Project

Harsh

25/07/2020

- Importing all required libraries

```
#Importing Libraries

#install.packages("rattle")
#install.packages("DataExplorer")
#install.packages("factoextra")

library(factoextra)
library(DataExplorer)
library(caret)
library(psych)
library(ggplot2)
library(gridExtra)
library(grid)
library(GGally)
library(reshape2)
library(C50)
library(gmodels)
library(rpart)
library(rpart.plot)
library(rattle)
library(neuralnet)
library(kernlab)
library(caretEnsemble)
library(pROC)
library(Metrics)
library(OneR)
library(tm)
library(wordcloud)
library(RColorBrewer)
```

1. Data Acquisition

- For Importing data, I have used read.csv function.
- Using head function, I observed first few rows of the data
- Since almost all the features are categorical, I have kept stringsAsFactors = True

```
#Importing Dataset
data <- read.csv("C:\\Users\\harsh\\Desktop\\Introduction to Machine learning and Data Mining\\Project\\

#Exploring Dataset
head(data)
```

```
##             Timestamp Age Gender        Country state self_employed
## 1 2014-08-27 11:29:31  37 Female  United States    IL          <NA>
## 2 2014-08-27 11:29:37  44      M  United States    IN          <NA>
## 3 2014-08-27 11:29:44  32   Male         Canada  <NA>          <NA>
## 4 2014-08-27 11:29:46  31   Male United Kingdom  <NA>          <NA>
## 5 2014-08-27 11:30:22  31   Male  United States    TX          <NA>
## 6 2014-08-27 11:31:22  33   Male  United States    TN          <NA>
##   family_history treatment work_interfere   no_employees remote_work
## 1             No       Yes          Often           6-25          No
## 2             No        No         Rarely More than 1000          No
## 3             No        No         Rarely           6-25          No
## 4            Yes       Yes          Often         26-100          No
## 5             No        No          Never        100-500         Yes
## 6            Yes        No      Sometimes           6-25          No
##   tech_company   benefits care_options wellness_program  seek_help  anonymity
## 1          Yes        Yes     Not sure               No        Yes        Yes
## 2           No Don't know           No       Don't know Don't know Don't know
## 3          Yes         No           No               No         No Don't know
## 4          Yes         No          Yes               No         No         No
## 5          Yes        Yes           No       Don't know Don't know Don't know
## 6          Yes        Yes     Not sure               No Don't know Don't know
##               leave mental_health_consequence phys_health_consequence
## 1     Somewhat easy                        No                      No
## 2        Don't know                     Maybe                      No
## 3 Somewhat difficult                       No                      No
## 4 Somewhat difficult                      Yes                     Yes
## 5        Don't know                        No                      No
## 6        Don't know                        No                      No
##      coworkers supervisor mental_health_interview phys_health_interview
## 1 Some of them        Yes                      No                 Maybe
## 2           No         No                      No                    No
## 3          Yes        Yes                     Yes                   Yes
## 4 Some of them         No                   Maybe                 Maybe
## 5 Some of them        Yes                     Yes                   Yes
## 6          Yes        Yes                      No                 Maybe
##   mental_vs_physical obs_consequence comments
## 1                Yes              No     <NA>
## 2         Don't know              No     <NA>
## 3                 No              No     <NA>
## 4                 No             Yes     <NA>
## 5         Don't know              No     <NA>
## 6         Don't know              No     <NA>
```

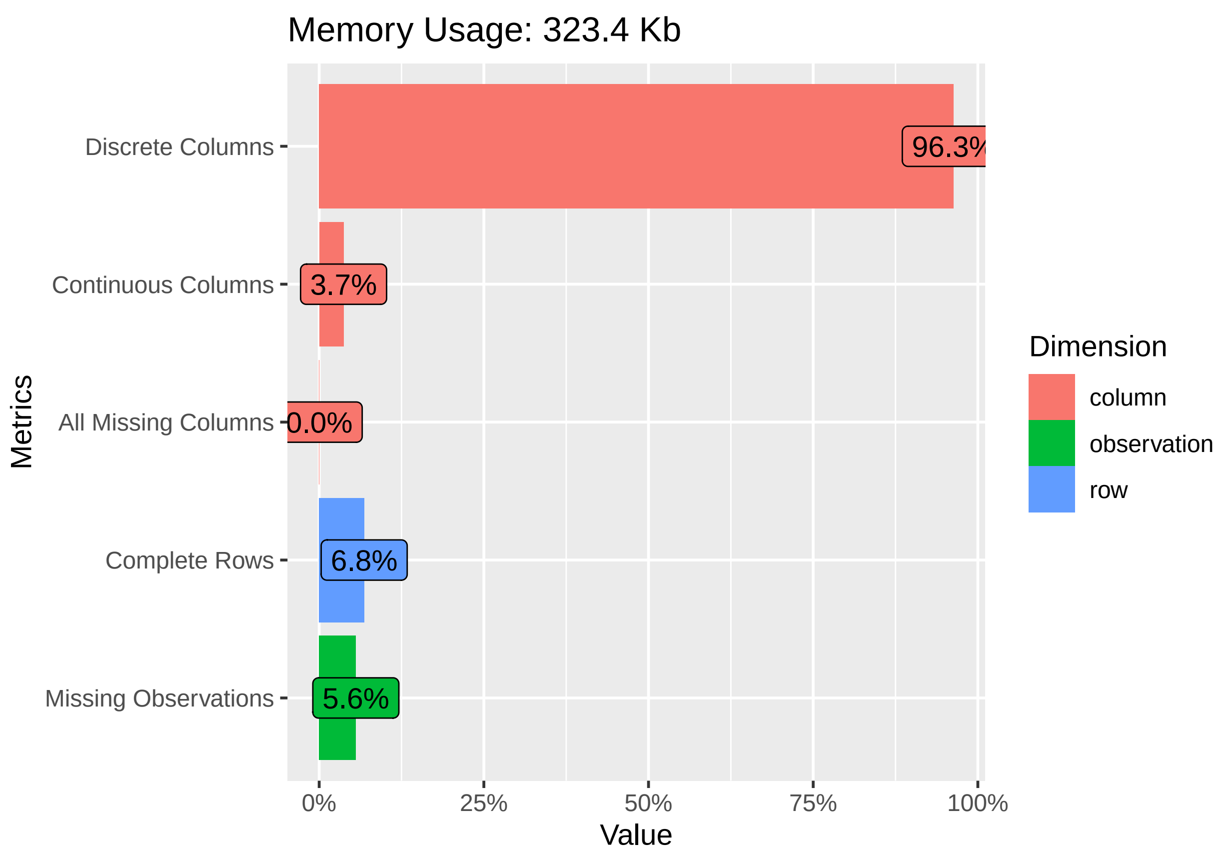
2. Data Exploration

Exploratory data plots

- I have used plot_intro function from DataExplorer package
- plot_intro provides an insight of what type of data is present along with that it provides the information about missing values
- Apart from that, I have used str and summary to understand the structure of the data present
- To calculate the number of NA present in the data I have created a function inside sapply it returns column name and NA present in it

- To get a better understanding of the distribution of the data, I have plotted each column in barplot
- In the plots we can observe Age, Gender, self_employed, work_interfere columns need to be cleaned

```r
###############################
### Exploratory data plots ###
###############################

#Visualizing structure of the dataset
plot_intro(data)
```




```r
#Exploratory Analysis
str(data)
```

```
## 'data.frame':    1259 obs. of  27 variables:
##  $ Timestamp                : Factor w/ 1246 levels "2014-08-27 11:29:31",..: 1 2 3 4 5 6 7 8 9 10 .
##  $ Age                      : num  37 44 32 31 31 33 35 39 42 23 ...
##  $ Gender                   : Factor w/ 49 levels "A little about you",..: 16 24 30 30 30 30 16 24 1
##  $ Country                  : Factor w/ 48 levels "Australia","Austria",..: 46 46 8 45 46 46 46 8 46
##  $ state                    : Factor w/ 45 levels "AL","AZ","CA",..: 11 12 NA NA 38 37 19 NA 11 NA .
##  $ self_employed            : Factor w/ 2 levels "No","Yes": NA NA NA NA NA NA NA NA NA NA ...
##  $ family_history           : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 2 1 2 1 ...
##  $ treatment                : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 2 1 2 1 ...
##  $ work_interfere           : Factor w/ 4 levels "Never","Often",..: 2 3 3 2 1 4 4 1 4 1 ...
##  $ no_employees             : Factor w/ 6 levels "1-5","100-500",..: 5 6 5 3 2 5 1 1 2 3 ...
##  $ remote_work              : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 2 1 1 ...
```

```
##  $ tech_company             : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 2 ...
##  $ benefits                 : Factor w/ 3 levels "Don't know","No",..: 3 1 2 2 3 3 2 2 3 1 ...
##  $ care_options             : Factor w/ 3 levels "No","Not sure",..: 2 1 1 3 1 2 1 3 3 1 ...
##  $ wellness_program         : Factor w/ 3 levels "Don't know","No",..: 2 1 2 2 1 2 2 2 2 1 ...
##  $ seek_help                : Factor w/ 3 levels "Don't know","No",..: 3 1 2 2 1 1 2 2 2 1 ...
##  $ anonymity                : Factor w/ 3 levels "Don't know","No",..: 3 1 1 2 1 1 2 3 2 1 ...
##  $ leave                    : Factor w/ 5 levels "Don't know","Somewhat difficult",..: 3 1 2 2 1 1 2
##  $ mental_health_consequence: Factor w/ 3 levels "Maybe","No","Yes": 2 1 2 3 2 2 1 2 1 2 ...
##  $ phys_health_consequence  : Factor w/ 3 levels "Maybe","No","Yes": 2 2 2 3 2 2 1 2 2 2 ...
##  $ coworkers                : Factor w/ 3 levels "No","Some of them",..: 2 1 3 2 2 3 2 1 3 3 ...
##  $ supervisor               : Factor w/ 3 levels "No","Some of them",..: 3 1 3 1 3 3 1 1 3 3 ...
##  $ mental_health_interview  : Factor w/ 3 levels "Maybe","No","Yes": 2 2 3 1 3 2 2 2 2 1 ...
##  $ phys_health_interview    : Factor w/ 3 levels "Maybe","No","Yes": 1 2 3 1 3 1 2 2 1 1 ...
##  $ mental_vs_physical       : Factor w/ 3 levels "Don't know","No",..: 3 1 2 2 1 1 1 2 2 3 ...
##  $ obs_consequence          : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ comments                 : Factor w/ 160 levels "-"," ","(yes but the situation was unusual and i
```

```r
summary(data)
```

```
##                Timestamp          Age                  Gender
##  2014-08-27 12:31:41:   2   Min.   :-1.726e+03   Male   :615
##  2014-08-27 12:37:50:   2   1st Qu.: 2.700e+01   male   :206
##  2014-08-27 12:43:28:   2   Median : 3.100e+01   Female :121
##  2014-08-27 12:44:51:   2   Mean   : 7.943e+07   M      :116
##  2014-08-27 12:54:11:   2   3rd Qu.: 3.600e+01   female : 62
##  2014-08-27 14:22:43:   2   Max.   : 1.000e+11   F      : 38
##  (Other)            :1247                        (Other):101
##            Country        state     self_employed family_history treatment
##  United States :751   CA     :138   No  :1095     No :767        No :622
##  United Kingdom:185   WA     : 70   Yes : 146     Yes:492        Yes:637
##  Canada        : 72   NY     : 57   NA's:  18
##  Germany       : 45   TN     : 45
##  Ireland       : 27   TX     : 44
##  Netherlands   : 27   (Other):390
##  (Other)       :152   NA's   :515
##    work_interfere         no_employees remote_work tech_company
##  Never    :213    1-5           :162   No :883     No : 228
##  Often    :144    100-500       :176   Yes:376     Yes:1031
##  Rarely   :173    26-100        :289
##  Sometimes:465    500-1000      : 60
##  NA's     :264    6-25          :290
##                   More than 1000:282
##
##       benefits     care_options   wellness_program      seek_help
##  Don't know:408   No      :501   Don't know:188     Don't know:363
##  No        :374   Not sure:314   No        :842     No        :646
##  Yes       :477   Yes     :444   Yes       :229     Yes       :250
##
##
##
##
##       anonymity                   leave     mental_health_consequence
##  Don't know:819   Don't know        :563   Maybe:477
##  No        : 65   Somewhat difficult:126   No   :490
```

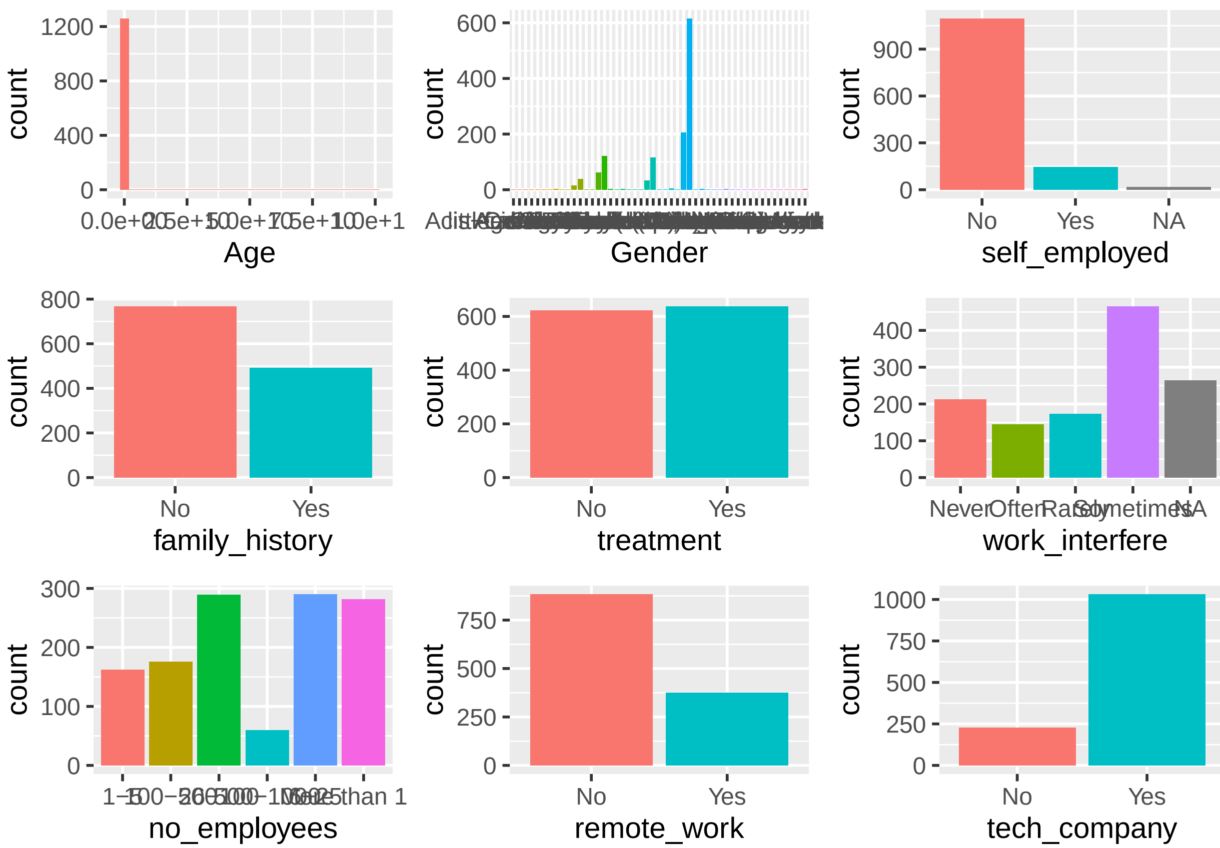
```
##  Yes       :375   Somewhat easy     :266   Yes  :292
##                   Very difficult    : 98
##                   Very easy         :206
##
##
##  phys_health_consequence        coworkers          supervisor
##  Maybe:273               No          :260   No          :393
##  No   :925               Some of them:774   Some of them:350
##  Yes  : 61               Yes         :225   Yes         :516
##
##
##
##
##  mental_health_interview phys_health_interview  mental_vs_physical
##  Maybe: 207              Maybe:557             Don't know:576
##  No   :1008              No   :500             No        :340
##  Yes  :  44              Yes  :202             Yes       :343
##
##
##
##
##  obs_consequence
##  No :1075
##  Yes: 184
##
##
##
##
##
##
##  * Small family business - YMMV.
##  -
##
##  (yes but the situation was unusual and involved a change in leadership at a very high level in the c
##  A close family member of mine struggles with mental health so I try not to stigmatize it. My employe
##  (Other)
##  NA's
```

```r
sapply(data, function(x) sum(is.na(x)))
```

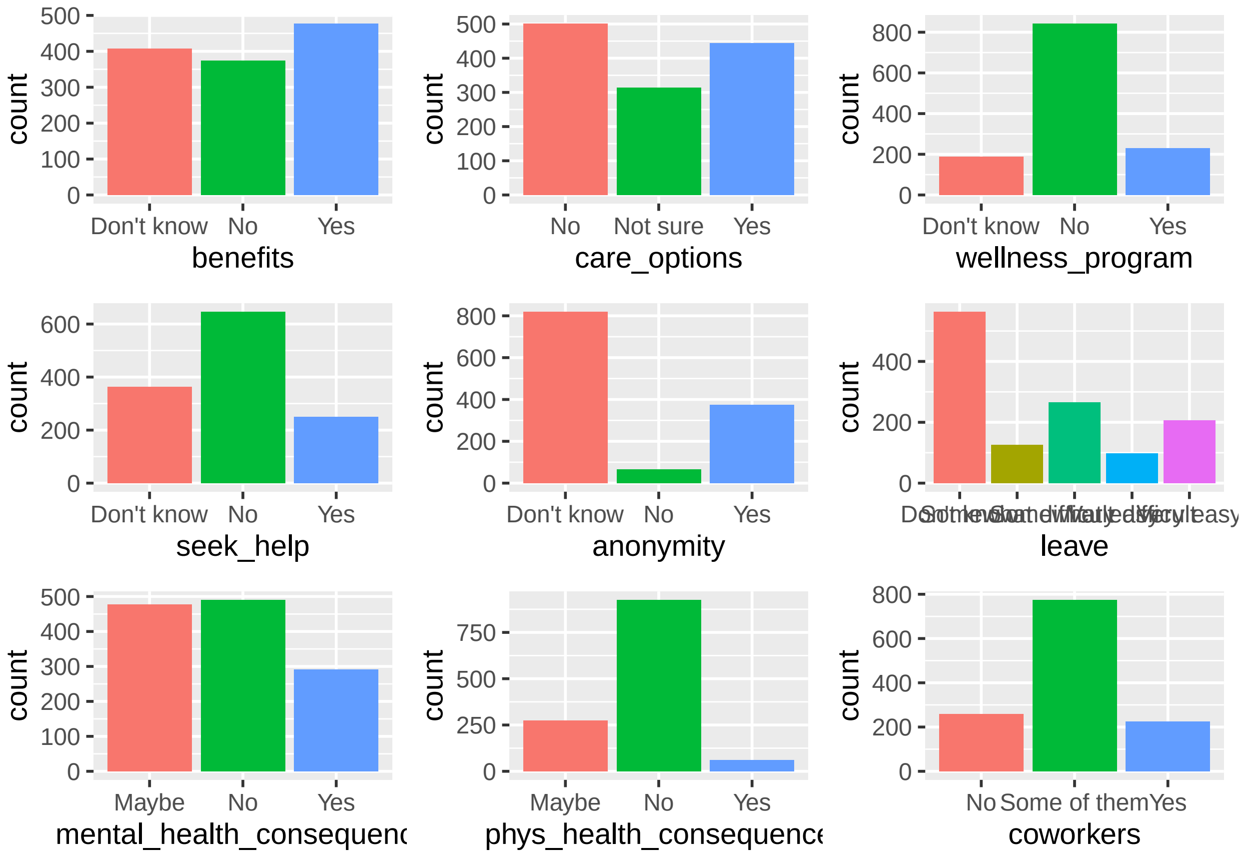
```
##                 Timestamp                       Age                    Gender
##                         0                         0                         0
##                   Country                     state             self_employed
##                         0                       515                        18
##            family_history                 treatment            work_interfere
##                         0                         0                       264
##              no_employees               remote_work              tech_company
##                         0                         0                         0
##                  benefits              care_options          wellness_program
##                         0                         0                         0
##                 seek_help                 anonymity                     leave
##                         0                         0                         0
## mental_health_consequence   phys_health_consequence                 coworkers
##                         0                         0                         0
```

```
##               supervisor   mental_health_interview     phys_health_interview
##                        0                         0                         0
##       mental_vs_physical           obs_consequence                  comments
##                        0                         0                      1095
```

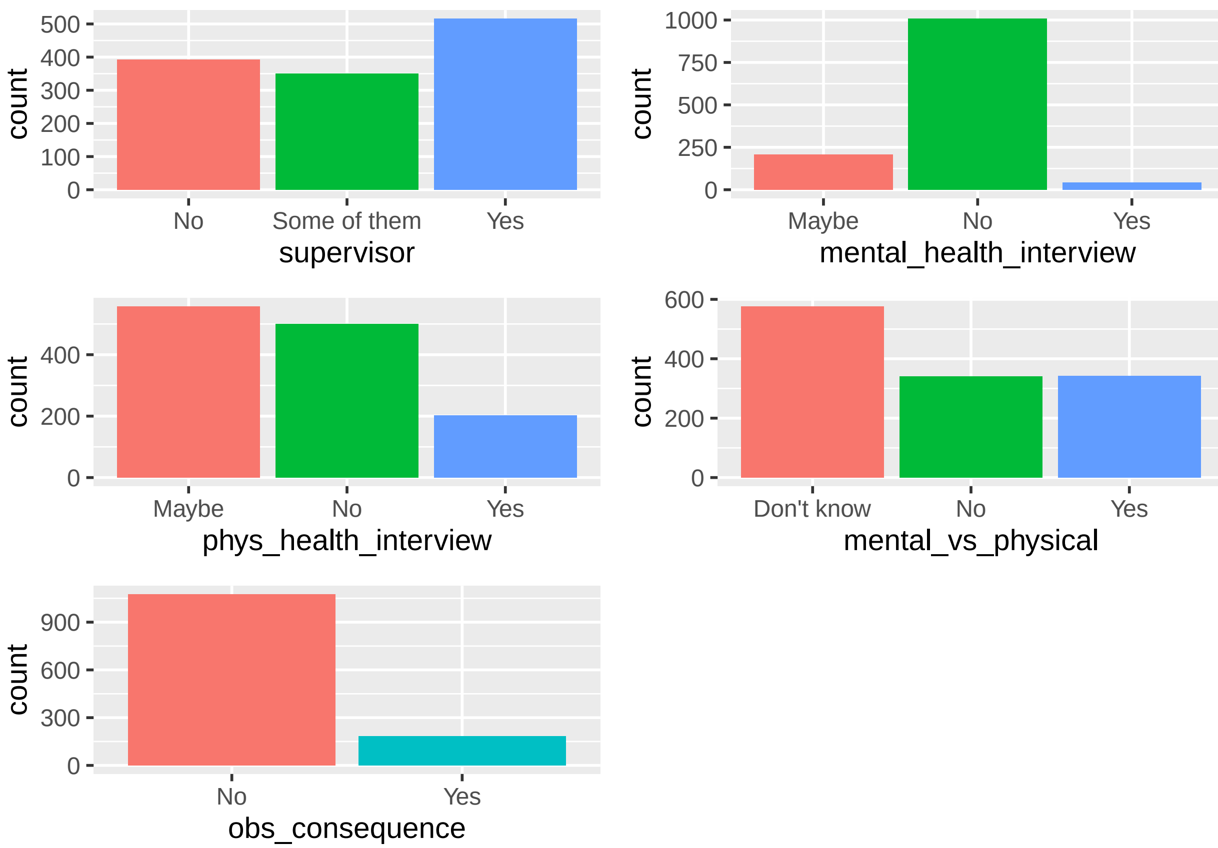
```r
#Plotting the distribution of the important features
g1 <- ggplot(data,aes(x=Age,fill="Steelblue"))+geom_histogram()+theme(legend.position = "none")
g2 <- ggplot(data,aes(x=Gender,fill=Gender))+geom_bar()+theme(legend.position = "none")
g3 <- ggplot(data,aes(x=self_employed,fill=self_employed))+geom_bar()+theme(legend.position = "none")
g4 <- ggplot(data,aes(x=family_history,fill=family_history))+geom_bar()+theme(legend.position = "none")
g5 <- ggplot(data,aes(x=treatment,fill=treatment))+geom_bar()+theme(legend.position = "none")
g6 <- ggplot(data,aes(x=work_interfere,fill=work_interfere))+geom_bar()+theme(legend.position = "none")
g7 <- ggplot(data,aes(x=no_employees,fill=no_employees))+geom_bar()+theme(legend.position = "none")
g8 <- ggplot(data,aes(x=remote_work,fill=remote_work))+geom_bar()+theme(legend.position = "none")
g9 <- ggplot(data,aes(x=tech_company,fill=tech_company))+geom_bar()+theme(legend.position = "none")
g10 <- ggplot(data,aes(x=benefits,fill=benefits))+geom_bar()+theme(legend.position = "none")
g11 <- ggplot(data,aes(x=care_options,fill=care_options))+geom_bar()+theme(legend.position = "none")
g12 <- ggplot(data,aes(x=wellness_program,fill=wellness_program))+geom_bar()+theme(legend.position = "no
g13 <- ggplot(data,aes(x=seek_help,fill=seek_help))+geom_bar()+theme(legend.position = "none")
g14 <- ggplot(data,aes(x=anonymity,fill=anonymity))+geom_bar()+theme(legend.position = "none")
g15 <- ggplot(data,aes(x=leave,fill=leave))+geom_bar()+theme(legend.position = "none")
g16 <- ggplot(data,aes(x=mental_health_consequence,fill=mental_health_consequence))+geom_bar()+theme(leg
g17 <- ggplot(data,aes(x=phys_health_consequence,fill=phys_health_consequence))+geom_bar()+theme(legend
g18 <- ggplot(data,aes(x=coworkers,fill=coworkers))+geom_bar()+theme(legend.position = "none")
g19 <- ggplot(data,aes(x=supervisor,fill=supervisor))+geom_bar()+theme(legend.position = "none")
g20 <- ggplot(data,aes(x=mental_health_interview,fill=mental_health_interview))+geom_bar()+theme(legend
g21 <- ggplot(data,aes(x=phys_health_interview,fill=phys_health_interview))+geom_bar()+theme(legend.posi
g22 <- ggplot(data,aes(x=mental_vs_physical,fill=mental_vs_physical))+geom_bar()+theme(legend.position =
g23 <- ggplot(data,aes(x=obs_consequence,fill=obs_consequence))+geom_bar()+theme(legend.position = "none

#Arranging the plots using grid.arrange function
grid.arrange(g1,g2,g3,g4,g5,g6,g7,g8,g9,nrow=3)
```




```r
grid.arrange(g10,g11,g12,g13,g14,g15,g16,g17,g18,nrow=3)
```




```
grid.arrange(g19,g20,g21,g22,g23,nrow=3)
```



Detection of outliers and data imputation

- I have checked only the age column since it is the only numerical column present in the whole dataset
- On observing the box plot and summary of Age column, I got to know that it has a few outliers
- This is because Age cannot contain negative values or values greater than 100
- I removed these outliers and imputed them with median value
- Apart from that, I have imputed mode values for the NA's present in self_employed and work_interfere columns
- Lastly Gender column was also cleaned, It contained many values for each type of gender so I generalized the column
- Plots for each cleaned columns have been shown below
- I have also stored this clean data in a new TableauDataCSV file for making a dashboard in Tableau.

```
##################################################
### Detection of outliers and Data imputation  ###
##################################################

#Creating a copy of data
MH_data <- data

#Mode function used to calculate mode
Mode <- function(x)
  {
    ux <- unique(x)
    ux[which.max(tabulate(match(x,ux)))]
```

```
  }

#Cleaning Age Column
summary(MH_data$Age)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -1.726e+03  2.700e+01  3.100e+01  7.943e+07  3.600e+01  1.000e+11
```

```
#Age Column has quite a few outliers present
#We can observe these incorrect values in the summary as well as in the plots
#Obvious outlier here are -1.726e+03 and 1.000e+11.
#Replacing with NA and then Imputing using median

mean_data <- mean(MH_data$Age)
sd_data <- sd(MH_data$Age)
zscore <- abs((MH_data$Age - mean_data)/sd_data)
print(MH_data[which((zscore>3)),2])
```

```
## [1] 1e+11
```

