Capstone Project Final Report

Jihye Ko

# Introduction

**Importance**

These days, the importance of mental health is being raised. It is estimated that one in four or five of people would experience some mental health problems in lifetime. For this reason, it is very important to recognize signs of mental health problems and provide help which people can easily access to.

One of the reasons which cause mental issue is stress. People are stressed out in many different reasons, such as school, works, relationships and so on. Especially, many workers in companies suffer from mental health in some way. Taking care of employees is one of responsibilities that the company should have. By solving this problem, the company can predict how effectively their mental health benefits are used by employees.

**Question**

Can we predict the employee’s accessibility of seeking help of mental health problem in companies based on the size of the company, knowledge of wellness program and care options that the company supply, and protection of anonymity of employee’s mental health issue?

# Description of Data

**Dataset**

The dataset used for this project was retrieved from [kaggle.com](https://www.kaggle.com/osmi/mental-health-in-tech-survey)

**Description of Data**

The data that I am using is survey type. This dataset is from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. This dataset contains 26 variables which are related to mental and physical health and their accessibility about health services in their company, especially in tech companies in this case.  
The varialbes as follows:

Description of variables

Age  
Gender  
Country  
State: If you live in the United States, which state or territory do you live in?  
Self\_employed: Are you self-employed?  
Family\_history: Do you have a family history of mental illness?  
Treatment: Have you sought treatment for a mental health condition?  
Work\_interfere: If you have a mental health condition, do you feel that it interferes with your work?  
No\_employees: How many employees does your company or organization have?  
Remote\_work: Do you work remotely (outside of an office) at least 50% of the time?  
Tech\_company: Is your employer provide mental health benefits?  
Benefits: Does your employer provide mental health benefits?  
Care\_options: Do you know the options for mental health care your employer provides?  
Wellness\_program: Has your employer ever discussed mental health as part of an employee wellness program?  
Seek\_help: Does your employer provide resources to learn more about mentlah health issue and how to seek help?  
Anonymity: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?  
Leave: How easy is it for you to take medical leave for a mental health condition?  
Mental\_health\_consequence: Do you think that discussing a mental health issue with your employer would have negative consequences?  
Phys\_health\_consequence: Do you think that discussing a physical health issue with your employer would have negative consequences?  
Coworkers: Would you be willing to discuss a mental health issue with your coworkers?  
Supervisor: Would you be willing to discuss a mental health issue with your direct supervisor(s)?  
Mental\_health\_interview: Would you bring up a mental health issue with a potential employer in an interview?  
Phys\_health\_interview: Would you bring up a physical health issue with a potential employer in an interview?  
Mental\_vs\_physical: Do you feel that your employer takes ental health as seriously as physical health?  
Obs\_consequence: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?

# Lists of libraries

* library(rattle)
* library(rpart.plot)
* library(RColorBrewer)
* library(rpart)
* library(tidyr)
* library(dplyr)
* library(ggplot2)
* library(plyr)
* library(corrplot)

# Analysis and Cleaning of variables in the Dataset

**Load the data from csv file**

## Set this to your current path of datasets  
setwd("C:\\");  
## Loading data  
raw = read.csv(file='raw.csv',header=TRUE,sep=',')  
summary(raw)

