Parkinson's disease

Mahsa Raeiati Banadkooki

8/14/2021

Test Code

This code just was written to analyize the Parkinson's disease as a sample code. It was used some visualization analysis. Also the decision tree and support vector machine were used for classifying the dataset.

The Data

https://wiki.socr.umich.edu/index.php/SOCR_Data_PD_BiomedBigMetadata (https://wiki.socr.umich.edu/index.php/SOCR_Data_PD_BiomedBigMetadata)

This case-study uses simulation to examine associations between clinical, demographic, imaging and genetics variables for Parkinson's disease. This is an example of Big Data for investigating important neurodegenerative disorders.

The dataset contains:

ID: Case subject identifier

Imaging Biomarkers: ComputeArea = surface area of 3D brain region of interest, Volume= the 3D volume/size of the region of interest, region name is encoded in morphometry measure name (e.g., L_putamen_ComputeArea represents the left putamen surface area): L_caudate_ComputeArea, L_caudate_Volume, R_caudate_ComputeArea, R_caudate_Volume, L_putamen_ComputeArea, L_putamen_Volume, R_putamen_ComputeArea, R_putamen_Volume, L_hippocampus_ComputeArea, L_hippocampus_Volume, R_hippocampus_ComputeArea, R_hippocampus_Volume, cerebellum_ComputeArea, cerebellum_Volume, L_lingual_gyrus_ComputeArea, L_lingual_gyrus_Volume, R_lingual_gyrus_ComputeArea, R_lingual_gyrus_Volume, L_fusiform_gyrus_ComputeArea, L_fusiform_gyrus_Volume, R_fusiform_gyrus_ComputeArea, R_fusiform_gyrus_Volume

Demographics variables: Sex, Weight, Age

Diagnosis: Dx, PD=Parkinson's, HC=Healthy Control, SWEDD = (tremor associated clinical parkinsonism features) scans without evidence of dopaminergic deficit

Genetics: chr12_rs34637584_GT, chr17_rs11868035_GT

Clinical: UPDRS_part_I, UPDRS_part_II, UPDRS_part_III (Movement Disorder Society-Sponsored Revision of the Unified. Parkinson's Disease Rating Scale (MDS-UPDRS)), (Normal=0)

