MGT 6203 Team Project:
Mental Health Disorder Analysis

Team 5: Cheng En Li, Eddy Trang, Takuya Wakayama, Taylor Lee, Victoria Kwong

Introduction

More people are paying attention to mental health than ever before. Top athletes, such as professional tennis player Naomi Osaka, legendary American swimmer Michael Phelps and NBA player Kevin Love, have shared their personal experiences and battles with mental health disorders, which includes Depression and Attention Deficit/Hyperactive Disorder (ADHD). Due to their fame status, they have managed to direct the public's attention to the prevalence and importance of mental health.

Over the years, there is an increase in people using meditation apps to relieve stress. This phenomenon is further exacerbated during the COVID-19 period. The global spending on mental health and well-being apps increased from $203 million in 2019 to $269 million in 2020. This spending is estimated to jump from $269 million in 2020 to $372 million in 2021. One study has also found that feelings of depression and anxiety increase by four fold in December 2020, as compared to the first half of 2019. According to the *Wall Street Journal*, meditation apps have a $1.2 billion market and will continue to grow, showing a huge business potential.

It is well-known worldwide that healthcare workers are suffering from burnout and other mental disorders due to the pandemic. About 20% of healthcare workers experience anxiety and depression. However, little is known about mental health in the tech industry. Thus, we want to do more research about it. We found mental health surveys for the tech industry collected by Open Source Mental Health (OSMH). The data has been collected since 2014. Here are some research questions that we hope to answer:

1. What are the factors that might explain cases of mental health disorders at the workplace?

2. Does the pandemic have an effect on mental health disorders?
3. Does the pandemic have an effect on the attitudes towards mental health disorders?

Description of the datasets

Raw data

Here are the dataset(s) that we will be using:

1) OSMH/OSMI Mental Health in Tech Survey(2019)
2) OSMH/OSMI Mental Health in Tech Survey(2020)
3) OSMH/OSMI Mental Health in Tech Survey(2021)

Description of dataset / key variables to analyze

We chose these 3 datasets because they are in consecutive years before (2019) and during the COVID-19 pandemic (2020-2021). The dataset provides data on attitudes towards mental health, as well as the frequency of mental health disorders in the tech workplace.

Dependent variable: Do you currently have a mental health disorder?
Independent variables: age, gender, family history, year (as an index of before or after the pandemic), etc.

Approach and Methodology

The dependent variable for our first research question is represented by the question "Do you currently have a mental health disorder?".

For the first research question, we first picked out the features that may be related to the cases of mental health disorders at the workplace. We then extract these features from the three dataset and merge them into one. After which, we did various data cleaning as follows:

- Rename the values for certain columns so that it is more understandable e.g. renaming “0” and “1” into “Self-Employed” and “Employed” respectively for the “Employment Type” column and renaming “0” and “1” into “Non-Tech” and “Tech” for the “Job Scope” column.
- Classify various expressions in the gender column into 4 distinct groups (M, F, Others and Unknown) .
- Keep the top 3 countries in ‘CountryLiveIn’ and ‘CountryWorkIn’ columns and combine the rest into “others”.
- Remove unreasonable age (e.g. age < 18 and > 100) and group the rest into 5 different age groups (< 30, 30-39, 40-49, 50-59, >60)
- Replace missing value with “unknown”.

Next, we found the probability of each feature and ran a chi-square test to test its significance level.

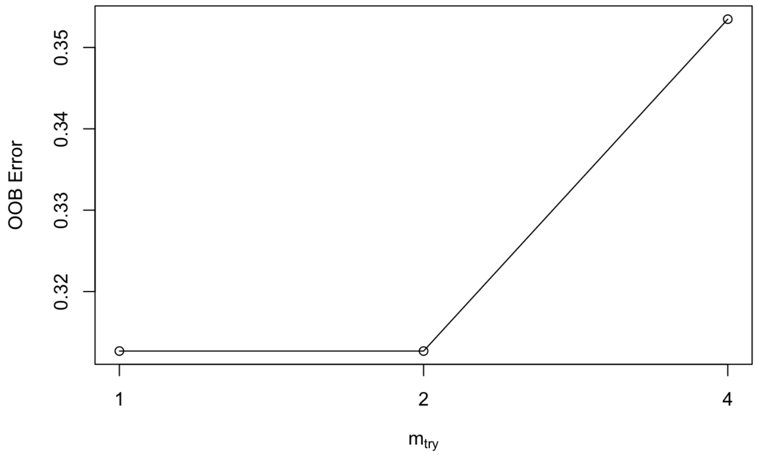
After which, we were interested in finding which feature is the most important feature in determining the presence of mental health disorders. As there are 4 possible answers to our outcome, we have decided to use a multi-class classification mode.