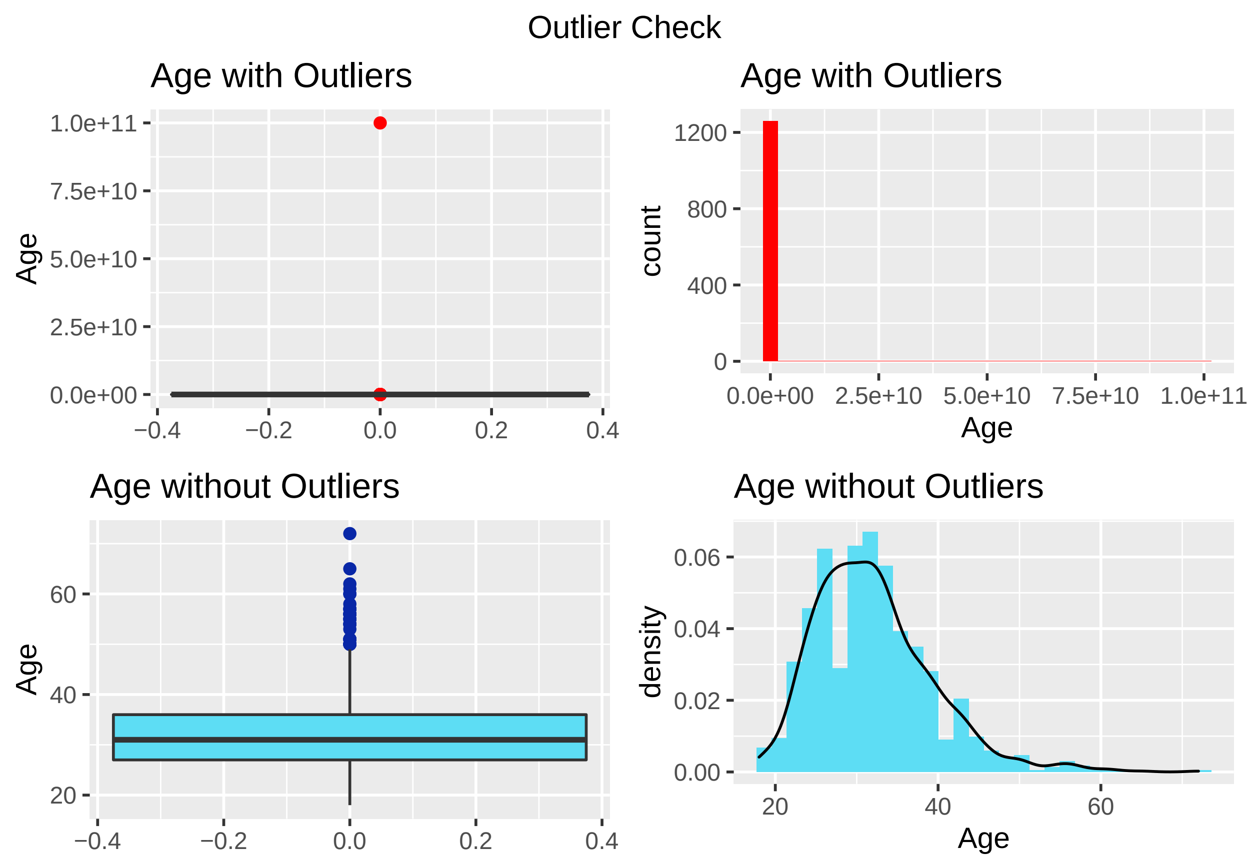
```
MH_data$Age <- sapply(MH_data$Age ,function(x) ifelse(x > 100 || x < 15, yes = NA,x))
sum(is.na(MH_data$Age))
```

```
## [1] 8
```

```
MH_data$Age[is.na(MH_data$Age)] <- median(MH_data$Age,na.rm = TRUE)
sum(is.na(MH_data$Age))
```

```
## [1] 0
```

```
#We can observe the difference in Age column before and after removing outliers
p1 <- ggplot(data,aes(y=Age),outcol="red")+geom_boxplot(outlier.colour="Red", outlier.shape=16,outlier.s
p2 <- ggplot(MH_data,aes(y=Age),outcol="blue")+geom_boxplot(outlier.colour="#0827A7", outlier.shape=16,o
p3 <- ggplot(data,aes(x=Age))+geom_histogram(fill="red")+ggtitle("Age with Outliers")
p4 <- ggplot(MH_data,aes(x=Age))+geom_histogram(aes(y=..density..),fill="#5DDDF4")+ggtitle("Age without
grid.arrange(p1,p3,p2,p4,nrow=2,top="Outlier Check")
```




```r
#Cleaning self-employed column
#On observing the summary of the data we see that self_emplyed
#column has many NA values present
summary(MH_data$self_employed)
```

```
##   No  Yes NA's
## 1095  146   18
```

```r
#Remove NA and impute mode values
#Since most of the columns are categorical
#variable imputation is done by Mode function
MH_data$self_employed[is.na(MH_data$self_employed)] <- Mode(MH_data$self_employed)
summary(MH_data$self_employed)
```

```
##   No  Yes
## 1113  146
```

```r
#Cleaning Gender column
summary(MH_data$Gender)
```

```
##                              A little about you
##                                               1
##                                         Agender
##                                               1
```

```
##                                            All
##                                              1
##                                      Androgyne
##                                              1
##                             cis-female/femme
##                                              1
##                                     Cis Female
##                                              1
##                                       cis male
##                                              1
##                                       Cis Male
##                                              2
##                                        Cis Man
##                                              1
##                                           Enby
##                                              1
##                                              f
##                                             15
##                                              F
##                                             38
##                                         femail
##                                              1
##                                         Femake
##                                              1
##                                         female
##                                             62
##                                         Female
##                                            121
##                                        Female 
##                                              2
##                                   Female (cis)
##                                              1
##                                 Female (trans)
##                                              2
##                                          fluid
##                                              1
##                                    Genderqueer
##                                              1
##                                 Guy (-ish) ^_^
##                                              1
##                                              m
##                                             34
##                                              M
##                                            116
##                                           Mail
##                                              1
##                                          maile
##                                              1
##                                           Make
##                                              4
##                                            Mal
##                                              1
##                                           male
##                                            206
```

```
##                                               Male
##                                                615
##                                           Male-ish
##                                                  1
##                                              Male 
##                                                  3
##                                        Male (CIS)
##                                                  1
##                           male leaning androgynous
##                                                  1
##                                               Malr
##                                                  1
##                                                Man
##                                                  2
##                                               msle
##                                                  1
##                                                Nah
##                                                  1
##                                             Neuter
##                                                  1
##                                         non-binary
##                                                  1
## ostensibly male, unsure what that really means
##                                                  1
##                                                  p
##                                                  1
##                                              queer
##                                                  1
##                                    queer/she/they
##                                                  1
##                              something kinda male?
##                                                  1
##                                       Trans-female
##                                                  1
##                                        Trans woman
##                                                  1
##                                              woman
##                                                  1
##                                              Woman
##                                                  3
```

```r
#Gender column has a lot of error values
#Using Unique function we can observe different types of gender values
Gender_list <- unique(MH_data$Gender)
Gender_list
```

```
##   [1] Female
##   [2] M
##   [3] Male
##   [4] male
##   [5] female
##   [6] m
##   [7] Male-ish
##   [8] maile
```

```
##  [9] Trans-female
## [10] Cis Female
## [11] F
## [12] something kinda male?
## [13] Cis Male
## [14] Woman
## [15] f
## [16] Mal
## [17] Male (CIS)
## [18] queer/she/they
## [19] non-binary
## [20] Femake
## [21] woman
## [22] Make
## [23] Nah
## [24] All
## [25] Enby
## [26] fluid
## [27] Genderqueer
## [28] Female
## [29] Androgyne
## [30] Agender
## [31] cis-female/femme
## [32] Guy (-ish) ^_^
## [33] male leaning androgynous
## [34] Male
## [35] Man
## [36] Trans woman
## [37] msle
## [38] Neuter
## [39] Female (trans)
## [40] queer
## [41] Female (cis)
## [42] Mail
## [43] cis male
## [44] A little about you
## [45] Malr
## [46] p
## [47] femail
## [48] Cis Man
## [49] ostensibly male, unsure what that really means
## 49 Levels: A little about you Agender All Androgyne ... Woman
```

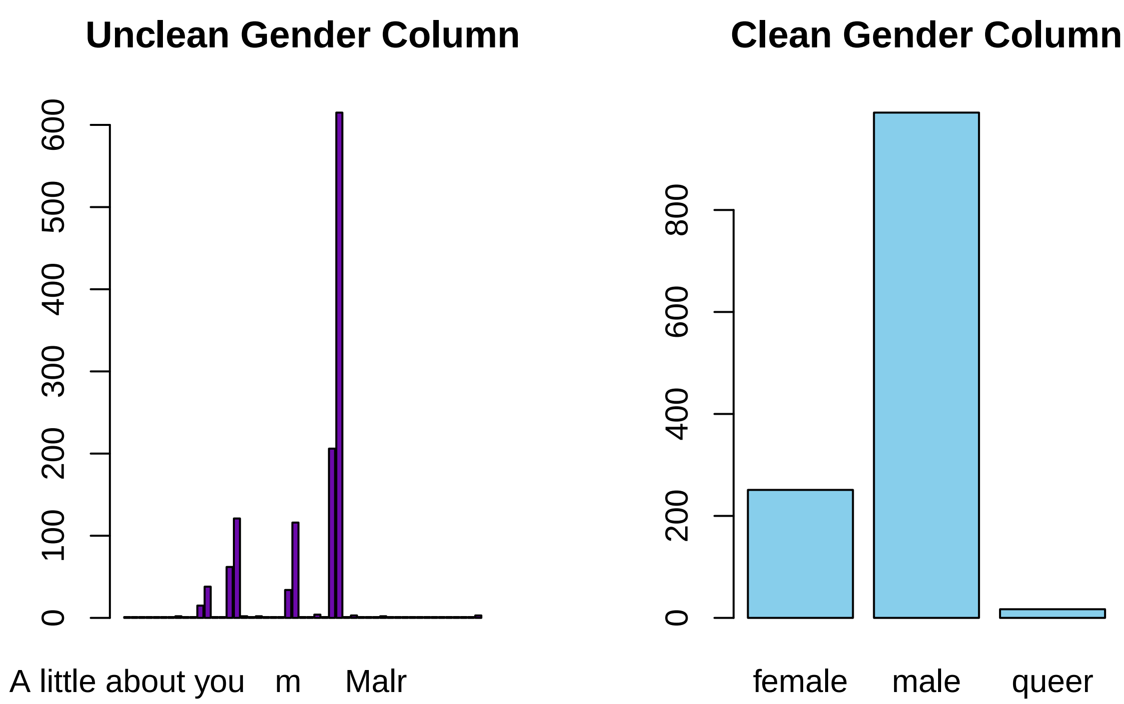
```
#We create a single vector for each type of gender and assign the different values present
Male <- c("Male ", "Mail", "maile","Cis Man", "Malr", "Man", "Male", "male", "M", "cis male", "m", "Mal
Female <- c("Female ","Female","femail","woman","Female","Female (cis)","cis-female/femme", "Cis Female
Queer <- c("Genderqueer","ostensibly male, unsure what that really means","p","A little about you","que

#Using the new vectors we make the proper distribution of gender
MH_data$Gender <- as.factor(ifelse(MH_data$Gender %in% Male,"male",ifelse(MH_data$Gender %in% Female,"fe

#Verifying Gender Column data after cleaning
str(MH_data$Gender)
```

```
##  Factor w/ 3 levels "female","male",..: 1 2 2 2 2 2 1 2 1 2 ...
```

```
table(MH_data$Gender)
```

```
##
## female   male  queer
##    251    991     17
```

```
par(mfrow=c(1,2))
barplot(table(data$Gender),col = "#6C0AAB",main = "Unclean Gender Column")
barplot(table(MH_data$Gender),col = "skyblue",main = "Clean Gender Column")
```



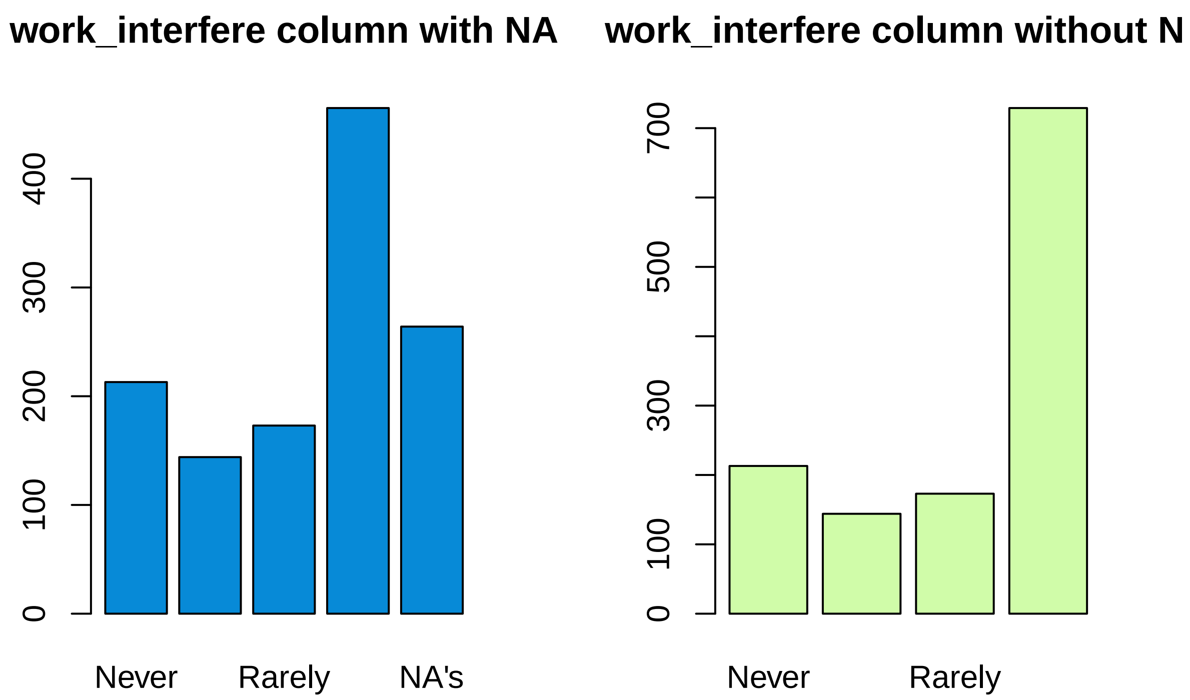

```
#Cleaning work_interfere
#Using Summary we can see that there are around 200 NA values present
summary(MH_data$work_interfere)
```

```
##      Never      Often     Rarely Sometimes       NA's
##        213        144        173        465        264
```

```
#Since it is a categorical variable we'll impute using mode function
MH_data$work_interfere[is.na(MH_data$work_interfere)] <- Mode(MH_data$work_interfere)
summary(MH_data$work_interfere)
```

```
##      Never      Often     Rarely Sometimes
##        213        144        173        729
```

```
#Observing the difference before and after imputationn
par(mfrow=c(1,2))
barplot(summary(data$work_interfere),col = "#078AD7",main = "work_interfere column with NA")
barplot(table(MH_data$work_interfere),col = "#D0FCA9",main = "work_interfere column without NA")
```




```
#Storing cleaned data for Tableau Visualization
write.table(MH_data,"TableauDataCSV.csv", sep = ",",col.names = !file.exists("myDF.csv"), append = T,ro
#Remove unwanted columns
#Comments,country,state,and timestamp are unwanted
#columns so we remove it from that dataset
MH_data <- MH_data[,c(-1,-4,-5,-27)]

#Verifying Cleaned data
summary(MH_data)
```

```
##       Age             Gender     self_employed family_history treatment
##  Min.   :18.00   female:251   No :1113      No :767        No :622  
##  1st Qu.:27.00   male  :991   Yes: 146      Yes:492        Yes:637  
##  Median :31.00   queer : 17                                         
##  Mean   :32.07                                                      
##  3rd Qu.:36.00                                                      
##  Max.   :72.00                                                      
##    work_interfere          no_employees remote_work tech_company
##  Never    :213    1-5            :162   No :883     No : 228    
##  Often    :144    100-500        :176   Yes:376     Yes:1031    
```

```
##  Rarely   :173    26-100        :289
##  Sometimes:729    500-1000      : 60
##                   6-25          :290
##                   More than 1000:282
##        benefits     care_options   wellness_program      seek_help
##  Don't know:408   No      :501   Don't know:188     Don't know:363
##  No        :374   Not sure:314   No        :842     No        :646
##  Yes       :477   Yes     :444   Yes       :229     Yes       :250
##
##
##
##       anonymity                  leave     mental_health_consequence
##  Don't know:819   Don't know        :563   Maybe:477
##  No        : 65   Somewhat difficult:126   No   :490
##  Yes       :375   Somewhat easy     :266   Yes  :292
##                   Very difficult    : 98
##                   Very easy         :206
##
##  phys_health_consequence        coworkers          supervisor
##  Maybe:273               No          :260   No          :393
##  No   :925               Some of them:774   Some of them:350
##  Yes  : 61               Yes         :225   Yes         :516
##
##
##
##  mental_health_interview phys_health_interview  mental_vs_physical
##  Maybe: 207              Maybe:557             Don't know:576
##  No   :1008              No   :500             No        :340
##  Yes  :  44              Yes  :202             Yes       :343
##
##
##
##  obs_consequence
##  No :1075
##  Yes: 184
##
##
##
##
```

```
#No NA values present after cleaning
sapply(MH_data, function(x) sum(is.na(x)))
```

```
##                       Age                    Gender             self_employed
##                         0                         0                         0
##            family_history                 treatment            work_interfere
##                         0                         0                         0
##              no_employees               remote_work              tech_company
##                         0                         0                         0
##                  benefits              care_options          wellness_program
##                         0                         0                         0
##                 seek_help                 anonymity                     leave
##                         0                         0                         0
## mental_health_consequence   phys_health_consequence                 coworkers
```

```
##                         0                          0                          0
##                supervisor    mental_health_interview      phys_health_interview
##                         0                          0                          0
##        mental_vs_physical            obs_consequence
##                         0                          0
```

```r
#Creating a copy of factor dataset for categorical classifiers
MH_data_factors <- MH_data
```

Feature Engineering - Dummy codes

- Since the whole data is categorical, I have already used factors datatype.
- So instead of dummy coding each column, I just converted the factors data to numeric which does the dummy coding part
- I have also stored the original factor dataset in a variable called MH_data_factors
- This numeric data is used only for neural network classifier and for correlation analysis, other than that all other algorithms make use of factor dataset

Correlation/Collinearity analysis - Numerical data is required for calculating correlation, so I have converted factor data to numerical data - Correlation plot is shown for whole data - I have also shown the plot of correlation between treatment and all other features

```r
########################################
### Feature engineering: dummy codes ###
########################################

#we have factor dataset, on converting it to numeric we get dummy codes
str(MH_data)
```

```
## 'data.frame':    1259 obs. of  23 variables:
##  $ Age                      : num  37 44 32 31 31 33 35 39 42 23 ...
##  $ Gender                   : Factor w/ 3 levels "female","male",..: 1 2 2 2 2 2 1 2 1 2 ...
##  $ self_employed            : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ family_history           : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 2 1 2 1 ...
##  $ treatment                : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 2 1 2 1 ...
##  $ work_interfere           : Factor w/ 4 levels "Never","Often",..: 2 3 3 2 1 4 4 1 4 1 ...
##  $ no_employees             : Factor w/ 6 levels "1-5","100-500",..: 5 6 5 3 2 5 1 1 2 3 ...
##  $ remote_work              : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 2 1 1 ...
##  $ tech_company             : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 2 ...
##  $ benefits                 : Factor w/ 3 levels "Don't know","No",..: 3 1 2 2 3 3 2 2 3 1 ...
##  $ care_options             : Factor w/ 3 levels "No","Not sure",..: 2 1 1 3 1 2 1 3 3 1 ...
##  $ wellness_program         : Factor w/ 3 levels "Don't know","No",..: 2 1 2 2 1 2 2 2 2 1 ...
##  $ seek_help                : Factor w/ 3 levels "Don't know","No",..: 3 1 2 2 1 1 2 2 2 1 ...
##  $ anonymity                : Factor w/ 3 levels "Don't know","No",..: 3 1 1 2 1 1 2 3 2 1 ...
##  $ leave                    : Factor w/ 5 levels "Don't know","Somewhat difficult",..: 3 1 2 2 1 1 2
##  $ mental_health_consequence: Factor w/ 3 levels "Maybe","No","Yes": 2 1 2 3 2 2 1 2 1 2 ...
##  $ phys_health_consequence  : Factor w/ 3 levels "Maybe","No","Yes": 2 2 2 3 2 2 1 2 2 2 ...
##  $ coworkers                : Factor w/ 3 levels "No","Some of them",..: 2 1 3 2 2 3 2 1 3 3 ...
##  $ supervisor               : Factor w/ 3 levels "No","Some of them",..: 3 1 3 1 3 3 1 1 3 3 ...
##  $ mental_health_interview  : Factor w/ 3 levels "Maybe","No","Yes": 2 2 3 1 3 2 2 2 2 1 ...
##  $ phys_health_interview    : Factor w/ 3 levels "Maybe","No","Yes": 1 2 3 1 3 1 2 2 1 1 ...
##  $ mental_vs_physical       : Factor w/ 3 levels "Don't know","No",..: 3 1 2 2 1 1 1 2 2 3 ...
##  $ obs_consequence          : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
```

```r
#Since Neural Network takes in only numeric data
#We convert the cleaned data to numeric type
#Converting to numeric will also do the dummy coding as the data was of
#factor type so converting to numeric makes it dummy coded
for (i in 1:ncol(MH_data)){
  if(is.factor(MH_data[,i] )){
    MH_data[,i] <- as.numeric(MH_data[,i])
  }
}

#Verifying the structure of the dataset
str(MH_data)
```

```
## 'data.frame':    1259 obs. of  23 variables:
##  $ Age                      : num  37 44 32 31 31 33 35 39 42 23 ...
##  $ Gender                   : num  1 2 2 2 2 2 1 2 1 2 ...
##  $ self_employed            : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ family_history           : num  1 1 1 2 1 2 2 1 2 1 ...
##  $ treatment                : num  2 1 1 2 1 1 2 1 2 1 ...
##  $ work_interfere           : num  2 3 3 2 1 4 4 1 4 1 ...
##  $ no_employees             : num  5 6 5 3 2 5 1 1 2 3 ...
##  $ remote_work              : num  1 1 1 1 2 1 2 2 1 1 ...
##  $ tech_company             : num  2 1 2 2 2 2 2 2 2 2 ...
##  $ benefits                 : num  3 1 2 2 3 3 2 2 3 1 ...
##  $ care_options             : num  2 1 1 3 1 2 1 3 3 1 ...
##  $ wellness_program         : num  2 1 2 2 1 2 2 2 2 1 ...
##  $ seek_help                : num  3 1 2 2 1 1 2 2 2 1 ...
##  $ anonymity                : num  3 1 1 2 1 1 2 3 2 1 ...
##  $ leave                    : num  3 1 2 2 1 1 2 1 4 1 ...
##  $ mental_health_consequence: num  2 1 2 3 2 2 1 2 1 2 ...
##  $ phys_health_consequence  : num  2 2 2 3 2 2 1 2 2 2 ...
##  $ coworkers                : num  2 1 3 2 2 3 2 1 3 3 ...
##  $ supervisor               : num  3 1 3 1 3 3 1 1 3 3 ...
##  $ mental_health_interview  : num  2 2 3 1 3 2 2 2 2 1 ...
##  $ phys_health_interview    : num  1 2 3 1 3 1 2 2 1 1 ...
##  $ mental_vs_physical       : num  3 1 2 2 1 1 1 2 2 3 ...
##  $ obs_consequence          : num  1 1 1 2 1 1 1 1 1 1 ...
```

```r
#########################################
### Correlation/collinearity analysis ###
#########################################

#Creating a correlation plot of whole dataset
cormat <- round(cor(MH_data),2)
cormat
```

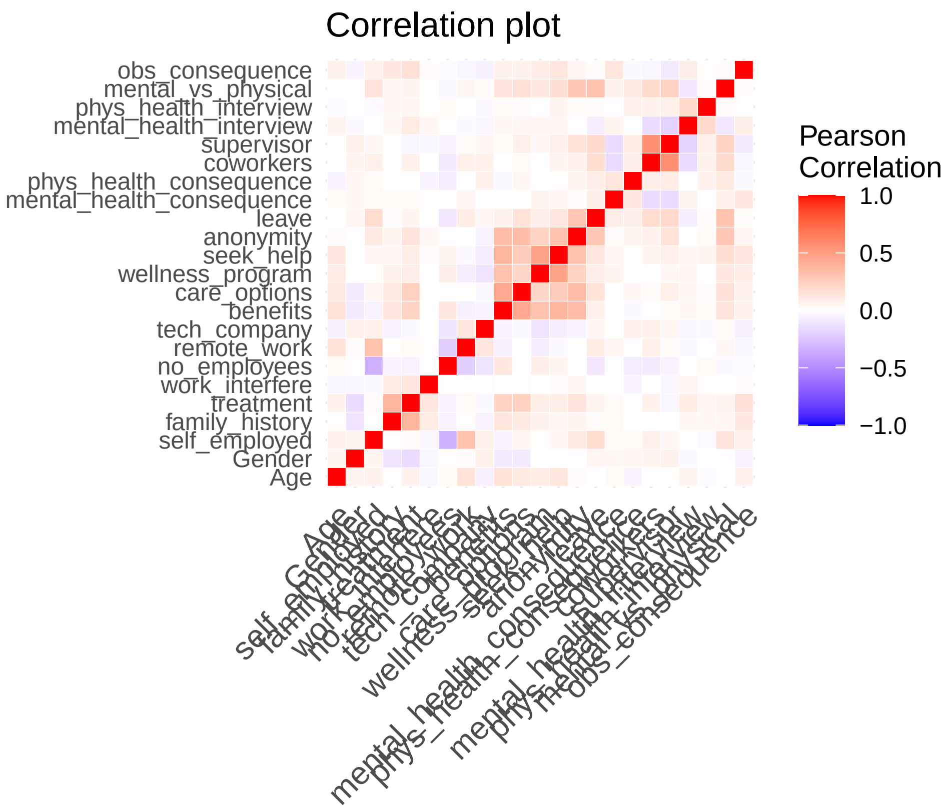
```
##                             Age Gender self_employed family_history treatment
## Age                        1.00   0.06          0.07           0.01      0.07
## Gender                     0.06   1.00          0.06          -0.12     -0.15
## self_employed              0.07   0.06          1.00           0.01      0.02
## family_history             0.01  -0.12          0.01           1.00      0.38
## treatment                  0.07  -0.15          0.02           0.38      1.00
## work_interfere            -0.04  -0.04         -0.03           0.10      0.13
```

```
## no_employees               0.03   0.01         -0.34          -0.05     -0.05
## remote_work                0.15   0.02          0.32           0.01      0.03
## tech_company              -0.06   0.08          0.08          -0.05     -0.03
## benefits                   0.15  -0.09         -0.05           0.13      0.23
## care_options               0.11  -0.09          0.05           0.11      0.24
## wellness_program           0.10   0.00          0.01           0.07      0.09
## seek_help                  0.13  -0.01          0.04           0.05      0.09
## anonymity                  0.02  -0.01          0.11           0.06      0.14
## leave                     -0.01   0.05          0.18           0.02      0.06
## mental_health_consequence  0.03   0.04          0.03           0.03      0.03
## phys_health_consequence   -0.05   0.05          0.03           0.00     -0.01
## coworkers                 -0.01   0.06          0.08           0.00      0.07
## supervisor                 0.01   0.07          0.04           0.00     -0.04
## mental_health_interview    0.06  -0.03         -0.01           0.04      0.10
## phys_health_interview     -0.02  -0.01         -0.02           0.04      0.05
## mental_vs_physical        -0.01  -0.01          0.14           0.04      0.06
## obs_consequence            0.07  -0.05          0.08           0.12      0.16
##                           work_interfere no_employees remote_work tech_company
## Age                                -0.04         0.03        0.15        -0.06
## Gender                             -0.04         0.01        0.02         0.08
## self_employed                      -0.03        -0.34        0.32         0.08
## family_history                      0.10        -0.05        0.01        -0.05
## treatment                           0.13        -0.05        0.03        -0.03
## work_interfere                      1.00         0.01        0.01         0.01
## no_employees                        0.01         1.00       -0.21        -0.11
## remote_work                         0.01        -0.21        1.00         0.13
## tech_company                        0.01        -0.11        0.13         1.00
## benefits                            0.00         0.12       -0.06        -0.05
## care_options                        0.01        -0.01        0.01        -0.03
## wellness_program                    0.00         0.09       -0.07        -0.12
## seek_help                           0.02         0.06       -0.03        -0.07
## anonymity                           0.04        -0.01        0.00        -0.05
## leave                               0.00        -0.10        0.10         0.05
## mental_health_consequence          -0.01        -0.01        0.05         0.00
## phys_health_consequence            -0.05        -0.08       -0.01         0.07
## coworkers                           0.00        -0.09        0.08         0.08
## supervisor                         -0.04        -0.05        0.03         0.05
## mental_health_interview             0.05         0.01       -0.03        -0.04
## phys_health_interview               0.01         0.03       -0.01        -0.03
## mental_vs_physical                  0.01        -0.03        0.04         0.03
## obs_consequence                     0.02        -0.02       -0.04        -0.06
##                           benefits care_options wellness_program seek_help
## Age                           0.15         0.11             0.10      0.13
## Gender                       -0.09        -0.09             0.00     -0.01
## self_employed                -0.05         0.05             0.01      0.04
## family_history                0.13         0.11             0.07      0.05
## treatment                     0.23         0.24             0.09      0.09
## work_interfere                0.00         0.01             0.00      0.02
## no_employees                  0.12        -0.01             0.09      0.06
## remote_work                  -0.06         0.01            -0.07     -0.03
## tech_company                 -0.05        -0.03            -0.12     -0.07
## benefits                      1.00         0.44             0.32      0.38
## care_options                  0.44         1.00             0.21      0.26
## wellness_program              0.32         0.21             1.00      0.47
```

```
## seek_help                       0.38         0.26             0.47       1.00
## anonymity                       0.34         0.35             0.23       0.32
## leave                           0.07         0.15             0.09       0.13
## mental_health_consequence      -0.01         0.00             0.06       0.05
## phys_health_consequence        -0.03         0.04            -0.01       0.01
## coworkers                      -0.01         0.03            -0.01       0.06
## supervisor                      0.03         0.08             0.04       0.08
## mental_health_interview         0.04         0.04             0.05       0.04
## phys_health_interview           0.03         0.02            -0.01       0.06
## mental_vs_physical              0.14         0.16             0.12       0.17
## obs_consequence                 0.07         0.07             0.10       0.13
##                           anonymity leave mental_health_consequence
## Age                            0.02 -0.01                      0.03
## Gender                        -0.01  0.05                      0.04
## self_employed                  0.11  0.18                      0.03
## family_history                 0.06  0.02                      0.03
## treatment                      0.14  0.06                      0.03
## work_interfere                 0.04  0.00                     -0.01
## no_employees                  -0.01 -0.10                     -0.01
## remote_work                    0.00  0.10                      0.05
## tech_company                  -0.05  0.05                      0.00
## benefits                       0.34  0.07                     -0.01
## care_options                   0.35  0.15                      0.00
## wellness_program               0.23  0.09                      0.06
## seek_help                      0.32  0.13                      0.05
## anonymity                      1.00  0.29                      0.02
## leave                          0.29  1.00                      0.09
## mental_health_consequence      0.02  0.09                      1.00
## phys_health_consequence        0.06  0.09                      0.13
## coworkers                      0.07  0.18                     -0.15
## supervisor                     0.15  0.20                     -0.15
## mental_health_interview        0.00 -0.07                      0.06
## phys_health_interview          0.03  0.02                     -0.01
## mental_vs_physical             0.29  0.31                      0.07
## obs_consequence                0.05  0.02                      0.13
##                           phys_health_consequence coworkers supervisor
## Age                                         -0.05     -0.01       0.01
## Gender                                       0.05      0.06       0.07
## self_employed                                0.03      0.08       0.04
## family_history                               0.00      0.00       0.00
## treatment                                   -0.01      0.07      -0.04
## work_interfere                              -0.05      0.00      -0.04
## no_employees                                -0.08     -0.09      -0.05
## remote_work                                 -0.01      0.08       0.03
## tech_company                                 0.07      0.08       0.05
## benefits                                    -0.03     -0.01       0.03
## care_options                                 0.04      0.03       0.08
## wellness_program                            -0.01     -0.01       0.04
## seek_help                                    0.01      0.06       0.08
## anonymity                                    0.06      0.07       0.15
## leave                                        0.09      0.18       0.20
## mental_health_consequence                    0.13     -0.15      -0.15
## phys_health_consequence                      1.00      0.09       0.10
## coworkers                                    0.09      1.00       0.57
```

```
## supervisor                                  0.10       0.57        1.00
## mental_health_interview                    -0.01      -0.15       -0.19
## phys_health_interview                       0.07       0.07        0.08
## mental_vs_physical                          0.11       0.19        0.23
## obs_consequence                            -0.03      -0.04       -0.09
##                           mental_health_interview phys_health_interview
## Age                                          0.06                 -0.02
## Gender                                      -0.03                 -0.01
## self_employed                               -0.01                 -0.02
## family_history                               0.04                  0.04
## treatment                                    0.10                  0.05
## work_interfere                               0.05                  0.01
## no_employees                                 0.01                  0.03
## remote_work                                 -0.03                 -0.01
## tech_company                                -0.04                 -0.03
## benefits                                     0.04                  0.03
## care_options                                 0.04                  0.02
## wellness_program                             0.05                 -0.01
## seek_help                                    0.04                  0.06
## anonymity                                    0.00                  0.03
## leave                                       -0.07                  0.02
## mental_health_consequence                    0.06                 -0.01
## phys_health_consequence                     -0.01                  0.07
## coworkers                                   -0.15                  0.07
## supervisor                                  -0.19                  0.08
## mental_health_interview                      1.00                  0.20
## phys_health_interview                        0.20                  1.00
## mental_vs_physical                          -0.10                  0.02
## obs_consequence                              0.09                  0.01
##                           mental_vs_physical obs_consequence
## Age                                    -0.01            0.07
## Gender                                 -0.01           -0.05
## self_employed                           0.14            0.08
## family_history                          0.04            0.12
## treatment                               0.06            0.16
## work_interfere                          0.01            0.02
## no_employees                           -0.03           -0.02
## remote_work                             0.04           -0.04
## tech_company                            0.03           -0.06
## benefits                                0.14            0.07
## care_options                            0.16            0.07
## wellness_program                        0.12            0.10
## seek_help                               0.17            0.13
## anonymity                               0.29            0.05
## leave                                   0.31            0.02
## mental_health_consequence               0.07            0.13
## phys_health_consequence                 0.11           -0.03
## coworkers                               0.19           -0.04
## supervisor                              0.23           -0.09
## mental_health_interview                -0.10            0.09
## phys_health_interview                   0.02            0.01
## mental_vs_physical                      1.00            0.02
## obs_consequence                         0.02            1.00
```

```r
melted_cormat <- melt(cormat)
ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Pearson\nCorrelation") +
  theme_minimal()+
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 12, hjust = 1))+
 coord_fixed()+xlab("")+ylab("")+ggtitle("Correlation plot")
```