Time: VisitTime: four time-points (baseline (0), 6, 12, and 18 month follow-ups).

```r
library(xml2)
library(rvest)
url_data <- read_html("http://wiki.socr.umich.edu/index.php/SOCR_Data_PD_BiomedBigMetadata")
data <- html_table(html_nodes(url_data, "table")[[1]])
data <- data[, -c(1,33)] # Remove the ID and Time
library(psych)
describe(data)
```

```
##                                vars     n      mean      sd   median   trimmed      mad
## L_caudate_ComputeArea             1  1128    600.39   26.42    600.0    600.63    27.43
## L_caudate_Volume                  2  1128    800.33   27.12    800.0    800.50    26.69
## R_caudate_ComputeArea             3  1128    894.54   29.95    897.0    895.57    29.65
## R_caudate_Volume                  4  1128   1001.16   32.56    998.0   1000.77    32.62
## L_putamen_ComputeArea             5  1128    898.93   31.05    897.5    898.40    31.13
## L_putamen_Volume                  6  1128   1400.40   37.78   1400.0   1400.89    38.55
## R_putamen_ComputeArea             7  1128   1299.71   32.41   1302.0   1299.68    31.88
## R_putamen_Volume                  8  1128   3000.30   58.14   3000.5   2999.92    58.56
## L_hippocampus_ComputeArea         9  1128   1302.09   37.49   1300.0   1300.78    35.58
## L_hippocampus_Volume             10  1128   3198.50   53.32   3199.5   3197.70    47.44
## R_hippocampus_ComputeArea        11  1128   1504.35   36.16   1504.5   1504.28    37.81
## R_hippocampus_Volume             12  1128   3799.29   62.61   3801.5   3798.86    54.11
## cerebellum_ComputeArea           13  1128  16699.60  126.68  16699.0  16699.51   126.02
## cerebellum_Volume                14  1128  14001.61  117.94  13995.5  14001.79   108.23
## L_lingual_gyrus_ComputeArea      15  1128   3300.20   57.70   3299.0   3299.29    51.89
## L_lingual_gyrus_Volume           16  1128  11010.39  111.11  11007.0  11008.59    99.33
## R_lingual_gyrus_ComputeArea      17  1128   3296.32   58.91   3294.0   3296.30    56.34
## R_lingual_gyrus_Volume           18  1128  12007.65  106.05  12001.0  12007.35   105.26
## L_fusiform_gyrus_ComputeArea     19  1128   3598.07   57.92   3594.5   3596.92    60.79
## L_fusiform_gyrus_Volume          20  1128  11011.18  110.89  11016.0  11013.34   104.52
## R_fusiform_gyrus_ComputeArea     21  1128   3298.63   58.05   3296.0   3298.44    53.37
## R_fusiform_gyrus_Volume          22  1128   9995.76   93.55   9993.5   9994.24    96.37
## Sex                              23  1128      0.59    0.49      1.0      0.61     0.00
## Weight                           24  1128     78.45   10.20     78.5     78.12     9.64
## Age                              25  1128     60.64   10.14     61.0     60.66    10.38
## Dx*                              26  1128      1.94    0.80      2.0      1.92     1.48
## chr12_rs34637584_GT              27  1128      0.54    0.50      1.0      0.55     0.00
## chr17_rs11868035_GT              28  1128      0.42    0.49      0.0      0.40     0.00
## UPDRS_part_I                     29  1128      0.77    0.73      1.0      0.72     1.48
## UPDRS_part_II                    30  1128      8.88    5.09      9.0      8.67     5.93
## UPDRS_part_III                   31  1128     13.02    7.80     13.0     12.74     8.90
##                                 min    max  range  skew  kurtosis    se
## L_caudate_ComputeArea           525    667    142 -0.09     -0.31  0.79
## L_caudate_Volume                719    890    171  0.01      0.22  0.81
## R_caudate_ComputeArea           795    977    182 -0.32      0.18  0.89
## R_caudate_Volume                916   1094    178  0.10     -0.30  0.97
## L_putamen_ComputeArea           815   1003    188  0.19      0.19  0.92
## L_putamen_Volume               1298   1507    209 -0.14     -0.19  1.12
```

```
## R_putamen_ComputeArea         1198  1392   194 -0.04    -0.06 0.96
## R_putamen_Volume              2846  3148   302  0.03    -0.38 1.73
## L_hippocampus_ComputeArea     1203  1422   219  0.35     0.15 1.12
## L_hippocampus_Volume          3036  3381   345  0.11     0.49 1.59
## R_hippocampus_ComputeArea     1414  1602   188  0.04    -0.31 1.08
## R_hippocampus_Volume          3634  4013   379  0.20     0.68 1.86
## cerebellum_ComputeArea       16378 17096   718  0.04    -0.16 3.77
## cerebellum_Volume            13680 14370   690  0.02     0.01 3.51
## L_lingual_gyrus_ComputeArea   3136  3469   333  0.14     0.16 1.72
## L_lingual_gyrus_Volume       10709 11488   779  0.27     0.66 3.31
## R_lingual_gyrus_ComputeArea   3135  3490   355  0.06     0.29 1.75
## R_lingual_gyrus_Volume       11679 12324   645  0.01    -0.11 3.16
## L_fusiform_gyrus_ComputeArea  3446  3763   317  0.15    -0.25 1.72
## L_fusiform_gyrus_Volume      10682 11394   712 -0.10     0.25 3.30
## R_fusiform_gyrus_ComputeArea  3094  3443   349 -0.09     0.30 1.73
## R_fusiform_gyrus_Volume       9736 10235   499  0.11    -0.16 2.79
## Sex                              0     1     1 -0.35    -1.88 0.01
## Weight                          51   109    58  0.22    -0.04 0.30
## Age                             31    87    56 -0.04    -0.35 0.30
## Dx*                              1     3     2  0.12    -1.43 0.02
## chr12_rs34637584_GT              0     1     1 -0.16    -1.98 0.01
## chr17_rs11868035_GT              0     1     1  0.33    -1.89 0.01
## UPDRS_part_I                     0     2     2  0.38    -1.07 0.02
## UPDRS_part_II                    1    20    19  0.21    -0.89 0.15
## UPDRS_part_III                   1    30    29  0.23    -0.86 0.23
```