## Age Gender Country state   
## Min. :-1.726e+03 F :251 United States :751 CA :138   
## 1st Qu.: 2.700e+01 M :995 United Kingdom:185 WA : 70   
## Median : 3.100e+01 OTHER: 13 Canada : 72 NY : 57   
## Mean : 7.943e+07 Germany : 45 TN : 45   
## 3rd Qu.: 3.600e+01 Ireland : 27 TX : 44   
## Max. : 1.000e+11 Netherlands : 27 (Other):390   
## (Other) :152 NA's :515   
## self\_employed family\_history treatment work\_interfere  
## No :1095 No :767 No :622 Never :213   
## Yes : 146 Yes:492 Yes:637 Often :144   
## NA's: 18 Rarely :173   
## Sometimes:465   
## NA's :264   
##   
##   
## no\_employees remote\_work tech\_company benefits   
## :451 No :883 No : 228 Don't know:408   
## 06월 25일 : 1 Yes:376 Yes:1031 No :374   
## 100-500 :176 Yes :477   
## 26-100 :289   
## 500-1000 : 60   
## More than 1000:282   
##   
## care\_options wellness\_program seek\_help anonymity   
## No :501 Don't know:188 Don't know:363 Don't know:819   
## Not sure:314 No :842 No :646 No : 65   
## Yes :444 Yes :229 Yes :250 Yes :375   
##   
##   
##   
##   
## leave mental\_health\_consequence  
## Don't know :563 Maybe:477   
## Somewhat difficult:126 No :490   
## 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   
##   
##   
##   
##   
##   
## comments   
## \* Small family business - YMMV. : 5   
## - : 1   
## : 1   
## (yes but the situation was unusual and involved a change in leadership at a very high level in the organization as well as an extended leave of absence) : 1   
## A close family member of mine struggles with mental health so I try not to stigmatize it. My employers/coworkers also seem compassionate toward any kind of health or family needs.: 1   
## (Other) : 155   
## NA's :1095

**Dataset**

1259 observations and 26 variables

**Data Wrangling**

1. Variable with missing values

The variable which had missing value was *no\_employees*, in other words, the size of the company. Since imputation was needed for dealing with missing values, the distribution of that varialbe was first analyzed. If it was normally distributed, then the mean can be used for missing values. If not, the median can be used for them.

Since the data of *no\_employees* were not normally distributed, the missing values were replaced by median (“100-500”). Also, the data of *no\_employees* were catagorized into two sizes: Medium or Large.

## Categorizing company into MEDIUM or LARGE (2 categories only)  
raw$no\_employees = plyr::revalue(raw$no\_employees, c("06월 25일"="MEDIUM", "100-500"="MEDIUM" , "26-100" = "MEDIUM" , "500-1000"="LARGE" , "More than 1000" = "LARGE"))  
## All unknown size companies default to MEDIUM  
raw$no\_employees[raw$no\_employees == ''] <- "MEDIUM"

1. Variable with variety answers

One of the variables, *Country*, had about more than 100 answers which made the further analysis difficult. Therefore, countires were catagorized with continents (South America, North America, Africa, Oceania, Asia and Europe).

## Categorizing countries into continents (6 categories)  
raw$Country = plyr::revalue(raw$Country, c("Mexico" = "North America", "Canada"="North America", "United States"="North America" , "United Kingdom" = "North America", "Costa Rica" = "North America", "Bahamas, The" = "North America"))  
raw$Country = plyr::revalue(raw$Country, c("Thailand" = "Asia", "Singapore" = "Asia", "Israel" = "Asia", "Japan"="Asia", "India" = "Asia", "China"="Asia", "Philippines" = "Asia"))  
raw$Country = plyr::revalue(raw$Country, c("Czech Republic" = "Europe", "Denmark" = "Europe", "Hungary" = "Europe", "Bosnia and Herzegovina"="Europe", "Greece" = "Europe", "France" = "Europe", "United Kingdom = Europe", "Portugal" = "Europe", "Switzerland" = "Europe", "Georgia" = "Europe", "Moldova" = "Europe", "Poland" = "Europe", "Austria" = "Europe", "Germany" = "Europe", "Slovenia" = "Europe", "Russia" = "Europe", "Ireland" = "Europe", "Italy" = "Europe", "Bulgaria" = "Europe", "Sweden" = "Europe", "Latvia" = "Europe", "Romania"="Europe", "Belgium" = "Europe", "Spain"="Europe", "Finland"="Europe", "Netherlands"="Europe", "Croatia" = "Europe", "Norway"="Europe"))  
raw$Country = plyr::revalue(raw$Country, c("Brazil"="South America", "Colombia" = "South America", "Uruguay"="South America"))  
raw$Country = plyr::revalue(raw$Country, c("Australia" = "Oceania", "New Zealand" = "Oceania"))  
raw$Country = plyr::revalue(raw$Country, c("Nigeria" = "Africa", "Zimbabwe"="Africa", "South Africa" = "Africa"))  
summary(raw$Country)