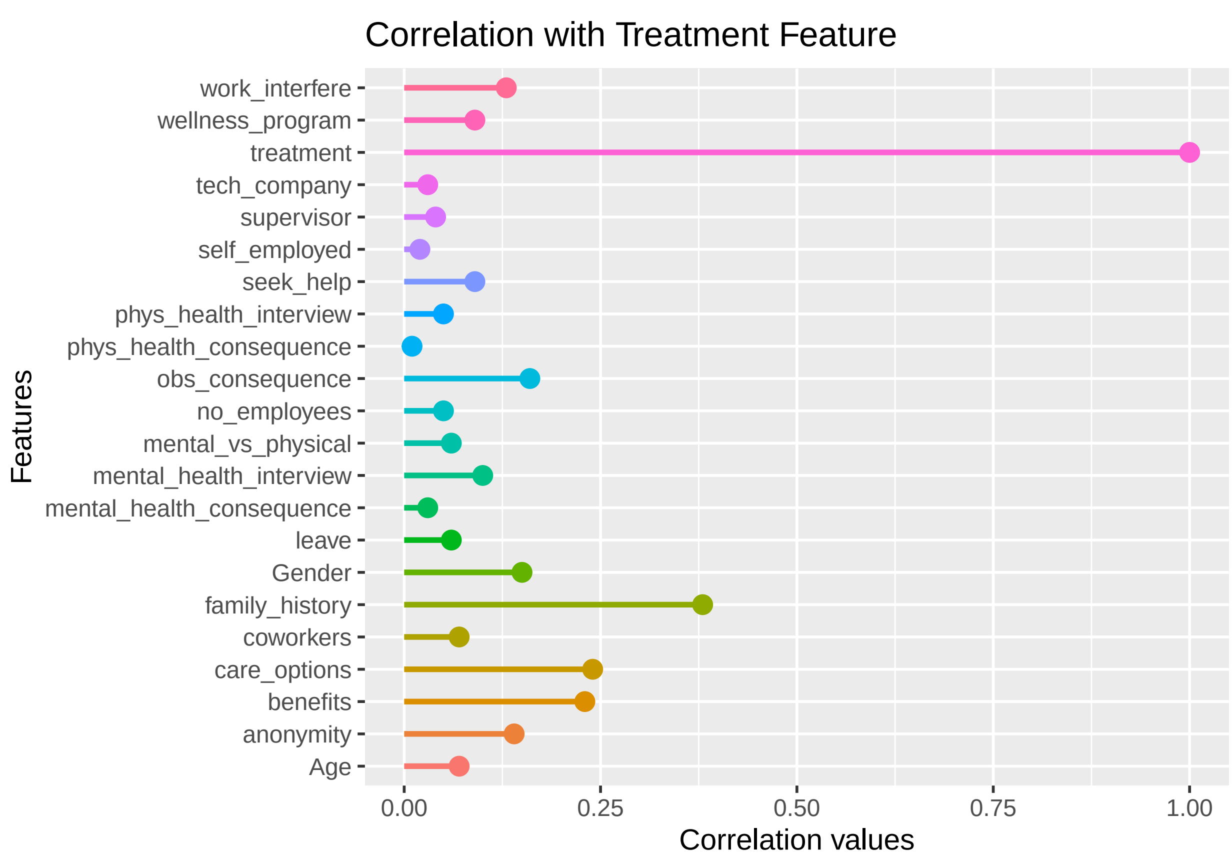



```r
#Observing correlation of all features with treatment feature
correlations <- as.data.frame(round(cor(MH_data[,-8],MH_data$treatment),2))
names <- rownames(correlations)
rownames(correlations) <- NULL
correlations <- cbind(names,correlations)
correlations <- correlations[order(-correlations$V1),]
correlations$V1 <- abs(correlations$V1)
correlations
```

```
##                         names   V1
## 5                   treatment 1.00
## 4              family_history 0.38
## 10               care_options 0.24
## 9                    benefits 0.23
```

```
## 22          obs_consequence 0.16
## 13                anonymity 0.14
## 6            work_interfere 0.13
## 19   mental_health_interview 0.10
## 11          wellness_program 0.09
## 12                seek_help 0.09
## 1                       Age 0.07
## 17                coworkers 0.07
## 14                    leave 0.06
## 21        mental_vs_physical 0.06
## 20     phys_health_interview 0.05
## 15 mental_health_consequence 0.03
## 3             self_employed 0.02
## 16   phys_health_consequence 0.01
## 8              tech_company 0.03
## 18               supervisor 0.04
## 7              no_employees 0.05
## 2                    Gender 0.15
```

```
#Plotting the correlation of all features with treatment feature
ggplot(data = correlations, aes(x = V1, y = names, color = names, group = names))+
  geom_segment(data = correlations,aes(x=0,xend = V1, y = names, yend = names),size = 1)+
  geom_point(size = 3)+ggtitle("Correlation with Treatment Feature")+
  theme(legend.position = "none")+xlab("Correlation values")+ylab("Features")
```



```
#I also tried pairs.panels function for
#correlation but since there are more than 15 features
#Plots are not clearly visible
#pairs.panels(MH_data)
```

3. Data Cleaning & Shaping

Data Imputation

- Data imputation is already done in previous chunks
- Imputation for age, self_employed and work_interfere is done

Proper Encoding of Data

- Encoding was done for only Age column
- Age is categorized into three types Fresher, Junior and Senior

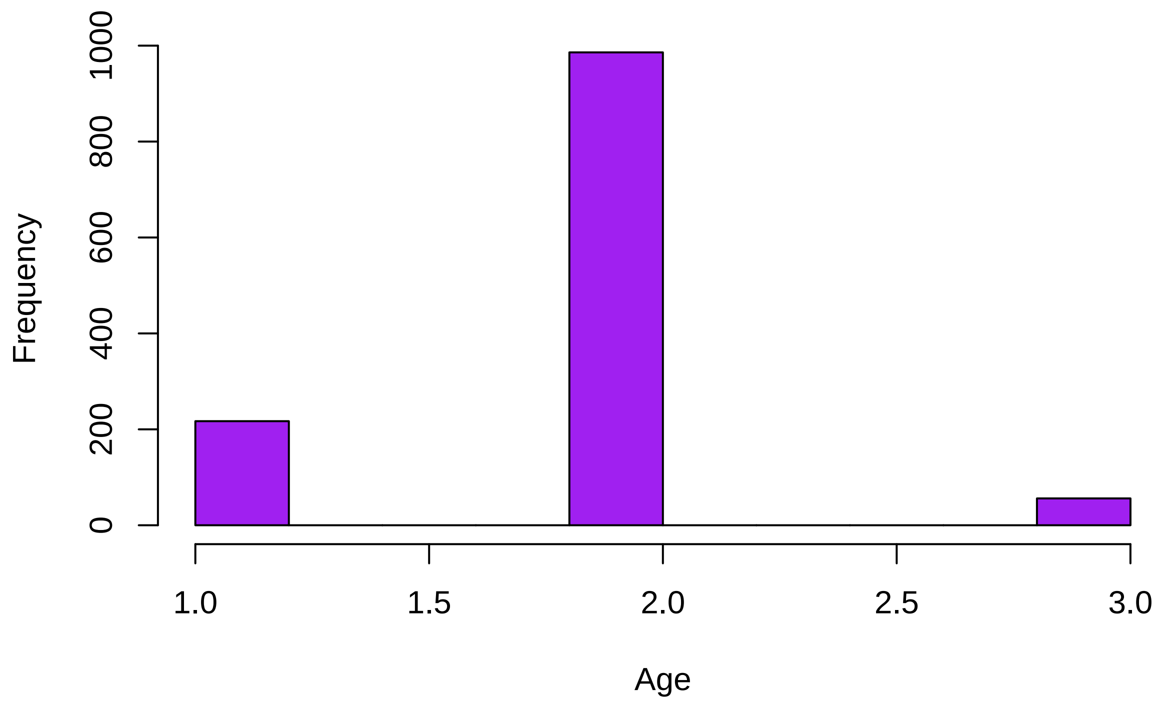
Normalization/Standardization

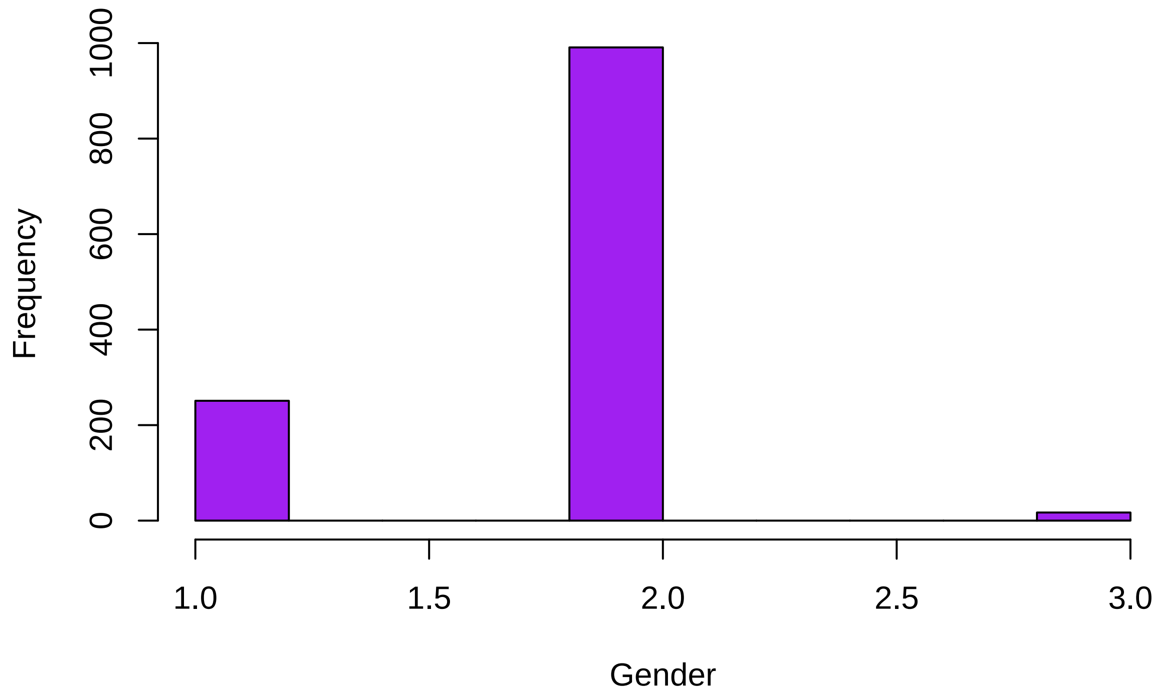
- Normalizing the data did not make any difference in predictions
- This is because the data is categorical and not continous
- So I have not used normalized data for my models

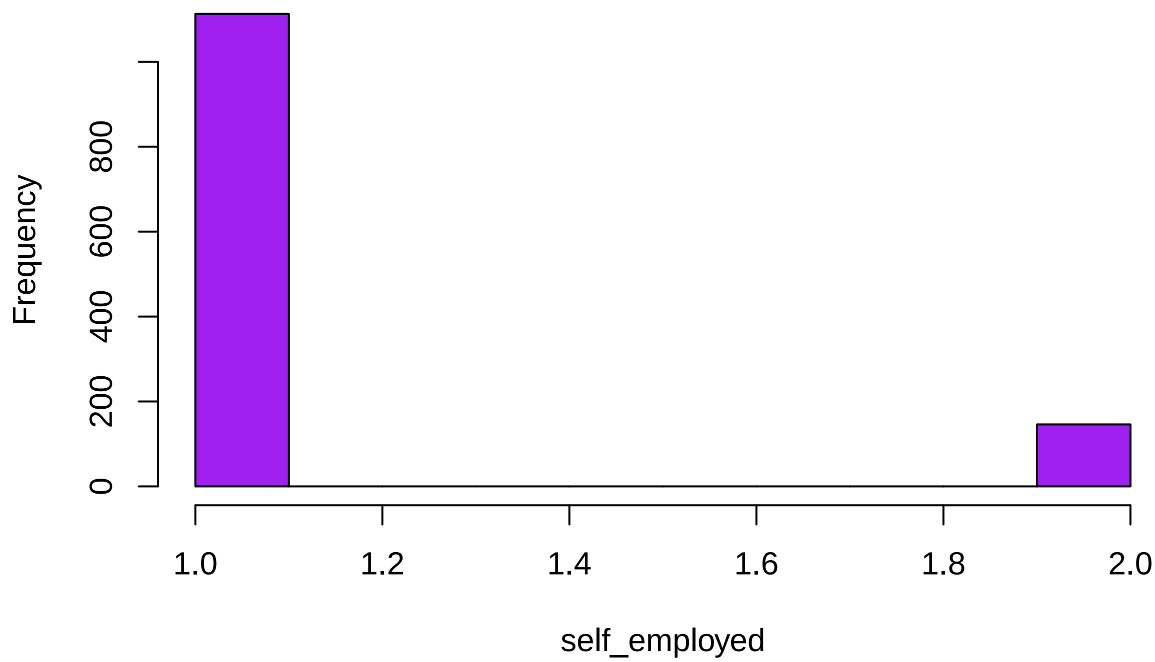
Feature engineering - PCA

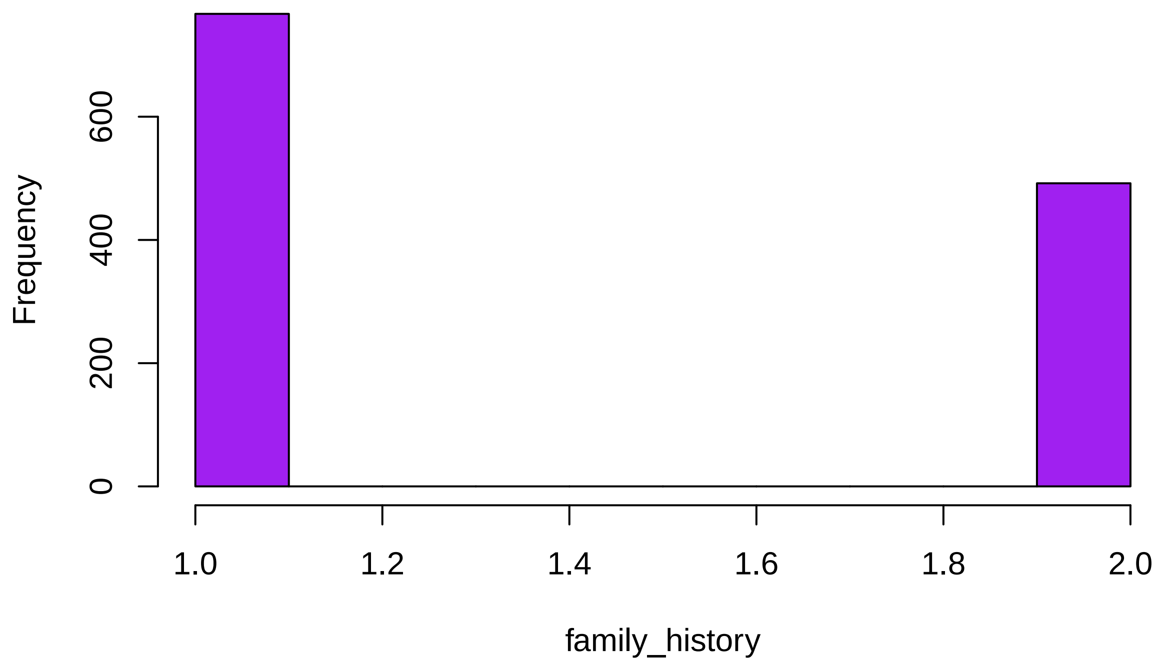
- Principal component analysis is also done using prComp function
- On observing the summary of the principal components, I got to know that reducing the features won't help much because there was very less amount of variance in the principal components
- Principal components are taken into consideration only when the cumulative variance is greater than 85%
- To get the cumulative variance of 85 or greater, I was forced to select 17 components which is almost the same as using 23 components
- Because of this I haven't used Principal components for my models

```
##################################################
### Proper encoding of data for algorithms used ###
##################################################

#Encoding of Age column to 3 types
MH_data$Age <- cut(MH_data$Age, breaks = c(15, 25, 45, 75), labels = c('Fresher', 'Junior', 'Senior'))
MH_data_factors$Age <- as.factor(cut(MH_data_factors$Age, breaks = c(15, 25, 45, 75), labels = c('Fresh

#Using as.numeric will convert the encoded data to dummy codes
MH_data$Age <- as.numeric(MH_data$Age)

#Observing the distribution of the data
for (i in 1:ncol(MH_data)) {
  hist(MH_data[,i],col="purple", xlab = colnames(MH_data[i]), main = NULL)
}
```

