```
drug_use <- drug_use%>%
  mutate(recent_cannabis_use=factor(ifelse(Cannabis>='CL3','Yes','No'),levels=c('No','Yes')))
class(drug_use$recent_cannabis_use)
```

```
## [1] "factor"
```

(b)

```
drug_use_subset <- drug_use %>% select(Age:SS, recent_cannabis_use)

train <- sample(1:nrow(drug_use_subset),1500)
drug_use_train <- drug_use_subset[train,]
drug_use_test <- drug_use_subset[-train,]

dim(drug_use_train)
```

```
## [1] 1500   13
```

```
dim(drug_use_test)
```

```
## [1] 385  13
```

(c)

```
rec.can.use.glm <- glm(recent_cannabis_use~.,data = drug_use_train,family = binomial)
summary(rec.can.use.glm)
```

```
##
```

```
## Call:
## glm(formula = recent_cannabis_use ~ ., family = binomial, data = drug_use_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8146  -0.5863   0.1350   0.5325   2.6949
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   1.11683    0.65361   1.709 0.087506 .
## Age                          -0.89226    0.09259  -9.636  < 2e-16 ***
## GenderFemale                 -0.72197    0.15775  -4.577 4.73e-06 ***
## Education                    -0.30538    0.07910  -3.861 0.000113 ***
## CountryCanada                13.13506  716.64358   0.018 0.985377
## CountryNew Zealand           -1.11240    0.32327  -3.441 0.000579 ***
## CountryOther                  0.21491    0.47243   0.455 0.649174
## CountryIreland               -0.08155    0.71929  -0.113 0.909731
## CountryUK                    -0.42960    0.37287  -1.152 0.249262
## CountryUSA                   -1.79660    0.19357  -9.281  < 2e-16 ***
## EthnicityAsian               -1.14782    0.96194  -1.193 0.232776
## EthnicityWhite                0.68445    0.64465   1.062 0.288353
## EthnicityMixed:White/Black    0.54179    1.02736   0.527 0.597944
## EthnicityOther                0.82331    0.76480   1.077 0.281698
## EthnicityMixed:White/Asian    0.64953    0.99709   0.651 0.514768
## EthnicityMixed:Black/Asian   14.26004  760.71134   0.019 0.985044
## Nscore                       -0.07431    0.09035  -0.822 0.410827
## Escore                       -0.17569    0.09761  -1.800 0.071889 .
## Oscore                        0.70846    0.09126   7.763 8.30e-15 ***
## Ascore                        0.06479    0.08080   0.802 0.422624
## Cscore                       -0.34271    0.08935  -3.836 0.000125 ***
## Impulsive                    -0.11983    0.10019  -1.196 0.231683
## SS                            0.61117    0.11105   5.504 3.72e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2076.5  on 1499  degrees of freedom
## Residual deviance: 1184.5  on 1477  degrees of freedom
## AIC: 1230.5
##
## Number of Fisher Scoring iterations: 14
```

Question 2

```
tree_parameters = tree.control(nobs=nrow(drug_use_train), minsize=10, mindev=1e-3)
```

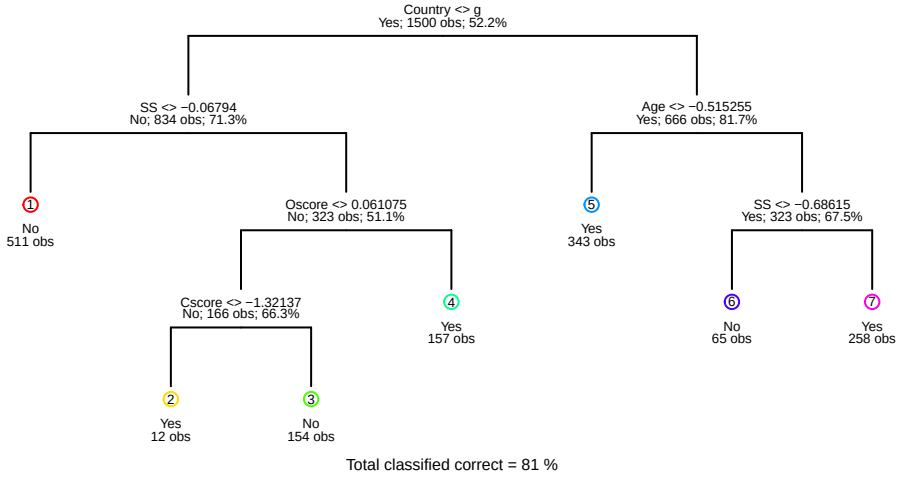
(a)

```
set.seed(2)
tree.drug_use <- tree(recent_cannabis_use~.,data=drug_use_train,control=tree_parameters)
cv=cv.tree(tree.drug_use,FUN=prune.misclass,K=10)
best.cv=cv$size[max(which(cv$dev==min(cv$dev)))]
best.cv
```

```
## [1] 7
```