Load the libraries

```
library(ggplot2)
library(reshape2)
library(viridis)
library(GGally)
library(rpart.plot)
library(rpart)
library(e1071)
set.seed(1234)
```

Data Visualization
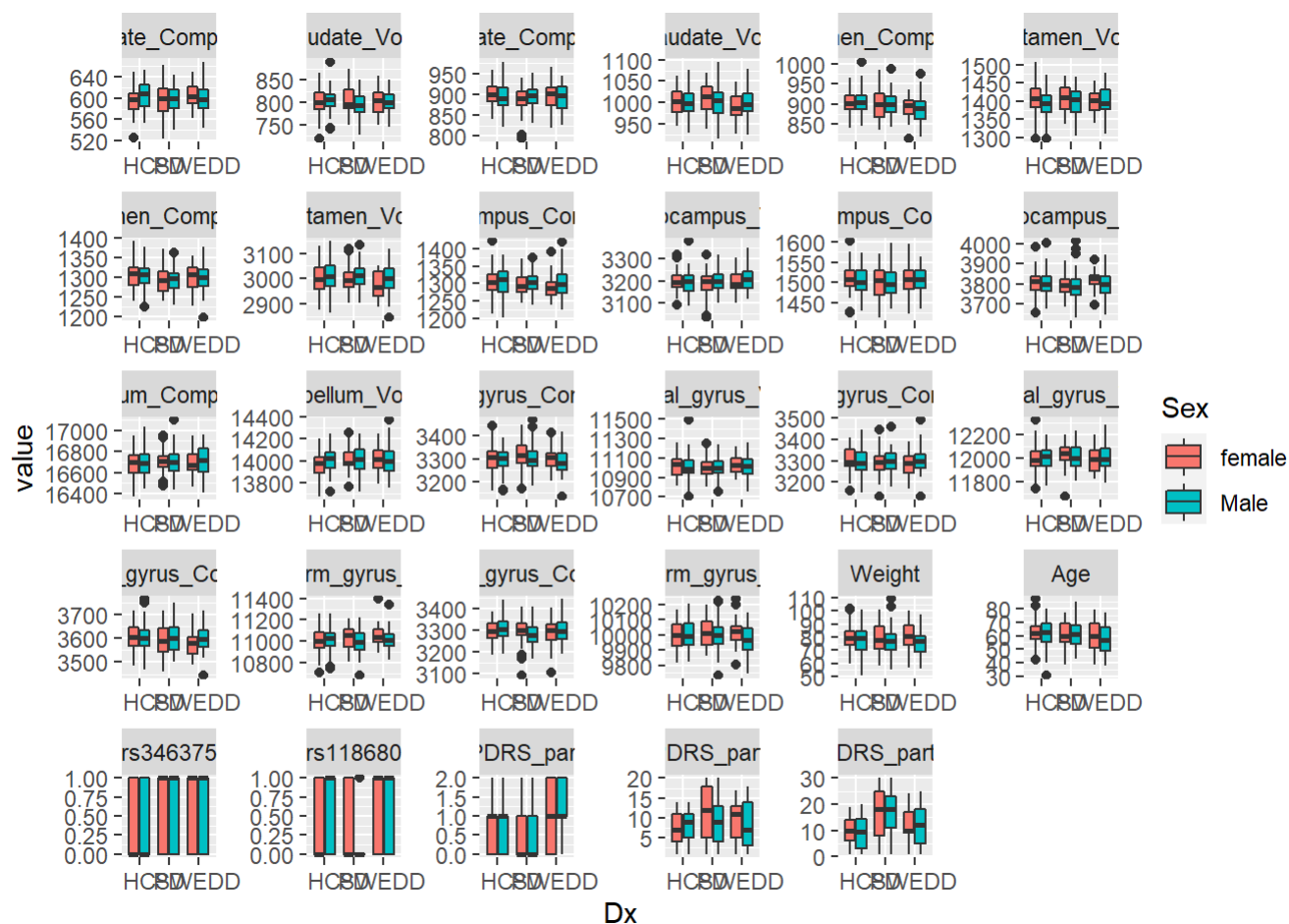
Create the sample boxplot for "weight" variable on two different grouping variables; "Age" on x-axis and "Sex" as fill color (legend variable).