We chose the random forest model to make the classifier as it has a high accuracy rate and offers a good feature selection indicator through the use of Gini importance or mean decrease in impurity (MDI) to calculate the importance of each feature.

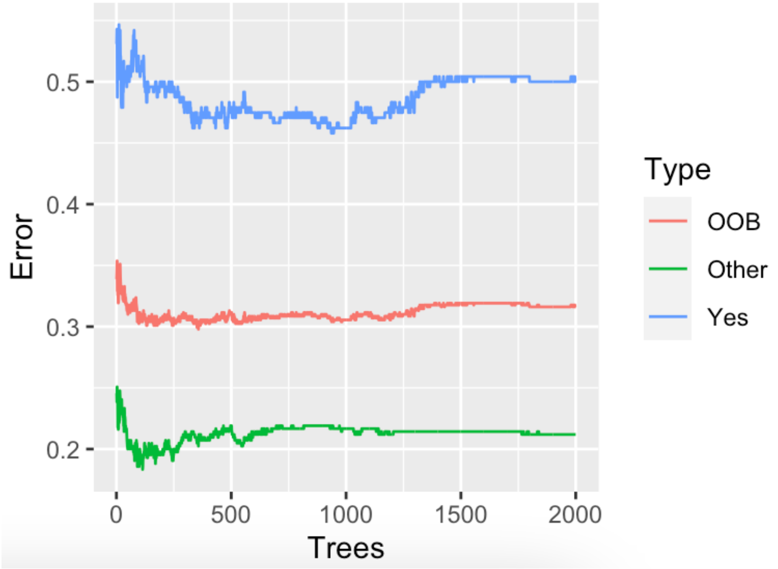
However, it did not work well on our dataset. It has high class errors for the “Don’t know” and “Possibly” categories. As a result, we merged the 4 possible outcomes into 2 outcomes, namely “Yes” and “Other”. The “Yes” outcome will include all the “Yes” answers. The “Other” will include the “Don’t Know”, ‘Possibly” and “No” outcomes.

To improve the accuracy of our model, we also performed parameter tuning using grid search and applied the best parameters in the random forest model. First, we try to find the number of features to consider at each split point(mtry) which resulted in the lowest error.

As you can see, the lowest OOB Error rate is achieved when the mtry is 2.



Then we search for the number of trees grown(ntree). The plot below shows the result we tried ntree=2000. We can see that the OOB Error rate converges around when ntree is around 500, the default number.



After which, we rebuild our model with this information.

For the other research questions, we are interested to see if the pandemic has an effect on the presence of mental health disorders and attitudes on mental health disorders. We rename the 2019 dataset into “before” and the 2020 and 2021 dataset into “after”. As all the variables are categorical data, we use the chi-square test to find its significance.

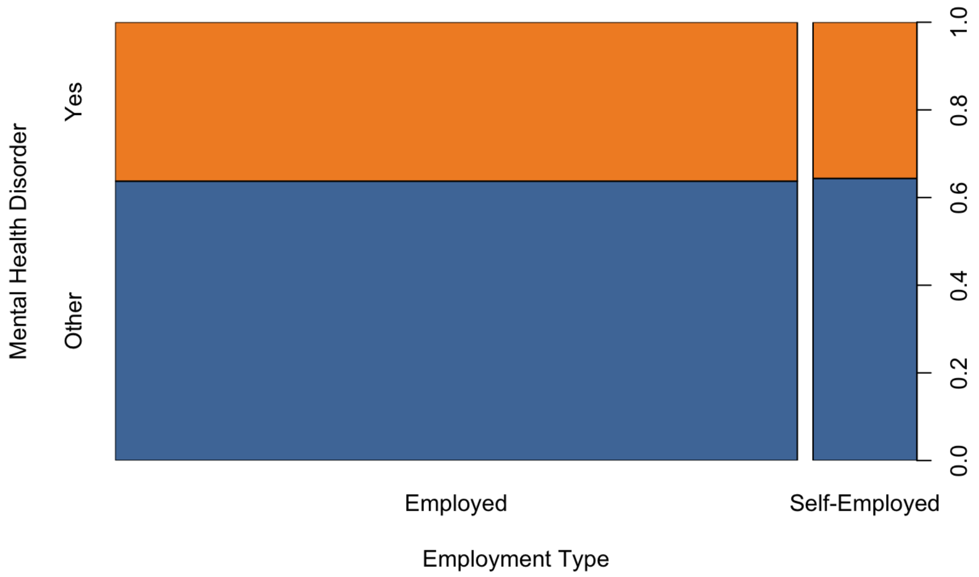
Results and Discussion

1. What are the factors that might explain cases of mental health disorders at the workplace?

We have decided to use mosaic plots to show our results as our features and outcome are both categorical data. Mosaic plot is able to show the percentage and a pictorial view of the amount of data available. To determine whether these factors are significant factors in determining one's likelihood of developing mental health disorders, we have decided to run the chi-squared test. The following chi-square test function is created:

```
Chi_func <- function(v1, v2){
  tb <- table(v1, v2)
  chi <- chisq.test(tb, correct = F)
  chi$p.value
}
```