(b)

```
prune.drug_use <- prune.misclass(tree.drug_use,best=best.cv)
draw.tree(prune.drug_use,nodeinfo = TRUE,cex=.4)
```



The first variable that is split is 'Country'

(c)

```
set.seed(2)
drug.pred <- predict(prune.drug_use,drug_use_test,type='class')
error <- table(drug.pred,drug_use_test$recent_cannabis_use)
error
```

```
##
## drug.pred  No Yes
##       No  133  50
##       Yes  36 166
```

```
TPR=error[2,2]/(error[2,2]+error[1,2])
FPR=error[2,1]/(error[2,1]+error[1,1])
TPR
```

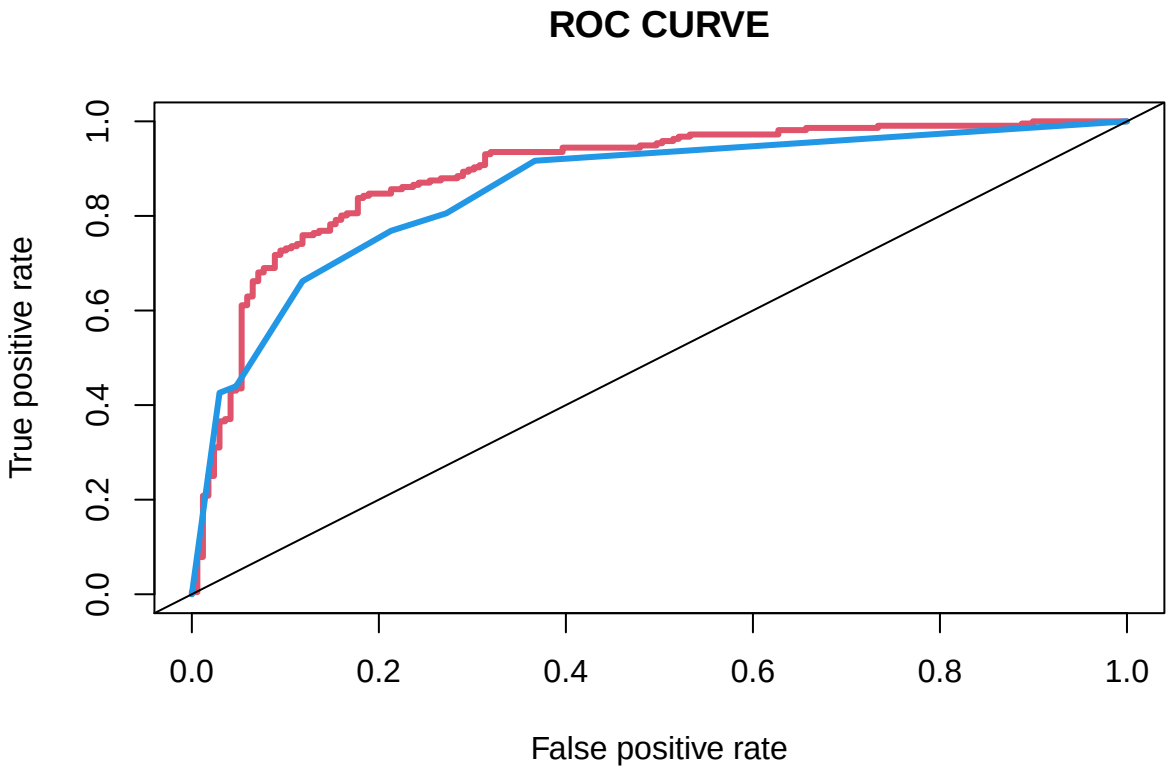
```
## [1] 0.7685185
```

```
FPR
```

```
## [1] 0.2130178
```