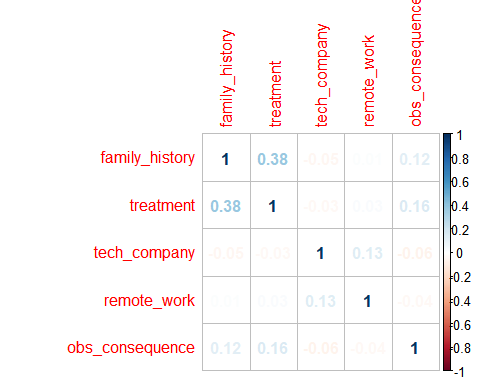
## Oceania Europe North America South America Asia   
## 29 177 1013 9 23   
## Africa   
## 8

# Correlation among variables

**Variables with two answers (Yes/No)**

Some variables (i.e. family\_history, treatment, tech\_company, remote\_work, obs\_consequence) had either ‘Yes’ or ‘No’ answers. To observe the relationship among variables, corrPlot was used. By default, corrPlot needs numeric values not string. Therefore, changing string to numeric values was necessary steps. Then, correlations between variables were plotted with correlation coefficients.

## Changing string to numeric values  
raw$family\_history = plyr::revalue(raw$family\_history, c("Yes" = 1, "No" = 3))  
raw$treatment = plyr::revalue(raw$treatment, c("Yes" = 1, "No" = 3))  
raw$tech\_company = plyr::revalue(raw$tech\_company, c("Yes" = 1, "No" =3))  
raw$remote\_work = plyr::revalue(raw$remote\_work, c("Yes" = 1, "No" = 3))  
raw$obs\_consequence = plyr::revalue(raw$obs\_consequence, c("Yes" = 1, "No" = 3))  
  
raw$family\_history <- as.numeric(as.character(raw$family\_history))  
raw$treatment <- as.numeric(as.character(raw$treatment))  
raw$tech\_company <- as.numeric(as.character(raw$tech\_company))  
raw$remote\_work <- as.numeric(as.character(raw$remote\_work))  
raw$obs\_consequence <- as.numeric(as.character(raw$obs\_consequence))  
  
## corrPlot to examine the relationships among variables  
MentalHealth\_Data\_1 <- subset(raw, select = c(family\_history, treatment, tech\_company, remote\_work, obs\_consequence))  
N <- cor(MentalHealth\_Data\_1)  
corrplot(N, method = "number")



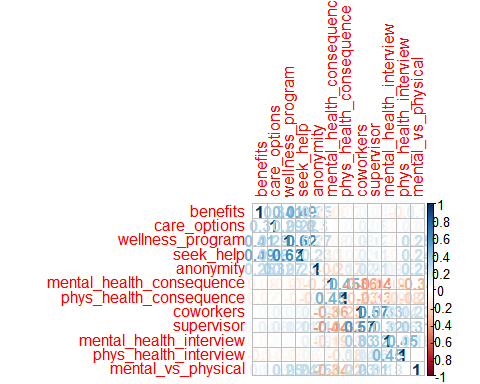
*Conclusion 1:*

From this corrPlot, there was no strong correlation between variables which had two answers.

**Variables with three answers (Yes/Don’t know/No)**

Other variables (i.e. benefits, care\_options, wellness\_program, seek\_hlep, anonymity, mental\_health\_consequence, phys\_health\_consequence, coworkers, supervisor, mental\_health\_interview, phys\_health\_interview, mental\_vs\_physical) had three answers as follows: ‘Yes’, ‘Don’t Know’, and ‘No’. Just like the above, strings were changed to numbers and correlations between variables were examined.