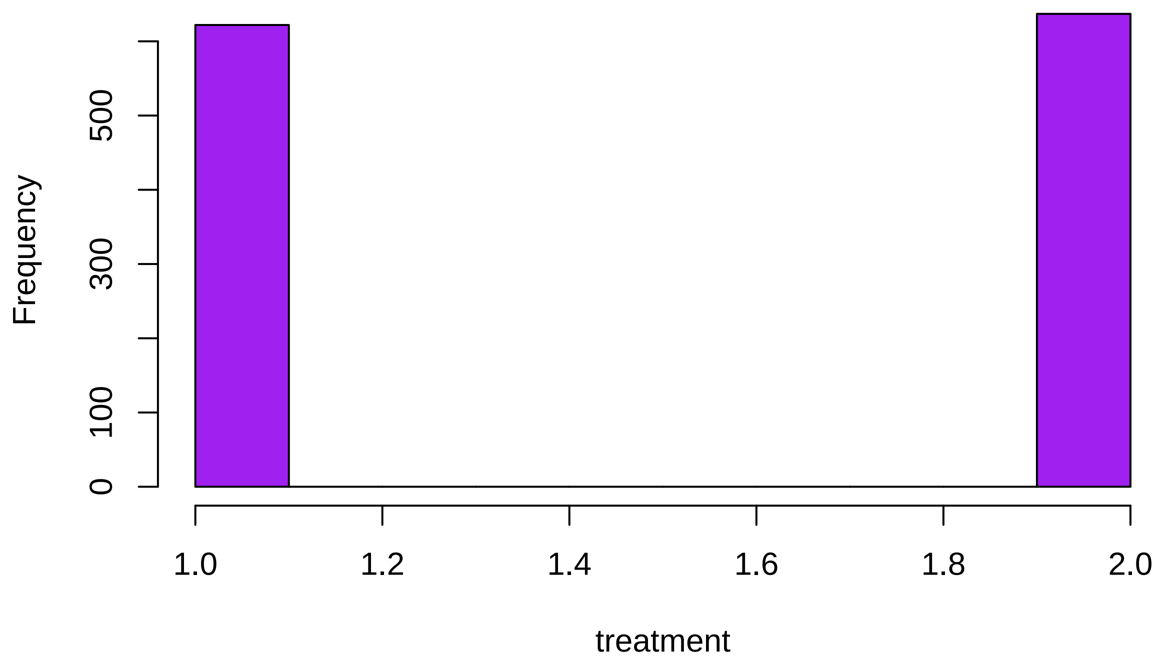

Frequency
0
200
400
600
800
1000
1.0
1.5
2.0
2.5
3.0
Age

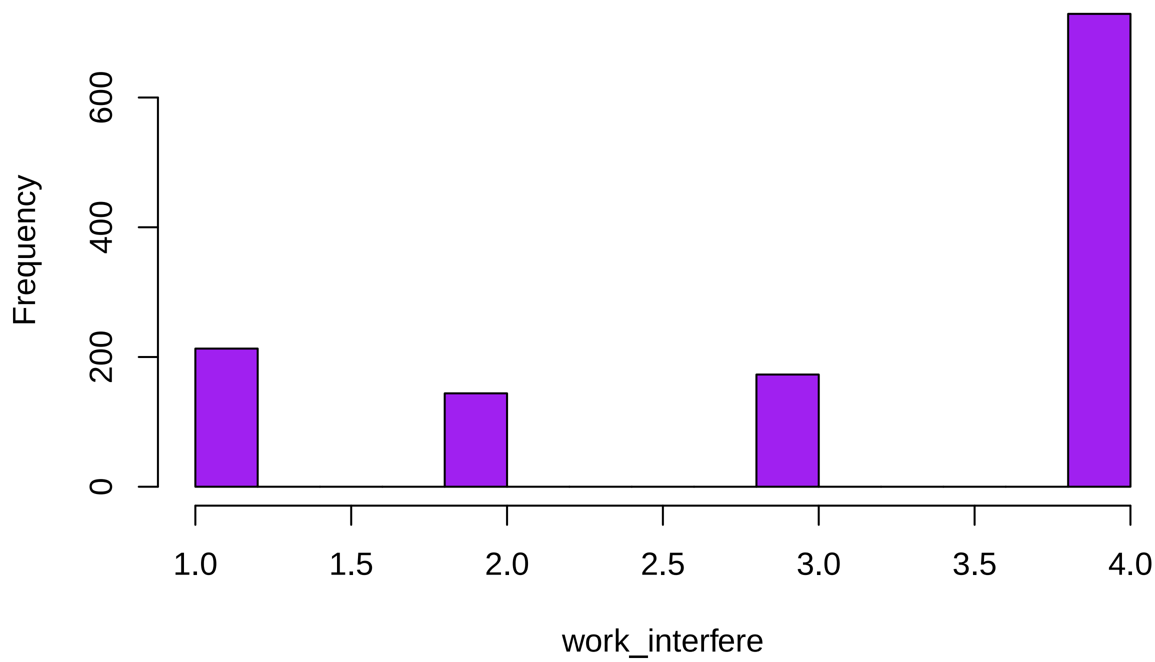

1000
800
600
400
200
0
Frequency
1.0
1.5
2.0
2.5
3.0
Gender

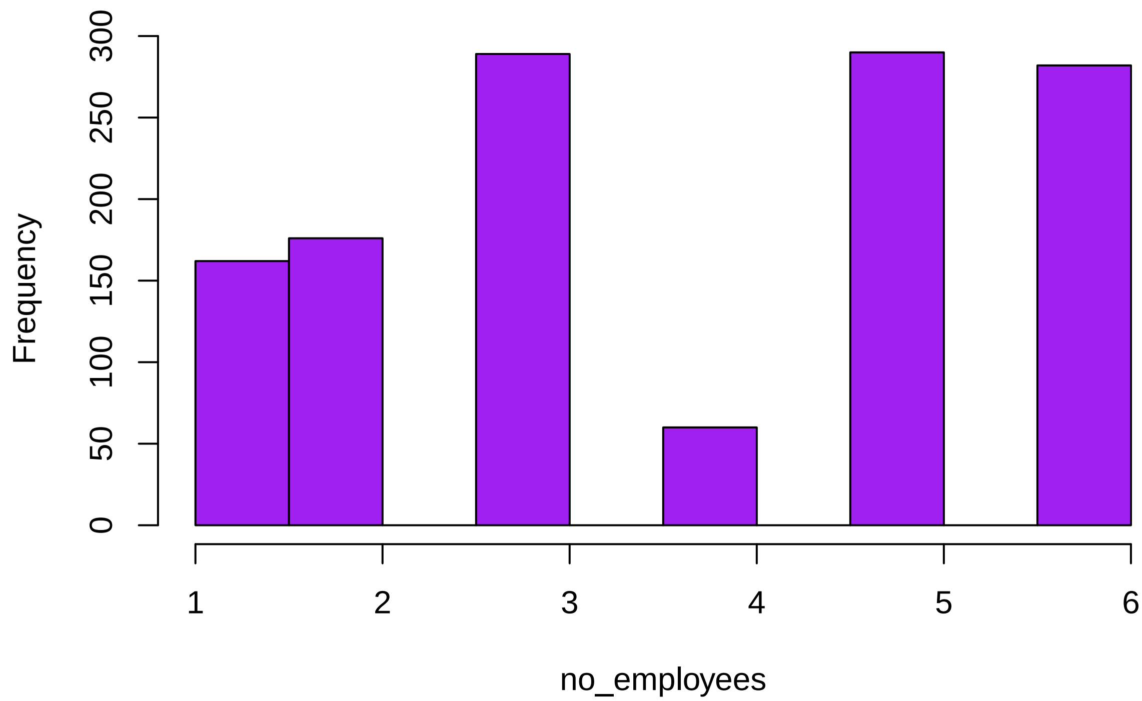

Frequency
0
200
400
600
800
1.0
1.2
1.4
1.6
1.8
2.0
self_employed

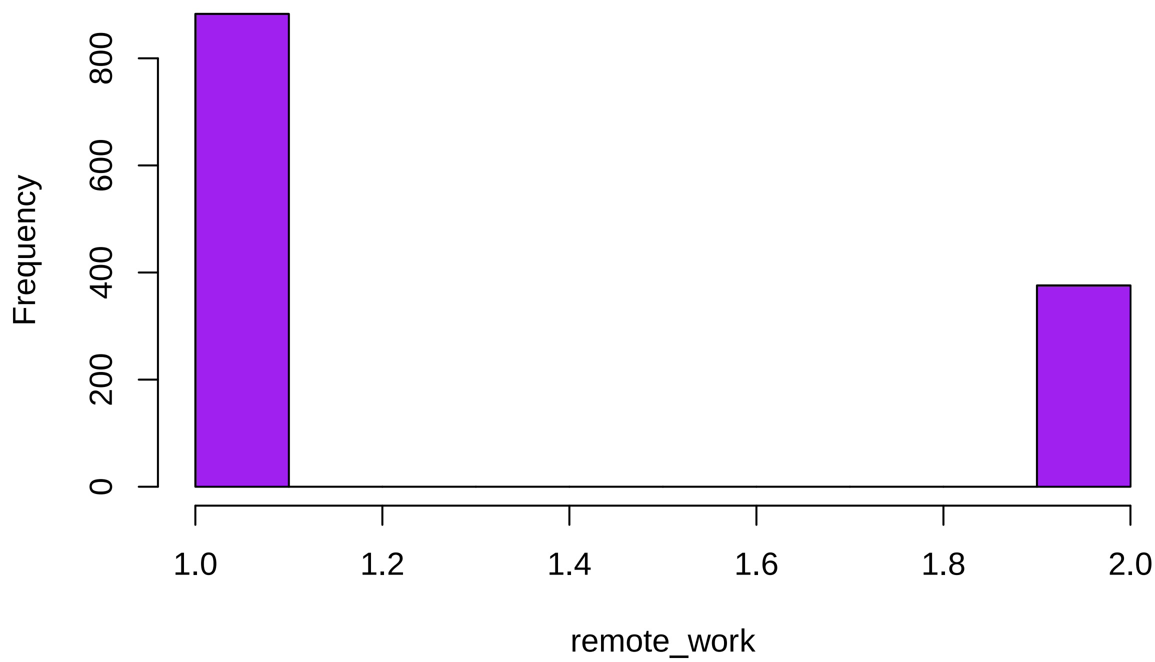


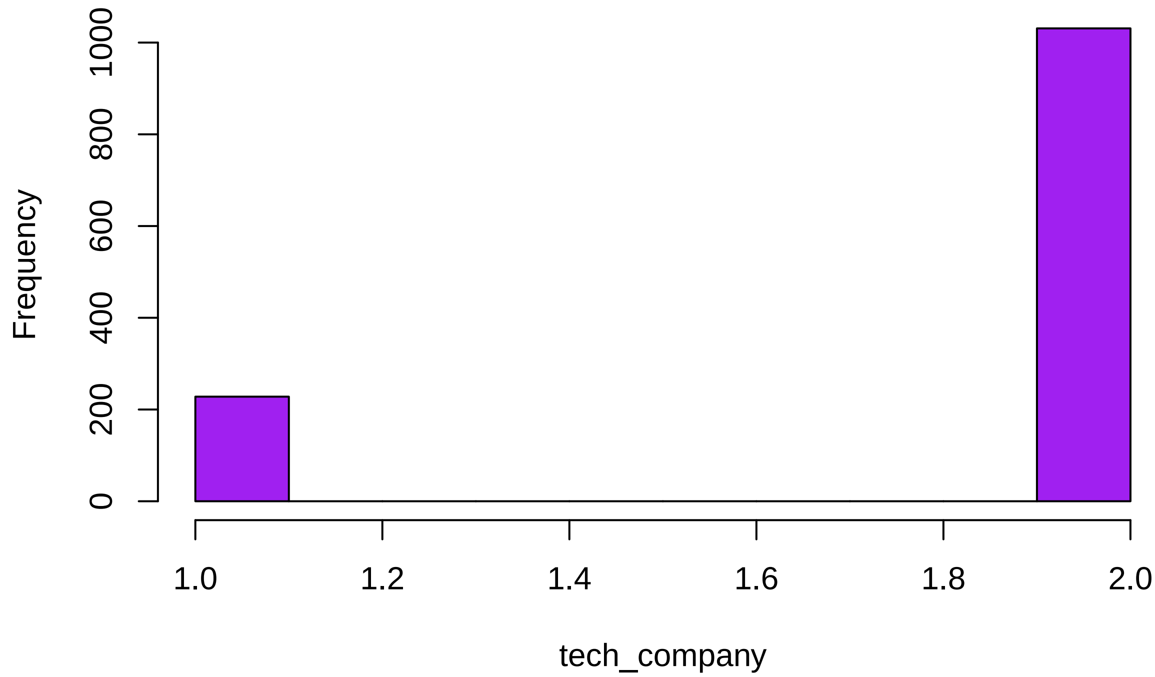
Frequency
0
200
400
600
1.0
1.2
1.4
1.6
1.8
2.0
family_history

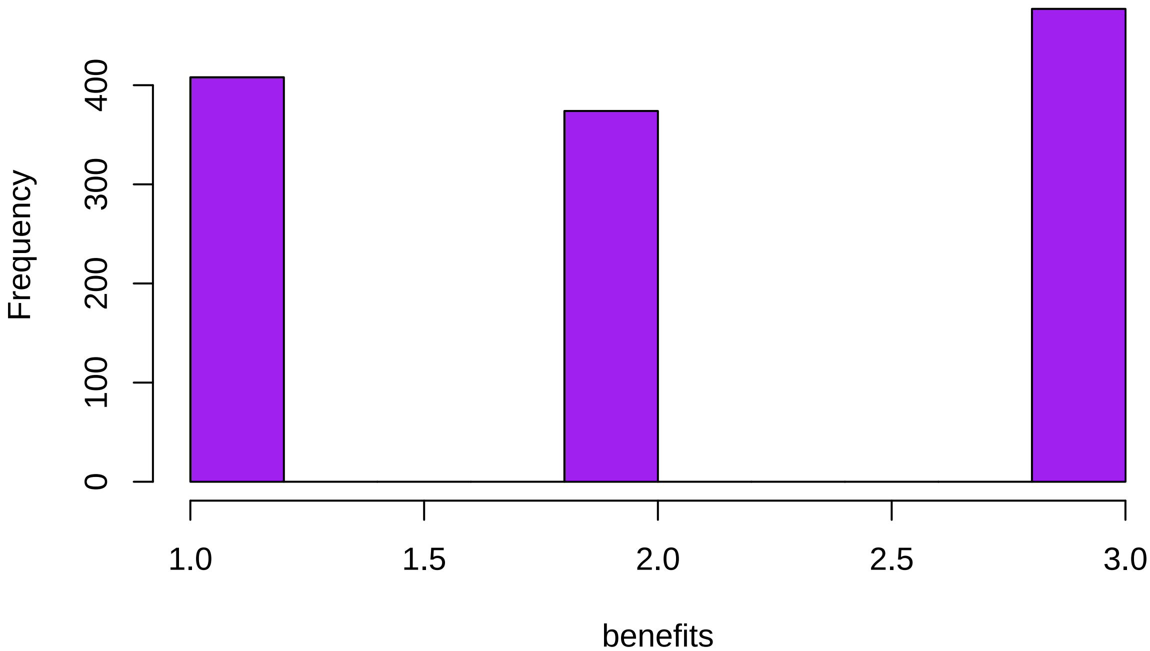

Frequency
0
100
300
500
1.0
1.2
1.4
1.6
1.8
2.0
treatment


Frequency
0
200
400
600
1.0
1.5
2.0
2.5
3.0
3.5
4.0
work_interfere

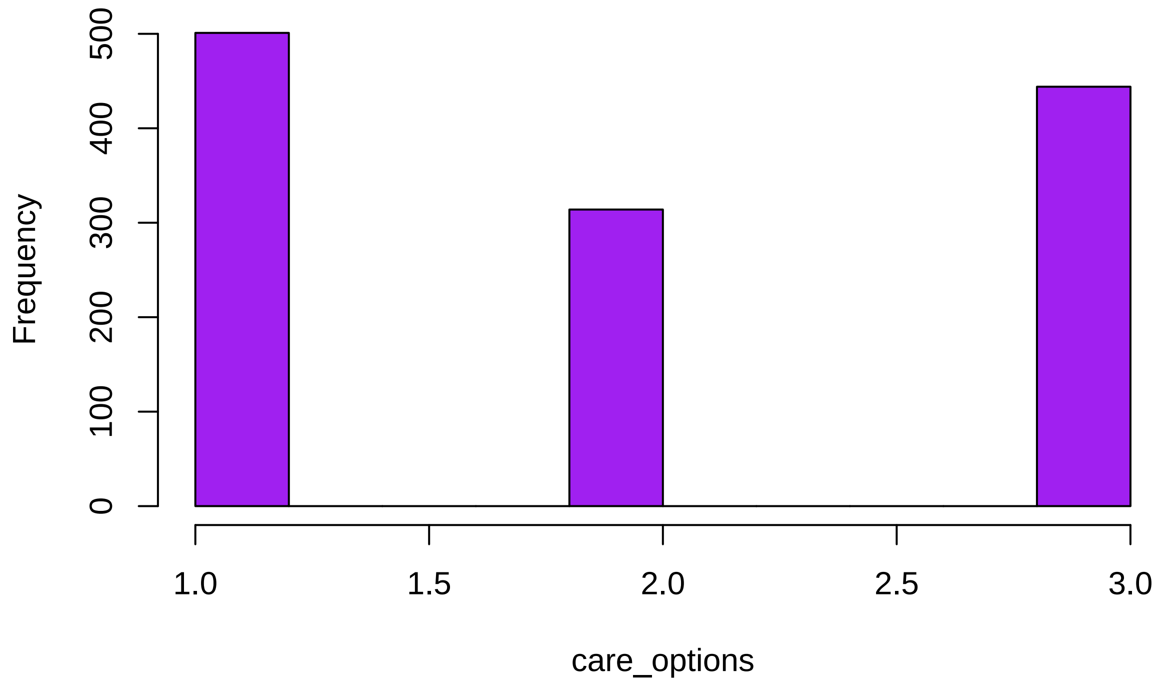

Frequency
0
50
100
150
200
250
300
1
2
3
4
5
6
no_employees

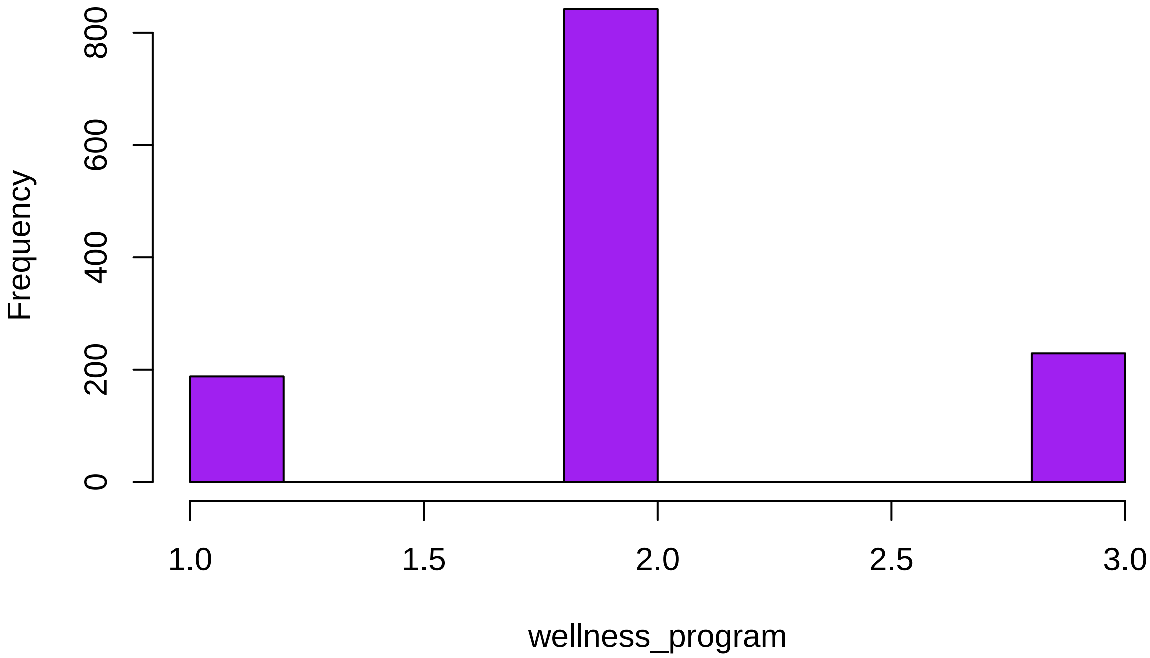

Frequency
0
200
400
600
800
1.0
1.2
1.4
1.6
1.8
2.0
remote_work

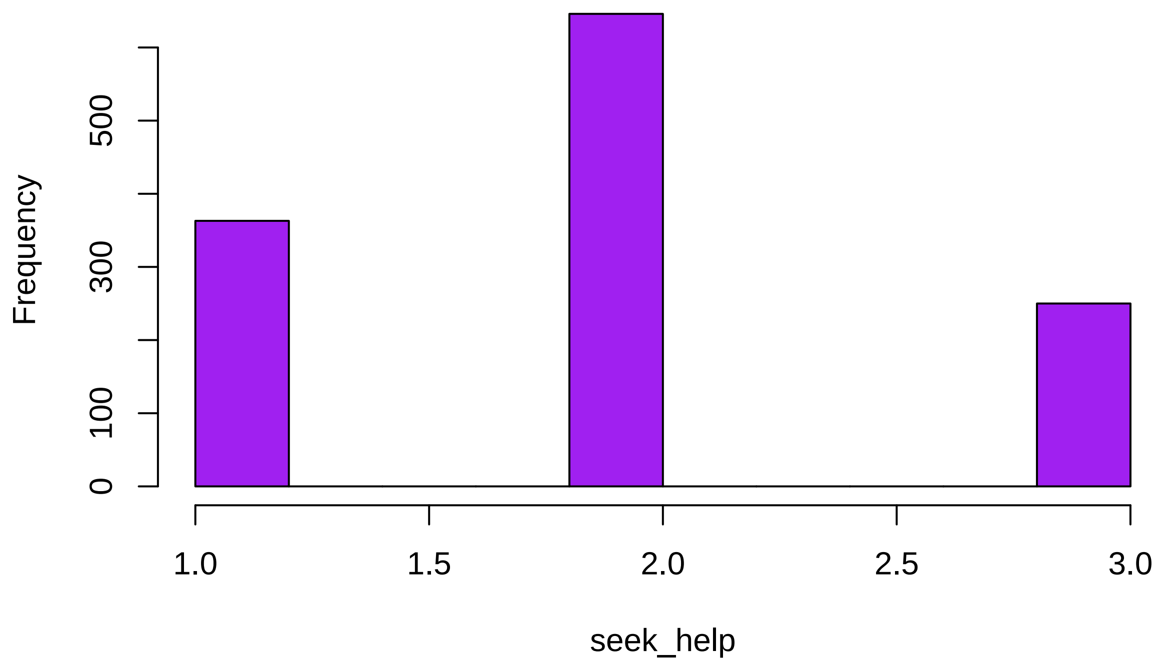

Frequency
0
200
400
600
800
1000
1.0
1.2
1.4
1.6
1.8
2.0
tech_company

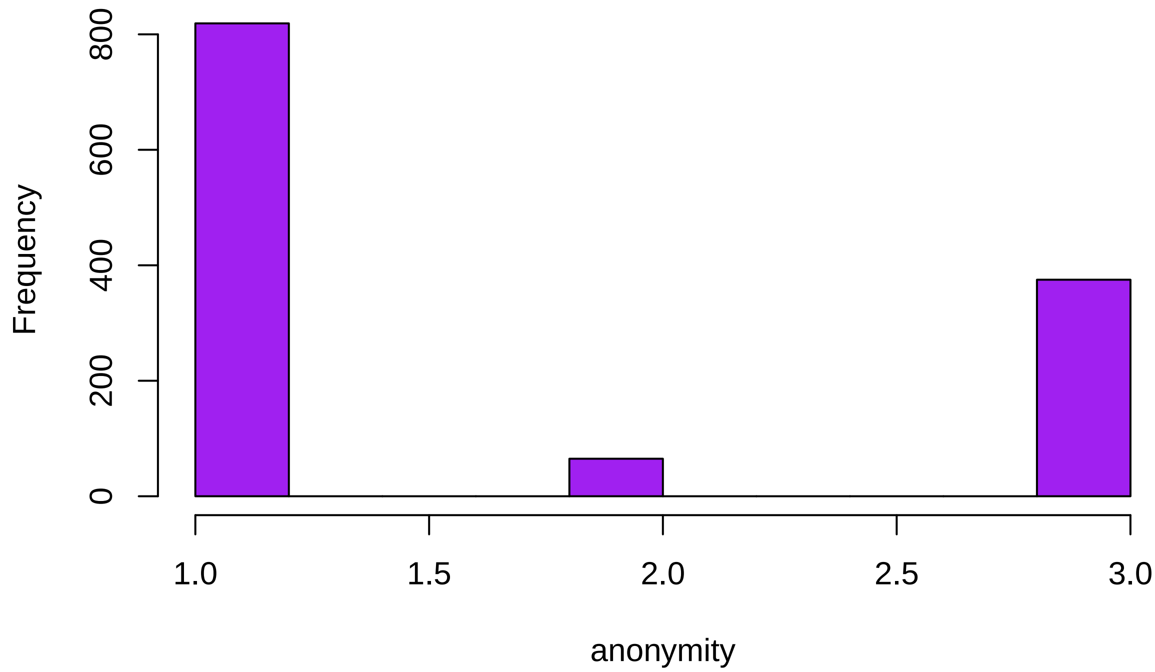


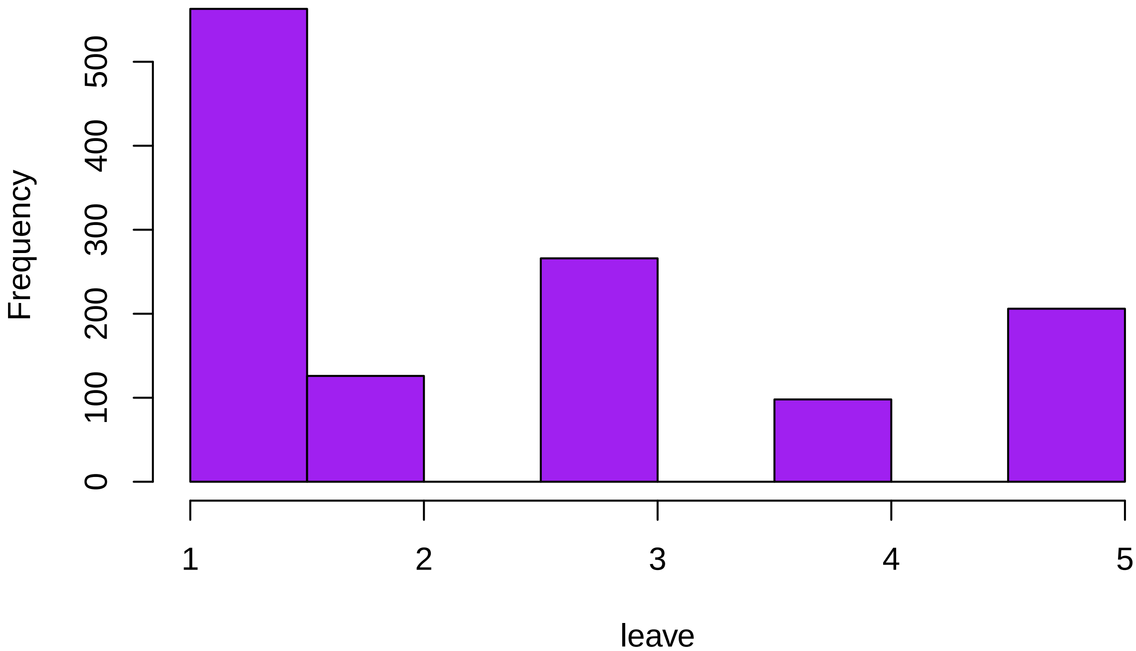
Frequency
0
100
200
300
400
1.0
1.5
2.0
2.5
3.0
benefits