Question 3 (a)

```
prob.glm <- predict(rec.can.use.glm,drug_use_test,type = 'response')
prob.tree <- predict(prune.drug_use,drug_use_test,type='vector')
pred.glm <- prediction(prob.glm,drug_use_test$recent_cannabis_use)
pred.tree <- prediction(prob.tree[,2],drug_use_test$recent_cannabis_use)
roc.glm <- performance(pred.glm,measure = 'tpr',x.measure = 'fpr')
roc.tree <- performance(pred.tree,measure = 'tpr',x.measure = 'fpr')
plot(roc.glm,col=2,lwd=3,main='ROC CURVE')
plot(roc.tree,col=4,lwd=3,main='ROC CURVE',add=TRUE)
abline(0,1)
```



(b)

```
auc.glm <- performance(pred.glm,'auc')@y.values
auc.tree <- performance(pred.tree,'auc')@y.values
auc.glm
```

```
## [[1]]
## [1] 0.8924227
```

```
auc.tree
```

```
## [[1]]
## [1] 0.8556734
```

The logistic regression model has a slightly larger AUC

Question 4

```
leukemia_data <- read_csv("leukemia_data.csv")
```

(a)

```
leukemia_data <- leukemia_data%>%mutate(Type=factor(Type))
table(leukemia_data$Type)
```

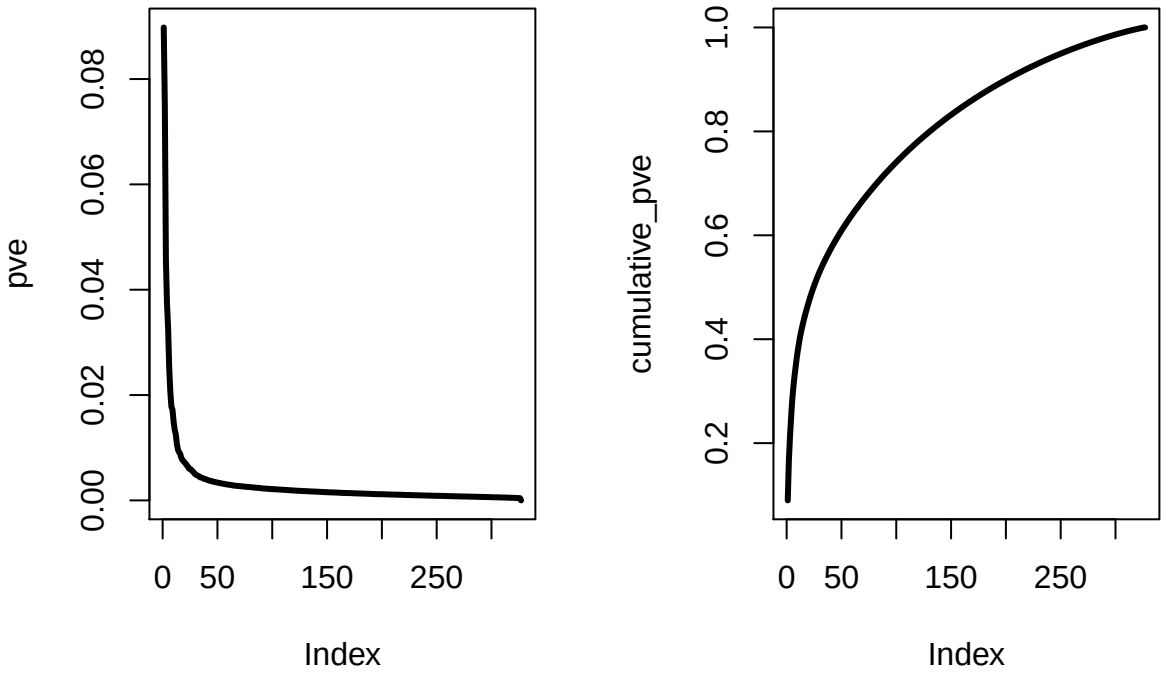
```
##
##    BCR-ABL   E2A-PBX1 Hyperdip50        MLL     OTHERS      T-ALL   TEL-AML1
##         15         27         64         20         79         43         79
```