## Changing string to numeric values  
raw$benefits = plyr::revalue(raw$benefits, c("Yes" = 1, "Don't know" = 2, "No" = 3))  
raw$care\_options = plyr::revalue(raw$care\_options, c("Yes" = 1, "Not sure" = 2, "No" = 3))  
raw$wellness\_program = plyr::revalue(raw$wellness\_program, c("Yes" = 1, "Don't know" = 2, "No" = 3))  
raw$seek\_help = plyr::revalue(raw$seek\_help, c("Yes" = 1, "Don't know" = 2, "No" =3))  
raw$anonymity = plyr::revalue(raw$anonymity, c("Yes" = 1, "Don't know" = 2, "No" = 3))  
raw$mental\_health\_consequence = plyr::revalue(raw$mental\_health\_consequence, c("Yes" = 1, "Maybe" = 3, "No" =3))  
raw$phys\_health\_consequence = plyr::revalue(raw$phys\_health\_consequence, c("Yes" = 1, "Maybe" = 2, "No" = 3))  
raw$coworkers = plyr::revalue(raw$coworkers, c("Yes" = 1, "Some of them" = 2, "No" =3))  
raw$supervisor = plyr::revalue(raw$supervisor, c("Yes" = 1, "Some of them" = 2, "No" = 3))  
raw$mental\_health\_interview = plyr::revalue(raw$mental\_health\_interview, c("Yes" = 1, "Maybe" = 2, "No" =3))  
raw$phys\_health\_interview = plyr::revalue(raw$phys\_health\_interview, c("Yes" = 1, "Maybe" = 2, "No" =3))  
raw$mental\_vs\_physical = plyr::revalue(raw$mental\_vs\_physical, c("Yes" = 1, "Don't know" = 2, "No" = 3))  
  
raw$benefits <- as.numeric(as.character(raw$benefits))  
raw$care\_options <- as.numeric(as.character(raw$care\_options))  
raw$wellness\_program <- as.numeric(as.character(raw$wellness\_program))  
raw$seek\_help <- as.numeric(as.character(raw$seek\_help))  
raw$anonymity <- as.numeric(as.character(raw$anonymity))  
raw$mental\_health\_consequence <- as.numeric(as.character(raw$mental\_health\_consequence))  
raw$phys\_health\_consequence <- as.numeric(as.character(raw$phys\_health\_consequence))  
raw$coworkers <- as.numeric(as.character(raw$coworkers))  
raw$supervisor <- as.numeric(as.character(raw$supervisor))  
raw$mental\_health\_interview <- as.numeric(as.character(raw$mental\_health\_interview))  
raw$phys\_health\_interview <- as.numeric(as.character(raw$phys\_health\_interview))  
raw$mental\_vs\_physical <- as.numeric(as.character(raw$mental\_vs\_physical))  
  
## corrPlot to examine the relationships among variables  
MentalHealth\_Data\_2 <- subset(raw, select = c(benefits, care\_options, wellness\_program, seek\_help, anonymity, mental\_health\_consequence, phys\_health\_consequence, coworkers,supervisor, mental\_health\_interview, phys\_health\_interview, mental\_vs\_physical))  
M <- cor(MentalHealth\_Data\_2)  
corrplot(M, method = "number")



*Conclusion 2:*

From this corrPlot, there was a quite strong relationship between *wellness\_program* and *seek\_help* whose correlation coefficient was 0.62. Since *seek\_help* had a positive correlation (not significant) with other variables (i.e. *benefits*), this variable would be proper for a dependent variable.

# Prediction

**Approach**

1. Make prediction models using *library(rpart)*, and choose the most proper one based on *CP values* and *variable importance*.
2. Predict the accessibility of seeking help related to mental health in a company using *classification trees*.
3. Test the model on test data (*limitation*: in this case, there is no test data, so the original data was used and the accuracy of the model was determined).

**Make prediction models**

In this project, *seek\_help* was used as a dependent variable, and other variables were used as independent variables. The complexity parameter (CP) is used to control the size of the decision tree and to select the optimal tree size.

Here are some models generated to find the most proper prediction model.