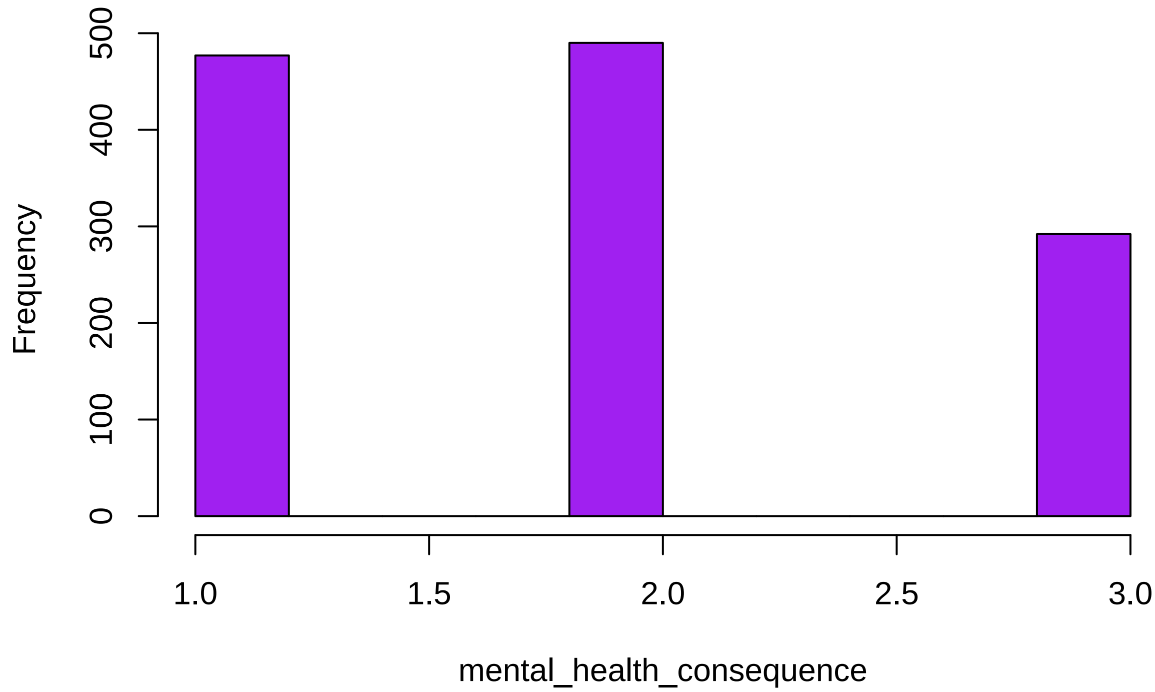

Frequency
0
100
200
300
400
500
1.0
1.5
2.0
2.5
3.0
care_options

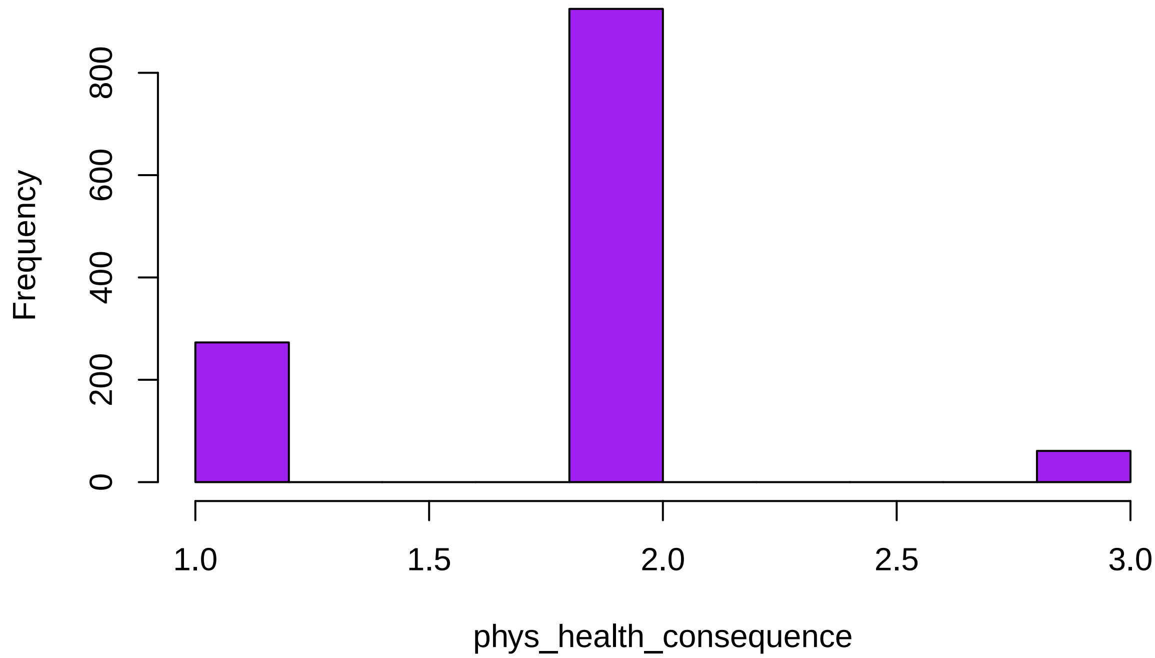

Frequency
0
200
400
600
800
1.0
1.5
2.0
2.5
3.0
wellness_program

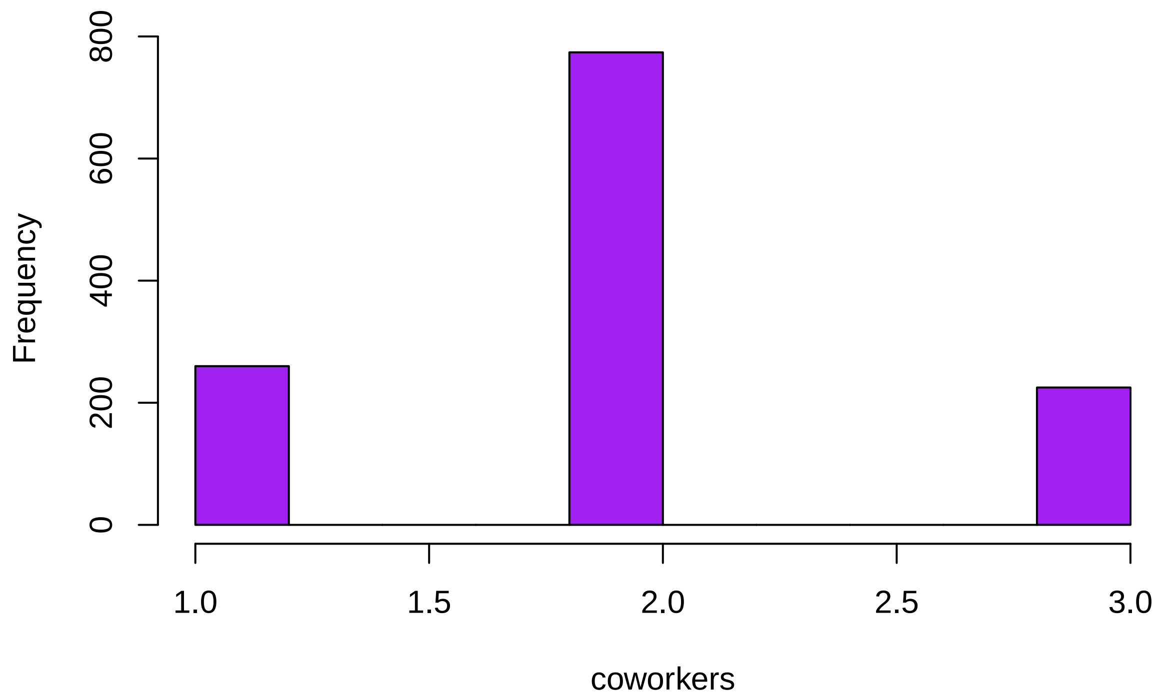

Frequency
0
100
300
500
1.0
1.5
2.0
2.5
3.0
seek_help

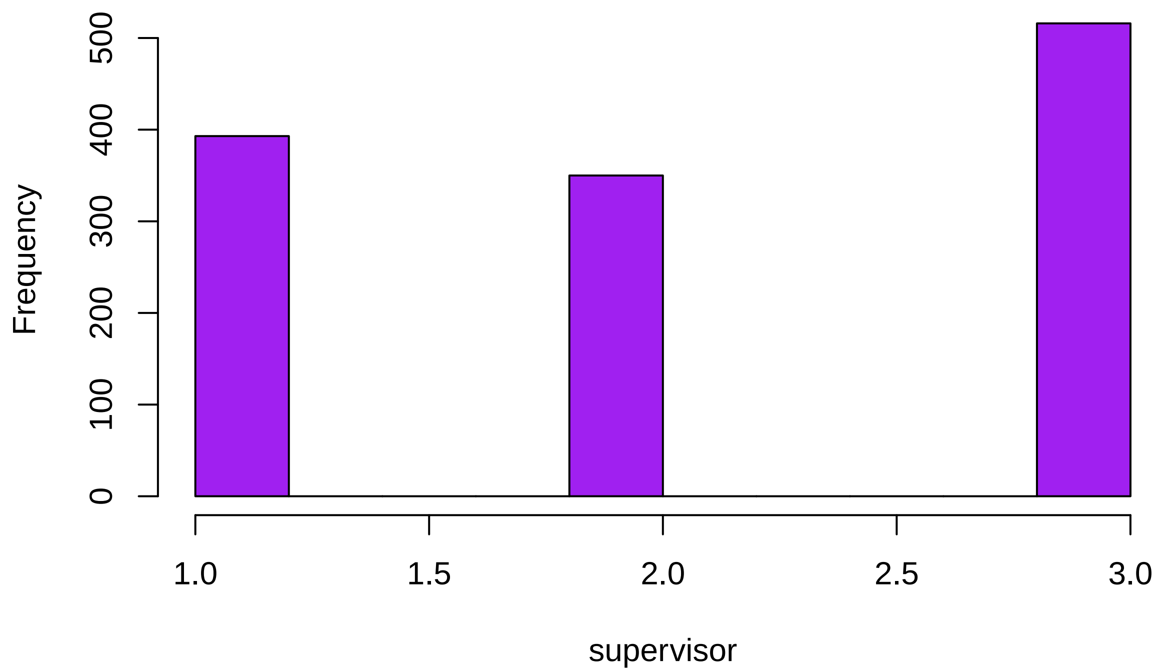

800
600
400
200
0
Frequency
1.0
1.5
2.0
2.5
3.0
anonymity

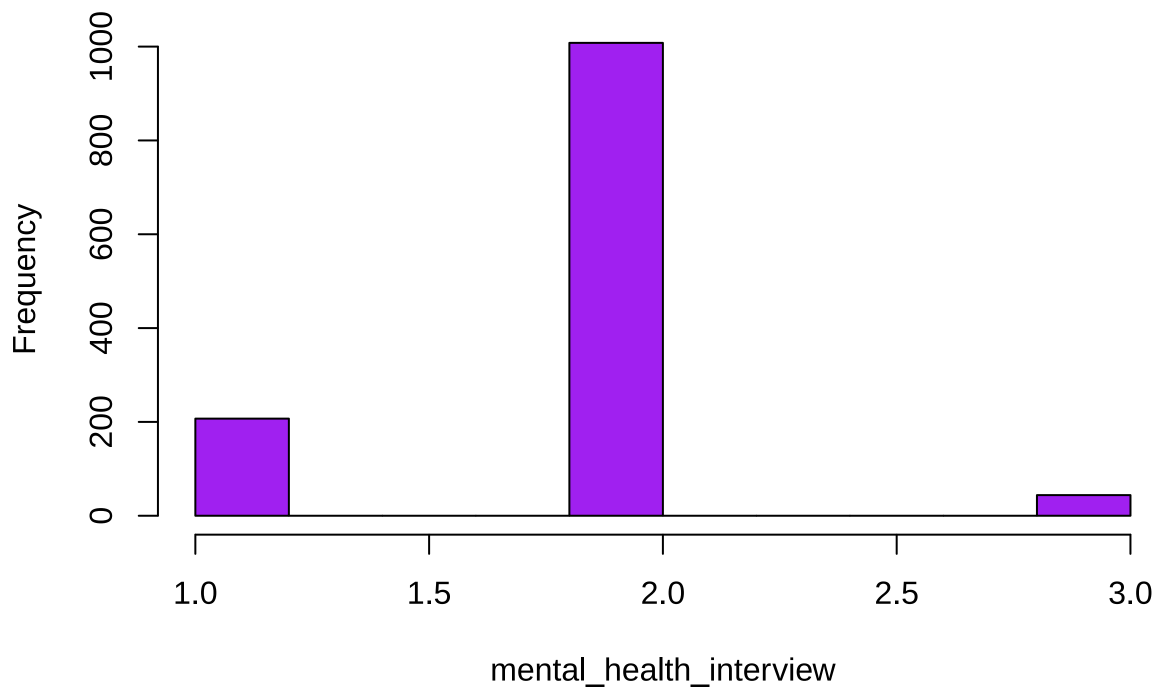

Frequency
0
100
200
300
400
500
1
2
3
4
5
leave


Frequency
0
100
200
300
400
500
1.0
1.5
2.0
2.5
3.0
mental_health_consequence

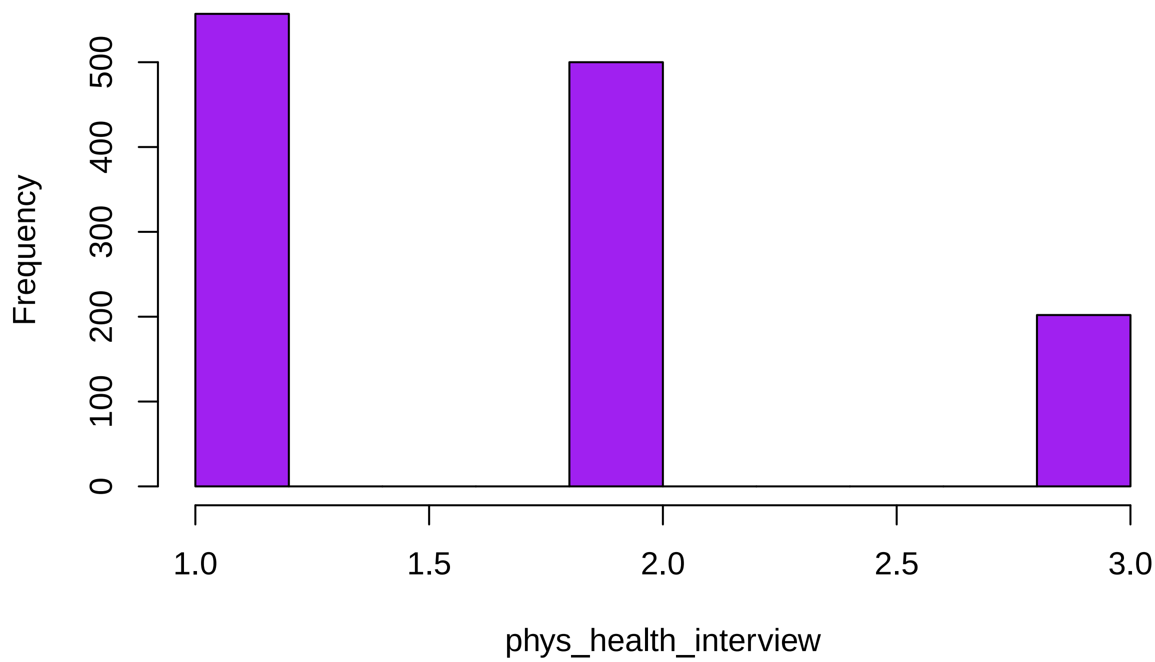

Frequency
0
200
400
600
800
1.0
1.5
2.0
2.5
3.0
phys_health_consequence

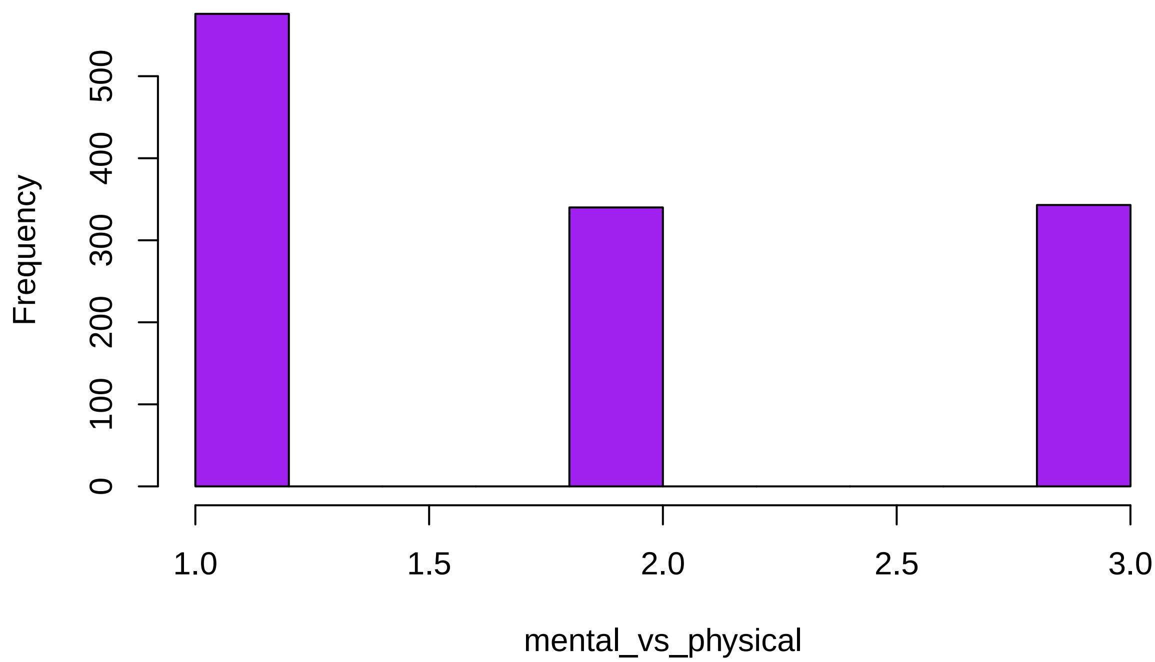

Frequency
0
200
400
600
800
1.0
1.5
2.0
2.5
3.0
coworkers

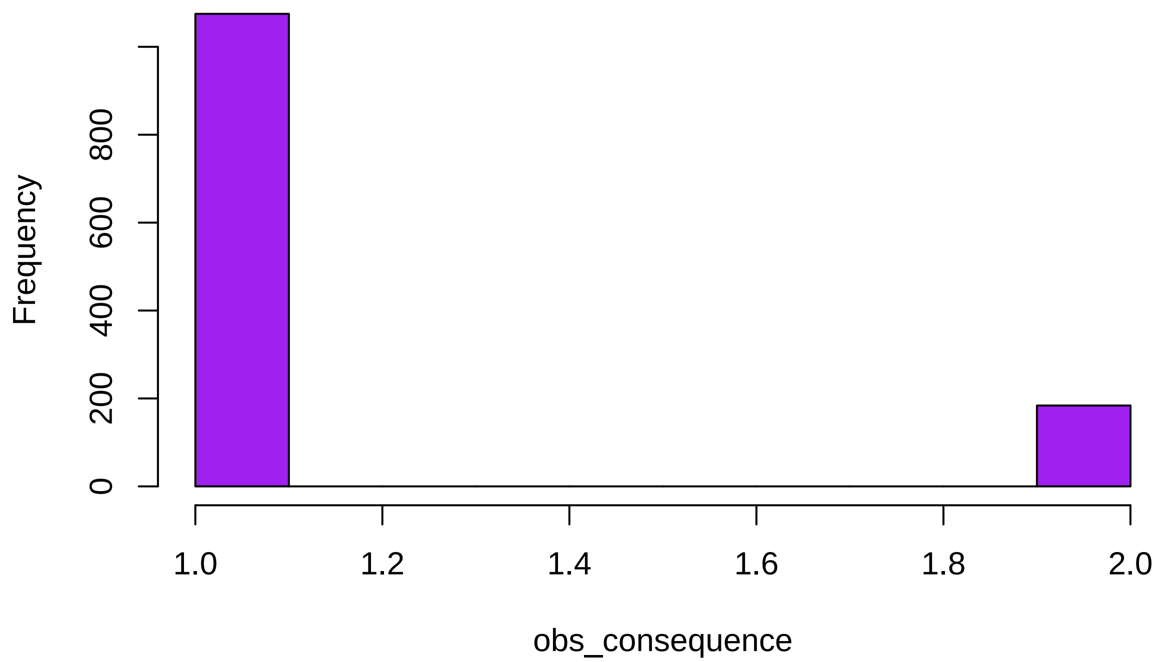

Frequency
0
100
200
300
400
500
1.0
1.5
2.0
2.5
3.0
supervisor



Frequency
0
200
400
600
800
1000
1.0
1.5
2.0
2.5
3.0
mental_health_interview


Frequency
0
100
200
300
400
500
1.0
1.5
2.0
2.5
3.0
phys_health_interview


Frequency
0
100
200
300
400
500
1.0
1.5
2.0
2.5
3.0
mental_vs_physical




```
#########################################################
### Normalization/standardization of feature values ###
#########################################################

#Normalization function
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

#Observing the new structure of the data after encoding Age feature
str(MH_data)
```

```
## 'data.frame':    1259 obs. of  23 variables:
##  $ Age                      : num  2 2 2 2 2 2 2 2 2 1 ...
##  $ Gender                   : num  1 2 2 2 2 2 1 2 1 2 ...
##  $ self_employed            : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ family_history           : num  1 1 1 2 1 2 2 1 2 1 ...
##  $ treatment                : num  2 1 1 2 1 1 2 1 2 1 ...
##  $ work_interfere           : num  2 3 3 2 1 4 4 1 4 1 ...
##  $ no_employees             : num  5 6 5 3 2 5 1 1 2 3 ...
##  $ remote_work              : num  1 1 1 1 2 1 2 2 1 1 ...
##  $ tech_company             : num  2 1 2 2 2 2 2 2 2 2 ...
##  $ benefits                 : num  3 1 2 2 3 3 2 2 3 1 ...
##  $ care_options             : num  2 1 1 3 1 2 1 3 3 1 ...
##  $ wellness_program         : num  2 1 2 2 1 2 2 2 2 1 ...
##  $ seek_help                : num  3 1 2 2 1 1 2 2 2 1 ...
```

```
##  $ anonymity                : num  3 1 1 2 1 1 2 3 2 1 ...
##  $ leave                    : num  3 1 2 2 1 1 2 1 4 1 ...
##  $ mental_health_consequence: num  2 1 2 3 2 2 1 2 1 2 ...
##  $ phys_health_consequence  : num  2 2 2 3 2 2 1 2 2 2 ...
##  $ coworkers                : num  2 1 3 2 2 3 2 1 3 3 ...
##  $ supervisor               : num  3 1 3 1 3 3 1 1 3 3 ...
##  $ mental_health_interview  : num  2 2 3 1 3 2 2 2 2 1 ...
##  $ phys_health_interview    : num  1 2 3 1 3 1 2 2 1 1 ...
##  $ mental_vs_physical       : num  3 1 2 2 1 1 1 2 2 3 ...
##  $ obs_consequence          : num  1 1 1 2 1 1 1 1 1 1 ...
```

```
#Creating a new normalized dataframe
MH_norm <- MH_data

#Normalizing the whole dataset
#Since the data is categorical it makes no sense to use the normalize data
#I tried model evaluation with normalized dataset it made no
#difference so I have just stored it
MH_norm[,-6] <- lapply(MH_data[,-6], normalize)
MH_norm[,6] <- as.factor(MH_data[,6])

################################
### Feature engineering: PCA ###
################################

#Performing PCA on the dataset
MH_PCA <- prcomp(MH_data[,-6], center = T, scale = T)

#Printing Principal components
print(MH_PCA)
```

```
## Standard deviations (1, .., p=22):
##  [1] 1.6840311 1.4596075 1.2851152 1.1837151 1.1109269 1.0815577 1.0226352
##  [8] 1.0067076 0.9501603 0.9483603 0.9290110 0.9135262 0.8784473 0.8523719
## [15] 0.8322443 0.8148475 0.7781787 0.7687332 0.7530433 0.7208846 0.7012437
## [22] 0.6229863
##
## Rotation (n x k) = (22 x 22):
##                                  PC1         PC2         PC3         PC4
## Age                       -0.088424372 -0.03359873  0.10397333 -0.21722259
## Gender                     0.049956472  0.16375107 -0.08221385 -0.34884462
## self_employed             -0.106610427  0.27493933  0.42886360 -0.15110859
## family_history            -0.154282565 -0.11900630  0.22382246  0.47037309
## treatment                 -0.232271270 -0.14321021  0.24244221  0.45000213
## no_employees               0.018418456 -0.27259094 -0.41128912 -0.05814209
## remote_work               -0.022443597  0.25115665  0.38461571 -0.07090230
## tech_company               0.042291755  0.22121691  0.12170552 -0.04996009
## benefits                  -0.375159682 -0.22254475 -0.08789274 -0.01601984
## care_options              -0.365951493 -0.09926413  0.02194666  0.04569248
## wellness_program          -0.318595719 -0.18539511 -0.10351649 -0.23663129
## seek_help                 -0.373950997 -0.12783348 -0.09805561 -0.22914090
## anonymity                 -0.390495085  0.02077367 -0.04152067 -0.10241060
## leave                     -0.257179624  0.26352547  0.05371667 -0.10652540
## mental_health_consequence -0.038633111 -0.08280877  0.26604974 -0.28630524
```

```
## phys_health_consequence   -0.063817887  0.15578805  0.01806694 -0.04645876
## coworkers                 -0.152362760  0.41175933 -0.22545427  0.27285769
## supervisor                -0.180075787  0.40493817 -0.32060865  0.20226311
## mental_health_interview   -0.008661403 -0.24213026  0.20387533  0.02257053
## phys_health_interview     -0.059164846 -0.01383769 -0.02342804  0.18689301
## mental_vs_physical        -0.289356527  0.21677291 -0.04291431 -0.07354014
## obs_consequence           -0.121613434 -0.15825486  0.23643315  0.02478391
##                                   PC5          PC6          PC7           PC8
## Age                        0.21319907 -0.403835101  0.002542382 -0.4459379665
## Gender                    -0.13519578 -0.175185347 -0.141633198 -0.3015686944
## self_employed              0.13208106 -0.096918033 -0.155977120  0.2343986191
## family_history            -0.03216240  0.138391758 -0.102924927 -0.2484304985
## treatment                  0.01325878  0.046171609  0.021619927 -0.2227876190
## no_employees              -0.09061892  0.045731791  0.035236677 -0.1422145393
## remote_work                0.22016956 -0.164311551  0.082384844  0.0155410470
## tech_company               0.04654817 -0.048133260  0.464631372 -0.3533132476
## benefits                   0.18481228 -0.063906923  0.247110820 -0.0827633438
## care_options               0.12913265  0.005439023  0.336610434  0.0479142899
## wellness_program           0.05925042 -0.083894146 -0.233869906 -0.0602956390
## seek_help                  0.02983619 -0.129982938 -0.158285143  0.0002838938
## anonymity                 -0.01230954  0.105090581  0.144472194  0.2517813814
## leave                     -0.15893291  0.187928885 -0.007492097  0.1806981933
## mental_health_consequence -0.40278971  0.290314880 -0.027087927 -0.2494841226
## phys_health_consequence   -0.50090518  0.106011765  0.277652587 -0.2521707149
## coworkers                 -0.01255615 -0.172819132 -0.206735373 -0.1693945411
## supervisor                -0.02758418 -0.126112789 -0.147519409 -0.1070653018
## mental_health_interview   -0.31286786 -0.461729099  0.088429167  0.1840907836
## phys_health_interview     -0.47656788 -0.498889794  0.065223175  0.2410197647
## mental_vs_physical        -0.14369482  0.247982610  0.014944987  0.1089047708
## obs_consequence           -0.13565072  0.069732134 -0.539408200 -0.1412344821
##                                    PC9        PC10         PC11        PC12
## Age                        0.479995490 -0.27648448  0.091115186  0.28525184
## Gender                    -0.452375139 -0.17206559 -0.581646251  0.17909961
## self_employed             -0.029290887  0.11959753 -0.084738539  0.04933878
## family_history            -0.073537708 -0.08935157 -0.340308826 -0.21148221
## treatment                 -0.056864835 -0.14039866 -0.135457486  0.01512637
## no_employees              -0.020792754 -0.41701112  0.134539220 -0.17374590
## remote_work                0.151378309 -0.16444090  0.004394644 -0.43440993
## tech_company              -0.556615755  0.07868287  0.417778248 -0.13865516
## benefits                  -0.030239348  0.04000501  0.007635907  0.03553468
## care_options              -0.009491502  0.06833832 -0.024249128  0.27878799
## wellness_program          -0.020453775  0.26638298 -0.088981860 -0.36634267
## seek_help                 -0.083708474  0.24338578  0.064349869 -0.29500549
## anonymity                 -0.090028465 -0.08226496 -0.117466955  0.19287469
## leave                     -0.058722591 -0.38938901  0.007116102  0.05121143
## mental_health_consequence  0.168725309 -0.13760833  0.095741754 -0.25712331
## phys_health_consequence    0.316882885  0.48081721 -0.144122615  0.13581288
## coworkers                  0.036480785  0.04279296  0.124490839 -0.02477578
## supervisor                 0.063338485  0.04040942  0.050266514 -0.01210642
## mental_health_interview   -0.115093327 -0.06905165  0.025123415  0.11056993
## phys_health_interview     -0.008065951 -0.10553373  0.076594212 -0.17702454
## mental_vs_physical         0.078728188 -0.26936009  0.162454029  0.01389731
## obs_consequence           -0.220395482  0.08897464  0.460414631  0.36321290
##                                   PC13        PC14         PC15        PC16
```

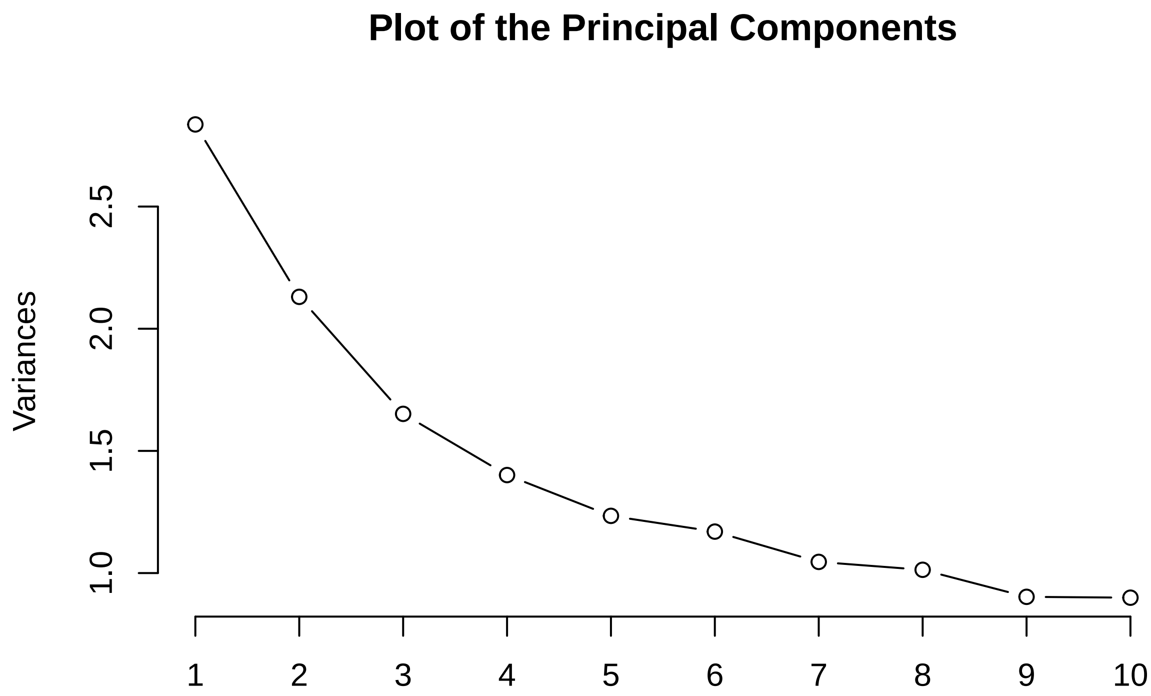
```
## Age                       -0.241065776  0.14378105 -0.144924841  0.08258944
## Gender                     0.173898125  0.05847802  0.019151656 -0.11515250
## self_employed              0.002991831  0.18362926  0.147902829 -0.09882234
## family_history            -0.188274452  0.22171619 -0.007997542  0.01554859
## treatment                 -0.026519198 -0.11585663  0.045908262  0.02650610
## no_employees               0.187790546  0.10417927  0.501776729  0.13824005
## remote_work                0.394734145 -0.02402717  0.365200248  0.11113786
## tech_company              -0.203443148  0.07389012 -0.055330826  0.04223887
## benefits                   0.175181173  0.04894260 -0.071123194 -0.10131032
## care_options               0.379874975 -0.15579368 -0.060979541 -0.10646713
## wellness_program          -0.254952893 -0.08803941  0.024625576  0.05288517
## seek_help                 -0.063422034  0.06158091 -0.037927058  0.09304876
## anonymity                  0.047077570  0.04035107  0.033829560  0.08203857
## leave                     -0.211540448 -0.21391318 -0.146776340  0.61311847
## mental_health_consequence  0.235387385 -0.32151553 -0.342161271 -0.27597016
## phys_health_consequence   -0.006492570  0.13492102  0.338960103  0.23563915
## coworkers                  0.080921120 -0.29011570  0.061898824 -0.09010151
## supervisor                 0.078354816 -0.18044548 -0.055213038 -0.10792525
## mental_health_interview   -0.299043144 -0.48285917  0.337383736 -0.18369513
## phys_health_interview      0.205559758  0.42372548 -0.343778826  0.05568941
## mental_vs_physical        -0.302523980  0.32705089  0.206461489 -0.56123328
## obs_consequence            0.246365404  0.13840892  0.139193210  0.13032505
##                                  PC17         PC18          PC19         PC20
## Age                        0.08224499 -0.047095148  0.0413030034  0.077849123
## Gender                    -0.09228119  0.017208288 -0.1441294609 -0.026705074
## self_employed             -0.11892520  0.100119767  0.6890803903 -0.098709752
## family_history             0.48764358  0.272485920  0.0528790564 -0.034916239
## treatment                 -0.57490634 -0.399457441  0.0388438942  0.059647792
## no_employees              -0.02111511  0.062923713  0.4118959267  0.033063221
## remote_work                0.09068147 -0.033113651 -0.3756482918  0.059394265
## tech_company               0.05453016 -0.028684370  0.0942034650  0.112284149
## benefits                  -0.00476076  0.104696316 -0.0037655771 -0.541184665
## care_options              -0.06770215  0.445727319 -0.0042909721  0.312447582
## wellness_program          -0.17791105  0.144302267  0.0003616504  0.519776051
## seek_help                  0.02314730 -0.154021372 -0.0967008736 -0.399719034
## anonymity                  0.43125867 -0.621302380  0.0800209751  0.245445452
## leave                     -0.11643643  0.269975133 -0.0489793523 -0.140323455
## mental_health_consequence  0.09140323 -0.075035806  0.1999712469 -0.018635461
## phys_health_consequence   -0.04535201  0.006318424 -0.0235847090 -0.051760164
## coworkers                 -0.07556283 -0.095051460  0.0761625784 -0.143219612
## supervisor                 0.25362356  0.072134453  0.1076910340  0.119971904
## mental_health_interview    0.17088677  0.068786936 -0.0363508735 -0.086764964
## phys_health_interview     -0.11311300  0.007592753  0.0112873862  0.111777312
## mental_vs_physical        -0.15575200  0.086783397 -0.2712565606  0.007959977
## obs_consequence            0.08708197  0.049560248 -0.1660779449  0.055488982
##                                   PC21         PC22
## Age                       -0.070045555 -0.029623255
## Gender                    -0.006039175  0.002438286
## self_employed              0.007598786  0.035045245
## family_history            -0.070787816 -0.135205894
## treatment                 -0.057232564  0.234449304
## no_employees              -0.079120568 -0.001146926
## remote_work                0.065769752  0.043257613
## tech_company              -0.003601179 -0.002884185
```

```
## benefits                     0.580529669  0.049582483
## care_options                -0.376872121 -0.108245858
## wellness_program             0.329132497 -0.088919901
## seek_help                   -0.600331184  0.098754775
## anonymity                    0.083729849 -0.127119066
## leave                        0.049789435  0.013081121
## mental_health_consequence    0.010194496 -0.016660088
## phys_health_consequence      0.035859356  0.015716707
## coworkers                    0.011015032 -0.646052111
## supervisor                   0.049187479  0.671084620
## mental_health_interview      0.005864944  0.050924550
## phys_health_interview        0.054297802 -0.050650854
## mental_vs_physical          -0.030332616 -0.004703564
## obs_consequence              0.104967888  0.033937527
```

```r
#Summary of Principal components
summary(MH_PCA)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4    PC5     PC6     PC7
## Standard deviation     1.6840 1.45961 1.28512 1.18372 1.1109 1.08156 1.02264
## Proportion of Variance 0.1289 0.09684 0.07507 0.06369 0.0561 0.05317 0.04754
## Cumulative Proportion  0.1289 0.22575 0.30082 0.36451 0.4206 0.47377 0.52131
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.00671 0.95016 0.94836 0.92901 0.91353 0.87845 0.85237
## Proportion of Variance 0.04607 0.04104 0.04088 0.03923 0.03793 0.03508 0.03302
## Cumulative Proportion  0.56738 0.60841 0.64929 0.68852 0.72646 0.76153 0.79456
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.83224 0.81485 0.77818 0.76873 0.75304 0.72088 0.70124
## Proportion of Variance 0.03148 0.03018 0.02753 0.02686 0.02578 0.02362 0.02235
## Cumulative Proportion  0.82604 0.85622 0.88375 0.91061 0.93639 0.96001 0.98236
##                           PC22
## Standard deviation     0.62299
## Proportion of Variance 0.01764
## Cumulative Proportion  1.00000
```

```r
#Plotting variance plot of the Principal components
screeplot(MH_PCA, type = "l", main = "Plot of the Principal Components")
```




```
#Based on the summary we can see that there is not much of a variance present.
#It is advisable to use PCA when the cumulative proportion is above 85%
#On observing the cumulative proportion we see that, a total of 17 components will
#be needed to make up 88% of the data which makes no sense because we will be reducing only 5 features
#Reducing the features won't increase the efficiency of the models based on  these components
#Hence I will not be using these principal components for evaluation of my models

write.table(MH_data,"shinyData.csv", sep = ",",col.names = !file.exists("shinyData.csv"), append = T,ro
```

4. Model Construction & Evaluation

Creation of training & validation subsets

- Data splitting is done in 75:25 ratio
- Partition is created using createDataPartition function

Construction of at least three related models

- I built 4 models which are as follows:
  - Logistic Regression (glm)
  - Neural Network (neuralnet)
  - Support Vector Machine (ksvm)
  - Recursive Partitioning - Decision Trees (rpart)

- For Neural Network model, I have used the numeric dataset whereas for all other models factor dataset is used