BCR-ABL occurs the least in this data

(b)

```
pr.leuk <- prcomp(leukemia_data[,-1],scale=TRUE,center=TRUE)
pr.var <- pr.leuk$sdev^2
```

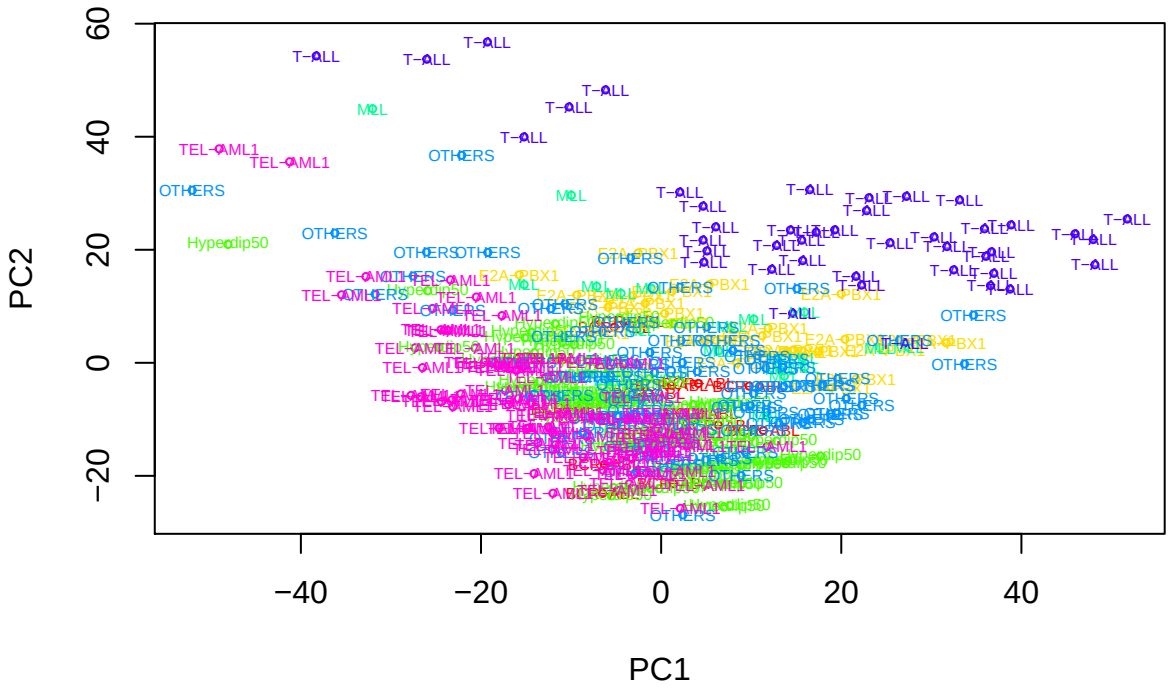
```
pve <- pr.var/sum(pr.var)
cumulative_pve <- cumsum(pve)
## This will put the next two plots side by side
par(mfrow=c(1, 2))
## Plot proportion of variance explained
plot(pve, type="l", lwd=3)
plot(cumulative_pve, type="l", lwd=3)
```