## Using the complete data to predict seek\_help a categorical variable. Using method="class"  
fit <- rpart(seek\_help ~ no\_employees + wellness\_program, method = "class", data = raw)  
fit1 <- rpart(seek\_help ~ no\_employees + wellness\_program + mental\_vs\_physical, method = "class", data = raw)  
fit2 <- rpart(seek\_help ~ no\_employees + wellness\_program + care\_options + anonymity + mental\_vs\_physical, method = "class", data = raw)  
fit3 <- rpart(seek\_help ~ no\_employees + wellness\_program + care\_options + anonymity + mental\_vs\_physical + mental\_health\_consequence + mental\_health\_interview, method = "class", data = raw)  
fit4 <- rpart(seek\_help ~ no\_employees + wellness\_program + care\_options, method = "class", data = raw)  
fit5 <- rpart(seek\_help ~ no\_employees + wellness\_program + Country, method = "class", data = raw)  
fit6 <- rpart(seek\_help ~ no\_employees + wellness\_program + care\_options + anonymity, method = "class", data = raw)

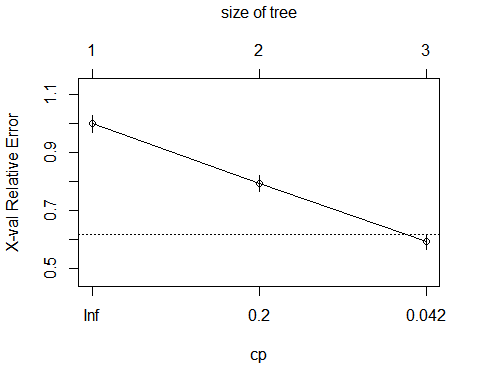
In this report, not all results of all models were reported. However, the interesting and common result was that only *wellness\_program* was used in tree construction in all models generated. Therefore, *CP values* were also the same among models.

The only difference was the *variable importance*. By comparing the importance of each variable in each model, it was determined that **fit6** (dependent variable: *seek\_help*, independent variables: *no\_employees, wellness\_program, care\_options, anonymity*) is the most proper prediction model (with variable importance as descending orders: *wellness\_program (77), care\_options (9), no\_employees (9), anonymity (6)*). Therefore, in this report, only the result of **fit6** is reported, and the model tree was generated as below.

## Tuning variable cp with best value of 0.01.   
printcp(fit6)

##   
## Classification tree:  
## rpart(formula = seek\_help ~ no\_employees + wellness\_program +   
## care\_options + anonymity, data = raw, method = "class")  
##   
## Variables actually used in tree construction:  
## [1] wellness\_program  
##   
## Root node error: 613/1259 = 0.48689  
##   
## n= 1259   
##   
## CP nsplit rel error xerror xstd  
## 1 0.23002 0 1.00000 1.00000 0.028932  
## 2 0.17781 1 0.76998 0.79282 0.028180  
## 3 0.01000 2 0.59217 0.59217 0.026220

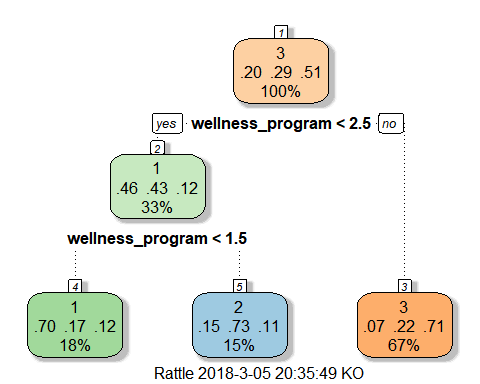
plotcp(fit6)