Evaluation of fit of models with holdout method

- For model evaluation I have created two functions, mean absolute error(MAE) and root mean squared error(RMSE)
- Along with that, I have calculated accuracy of each model using the confusionMatrix function
- I have also calculated AUC for each model

```
#Function for evaluating mean absolute error
MAE <- function(actual, predicted)
{
  mean(abs(actual - predicted))
}

#Function for evaluating root mean squared error
RMSE <- function(actual, pred)
{
  return(sqrt(sum((actual-pred)^2)/length(actual)))
}
```

```
##################################################
### Creation of training & validation subsets ###
##################################################

#Converting the predictor variable to factor
MH_data$treatment <- as.factor(MH_data$treatment)

#Setting the seed values for randomness
set.seed(101)

#Splitting the dataset into 75:25 ratio
index <- createDataPartition(MH_data$treatment, p=0.75, list = FALSE, times = 1)
#Using numeric dataset for Neural Network
training_data <- MH_data[index, ]
testing_data <- MH_data[-index, ]
#Using categorical dataset for glm,SVM and rpart
training_data_factor <- MH_data_factors[index, ]
testing_data_factor <- MH_data_factors[-index, ]
```

```
###########################
### Logistic Regression ###
###########################

#Building the logistic regression model using glm function
lm <- glm( treatment~., data = training_data_factor, family = "binomial" )

#Observing the summary of the model
summary(lm)
```

```
##
```

```
## Call:
## glm(formula = treatment ~ ., family = "binomial", data = training_data_factor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8239  -0.8090   0.1860   0.8026   2.6723
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -3.24310    0.67137  -4.831 1.36e-06 ***
## AgeJunior                     0.26337    0.22208   1.186 0.235654
## AgeSenior                     0.99091    0.49030   2.021 0.043277 *
## Gendermale                   -0.82016    0.21261  -3.858 0.000114 ***
## Genderqueer                  -0.13513    0.81308  -0.166 0.868008
## self_employedYes             -0.25971    0.31561  -0.823 0.410569
## family_historyYes             1.28243    0.17086   7.506 6.10e-14 ***
## work_interfereOften           3.18225    0.39499   8.057 7.84e-16 ***
## work_interfereRarely          2.25511    0.33019   6.830 8.51e-12 ***
## work_interfereSometimes       1.56644    0.26753   5.855 4.77e-09 ***
## no_employees100-500           0.14827    0.37130   0.399 0.689655
## no_employees26-100            0.19910    0.33381   0.596 0.550881
## no_employees500-1000         -0.60993    0.49880  -1.223 0.221406
## no_employees6-25             -0.03684    0.31411  -0.117 0.906632
## no_employeesMore than 1000   -0.11286    0.37633  -0.300 0.764262
## remote_workYes                0.13436    0.19302   0.696 0.486373
## tech_companyYes               0.24236    0.22514   1.076 0.281706
## benefitsNo                    0.04840    0.25441   0.190 0.849107
## benefitsYes                   0.37003    0.24989   1.481 0.138656
## care_optionsNot sure         -0.15895    0.22409  -0.709 0.478132
## care_optionsYes               0.63793    0.22564   2.827 0.004695 **
## wellness_programNo            0.16357    0.27837   0.588 0.556790
## wellness_programYes          -0.16549    0.33351  -0.496 0.619746
## seek_helpNo                  -0.47806    0.24336  -1.964 0.049477 *
## seek_helpYes                 -0.17256    0.29880  -0.578 0.563582
## anonymityNo                  -0.35534    0.39184  -0.907 0.364487
## anonymityYes                  0.55958    0.21903   2.555 0.010624 *
## leaveSomewhat difficult       0.56852    0.29465   1.930 0.053668 .
## leaveSomewhat easy            0.11588    0.22490   0.515 0.606363
## leaveVery difficult           0.79807    0.36532   2.185 0.028920 *
## leaveVery easy                0.15056    0.25560   0.589 0.555825
## mental_health_consequenceNo  -0.47945    0.22841  -2.099 0.035810 *
## mental_health_consequenceYes  0.10068    0.26054   0.386 0.699181
## phys_health_consequenceNo     0.16044    0.22275   0.720 0.471362
## phys_health_consequenceYes    0.19622    0.44663   0.439 0.660424
## coworkersSome of them         0.59239    0.23430   2.528 0.011460 *
## coworkersYes                  1.17566    0.33516   3.508 0.000452 ***
## supervisorSome of them       -0.07965    0.23608  -0.337 0.735822
## supervisorYes                -0.28324    0.26485  -1.069 0.284875
## mental_health_interviewNo     0.52649    0.26303   2.002 0.045321 *
## mental_health_interviewYes    0.71852    0.57126   1.258 0.208470
## phys_health_interviewNo       0.09658    0.18895   0.511 0.609276
## phys_health_interviewYes      0.17631    0.27594   0.639 0.522861
## mental_vs_physicalNo         -0.03356    0.21925  -0.153 0.878329
## mental_vs_physicalYes        -0.03553    0.22877  -0.155 0.876593
```

```
## obs_consequenceYes            0.37952    0.25316   1.499 0.133848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1309.92  on 944  degrees of freedom
## Residual deviance:  945.08  on 899  degrees of freedom
## AIC: 1037.1
##
## Number of Fisher Scoring iterations: 5
```

```
#Predicting the output for testing dataset
predict_prob <- predict(lm, testing_data_factor, type = "response")

#Since we recieve output as probability values we convert it
pred_glm <- as.factor(ifelse(predict_prob < 0.5, "No", "Yes"))

#Model evaluation using confusionMatrix, Accuracy, RMSE, MAE, and AUC
confusionMatrix(pred_glm,testing_data_factor$treatment)
```

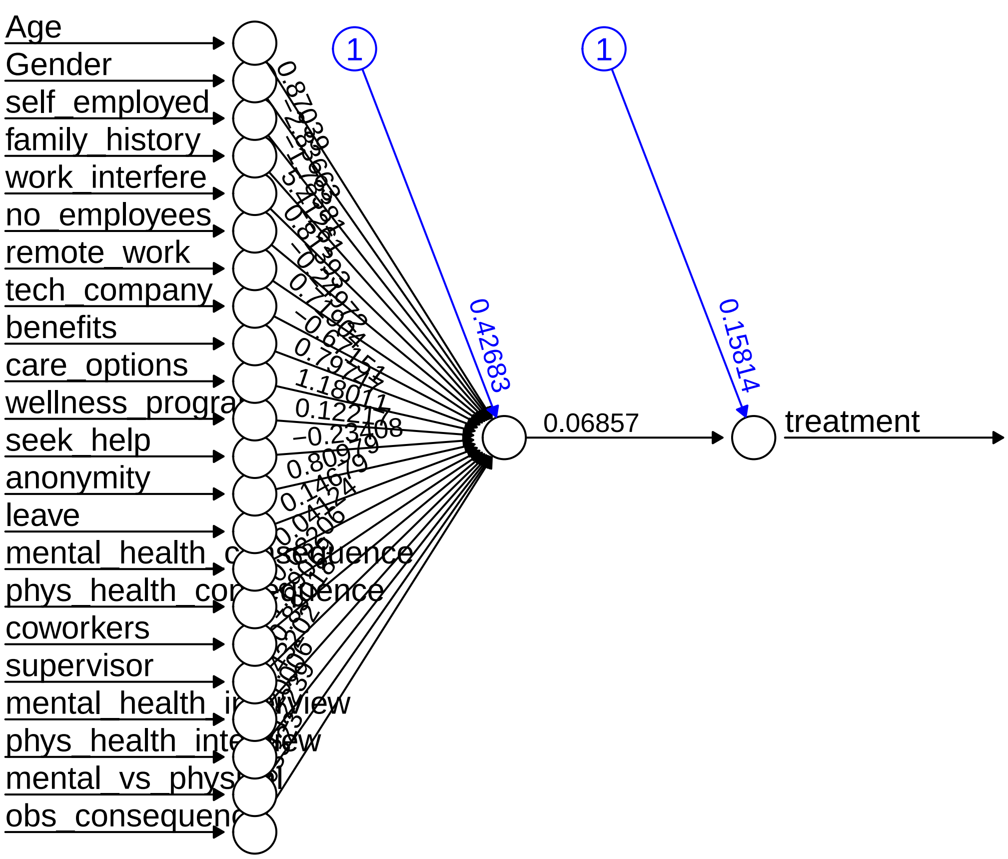
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  123  35
##        Yes  32 124
##
##                Accuracy : 0.7866
##                  95% CI : (0.7371, 0.8306)
##     No Information Rate : 0.5064
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.5733
##
##  Mcnemar's Test P-Value : 0.807
##
##             Sensitivity : 0.7935
##             Specificity : 0.7799
##          Pos Pred Value : 0.7785
##          Neg Pred Value : 0.7949
##              Prevalence : 0.4936
##          Detection Rate : 0.3917
##    Detection Prevalence : 0.5032
##       Balanced Accuracy : 0.7867
##
##        'Positive' Class : No
##
```

```
accuracy_glm <- accuracy(pred_glm,testing_data_factor$treatment)
RMSE_glm <- RMSE(as.numeric(testing_data_factor$treatment), as.numeric(pred_glm))
MAE_glm <- MAE(as.numeric(testing_data_factor$treatment), as.numeric(pred_glm))
roc_glm <- roc(as.numeric(testing_data_factor$treatment), as.numeric(pred_glm))
```

```r
#Converting predictor feature to numeric for neural networks
training_data$treatment <- as.numeric(training_data$treatment)
testing_data$treatment <- as.numeric(testing_data$treatment)

######################
### Neural Network ###
######################

#Building neural network model with 1 hidden layer along with softplus function
softplus <- function(x) log(1+exp(x))
neuralnet_model <- neuralnet(treatment~., data = training_data,stepmax=1e+08,threshold = 0.5,rep = 1,li

#Using compute() function to predict the outcome of testing dataset
nn_predictions <- compute(neuralnet_model, testing_data[,-6])
net_results <- nn_predictions$net.result

#Checking the correlation of both predictor and predicted values
cor(net_results,as.numeric(testing_data$treatment))
```

```
##            [,1]
## [1,] 0.5601729
```

```r
#Plotting the neural network
plot(neuralnet_model, rep="best")
```



```r
#Converting the numeric prediction to category
pred_nn <- net_results
pred_nn <- as.factor(ifelse(pred_nn>1.5, 2, 1))

#Model evaluation using confusionMatrix, Accuracy, RMSE, MAE, and AUC
confusionMatrix(pred_nn,as.factor(testing_data$treatment))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2
##          1 135  62
##          2  20  97
##
##                Accuracy : 0.7389
##                  95% CI : (0.6866, 0.7866)
##     No Information Rate : 0.5064
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4794
##
##  Mcnemar's Test P-Value : 5.963e-06
##
##             Sensitivity : 0.8710
##             Specificity : 0.6101
##          Pos Pred Value : 0.6853
##          Neg Pred Value : 0.8291
##              Prevalence : 0.4936
##          Detection Rate : 0.4299
##    Detection Prevalence : 0.6274
##       Balanced Accuracy : 0.7405
##
##        'Positive' Class : 1
##
```

```r
accuracy_nn <- accuracy(pred_nn,as.factor(testing_data$treatment))
RMSE_nn <- RMSE(as.numeric(testing_data$treatment), as.numeric(pred_nn))
MAE_nn <- MAE(as.numeric(testing_data$treatment), as.numeric(pred_nn))
roc_nn <- roc(as.numeric(testing_data$treatment), as.numeric(pred_nn))
```

```r
##############################
### Support Vector Machine ###
##############################

#Building SVM model with categorical data
svm_model <- ksvm(treatment ~ ., data = training_data_factor,prob.model=TRUE,kernel="rbfdot")

#Predicting outcome of the testing dataset
pred_svm <- predict(svm_model, testing_data_factor)

#Observing first few predictions
head(pred_svm)
```

```
## [1] Yes Yes No  Yes Yes Yes
## Levels: No Yes
```

```
#pred_svm <- as.factor(ifelse(pred_svm>1.5, 2, 1))

#Model evaluation using confusionMatrix, Accuracy, RMSE, MAE, and AUC
confusionMatrix(as.factor(pred_svm),as.factor(testing_data_factor$treatment))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  121  45
##        Yes  34 114
##
##                Accuracy : 0.7484
##                  95% CI : (0.6966, 0.7955)
##     No Information Rate : 0.5064
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.4972
##
##  Mcnemar's Test P-Value : 0.2606
##
##             Sensitivity : 0.7806
##             Specificity : 0.7170
##          Pos Pred Value : 0.7289
##          Neg Pred Value : 0.7703
##              Prevalence : 0.4936
##          Detection Rate : 0.3854
##    Detection Prevalence : 0.5287
##       Balanced Accuracy : 0.7488
##
##        'Positive' Class : No
##
```

```
accuracy_svm <- accuracy(as.factor(pred_svm),as.factor(testing_data_factor$treatment))
RMSE_svm <- RMSE(as.numeric(testing_data_factor$treatment), as.numeric(pred_svm))
MAE_svm <- MAE(as.numeric(testing_data_factor$treatment), as.numeric(pred_svm))
roc_svm <- roc(as.numeric(testing_data_factor$treatment), as.numeric(pred_svm))
```

```
###############################################
### Recursive Partitioning - Decision Trees ###
###############################################

#Building Decision tree model using rpart function
rpart_model <- rpart(treatment ~ ., data = training_data_factor[,-3],method = "class")
rpart_model
```

```
## n= 945
##
## node), split, n, loss, yval, (yprob)
```

```
##       * denotes terminal node
##
##  1) root 945 467 Yes (0.4941799 0.5058201)
##    2) family_history=No 573 208 No (0.6369983 0.3630017)
##      4) work_interfere=Never 127  14 No (0.8897638 0.1102362) *
##      5) work_interfere=Often,Rarely,Sometimes 446 194 No (0.5650224 0.4349776)
##       10) care_options=No,Not sure 303 103 No (0.6600660 0.3399340)
##         20) work_interfere=Rarely,Sometimes 270  80 No (0.7037037 0.2962963) *
##         21) work_interfere=Often 33  10 Yes (0.3030303 0.6969697) *
##       11) care_options=Yes 143  52 Yes (0.3636364 0.6363636) *
##    3) family_history=Yes 372 102 Yes (0.2741935 0.7258065)
##      6) work_interfere=Never 32  10 No (0.6875000 0.3125000) *
##      7) work_interfere=Often,Rarely,Sometimes 340  80 Yes (0.2352941 0.7647059) *
```

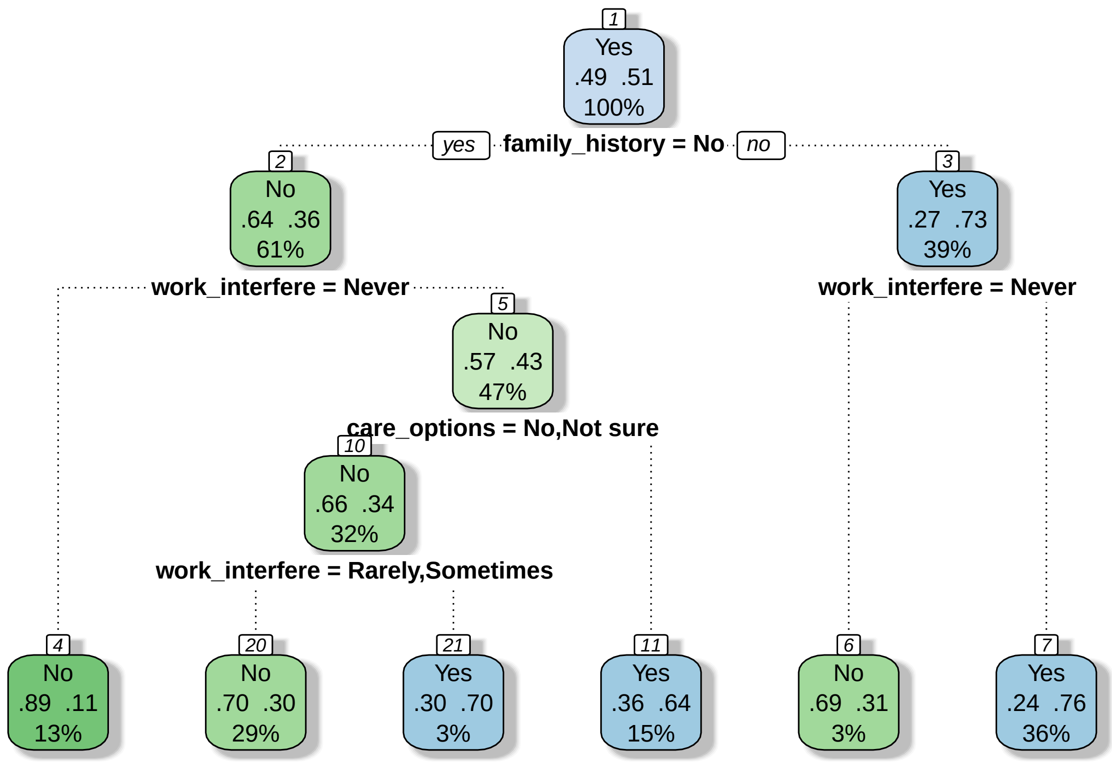
```r
#Observing the importance of each variable using summary
#We can see 45% of the predictions is dependent on family history feature
summary(rpart_model)
```

```
## Call:
## rpart(formula = treatment ~ ., data = training_data_factor[,
##     -3], method = "class")
##   n= 945
##
##           CP nsplit rel error    xerror       xstd
## 1 0.33618844      0 1.0000000 1.0406852 0.03289968
## 2 0.04175589      1 0.6638116 0.6638116 0.03090544
## 3 0.02783726      3 0.5802998 0.6102784 0.03021076
## 4 0.02569593      4 0.5524625 0.6124197 0.03024078
## 5 0.01000000      5 0.5267666 0.5781585 0.02973726
##
## Variable importance
##   family_history   work_interfere     care_options         benefits
##               45               33               13                2
##        anonymity  obs_consequence wellness_program        seek_help
##                2                2                1                1
##           Gender
##                1
##
## Node number 1: 945 observations,    complexity param=0.3361884
##   predicted class=Yes  expected loss=0.4941799  P(node) =1
##     class counts:   467   478
##    probabilities: 0.494 0.506
##   left son=2 (573 obs) right son=3 (372 obs)
##   Primary splits:
##       family_history splits as  LR,   improve=59.38019, (0 missing)
##       work_interfere splits as  LRRR, improve=48.14946, (0 missing)
##       care_options   splits as  LLR,  improve=31.51524, (0 missing)
##       Gender         splits as  RLR,  improve=20.36115, (0 missing)
##       benefits       splits as  LLR,  improve=16.85070, (0 missing)
##   Surrogate splits:
##       obs_consequence splits as  LR,   agree=0.621, adj=0.038, (0 split)
##       Gender          splits as  RLR,  agree=0.615, adj=0.022, (0 split)
##       work_interfere  splits as  LRLL, agree=0.615, adj=0.022, (0 split)
##
```

```
## Node number 2: 573 observations,    complexity param=0.04175589
##   predicted class=No   expected loss=0.3630017  P(node) =0.6063492
##     class counts:   365   208
##    probabilities: 0.637 0.363
##   left son=4 (127 obs) right son=5 (446 obs)
##   Primary splits:
##       work_interfere  splits as  LRRR, improve=20.849190, (0 missing)
##       care_options    splits as  LLR,  improve=18.900100, (0 missing)
##       Gender          splits as  RLR,  improve=10.944120, (0 missing)
##       benefits        splits as  LLR,  improve=10.335870, (0 missing)
##       obs_consequence splits as  LR,   improve= 7.672266, (0 missing)
##
## Node number 3: 372 observations,    complexity param=0.02569593
##   predicted class=Yes  expected loss=0.2741935  P(node) =0.3936508
##     class counts:   102   270
##    probabilities: 0.274 0.726
##   left son=6 (32 obs) right son=7 (340 obs)
##   Primary splits:
##       work_interfere splits as  LRRR,   improve=11.961570, (0 missing)
##       care_options   splits as  LLR,    improve= 5.137553, (0 missing)
##       anonymity      splits as  LRR,    improve= 4.139082, (0 missing)
##       no_employees   splits as  RRRLRR, improve= 3.719205, (0 missing)
##       leave          splits as  LRLRL,  improve= 3.143968, (0 missing)
##
## Node number 4: 127 observations
##   predicted class=No   expected loss=0.1102362  P(node) =0.1343915
##     class counts:   113    14
##    probabilities: 0.890 0.110
##
## Node number 5: 446 observations,    complexity param=0.04175589
##   predicted class=No   expected loss=0.4349776  P(node) =0.4719577
##     class counts:   252   194
##    probabilities: 0.565 0.435
##   left son=10 (303 obs) right son=11 (143 obs)
##   Primary splits:
##       care_options   splits as  LLR,  improve=17.073280, (0 missing)
##       work_interfere splits as  -RLL, improve=12.396530, (0 missing)
##       benefits       splits as  LLR,  improve=10.580220, (0 missing)
##       Gender         splits as  RLR,  improve= 7.447124, (0 missing)
##       anonymity      splits as  LLR,  improve= 6.576086, (0 missing)
##   Surrogate splits:
##       benefits         splits as  LLR, agree=0.740, adj=0.189, (0 split)
##       anonymity        splits as  LLR, agree=0.729, adj=0.154, (0 split)
##       wellness_program splits as  LLR, agree=0.713, adj=0.105, (0 split)
##       seek_help        splits as  LLR, agree=0.706, adj=0.084, (0 split)
##       Age              splits as  LLR, agree=0.686, adj=0.021, (0 split)
##
## Node number 6: 32 observations
##   predicted class=No   expected loss=0.3125  P(node) =0.03386243
##     class counts:    22    10
##    probabilities: 0.688 0.312
##
## Node number 7: 340 observations
##   predicted class=Yes  expected loss=0.2352941  P(node) =0.3597884
```

```
##     class counts:    80   260
##    probabilities: 0.235 0.765
##
## Node number 10: 303 observations,    complexity param=0.02783726
##   predicted class=No   expected loss=0.339934  P(node) =0.3206349
##     class counts:   200   103
##    probabilities: 0.660 0.340
##   left son=20 (270 obs) right son=21 (33 obs)
##   Primary splits:
##       work_interfere        splits as  -RLL,  improve=9.441611, (0 missing)
##       leave                 splits as  LRLRL, improve=3.953442, (0 missing)
##       Age                   splits as  LLR,   improve=2.008346, (0 missing)
##       obs_consequence       splits as  LR,    improve=1.890667, (0 missing)
##       phys_health_interview splits as  RLL,   improve=1.877892, (0 missing)
##
## Node number 11: 143 observations
##   predicted class=Yes  expected loss=0.3636364  P(node) =0.1513228
##     class counts:    52    91
##    probabilities: 0.364 0.636
##
## Node number 20: 270 observations
##   predicted class=No   expected loss=0.2962963  P(node) =0.2857143
##     class counts:   190    80
##    probabilities: 0.704 0.296
##
## Node number 21: 33 observations
##   predicted class=Yes  expected loss=0.3030303  P(node) =0.03492063
##     class counts:    10    23
##    probabilities: 0.303 0.697
```

```r
#plotting the tree using fancyRpartPlot function
fancyRpartPlot(rpart_model)
```



Rattle 2020−Aug−08 14:23:41 harsh

```r
#Predicting the outcome using testing dataset
pred_rpart <- predict(rpart_model, testing_data_factor)

#Since the output is in terms of probabilities we convert it to categorical values
pred_rpart <- as.factor(ifelse(pred_rpart[,2] < 0.5, "No", "Yes"))

#Model evaluation using confusionMatrix, Accuracy, RMSE, MAE, and AUC
confusionMatrix(pred_rpart,testing_data_factor$treatment)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  120  27
##        Yes  35 132
##
##                Accuracy : 0.8025
##                  95% CI : (0.7542, 0.8451)
##     No Information Rate : 0.5064
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6048
##
##  Mcnemar's Test P-Value : 0.374
##
##             Sensitivity : 0.7742
```

```
##             Specificity : 0.8302
##          Pos Pred Value : 0.8163
##          Neg Pred Value : 0.7904
##              Prevalence : 0.4936
##          Detection Rate : 0.3822
##    Detection Prevalence : 0.4682
##       Balanced Accuracy : 0.8022
##
##        'Positive' Class : No
##
```

```
accuracy_rpart <- accuracy(pred_rpart,testing_data_factor$treatment)
RMSE_rpart <- RMSE(as.numeric(testing_data_factor$treatment), as.numeric(pred_rpart))
MAE_rpart <- MAE(as.numeric(testing_data_factor$treatment), as.numeric(pred_rpart))
roc_rpart <- roc(as.numeric(testing_data_factor$treatment), as.numeric(pred_rpart))
```

Evaluation with k-fold cross-validation

- k-Fold Cross Validation is done for the whole dataset
- I have used k = 10 which means 10 folds take place along with 3 repetitions
- For testing the data, I have used 3 models to test the k-fold CV
- Accuracy of each model is printed and based on the observation average accuracy is around 72-73%

```
###############################
### K-fold Cross Validation ###
###############################

#Creating a train function for cross validation
#We use k = 10 folds with repeated validation
fitControl <- trainControl(## 10-fold CV
                           method = "repeatedcv",
                           number = 10,repeats = 3,savePredictions = TRUE)

#Cross validation is done using 3 models glm, SVM with
#Radial function, and rpart function
cv_glm <- train(treatment ~ ., data = MH_data_factors,
                 method = "glm",
                 trControl = fitControl)

cv_svm <- train(treatment ~ ., data = MH_data_factors,
                 method = "svmRadial",
                 trControl = fitControl)

cv_rpart <- train(treatment ~ ., data = MH_data_factors,
                 method = "rpart",
                 trControl = fitControl)

#Printing the accuracies of each model with cross validation
#cv_glm
print(cv_glm)
```

```
## Generalized Linear Model
```

```
##
## 1259 samples
##   22 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 1133, 1134, 1133, 1133, 1133, 1134, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7381586  0.4762473
```

```r
#cv_svm
print(cv_svm)
```

```
## Support Vector Machines with Radial Basis Function Kernel
##
## 1259 samples
##   22 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 1134, 1133, 1133, 1133, 1133, 1134, ...
## Resampling results across tuning parameters:
##
##   C     Accuracy   Kappa
##   0.25  0.7355133  0.4711580
##   0.50  0.7434456  0.4869914
##   1.00  0.7442394  0.4886108
##
## Tuning parameter 'sigma' was held constant at a value of 0.01287498
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were sigma = 0.01287498 and C = 1.
```

```r
#cv_rpart
print(cv_rpart)
```

```
## CART
##
## 1259 samples
##   22 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 1134, 1133, 1133, 1134, 1133, 1133, ...
## Resampling results across tuning parameters:
##
##   cp          Accuracy   Kappa
##   0.01982851  0.7039752  0.4077652
##   0.06109325  0.6875810  0.3766278
```

```
##   0.35852090  0.5720861  0.1390566
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.01982851.
```

Tuning of models

- I have tuned all the models as follows:
  - Logistic Regression: Stepwise backward elimination method is used to evaluate the new formula with reduced features
  - Neural Network : Increased the number of hidden layer to 3
  - Support Vector Machine: Changed the kernel function to vanilladot
  - Recursive Partitioning: Changed the complexity parameter to 0.025
- Tuning the models did not result in improved accuracies, only improvement was observed in SVM model
- Apart from SVM model, all other models accuracy remained the same or reduced

```r
########################
### Tuning of Models ###
########################

#Using stepwise backward method for glm()
step(lm, direction="backward")
```

```
## Start:  AIC=1037.08
## treatment ~ Age + Gender + self_employed + family_history + work_interfere +
##     no_employees + remote_work + tech_company + benefits + care_options +
##     wellness_program + seek_help + anonymity + leave + mental_health_consequence +
##     phys_health_consequence + coworkers + supervisor + mental_health_interview +
##     phys_health_interview + mental_vs_physical + obs_consequence
##
##                             Df Deviance    AIC
## - no_employees               5   949.69 1031.7
## - mental_vs_physical         2   945.11 1033.1
## - phys_health_interview      2   945.63 1033.6
## - phys_health_consequence    2   945.68 1033.7
## - supervisor                 2   946.33 1034.3
## - wellness_program           2   946.44 1034.4
## - remote_work                1   945.56 1035.6
## - benefits                   2   947.56 1035.6
## - self_employed              1   945.75 1035.8
## - leave                      4   951.94 1035.9
## - tech_company               1   946.24 1036.2
## - seek_help                  2   949.03 1037.0
## <none>                           945.08 1037.1
## - obs_consequence            1   947.35 1037.3
## - Age                        2   949.40 1037.4
## - mental_health_interview    2   949.81 1037.8
## - mental_health_consequence  2   949.93 1037.9
## - anonymity                  2   953.29 1041.3
## - care_options               2   955.98 1044.0
## - coworkers                  2   957.73 1045.7
```