1. Employment Type vs. Presence of Mental Health Disorders



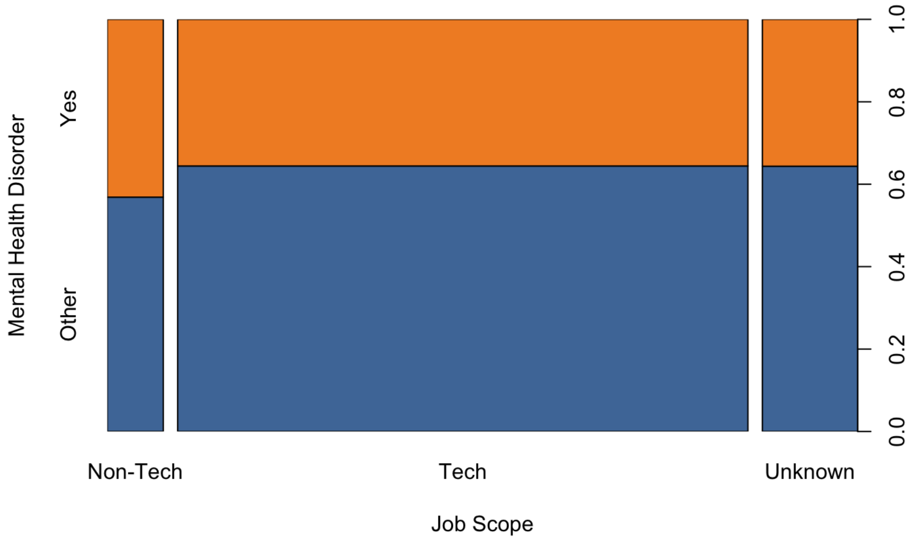
From the graph above, we can conclude that there is no difference in the likelihood of getting mental health disorders between employment types.

```
Chi_func(dat$EmploymentType, dat$CurrentlyMentalHealthDisorder)
```

```
# [1] 0.9107304
```

We also did a chi-squared test and found that we failed to reject the null hypothesis that there is no relationship between employment type and mental health disorders as the p-value is >0.05.

2. Job Scope vs. Presence of Mental Health disorders

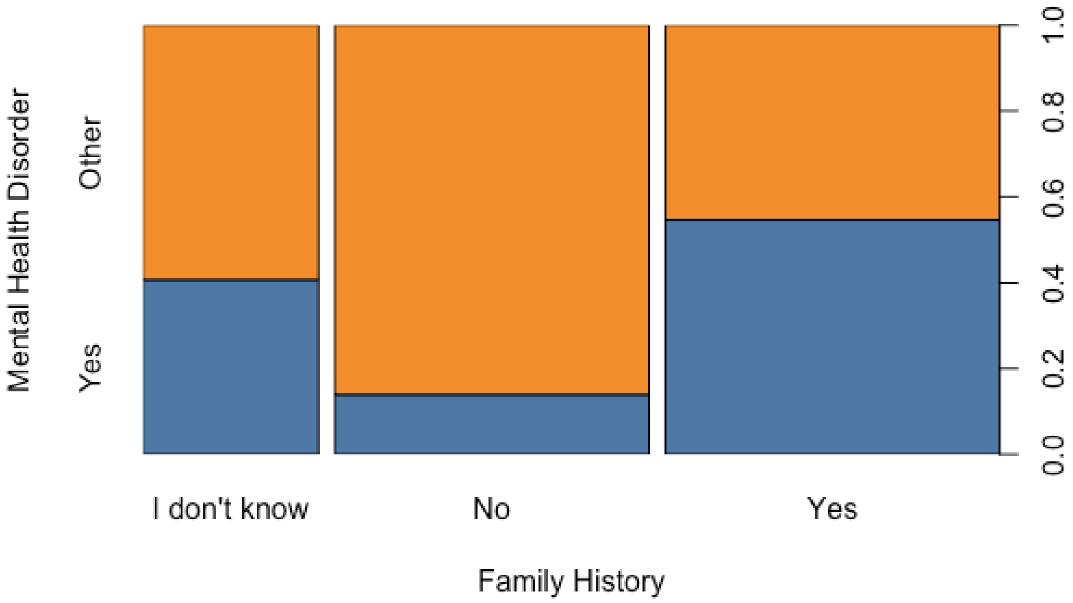


From the graph above, we can conclude that there is no difference in the likelihood of getting mental health disorders between different job scopes.

```
Chi_func(dat$JobScope, dat$CurrentlyMentalHealthDisorder)
# [1] 0.5592186
```

We ran the chi-squared test and found that we failed to reject the null hypothesis that there is no relationship between job scope and mental health disorders as the p-value is >0.05.