```
gplot <- ggplot(data, aes( x= Age, y=Weight, fill=factor(Sex))) +
        geom_boxplot() +
        geom_jitter(width=0.1,alpha=0.2) +
        facet_wrap(~Dx) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
print(gplot)
```
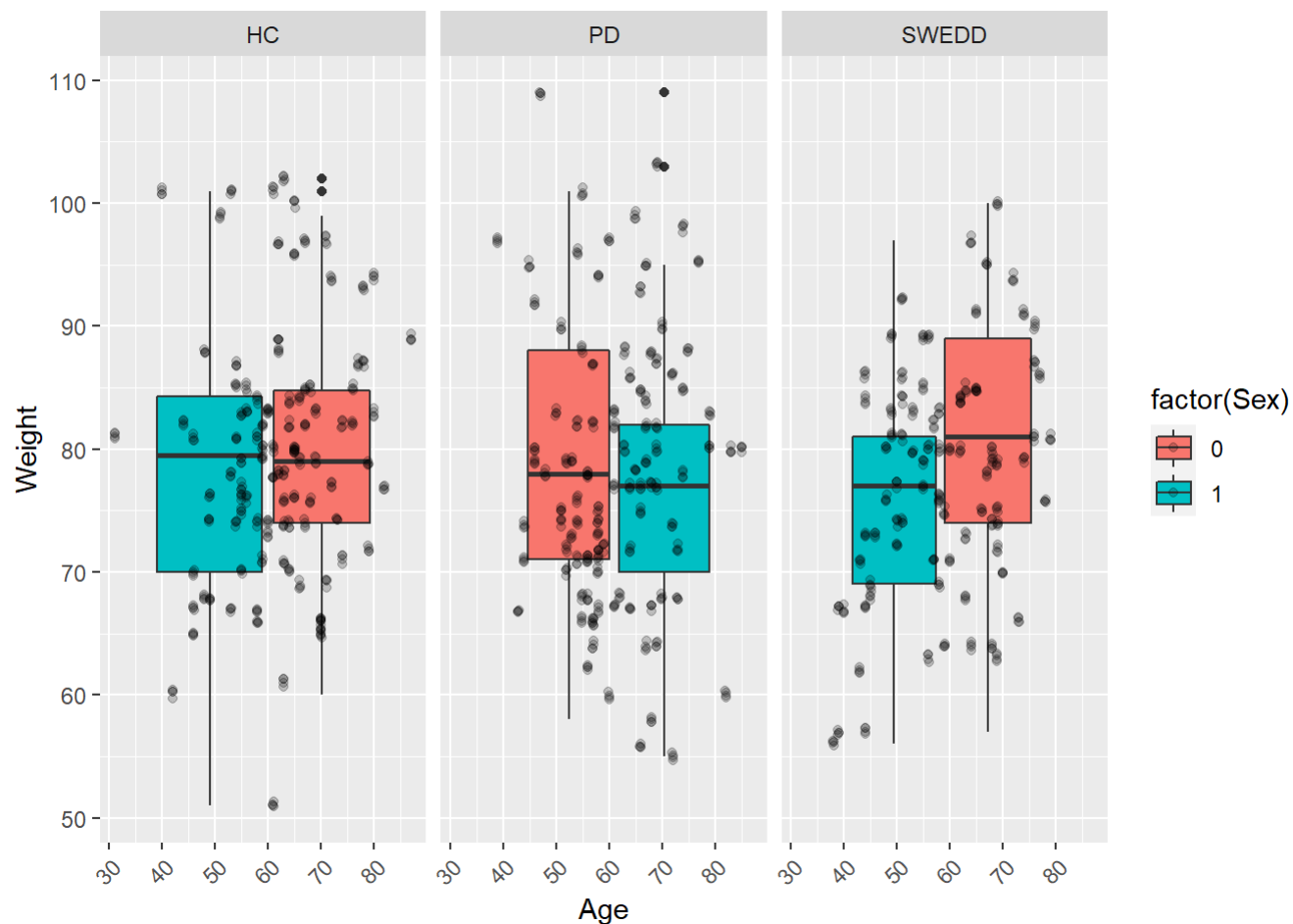