```
## - Gender                   2   960.75 1048.8
## - family_history           1  1003.87 1093.9
## - work_interfere           3  1040.22 1126.2
##
## Step:  AIC=1031.69
## treatment ~ Age + Gender + self_employed + family_history + work_interfere +
##     remote_work + tech_company + benefits + care_options + wellness_program +
##     seek_help + anonymity + leave + mental_health_consequence +
##     phys_health_consequence + coworkers + supervisor + mental_health_interview +
##     phys_health_interview + mental_vs_physical + obs_consequence
##
##                             Df Deviance    AIC
## - mental_vs_physical         2   949.78 1027.8
## - phys_health_interview      2   950.02 1028.0
## - phys_health_consequence    2   950.43 1028.4
## - supervisor                 2   950.86 1028.9
## - wellness_program           2   951.15 1029.2
## - benefits                   2   951.90 1029.9
## - remote_work                1   950.26 1030.3
## - self_employed              1   950.59 1030.6
## - leave                      4   956.78 1030.8
## - seek_help                  2   952.97 1031.0
## - tech_company               1   951.65 1031.7
## <none>                           949.69 1031.7
## - obs_consequence            1   951.91 1031.9
## - Age                        2   953.91 1031.9
## - mental_health_interview    2   954.46 1032.5
## - mental_health_consequence  2   954.67 1032.7
## - anonymity                  2   957.93 1035.9
## - care_options               2   961.87 1039.9
## - coworkers                  2   963.15 1041.2
## - Gender                     2   965.07 1043.1
## - family_history             1  1010.32 1090.3
## - work_interfere             3  1046.25 1122.2
##
## Step:  AIC=1027.78
## treatment ~ Age + Gender + self_employed + family_history + work_interfere +
##     remote_work + tech_company + benefits + care_options + wellness_program +
##     seek_help + anonymity + leave + mental_health_consequence +
##     phys_health_consequence + coworkers + supervisor + mental_health_interview +
##     phys_health_interview + obs_consequence
##
##                             Df Deviance    AIC
## - phys_health_interview      2   950.11 1024.1
## - phys_health_consequence    2   950.50 1024.5
## - supervisor                 2   951.00 1025.0
## - wellness_program           2   951.28 1025.3
## - benefits                   2   952.06 1026.1
## - remote_work                1   950.38 1026.4
## - self_employed              1   950.71 1026.7
## - leave                      4   956.90 1026.9
## - seek_help                  2   953.06 1027.1
## - tech_company               1   951.74 1027.7
## <none>                           949.78 1027.8
```

```
## - Age                        2   954.01 1028.0
## - obs_consequence            1   952.07 1028.1
## - mental_health_interview    2   954.58 1028.6
## - mental_health_consequence  2   955.08 1029.1
## - anonymity                  2   957.97 1032.0
## - care_options               2   961.90 1035.9
## - coworkers                  2   963.18 1037.2
## - Gender                     2   965.18 1039.2
## - family_history             1  1010.39 1086.4
## - work_interfere             3  1046.50 1118.5
##
## Step:  AIC=1024.11
## treatment ~ Age + Gender + self_employed + family_history + work_interfere +
##     remote_work + tech_company + benefits + care_options + wellness_program +
##     seek_help + anonymity + leave + mental_health_consequence +
##     phys_health_consequence + coworkers + supervisor + mental_health_interview +
##     obs_consequence
##
##                             Df Deviance    AIC
## - phys_health_consequence    2   950.89 1020.9
## - supervisor                 2   951.33 1021.3
## - wellness_program           2   951.56 1021.6
## - benefits                   2   952.41 1022.4
## - remote_work                1   950.71 1022.7
## - self_employed              1   951.07 1023.1
## - leave                      4   957.09 1023.1
## - seek_help                  2   953.32 1023.3
## - tech_company               1   952.07 1024.1
## <none>                           950.11 1024.1
## - Age                        2   954.23 1024.2
## - obs_consequence            1   952.40 1024.4
## - mental_health_consequence  2   955.51 1025.5
## - mental_health_interview    2   955.86 1025.9
## - anonymity                  2   958.25 1028.2
## - care_options               2   962.26 1032.3
## - coworkers                  2   963.47 1033.5
## - Gender                     2   965.57 1035.6
## - family_history             1  1011.36 1083.4
## - work_interfere             3  1047.60 1115.6
##
## Step:  AIC=1020.89
## treatment ~ Age + Gender + self_employed + family_history + work_interfere +
##     remote_work + tech_company + benefits + care_options + wellness_program +
##     seek_help + anonymity + leave + mental_health_consequence +
##     coworkers + supervisor + mental_health_interview + obs_consequence
##
##                             Df Deviance    AIC
## - supervisor                 2   952.07 1018.1
## - wellness_program           2   952.33 1018.3
## - benefits                   2   953.21 1019.2
## - remote_work                1   951.48 1019.5
## - self_employed              1   951.88 1019.9
## - leave                      4   958.05 1020.0
## - seek_help                  2   954.16 1020.2
```

```
## - Age                        2   954.80 1020.8
## <none>                           950.89 1020.9
## - tech_company               1   952.95 1021.0
## - obs_consequence            1   953.10 1021.1
## - mental_health_consequence  2   955.82 1021.8
## - mental_health_interview    2   956.81 1022.8
## - anonymity                  2   959.18 1025.2
## - care_options               2   963.49 1029.5
## - coworkers                  2   964.39 1030.4
## - Gender                     2   966.05 1032.0
## - family_history             1  1012.38 1080.4
## - work_interfere             3  1048.67 1112.7
##
## Step:  AIC=1018.07
## treatment ~ Age + Gender + self_employed + family_history + work_interfere +
##     remote_work + tech_company + benefits + care_options + wellness_program +
##     seek_help + anonymity + leave + mental_health_consequence +
##     coworkers + mental_health_interview + obs_consequence
##
##                             Df Deviance    AIC
## - wellness_program           2   953.52 1015.5
## - benefits                   2   954.41 1016.4
## - remote_work                1   952.66 1016.7
## - self_employed              1   952.87 1016.9
## - seek_help                  2   955.33 1017.3
## - leave                      4   959.34 1017.3
## - Age                        2   955.81 1017.8
## <none>                           952.07 1018.1
## - obs_consequence            1   954.24 1018.2
## - tech_company               1   954.25 1018.2
## - mental_health_interview    2   958.36 1020.4
## - anonymity                  2   960.10 1022.1
## - mental_health_consequence  2   960.12 1022.1
## - coworkers                  2   964.62 1026.6
## - care_options               2   964.73 1026.7
## - Gender                     2   967.73 1029.7
## - family_history             1  1012.49 1076.5
## - work_interfere             3  1049.77 1109.8
##
## Step:  AIC=1015.52
## treatment ~ Age + Gender + self_employed + family_history + work_interfere +
##     remote_work + tech_company + benefits + care_options + seek_help +
##     anonymity + leave + mental_health_consequence + coworkers +
##     mental_health_interview + obs_consequence
##
##                             Df Deviance    AIC
## - benefits                   2   955.42 1013.4
## - remote_work                1   954.17 1014.2
## - seek_help                  2   956.36 1014.4
## - self_employed              1   954.55 1014.5
## - Age                        2   957.12 1015.1
## - leave                      4   961.40 1015.4
## <none>                           953.52 1015.5
## - obs_consequence            1   955.61 1015.6
```

```
## - tech_company              1   955.91 1015.9
## - mental_health_interview    2   959.73 1017.7
## - anonymity                  2   961.03 1019.0
## - mental_health_consequence  2   961.73 1019.7
## - coworkers                  2   965.53 1023.5
## - care_options               2   966.29 1024.3
## - Gender                     2   968.77 1026.8
## - family_history             1  1015.10 1075.1
## - work_interfere             3  1051.69 1107.7
##
## Step:  AIC=1013.42
## treatment ~ Age + Gender + self_employed + family_history + work_interfere +
##     remote_work + tech_company + care_options + seek_help + anonymity +
##     leave + mental_health_consequence + coworkers + mental_health_interview +
##     obs_consequence
##
##                             Df Deviance    AIC
## - remote_work               1   955.99 1012.0
## - leave                     4   962.84 1012.8
## - self_employed             1   957.02 1013.0
## - seek_help                 2   959.15 1013.1
## - Age                       2   959.30 1013.3
## <none>                          955.42 1013.4
## - obs_consequence           1   957.47 1013.5
## - tech_company              1   957.83 1013.8
## - mental_health_interview    2   961.71 1015.7
## - mental_health_consequence  2   964.43 1018.4
## - anonymity                  2   964.71 1018.7
## - coworkers                  2   967.57 1021.6
## - Gender                     2   971.63 1025.6
## - care_options               2   974.41 1028.4
## - family_history             1  1020.34 1076.3
## - work_interfere             3  1053.42 1105.4
##
## Step:  AIC=1011.99
## treatment ~ Age + Gender + self_employed + family_history + work_interfere +
##     tech_company + care_options + seek_help + anonymity + leave +
##     mental_health_consequence + coworkers + mental_health_interview +
##     obs_consequence
##
##                             Df Deviance    AIC
## - self_employed             1   957.16 1011.2
## - leave                     4   963.46 1011.5
## - seek_help                 2   959.80 1011.8
## - obs_consequence           1   957.96 1012.0
## <none>                          955.99 1012.0
## - Age                       2   960.19 1012.2
## - tech_company              1   958.66 1012.7
## - mental_health_interview    2   962.05 1014.0
## - mental_health_consequence  2   965.20 1017.2
## - anonymity                  2   965.28 1017.3
## - coworkers                  2   968.50 1020.5
## - Gender                     2   972.24 1024.2
## - care_options               2   974.78 1026.8
```

```
## - family_history             1  1021.21 1075.2
## - work_interfere             3  1054.79 1104.8
##
## Step:  AIC=1011.16
## treatment ~ Age + Gender + family_history + work_interfere +
##     tech_company + care_options + seek_help + anonymity + leave +
##     mental_health_consequence + coworkers + mental_health_interview +
##     obs_consequence
##
##                             Df Deviance    AIC
## - leave                      4   964.12 1010.1
## - obs_consequence            1   958.95 1011.0
## - Age                        2   961.10 1011.1
## <none>                           957.16 1011.2
## - seek_help                  2   961.24 1011.2
## - tech_company               1   959.62 1011.6
## - mental_health_interview    2   963.24 1013.2
## - anonymity                  2   966.26 1016.3
## - mental_health_consequence  2   966.87 1016.9
## - coworkers                  2   969.58 1019.6
## - Gender                     2   973.66 1023.7
## - care_options               2   975.96 1026.0
## - family_history             1  1022.37 1074.4
## - work_interfere             3  1055.34 1103.3
##
## Step:  AIC=1010.12
## treatment ~ Age + Gender + family_history + work_interfere +
##     tech_company + care_options + seek_help + anonymity + mental_health_consequence +
##     coworkers + mental_health_interview + obs_consequence
##
##                             Df Deviance    AIC
## - seek_help                  2   966.85 1008.9
## <none>                           964.12 1010.1
## - Age                        2   968.26 1010.3
## - tech_company               1   966.49 1010.5
## - obs_consequence            1   967.22 1011.2
## - mental_health_interview    2   970.61 1012.6
## - anonymity                  2   972.26 1014.3
## - mental_health_consequence  2   976.39 1018.4
## - coworkers                  2   977.07 1019.1
## - Gender                     2   980.22 1022.2
## - care_options               2   983.74 1025.7
## - family_history             1  1028.93 1072.9
## - work_interfere             3  1064.90 1104.9
##
## Step:  AIC=1008.85
## treatment ~ Age + Gender + family_history + work_interfere +
##     tech_company + care_options + anonymity + mental_health_consequence +
##     coworkers + mental_health_interview + obs_consequence
##
##                             Df Deviance    AIC
## - tech_company               1   968.78 1008.8
## <none>                           966.85 1008.9
## - Age                        2   971.15 1009.1
```

```
## - obs_consequence           1   969.75 1009.8
## - mental_health_interview    2   973.39 1011.4
## - anonymity                  2   975.30 1013.3
## - mental_health_consequence  2   978.19 1016.2
## - coworkers                  2   979.11 1017.1
## - Gender                     2   983.31 1021.3
## - care_options               2   985.63 1023.6
## - family_history             1  1032.96 1073.0
## - work_interfere             3  1065.56 1101.6
##
## Step:  AIC=1008.78
## treatment ~ Age + Gender + family_history + work_interfere +
##     care_options + anonymity + mental_health_consequence + coworkers +
##     mental_health_interview + obs_consequence
##
##                             Df Deviance    AIC
## <none>                          968.78 1008.8
## - Age                        2   973.01 1009.0
## - obs_consequence            1   971.58 1009.6
## - mental_health_interview    2   975.27 1011.3
## - anonymity                  2   976.68 1012.7
## - mental_health_consequence  2   979.29 1015.3
## - coworkers                  2   981.20 1017.2
## - Gender                     2   984.68 1020.7
## - care_options               2   988.02 1024.0
## - family_history             1  1035.13 1073.1
## - work_interfere             3  1066.86 1100.9
```

```
##
## Call:  glm(formula = treatment ~ Age + Gender + family_history + work_interfere +
##     care_options + anonymity + mental_health_consequence + coworkers +
##     mental_health_interview + obs_consequence, family = "binomial",
##     data = training_data_factor)
##
## Coefficients:
##                 (Intercept)                    AgeJunior
##                    -2.96149                      0.31295
##                   AgeSenior                   Gendermale
##                     0.88006                     -0.79112
##                 Genderqueer            family_historyYes
##                    -0.16568                      1.30635
##         work_interfereOften          work_interfereRarely
##                     3.12159                      2.19535
##     work_interfereSometimes         care_optionsNot sure
##                     1.53460                     -0.06248
##             care_optionsYes                  anonymityNo
##                     0.78126                     -0.33653
##                anonymityYes   mental_health_consequenceNo
##                     0.49272                     -0.50720
## mental_health_consequenceYes         coworkersSome of them
##                     0.24478                      0.52729
##                coworkersYes     mental_health_interviewNo
##                     1.03599                      0.55990
##   mental_health_interviewYes            obs_consequenceYes
```

```
##                        0.86060                        0.40007
##
## Degrees of Freedom: 944 Total (i.e. Null);  925 Residual
## Null Deviance:       1310
## Residual Deviance: 968.8     AIC: 1009
```

```r
#Tuning logistic regression model based on the
#formula generated from step function
new_lm <- glm(formula = treatment ~ Age + Gender + family_history + work_interfere +
                tech_company + care_options + anonymity + mental_health_consequence +
                coworkers + mental_health_interview + obs_consequence, family = "binomial",
              data = training_data_factor)

#Predicting the outcome of new model
predict_prob <- predict(new_lm, testing_data_factor, type = "response")

#converting the probability values to categorical values
pred_glm_tuned <- (as.factor(ifelse(predict_prob < 0.5, "No", "Yes")))

#Model evaluation using confusionMatrix, Accuracy, RMSE, MAE, and AUC
confusionMatrix(as.factor(pred_glm_tuned),as.factor(testing_data_factor$treatment))
```

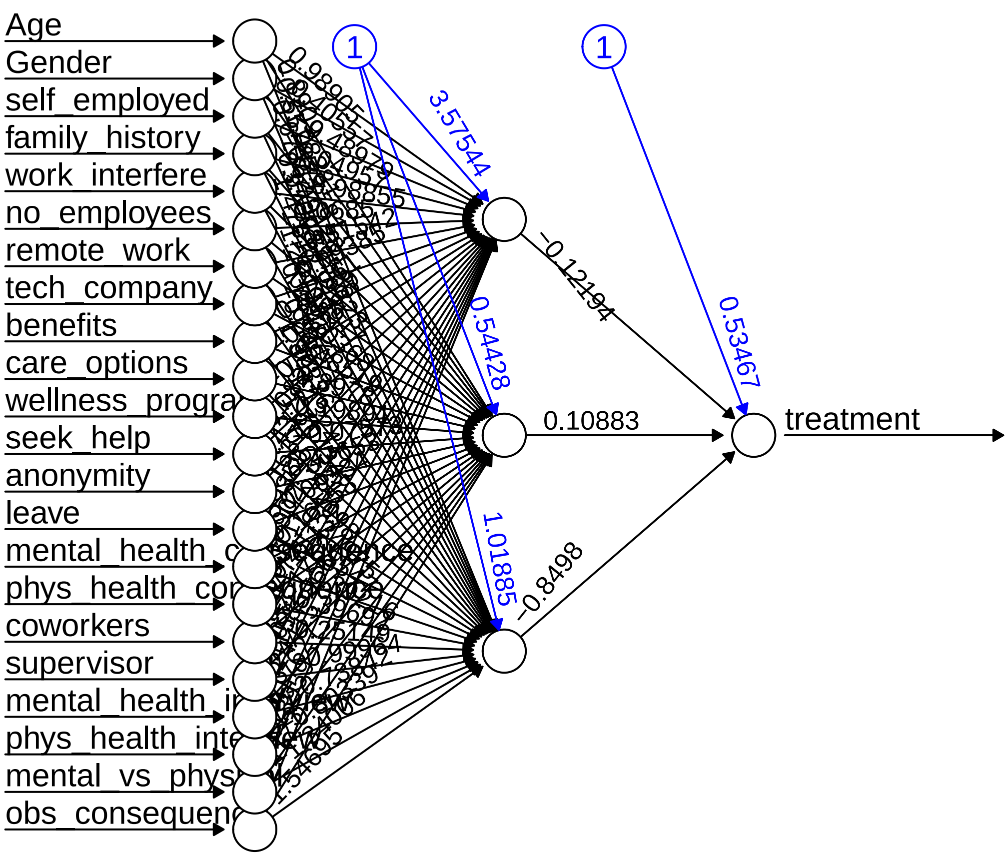
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  123  40
##        Yes  32 119
##
##                Accuracy : 0.7707
##                  95% CI : (0.7202, 0.816)
##     No Information Rate : 0.5064
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.5416
##
##  Mcnemar's Test P-Value : 0.4094
##
##             Sensitivity : 0.7935
##             Specificity : 0.7484
##          Pos Pred Value : 0.7546
##          Neg Pred Value : 0.7881
##              Prevalence : 0.4936
##          Detection Rate : 0.3917
##    Detection Prevalence : 0.5191
##       Balanced Accuracy : 0.7710
##
##        'Positive' Class : No
##
```

```r
accuracy_glm_tuned <- accuracy(pred_glm_tuned,testing_data_factor$treatment)
RMSE_glm_tuned <- RMSE(as.numeric(testing_data_factor$treatment), as.numeric(pred_glm_tuned))
MAE_glm_tuned <- MAE(as.numeric(testing_data_factor$treatment), as.numeric(pred_glm_tuned))
```

```r
roc_glm_tuned <- roc(as.numeric(testing_data_factor$treatment), as.numeric(pred_glm_tuned))

training_data$treatment <- as.numeric(training_data$treatment)
testing_data$treatment <- as.numeric(testing_data$treatment)

#Tuning neural network model by adding hidden layers to it
softplus <- function(x) log(1+exp(x))
neuralnet_model <- neuralnet(treatment~., data = training_data,stepmax=1e+08, hidden = 3, threshold = 0

#Using compute() function to predict the outcome of testing dataset
nn_predictions_tuned <- compute(neuralnet_model, testing_data[,-6])
net_results_tuned <- nn_predictions_tuned$net.result

#Checking the correlation of both predictor and predicted values
cor(net_results_tuned,as.numeric(testing_data$treatment))
```

```
##           [,1]
## [1,] 0.8495929
```

```r
#Plotting the neural network
plot(neuralnet_model,rep="best")
```




```r
#Converting numeric predictions to categorical values
pred_nn_tuned <- net_results
```

```r
pred_nn_tuned <- as.factor(ifelse(pred_nn_tuned > 1.5, 2, 1))

#Model evaluation using confusionMatrix, Accuracy, RMSE, MAE, and AUC
confusionMatrix(pred_nn_tuned,as.factor(testing_data$treatment))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2
##          1 135  62
##          2  20  97
##
##                Accuracy : 0.7389
##                  95% CI : (0.6866, 0.7866)
##     No Information Rate : 0.5064
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4794
##
##  Mcnemar's Test P-Value : 5.963e-06
##
##             Sensitivity : 0.8710
##             Specificity : 0.6101
##          Pos Pred Value : 0.6853
##          Neg Pred Value : 0.8291
##              Prevalence : 0.4936
##          Detection Rate : 0.4299
##    Detection Prevalence : 0.6274
##       Balanced Accuracy : 0.7405
##
##        'Positive' Class : 1
##
```

```r
accuracy_nn_tuned <- accuracy(pred_nn_tuned,as.factor(testing_data$treatment))
RMSE_nn_tuned <- RMSE(as.numeric(testing_data$treatment), as.numeric(pred_nn_tuned))
MAE_nn_tuned <- MAE(as.numeric(testing_data$treatment), as.numeric(pred_nn_tuned))
roc_nn_tuned <- roc(as.numeric(testing_data$treatment), as.numeric(pred_nn_tuned))

#Tuning SVM model by using Linear function instead of RBF function
svm_model <- ksvm(treatment ~ ., data = training_data_factor,prob.model=TRUE,kernel="vanilladot")
```

```
##  Setting default kernel parameters
```

```r
##Predicting the outcome of tuned model
pred_svm_tuned <- predict(svm_model, testing_data_factor)
head(pred_svm_tuned)
```

```
## [1] Yes Yes No  Yes Yes Yes
## Levels: No Yes
```

```r
#Model evaluation using confusionMatrix, Accuracy, RMSE, MAE, and AUC
confusionMatrix(as.factor(pred_svm_tuned),as.factor(testing_data_factor$treatment))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  123  42
##        Yes  32 117
##
##                Accuracy : 0.7643
##                  95% CI : (0.7134, 0.8102)
##     No Information Rate : 0.5064
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.529
##
##  Mcnemar's Test P-Value : 0.2955
##
##             Sensitivity : 0.7935
##             Specificity : 0.7358
##          Pos Pred Value : 0.7455
##          Neg Pred Value : 0.7852
##              Prevalence : 0.4936
##          Detection Rate : 0.3917
##    Detection Prevalence : 0.5255
##       Balanced Accuracy : 0.7647
##
##        'Positive' Class : No
##
```

```r
accuracy_svm_tuned <- accuracy(as.factor(pred_svm_tuned),as.factor(testing_data_factor$treatment))
RMSE_svm_tuned <- RMSE(as.numeric(testing_data_factor$treatment), as.numeric(pred_svm_tuned))
MAE_svm_tuned <- MAE(as.numeric(testing_data_factor$treatment), as.numeric(pred_svm_tuned))
roc_svm_tuned <- roc(as.numeric(testing_data_factor$treatment), as.numeric(pred_svm_tuned))


#Tuning Decision Trees by using complexity parameter value as 0.025
rpart_model <- rpart(treatment ~ ., data = training_data_factor[,-3],method = "class",cp=0.025)
rpart_model
```

```
## n= 945
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 945 467 Yes (0.4941799 0.5058201)
##    2) family_history=No 573 208 No (0.6369983 0.3630017)
##      4) work_interfere=Never 127  14 No (0.8897638 0.1102362) *
##      5) work_interfere=Often,Rarely,Sometimes 446 194 No (0.5650224 0.4349776)
##       10) care_options=No,Not sure 303 103 No (0.6600660 0.3399340)
##         20) work_interfere=Rarely,Sometimes 270  80 No (0.7037037 0.2962963) *
```

```
##         21) work_interfere=Often 33  10 Yes (0.3030303 0.6969697) *
##       11) care_options=Yes 143  52 Yes (0.3636364 0.6363636) *
##    3) family_history=Yes 372 102 Yes (0.2741935 0.7258065)
##      6) work_interfere=Never 32  10 No (0.6875000 0.3125000) *
##      7) work_interfere=Often,Rarely,Sometimes 340  80 Yes (0.2352941 0.7647059) *
```

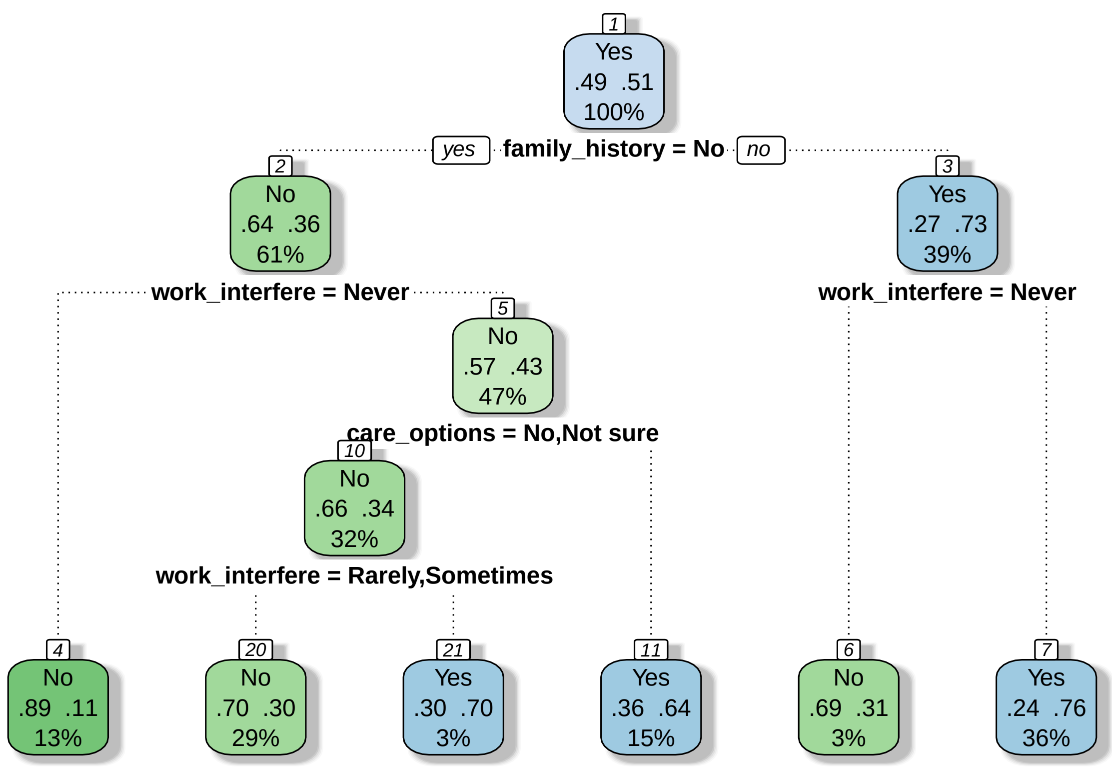
```r
##Observing the importance of each variable using summary
summary(rpart_model)
```

```
## Call:
## rpart(formula = treatment ~ ., data = training_data_factor[,
##     -3], method = "class", cp = 0.025)
##   n= 945
##
##           CP nsplit rel error    xerror       xstd
## 1 0.33618844      0 1.0000000 1.0385439 0.03290160
## 2 0.04175589      1 0.6638116 0.6638116 0.03090544
## 3 0.02783726      3 0.5802998 0.6252677 0.03041692
## 4 0.02569593      4 0.5524625 0.6209850 0.03035896
## 5 0.02500000      5 0.5267666 0.6038544 0.03011955
##
## Variable importance
##   family_history   work_interfere     care_options         benefits
##               45               33               13                2
##        anonymity  obs_consequence wellness_program        seek_help
##                2                2                1                1
##           Gender
##                1
##
## Node number 1: 945 observations,    complexity param=0.3361884
##   predicted class=Yes  expected loss=0.4941799  P(node) =1
##     class counts:   467   478
##    probabilities: 0.494 0.506
##   left son=2 (573 obs) right son=3 (372 obs)
##   Primary splits:
##       family_history splits as  LR,   improve=59.38019, (0 missing)
##       work_interfere splits as  LRRR, improve=48.14946, (0 missing)
##       care_options   splits as  LLR,  improve=31.51524, (0 missing)
##       Gender         splits as  RLR,  improve=20.36115, (0 missing)
##       benefits       splits as  LLR,  improve=16.85070, (0 missing)
##   Surrogate splits:
##       obs_consequence splits as  LR,   agree=0.621, adj=0.038, (0 split)
##       Gender          splits as  RLR,  agree=0.615, adj=0.022, (0 split)
##       work_interfere  splits as  LRLL, agree=0.615, adj=0.022, (0 split)
##
## Node number 2: 573 observations,    complexity param=0.04175589
##   predicted class=No   expected loss=0.3630017  P(node) =0.6063492
##     class counts:   365   208
##    probabilities: 0.637 0.363
##   left son=4 (127 obs) right son=5 (446 obs)
##   Primary splits:
##       work_interfere  splits as  LRRR, improve=20.849190, (0 missing)
##       care_options    splits as  LLR,  improve=18.900100, (0 missing)
##       Gender          splits as  RLR,  improve=10.944120, (0 missing)
```

```
##       benefits        splits as  LLR,  improve=10.335870, (0 missing)
##       obs_consequence splits as  LR,   improve= 7.672266, (0 missing)
##
## Node number 3: 372 observations,    complexity param=0.02569593
##   predicted class=Yes  expected loss=0.2741935  P(node) =0.3936508
##     class counts:   102   270
##    probabilities: 0.274 0.726
##   left son=6 (32 obs) right son=7 (340 obs)
##   Primary splits:
##       work_interfere splits as  LRRR,   improve=11.961570, (0 missing)
##       care_options   splits as  LLR,    improve= 5.137553, (0 missing)
##       anonymity      splits as  LRR,    improve= 4.139082, (0 missing)
##       no_employees   splits as  RRRLRR, improve= 3.719205, (0 missing)
##       leave          splits as  LRLRL,  improve= 3.143968, (0 missing)
##
## Node number 4: 127 observations
##   predicted class=No   expected loss=0.1102362  P(node) =0.1343915
##     class counts:   113    14
##    probabilities: 0.890 0.110
##
## Node number 5: 446 observations,    complexity param=0.04175589
##   predicted class=No   expected loss=0.4349776  P(node) =0.4719577
##     class counts:   252   194
##    probabilities: 0.565 0.435
##   left son=10 (303 obs) right son=11 (143 obs)
##   Primary splits:
##       care_options   splits as  LLR,  improve=17.073280, (0 missing)
##       work_interfere splits as  -RLL, improve=12.396530, (0 missing)
##       benefits       splits as  LLR,  improve=10.580220, (0 missing)
##       Gender         splits as  RLR,  improve= 7.447124, (0 missing)
##       anonymity      splits as  LLR,  improve= 6.576086, (0 missing)
##   Surrogate splits:
##       benefits         splits as  LLR, agree=0.740, adj=0.189, (0 split)
##       anonymity        splits as  LLR, agree=0.729, adj=0.154, (0 split)
##       wellness_program splits as  LLR, agree=0.713, adj=0.105, (0 split)
##       seek_help        splits as  LLR, agree=0.706, adj=0.084, (0 split)
##       Age              splits as  LLR, agree=0.686, adj=0.021, (0 split)
##
## Node number 6: 32 observations
##   predicted class=No   expected loss=0.3125  P(node) =0.03386243
##     class counts:    22    10
##    probabilities: 0.688 0.312
##
## Node number 7: 340 observations
##   predicted class=Yes  expected loss=0.2352941  P(node) =0.3597884
##     class counts:    80   260
##    probabilities: 0.235 0.765
##
## Node number 10: 303 observations,    complexity param=0.02783726
##   predicted class=No   expected loss=0.339934  P(node) =0.3206349
##     class counts:   200   103
##    probabilities: 0.660 0.340
##   left son=20 (270 obs) right son=21 (33 obs)
##   Primary splits:
```

```
##       work_interfere        splits as  -RLL,  improve=9.441611, (0 missing)
##       leave                 splits as  LRLRL, improve=3.953442, (0 missing)
##       Age                   splits as  LLR,   improve=2.008346, (0 missing)
##       obs_consequence       splits as  LR,    improve=1.890667, (0 missing)
##       phys_health_interview splits as  RLL,   improve=1.877892, (0 missing)
##
## Node number 11: 143 observations
##    predicted class=Yes  expected loss=0.3636364  P(node) =0.1513228
##      class counts:    52    91
##     probabilities: 0.364 0.636
##
## Node number 20: 270 observations
##    predicted class=No   expected loss=0.2962963  P(node) =0.2857143
##      class counts:   190    80
##     probabilities: 0.704 0.296
##
## Node number 21: 33 observations
##    predicted class=Yes  expected loss=0.3030303  P(node) =0.03492063
##      class counts:    10    23
##     probabilities: 0.303 0.697
```

```r
#plotting the tree using fancyRpartPlot function
fancyRpartPlot(rpart_model)
```