(c)

```
rainbow_colors <- rainbow(7)
plot_colors <- rainbow_colors[leukemia_data$Type]
```

```
plot(pr.leuk$x[,1:2],col=plot_colors,cex=.5)
text(pr.leuk$x[,1:2],labels = leukemia_data$Type,cex=.5,col=plot_colors)
```




```
head(pr.leuk$x[,1:2])
```

```
##              PC1        PC2
## [1,] -10.414898  -8.107584
## [2,]  -1.377304  -5.386586
## [3,]  -3.720294   7.290351
## [4,]   1.159456  -3.953322
## [5,]  -5.177178   6.313023
## [6,]  11.346689 -11.979690
```

```
head(sort(abs(pr.leuk$rotation[,1]),decreasing = TRUE))
```

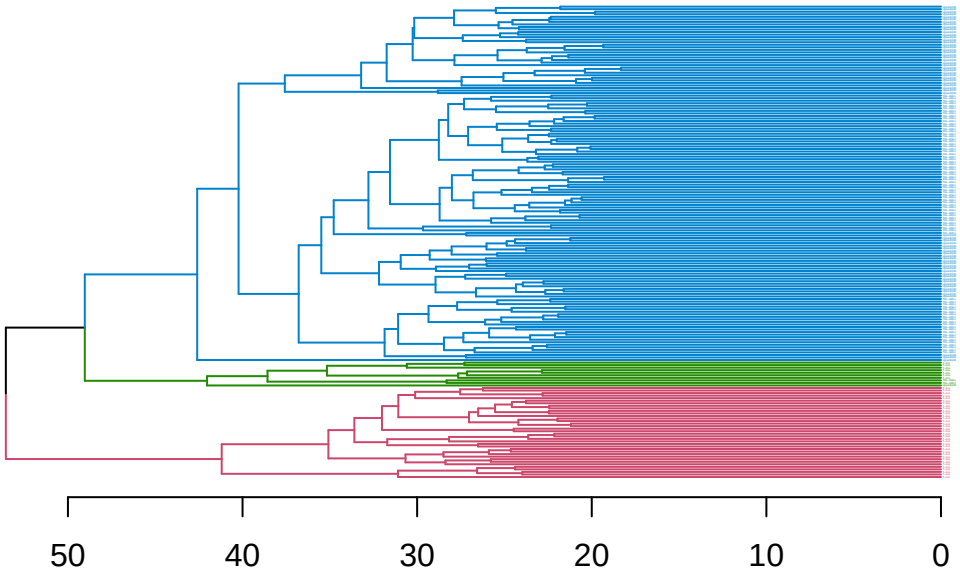
```
##     SEMA3F       CCT2       LDHB       COX6C     SNRPD2       ELK3
## 0.04517148 0.04323818 0.04231619 0.04183480 0.04179822 0.04155821
```

T-ALL looks to be the most separated. SEMA3F has the highest absolute loading value.

(f)

```
leukemia_subset <- leukemia_data%>%filter(Type=='T-ALL'|Type=='TEL-AML1'|Type=='Hyperdip50')
dis <- dist(leukemia_subset[,-1],method='euclidean')
leukemia.hc <- hclust(dis,method='complete')
dend1 <- as.dendrogram(leukemia.hc)
dend1 <- color_branches(dend1,k=3)
dend1 <- color_labels(dend1,k=3)
dend1 <- set(dend1,'labels_cex',0.1)
dend1 <- set_labels(dend1,labels=leukemia_subset$Type[order.dendrogram(dend1)])
plot(dend1,horiz=T,main='Dendogram colored by three clusters')
```

Dendogram colored by three clusters




```
dend2 <- as.dendrogram(leukemia.hc)
dend2 <- color_branches(dend2,k=5)
dend2 <- color_labels(dend2,k=5)
dend2 <- set(dend2,'labels_cex',0.1)
dend2 <- set_labels(dend2,labels=leukemia_subset$Type[order.dendrogram(dend2)])
plot(dend2,horiz=T,main='Dendogram colored by five clusters')
```

Dendogram colored by five clusters