## Understanding the summary of fit  
summary(fit6)

## Call:  
## rpart(formula = seek\_help ~ no\_employees + wellness\_program +   
## care\_options + anonymity, data = raw, method = "class")  
## n= 1259   
##   
## CP nsplit rel error xerror xstd  
## 1 0.2300163 0 1.0000000 1.0000000 0.02893165  
## 2 0.1778140 1 0.7699837 0.7928222 0.02817958  
## 3 0.0100000 2 0.5921697 0.5921697 0.02622008  
##   
## Variable importance  
## wellness\_program care\_options no\_employees anonymity   
## 77 9 9 6   
##   
## Node number 1: 1259 observations, complexity param=0.2300163  
## predicted class=3 expected loss=0.4868944 P(node) =1  
## class counts: 250 363 646  
## probabilities: 0.199 0.288 0.513   
## left son=2 (417 obs) right son=3 (842 obs)  
## Primary splits:  
## wellness\_program < 2.5 to the left, improve=150.75030, (0 missing)  
## no\_employees splits as -RL, improve= 66.73345, (0 missing)  
## care\_options < 1.5 to the left, improve= 36.97559, (0 missing)  
## anonymity < 1.5 to the left, improve= 31.64966, (0 missing)  
## Surrogate splits:  
## no\_employees splits as -RL, agree=0.713, adj=0.134, (0 split)  
##   
## Node number 2: 417 observations, complexity param=0.177814  
## predicted class=1 expected loss=0.5443645 P(node) =0.3312153  
## class counts: 190 178 49  
## probabilities: 0.456 0.427 0.118   
## left son=4 (229 obs) right son=5 (188 obs)  
## Primary splits:  
## wellness\_program < 1.5 to the left, improve=63.410290, (0 missing)  
## care\_options < 1.5 to the left, improve=44.393380, (0 missing)  
## anonymity < 1.5 to the left, improve=26.136410, (0 missing)  
## no\_employees splits as -RL, improve= 9.995659, (0 missing)  
## Surrogate splits:  
## care\_options < 1.5 to the left, agree=0.727, adj=0.394, (0 split)  
## anonymity < 1.5 to the left, agree=0.664, adj=0.255, (0 split)  
## no\_employees splits as -RL, agree=0.578, adj=0.064, (0 split)  
##   
## Node number 3: 842 observations  
## predicted class=3 expected loss=0.2909739 P(node) =0.6687847  
## class counts: 60 185 597  
## probabilities: 0.071 0.220 0.709   
##   
## Node number 4: 229 observations  
## predicted class=1 expected loss=0.2969432 P(node) =0.1818904  
## class counts: 161 40 28  
## probabilities: 0.703 0.175 0.122   
##   
## Node number 5: 188 observations  
## predicted class=2 expected loss=0.2659574 P(node) =0.1493249  
## class counts: 29 138 21  
## probabilities: 0.154 0.734 0.112

## Plotting the model tree  
fancyRpartPlot(fit6)



**Prediction**

Based on the prediction model, the test data should be predicted to see whether the prediction model is reliable or not. However, in this case, there is no test data, so the original data was reused to see the accuracy of the model.

## This is not the ideal way. We are using the same inputdata for testing accuracy.  
## Just for demonstration purpose. Ideally we should spit into training , testing and validation data sets  
  
## Calculating accuracy of the model  
t\_pred = predict(fit6,raw,type="class")  
t = raw$seek\_help  
accuracy = sum(t\_pred == t)/length(t)  
print(accuracy)

## [1] 0.7116759

Accuracy: 0.7117

This means that the model has about 71% of accuracy in prediction.

*Conclusion:*

The question for this project was whether we can predict the employee’s willingness of seeking help of mental health problem in companies based on the size of the company, knowledge of wellness program and care options that the company supply, and protection of anonymity of employee’s mental health issue. Based on the *classification tree*, it was found out that there is a strong correlation between *seek\_help (the employee’s willingness of seeking help of mental health problem and the companies ability to provide resources)* and *wellness\_program (the knowledge of wellness program and care options that the company supply)*. Even though other variables, such as *care\_options, no\_employees, anonymity* did not show a strong correlation to dependent variable, they still can be good predictors and improve prediction model.

# Further research

* To build a better prediction model, it is necessary to have a test data (i.e. survey conducted in following year).
* For future survey, it would be better to lessen the number of questions. In other words, variables, like *benefits* and *care\_options* can be combined as one question because both are asking whether employees know the company provides mental health benefits or not. By combinding variables which have similar questions, the better prediction model can be constructed.

# Recommendations

* Since an employer who discusses mental health to employees provides more resources and chances to seek help for them who have mental health issue, it is necessary that the employer needs to know not only physical health care systems but also mental health care systems and programs. By doing so, employees can easily access to ask help about their mental health problems, and probably their work efficiency would increase.
* The case when an employer gives information about mental health program and approaches to employees gives more accessibility and awareness about the care options for them. So, it is recommended for the employer to approach to employees first about their mental health.
* The larger companies are more likely aware and provide resources to learn more about mental health issues and how to seek help. Therefore, it is necessary that small to medium sized companies should care and know more about mental health care programs for their employees.