Rattle 2020−Aug−08 14:24:15 harsh

```r
##Predicting the outcome of tuned model
pred_rpart_tuned <- predict(rpart_model, testing_data_factor)
pred_rpart_tuned <- as.factor(ifelse(pred_rpart_tuned[,2] < 0.5, "No", "Yes"))

#Model evaluation using confusionMatrix, Accuracy, RMSE, MAE, and AUC
confusionMatrix(pred_rpart_tuned,testing_data_factor$treatment)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  120  27
##        Yes  35 132
##
##                Accuracy : 0.8025
##                  95% CI : (0.7542, 0.8451)
##     No Information Rate : 0.5064
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6048
##
##  Mcnemar's Test P-Value : 0.374
##
##             Sensitivity : 0.7742
##             Specificity : 0.8302
##          Pos Pred Value : 0.8163
##          Neg Pred Value : 0.7904
##              Prevalence : 0.4936
##          Detection Rate : 0.3822
##    Detection Prevalence : 0.4682
##       Balanced Accuracy : 0.8022
##
##        'Positive' Class : No
##
```

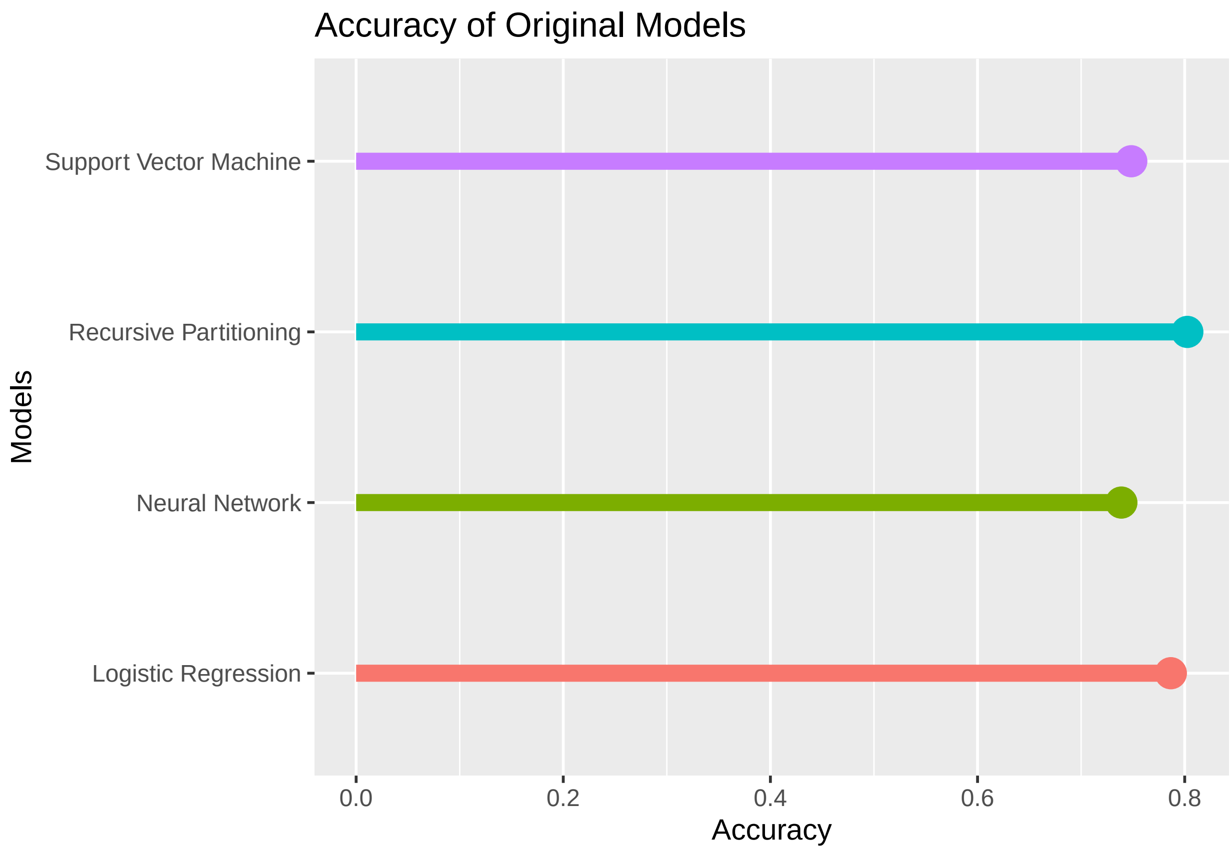
```r
accuracy_rpart_tuned <- accuracy(pred_rpart_tuned,testing_data_factor$treatment)
RMSE_rpart_tuned <- RMSE(as.numeric(testing_data_factor$treatment), as.numeric(pred_rpart_tuned))
MAE_rpart_tuned <- MAE(as.numeric(testing_data_factor$treatment), as.numeric(pred_rpart_tuned))
roc_rpart_tuned <- roc(as.numeric(testing_data_factor$treatment), as.numeric(pred_rpart_tuned))
```

Comparison of models

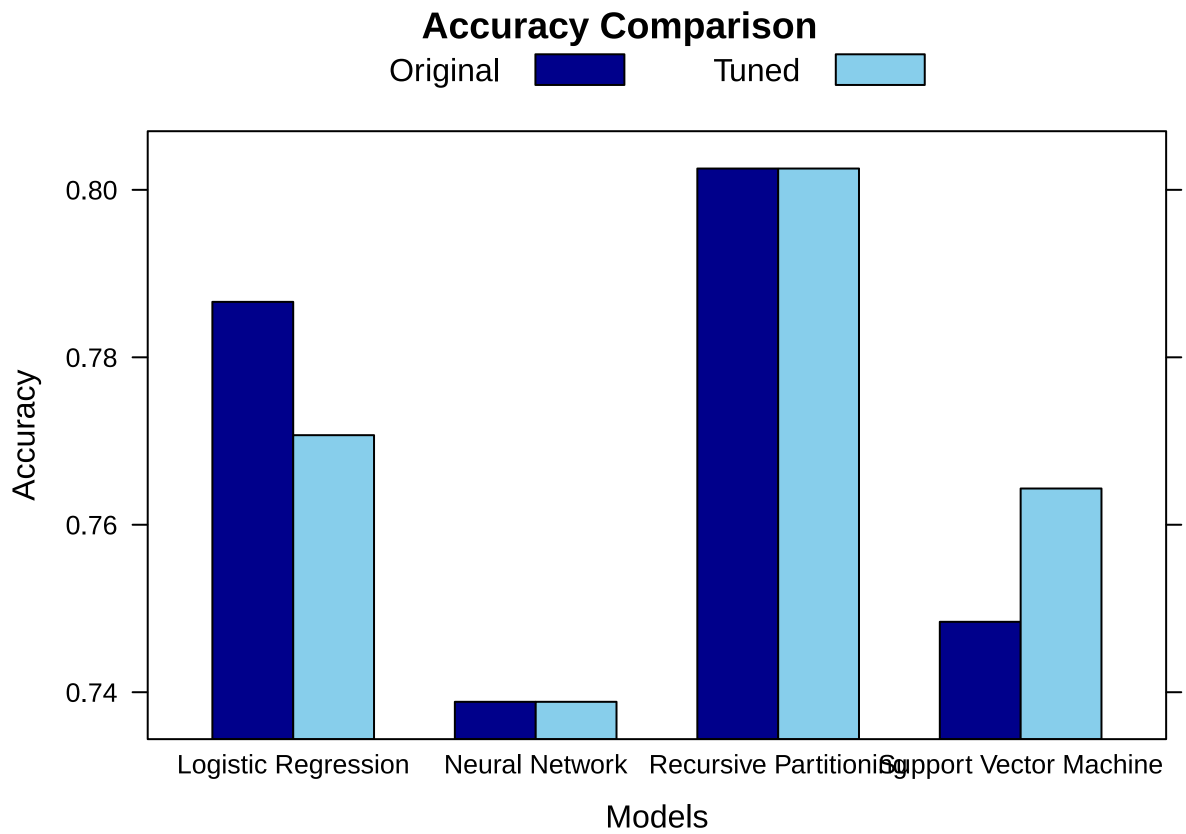
- For Comparison, plot of accuracy and other metric are shown below
- It is observed from the plot that recursive partition models performs the best amongst the others
- Recursive partition has the best accuracy along with the lowest RMSE and MAE error compared to other models
- Separate dataframe is created for MAE, RMSE and AUC values of each model

```r
############################
### Comparison of models ###
############################
```

```
#Creating a dataframe of model accuracy for comparison
comparison_acc <- data.frame(Models = c("Logistic Regression","Neural Network","Support Vector Machine"
                         Original = c(accuracy_glm,accuracy_nn,accuracy_svm,accuracy_rpart),
                         Tuned = c(accuracy_glm_tuned,accuracy_nn_tuned,accuracy_svm_tuned,accuracy_rpa

#Plotting the accuracies of the models
ggplot(data = comparison_acc, aes(x = Original, y = Models, color = Models, group = Models))+
  geom_segment(data = comparison_acc,aes(x=0,xend = Original, y = Models, yend = Models),size = 3)+
  geom_point(size = 5)+ggtitle("Accuracy of Original Models")+
  theme(legend.position = "none")+xlab("Accuracy")+ylab("Models")
```



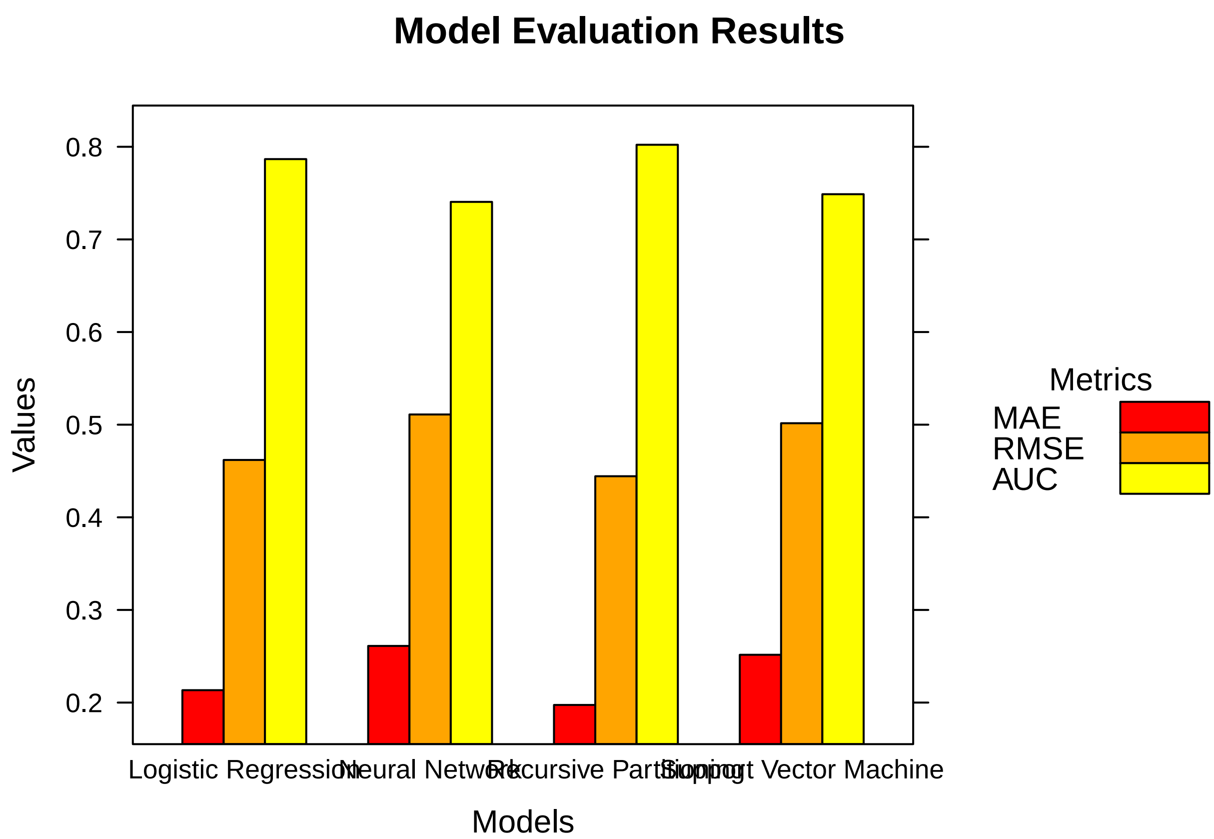

```
#Comparing the accuracy of original and tuned models
colors = c('Darkblue', 'skyblue')
barchart(Original+Tuned~Models,data=comparison_acc,run=best,
         ylab = "Accuracy",
         xlab = "Models",
         scales=list(alternating=1),
         auto.key=list(space="top", columns=2,points=FALSE,
                       rectangles=TRUE, cex.title=1),
         par.settings=list(superpose.polygon=list(col=colors)),main = "Accuracy Comparison")
```




```
#Comparing the model evaluation metrics of all the models
comparison <- data.frame(Models = c("Logistic Regression","Neural Network","Support Vector Machine","Re
                         MAE = c(MAE_glm,MAE_nn,MAE_svm,MAE_rpart), RMSE = c(RMSE_glm,RMSE_nn,RMSE_svm,
                         AUC = c(roc_glm$auc,roc_nn$auc,roc_svm$auc,roc_rpart$auc))
#Comparison Dataframe
comparison
```

```
##                    Models       MAE      RMSE       AUC
## 1    Logistic Regression 0.2133758 0.4619262 0.7867113
## 2         Neural Network 0.2611465 0.5110249 0.7405153
## 3 Support Vector Machine 0.2515924 0.5015898 0.7488131
## 4 Recursive Partitioning 0.1974522 0.4443560 0.8021911
```

```
#Plotting the comparison of model evaluation metrics of all the models
colors = c('red', 'orange', 'yellow')
barchart(MAE+RMSE+AUC~Models,data=comparison,run=best,
         ylab = "Values",
         xlab = "Models",scales=list(alternating=1),
         auto.key=list(space='right', rows=3,points=FALSE,
                        rectangles=TRUE,title="Metrics", cex.title=1),
         par.settings=list(superpose.polygon=list(col=colors)),main="Model Evaluation Results")
```



Construction of ensemble model

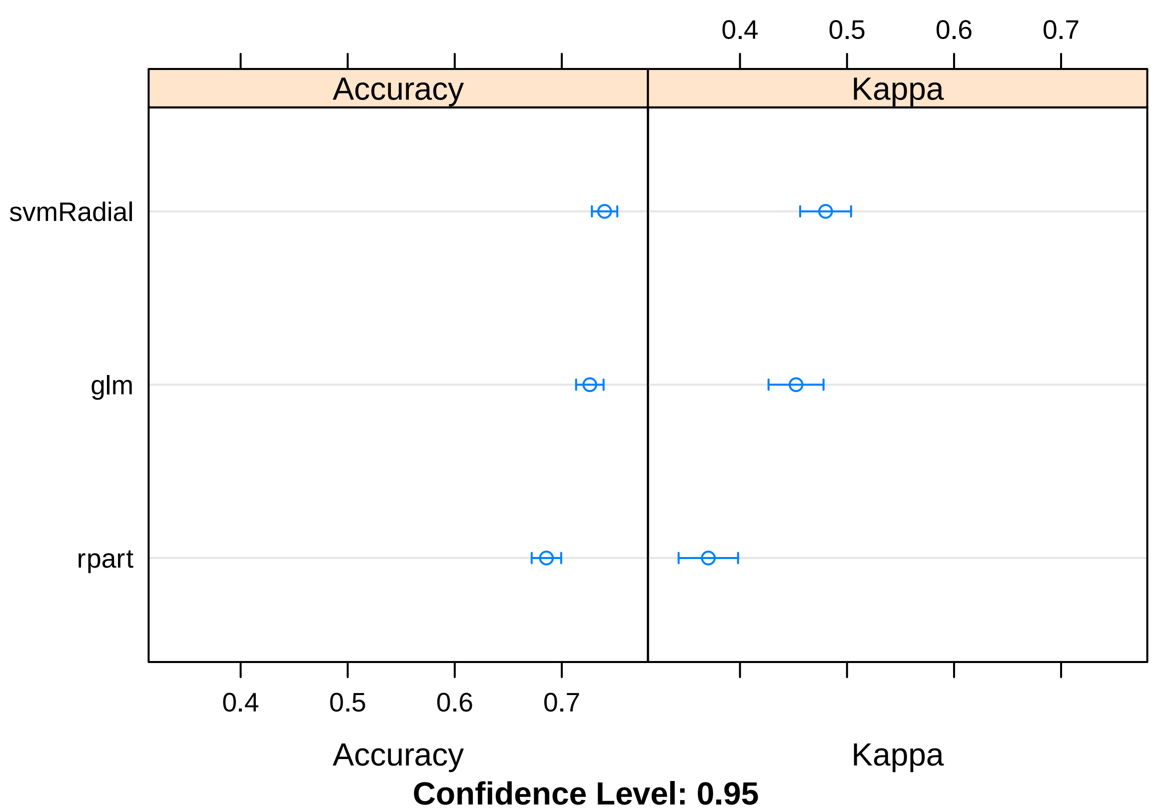
- Stack learner from caretEnsemble is used to build a stacked ensemble model
- For base model, I have used rpart, glm and svmRadial algorithms
- For the final stack learner, I have used glm i.e logistic regression
- Each model from base make individual predictions and final predictions is done from these outcomes using logistic regression
- Comparison of ensemble model with other model is also shown using a bar chart
- It is observed that accuracy of ensemble model is lesser compared to decision trees

```
######################################
### Construction of ensemble model ###
######################################

#Using train function from caret package to
#create a base model which consist 3 models
control <- trainControl(method="repeatedcv", number=10, repeats=3, savePredictions="all", classProbs=TRU
algorithmList <- c('rpart', 'glm', 'svmRadial')
set.seed(101)

#Training the models using training dataset
models <- caretList(treatment~., data=training_data_factor, trControl=control, methodList=algorithmList)
results <- resamples(models)
```

```r
#Observing the results using summary and dotplot
summary(results)
```

```
##
## Call:
## summary.resamples(object = results)
##
## Models: rpart, glm, svmRadial
## Number of resamples: 30
##
## Accuracy
##                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## rpart     0.6063830 0.6649216 0.6808511 0.6856636 0.7150336 0.7473684    0
## glm       0.6315789 0.7127660 0.7301792 0.7262126 0.7466965 0.7894737    0
## svmRadial 0.6702128 0.7263158 0.7368421 0.7399940 0.7572508 0.8000000    0
##
## Kappa
##                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## rpart     0.2091860 0.3285150 0.3627103 0.3704482 0.4280035 0.4937833    0
## glm       0.2625859 0.4255319 0.4604493 0.4523922 0.4933561 0.5790873    0
## svmRadial 0.3416177 0.4523889 0.4737424 0.4800114 0.5148224 0.6002215    0
```

```r
dotplot(results)
```



```r
#Creating a new traincontrol method for final stage of the stack learner
stackControl <- trainControl(method="repeatedcv", number=10, repeats=3, savePredictions="all", classPro
set.seed(101)

# Using glm at the final stage of the stack learner
stack.glm <- caretStack(models, method="glm", metric="Accuracy", trControl=stackControl)

#Printing the accuracy of the model
print(stack.glm)
```

```
## A glm ensemble of 3 base models: rpart, glm, svmRadial
##
## Ensemble results:
## Generalized Linear Model
##
## 2835 samples
##    3 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 2552, 2551, 2552, 2552, 2552, 2552, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7412185  0.4823754
```

```r
#Predicting the outcome for the stack ensemble learner
pred_ensemble <- predict(stack.glm, testing_data_factor)

#Model evaluation using confusionMatrix, Accuracy, RMSE, MAE, and AUC
confusionMatrix(pred_ensemble,testing_data_factor$treatment)
```

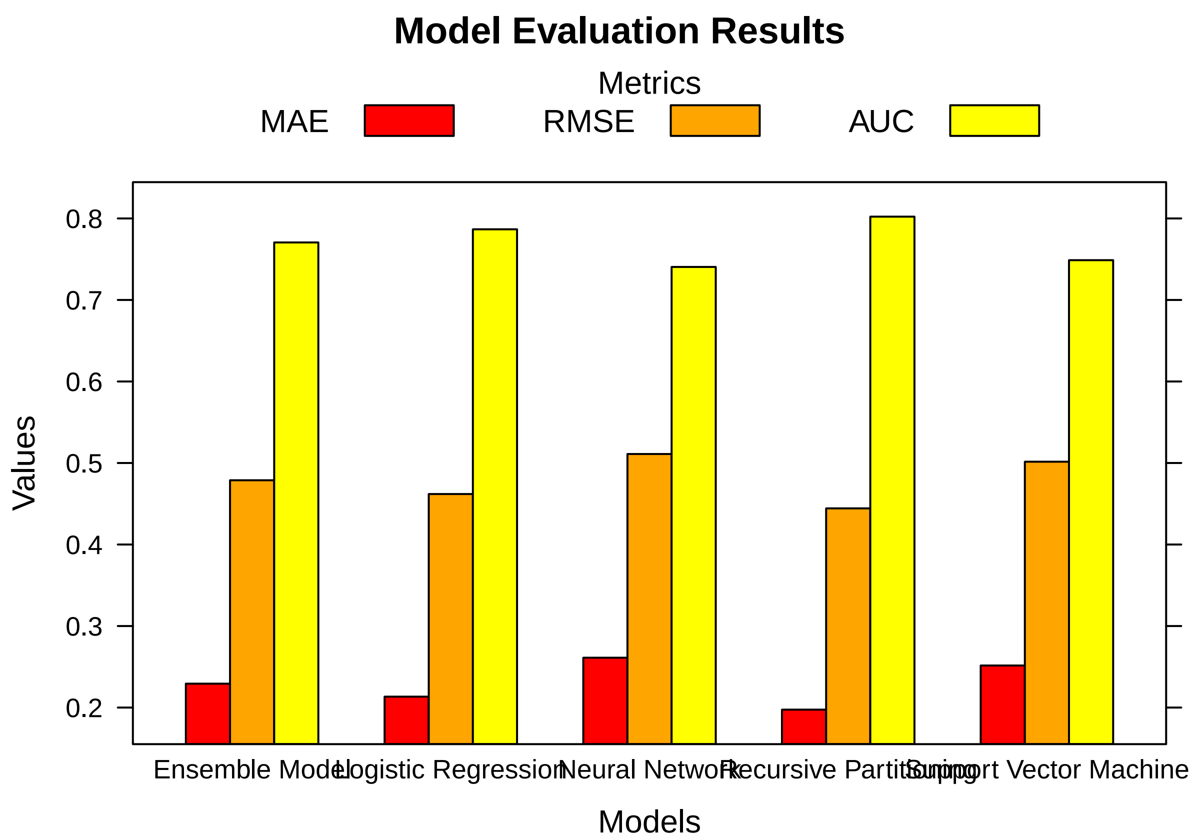
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  118  35
##        Yes  37 124
##
##                Accuracy : 0.7707
##                  95% CI : (0.7202, 0.816)
##     No Information Rate : 0.5064
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.5413
##
##  Mcnemar's Test P-Value : 0.9062
##
##             Sensitivity : 0.7613
##             Specificity : 0.7799
##          Pos Pred Value : 0.7712
##          Neg Pred Value : 0.7702
```

```
##              Prevalence : 0.4936
##          Detection Rate : 0.3758
##    Detection Prevalence : 0.4873
##       Balanced Accuracy : 0.7706
##
##        'Positive' Class : No
##
```

```r
accuracy_ensemble <- accuracy(pred_ensemble,testing_data_factor$treatment)
RMSE_ensemble <- RMSE(as.numeric(testing_data_factor$treatment), as.numeric(pred_ensemble))
MAE_ensemble <- MAE(as.numeric(testing_data_factor$treatment), as.numeric(pred_ensemble))
roc_ensemble <- roc(as.numeric(testing_data_factor$treatment), as.numeric(pred_ensemble))

#Comparing the model evaluation metrics of all the models
comparison_new <- data.frame(Models = c("Logistic Regression","Neural Network","Support Vector Machine"
                                        "Recursive Partitioning","Ensemble Model"),
                             MAE = c(MAE_glm,MAE_nn,MAE_svm,MAE_rpart,MAE_ensemble),
                             RMSE = c(RMSE_glm,RMSE_nn,RMSE_svm,RMSE_rpart,RMSE_ensemble),
                             AUC = c(roc_glm$auc,roc_nn$auc,roc_svm$auc,roc_rpart$auc,roc_ensemble$auc))
#Comparison Dataframe
comparison_new
```

```
##                   Models       MAE      RMSE       AUC
## 1    Logistic Regression 0.2133758 0.4619262 0.7867113
## 2         Neural Network 0.2611465 0.5110249 0.7405153
## 3 Support Vector Machine 0.2515924 0.5015898 0.7488131
## 4 Recursive Partitioning 0.1974522 0.4443560 0.8021911
## 5         Ensemble Model 0.2292994 0.4788521 0.7705823
```

```r
#Plotting the comparison of model evaluation metrics of all the models
colors = c('red', 'orange', 'yellow')
barchart(MAE+RMSE+AUC~Models,data=comparison_new,run=best,
         ylab = "Values",
         xlab = "Models",scales=list(alternating=1),
         auto.key=list(space='top', columns=3,points=FALSE,
                       rectangles=TRUE,title="Metrics", cex.title=1),
         par.settings=list(superpose.polygon=list(col=colors)),main="Model Evaluation Results")
```



Model Deployment

- For model deployment, I have used neural network model and stored it in a rds file
- This .rds file is used in RShiny app to make predictions
- I have deployed the RShiny app using Heroku

```
#RDS File for Shiny R app
saveRDS(neuralnet_model,'model.rds')
```

- A wordcloud is built using the comments feature to see what most of the professional felt like sharing

```
#Getting comments from the survey
comments <- data[,27]
comments1 <-  comments[!is.na(comments)]
comments_corpus <- Corpus(VectorSource(comments1))
#We can observe total documents using print
print(comments_corpus)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 164
```

```r
#To observe the content we use inspect() function
inspect(comments_corpus[1:2])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 2
##
## [1] I'm not on my company's health insurance which could be part of the reason I answered Don't know
## [2] I have chronic low-level neurological issues that have mental health side effects. One of my supe
```

```r
#We remove all the numbers and punctuations using tm_map() function. It is used to transform data.
corpus_clean <- tm_map(comments_corpus, tolower)
```

```
## Warning in tm_map.SimpleCorpus(comments_corpus, tolower): transformation drops
## documents
```

```r
corpus_clean <- tm_map(corpus_clean, removeNumbers)
```

```
## Warning in tm_map.SimpleCorpus(corpus_clean, removeNumbers): transformation
## drops documents
```

```r
corpus_clean <- tm_map(corpus_clean, removeWords, stopwords())
```

```
## Warning in tm_map.SimpleCorpus(corpus_clean, removeWords, stopwords()):
## transformation drops documents
```

```r
corpus_clean <- tm_map(corpus_clean, removePunctuation)
```

```
## Warning in tm_map.SimpleCorpus(corpus_clean, removePunctuation): transformation
## drops documents
```

```r
corpus_clean <- tm_map(corpus_clean, stripWhitespace)
```

```
## Warning in tm_map.SimpleCorpus(corpus_clean, stripWhitespace): transformation
## drops documents
```

```r
#We verify using inspect whether all unwanted characters are removed
inspect(corpus_clean[1:2])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 2
##
## [1]  companys health insurance part reason answered know many questions
## [2]  chronic lowlevel neurological issues mental health side effects one supervisors also experience
```

```r
#Using wordcloud to see most common words used in comments of the survey
wordcloud(corpus_clean,max.words=100 ,random.order=FALSE,rot.per=0.35,colors=brewer.pal(8,"RdYlBu"))
```