Correlation plot for dataset

It can be seen from the result that there is no relation between features. They are completely uncorrelated.

```
cormat <- round(cor(as.matrix(data[, -26])),2) # Remove the Dx, it is categorical data
melted_cor_data <- melt(cormat)
ggplot(data = melted_cor_data, aes(x=Var1, y=Var2, fill=value)) +
    geom_tile(color="white", size=0.1)+
    scale_fill_viridis(discrete=FALSE)
```
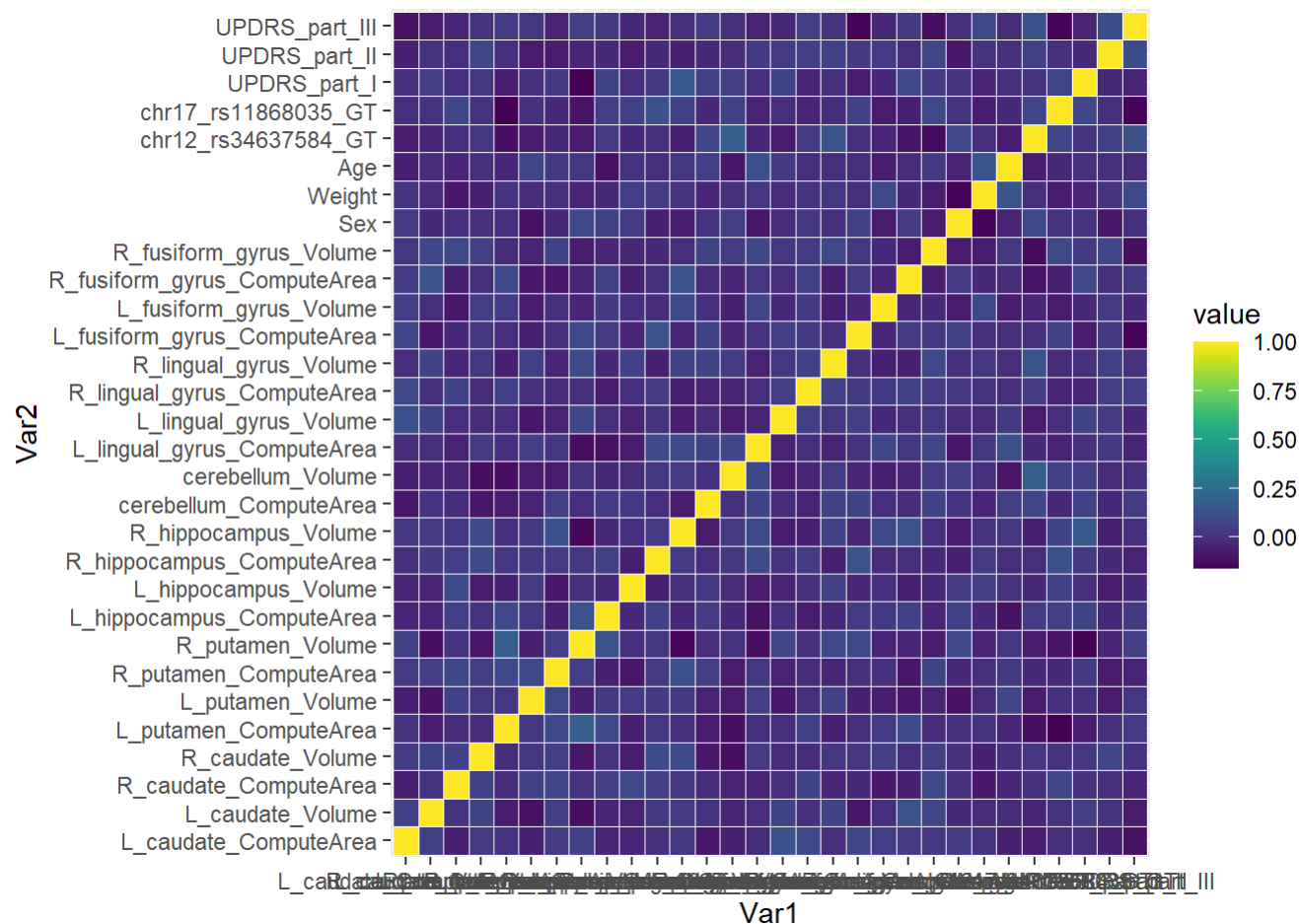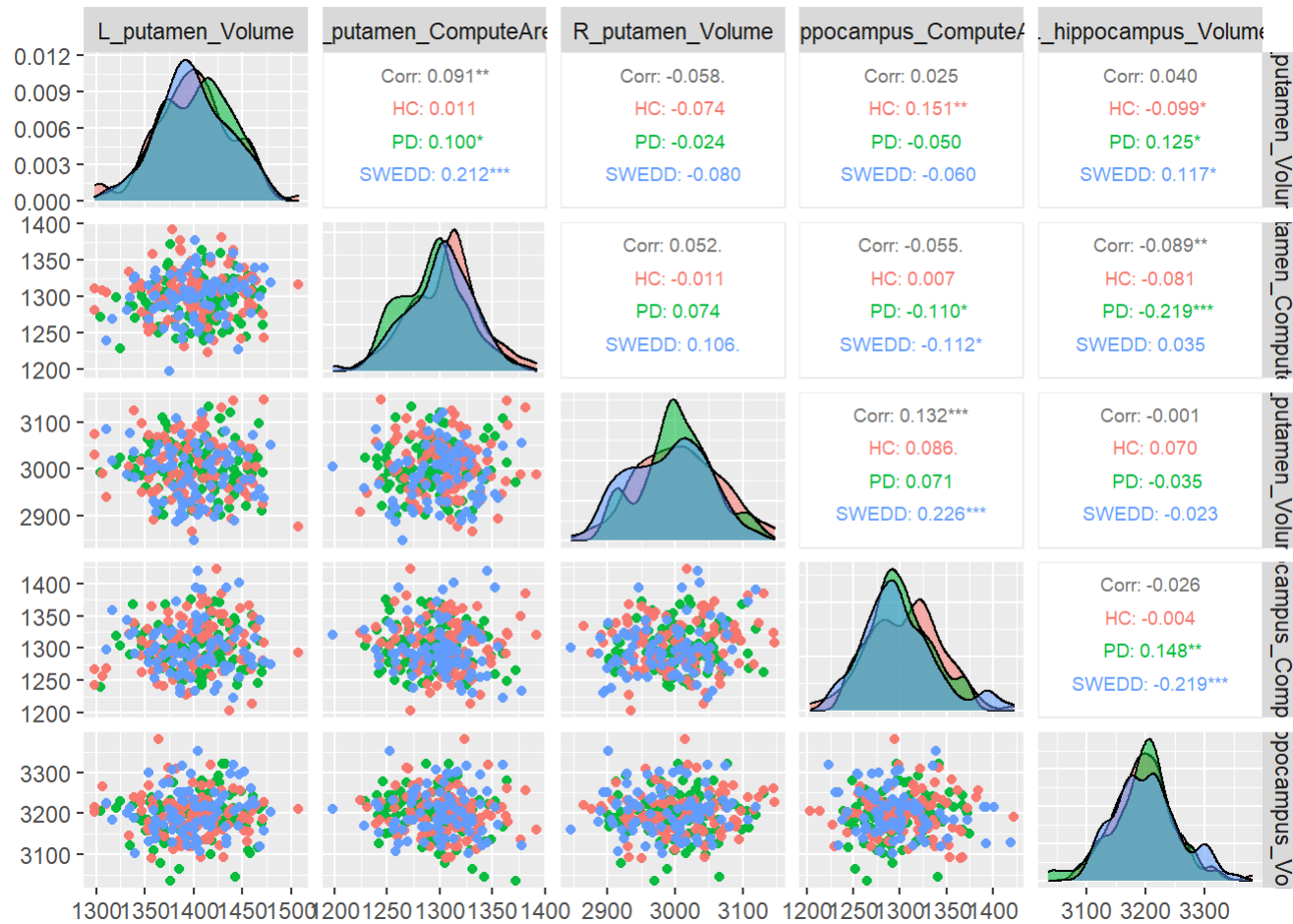
```
ggpairs(data, columns = 1:5, aes(color = Dx, alpha = 0.5),
        upper = list(continuous = wrap("cor", size = 2.5)))
```
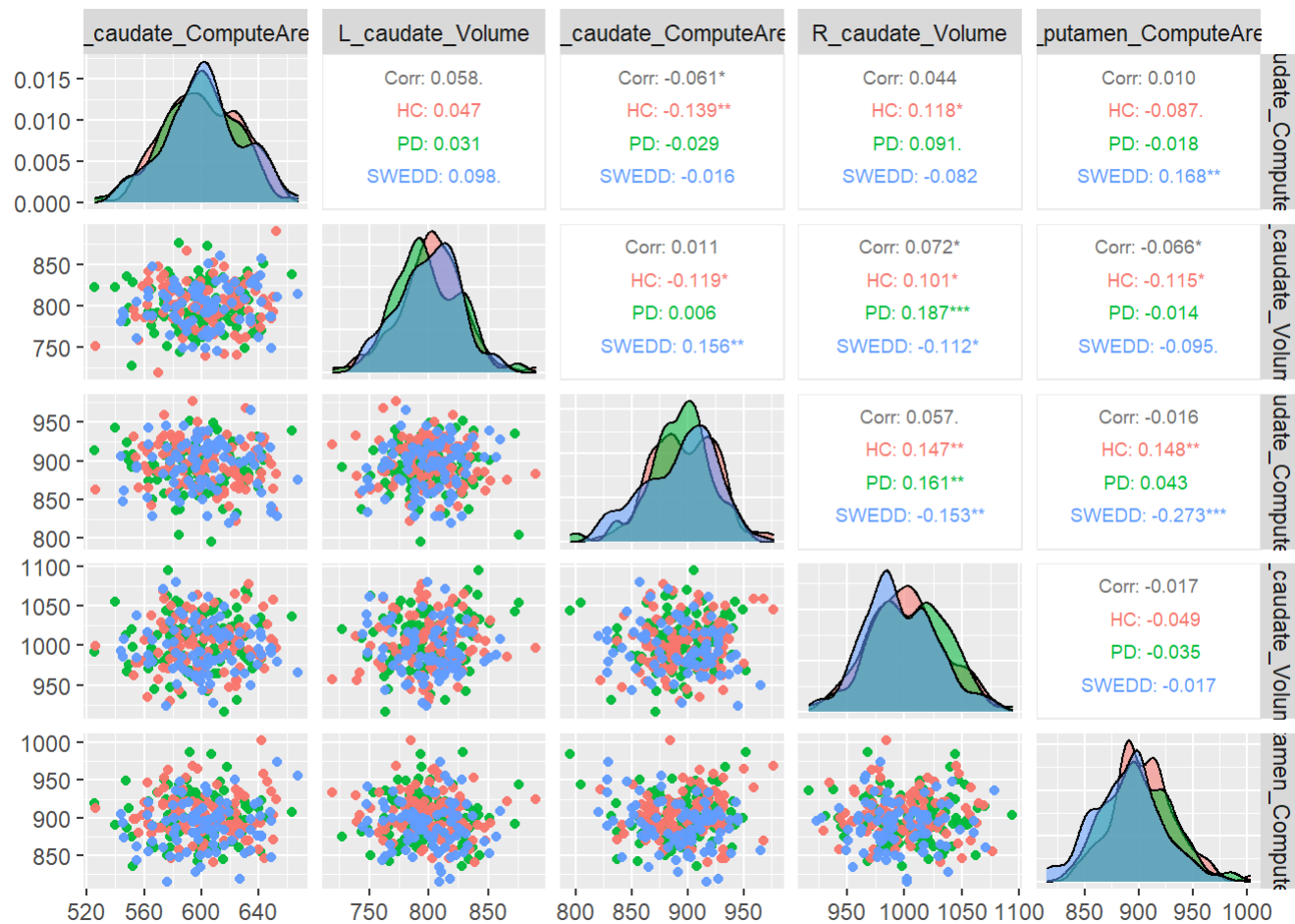


```
ggpairs(data, columns = 6:10, aes(color = Dx, alpha = 0.5),
        upper = list(continuous = wrap("cor", size = 2.5)))
```
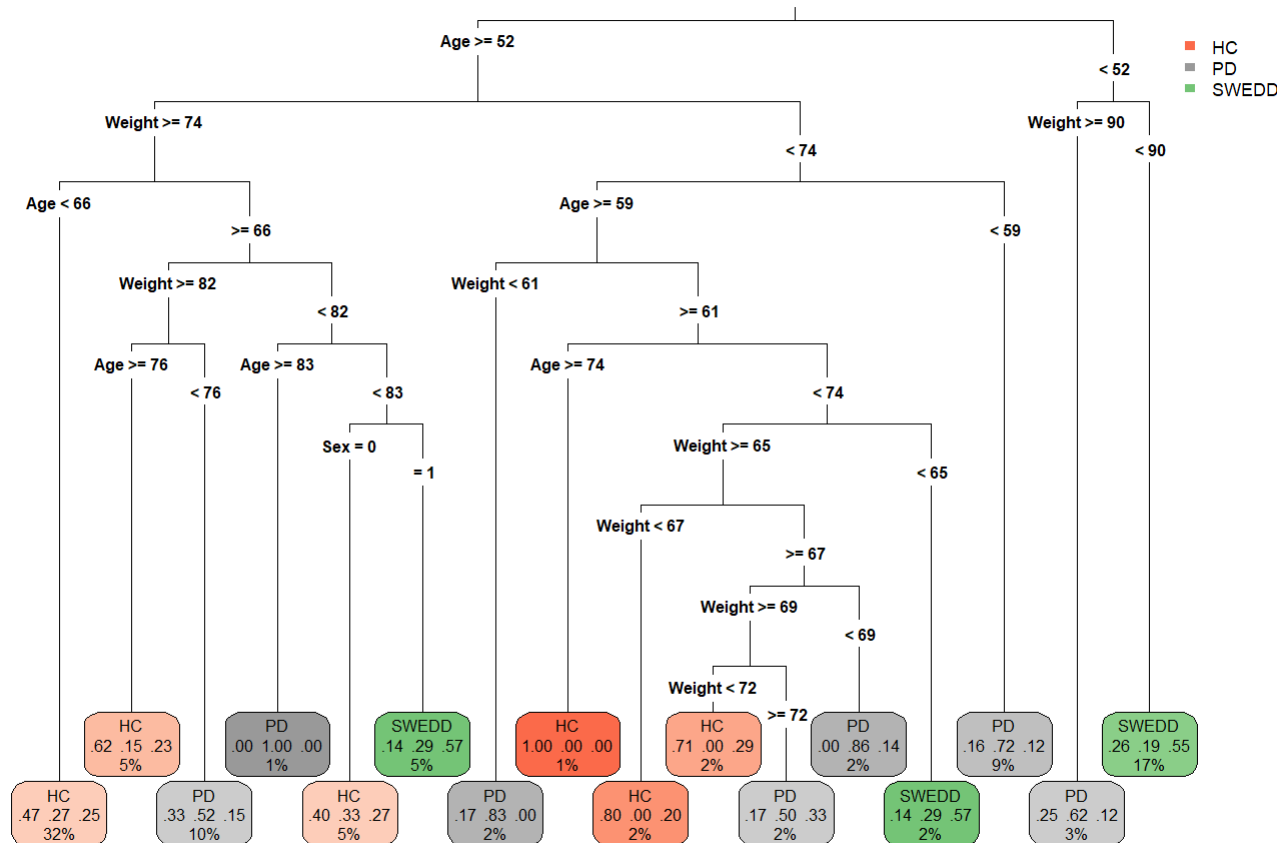
Decision Tree

```
data.tree <- rpart(Dx~Age + Sex+ Weight, data = data)

rpart.plot(data.tree, fallen.leaves = T,  digit = 2,
           type = 3,  cex=0.5, legend.x = 0.98, legend.y = 1)
```

Support Vector Machine

SVM was used for classifying the data into three different groups with 69% of accuracy.

```r
# Normalized the Data
set.seed(299)
normalize <- function(x) { return ((x - min(x)) / (max(x) - min(x)))}
data_norm <- as.data.frame(lapply(data[, -26], normalize))
data_norm <- data.frame(data_norm, "Dx" = data$Dx)

# Seprate the 70 % of data for Train data and 30% of data for Test
sub_sample <- sample(nrow(data_norm), floor(nrow(data_norm)*0.70))
data_train <- data_norm[sub_sample, ]
data_test <- data_norm[-sub_sample, ]

# Support vector Machine for classify data into three class

svm_model <- svm(Dx~., data = data_train, kernel="linear")
table(data_test$Dx,  predict(svm_model, data_test))
```

```
##
##          HC PD SWEDD
##    HC    84 18    17
##    PD     8 99    13
##    SWEDD 24 25    51
```

```r
result <- predict(svm_model, data_test)==data_test$Dx
prop.table(table(result))
```

```
## result
##      FALSE      TRUE
## 0.3097345 0.6902655
```