3. Family History vs. Presence of Mental Health Disorders

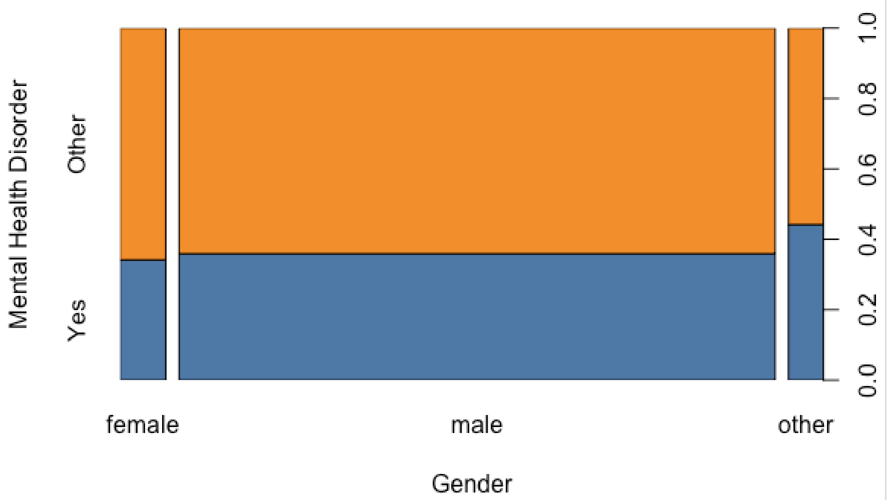


We can conclude that having one who has a family history with mental health disorders are more likely to develop mental health disorders as well.

```
# FamilyHistory ***
Chi_func(dat$FamilyHistory, dat$CurrentlyMentalHealthDisorder)
# [1] 2.888205e-21
```

We ran the chi-squared test and found that we can reject the null hypothesis that there is no relationship between family history and mental health disorders as the p value is <0.05.

4. Gender vs. Presence of Mental Health Disorders

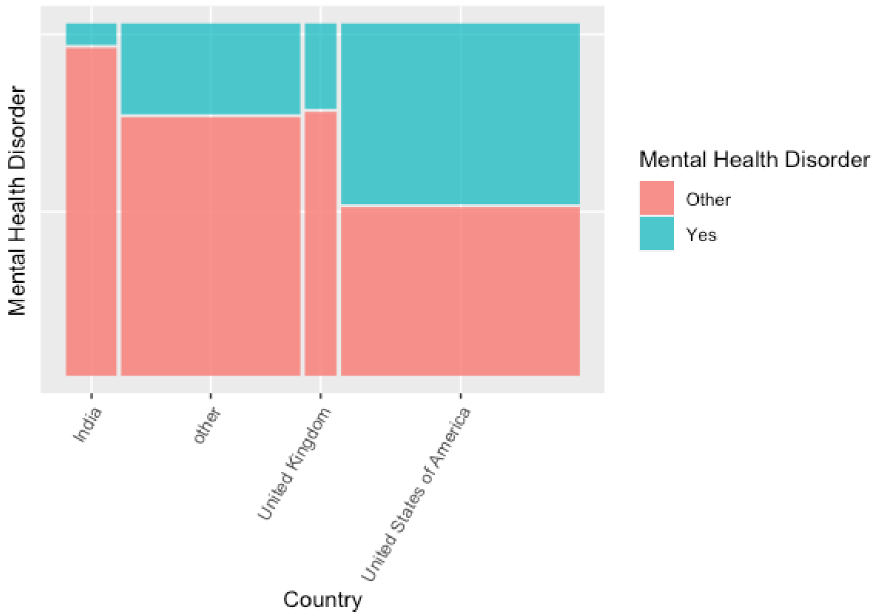


Others include people who do not identify themselves as cisgender. From the graph above, the likelihood of cis-female and cis-male who develop mental health disorder is about the same. However, the likelihood of developing mental health disorders for other genders is much higher. However, this could also be attributed to the low sample size of the “other” category as depicted in the graph.

```
Chi_func(dat$Gender, dat$CurrentlyMentalHealthDisorder)
# [1] 0.5955878
```

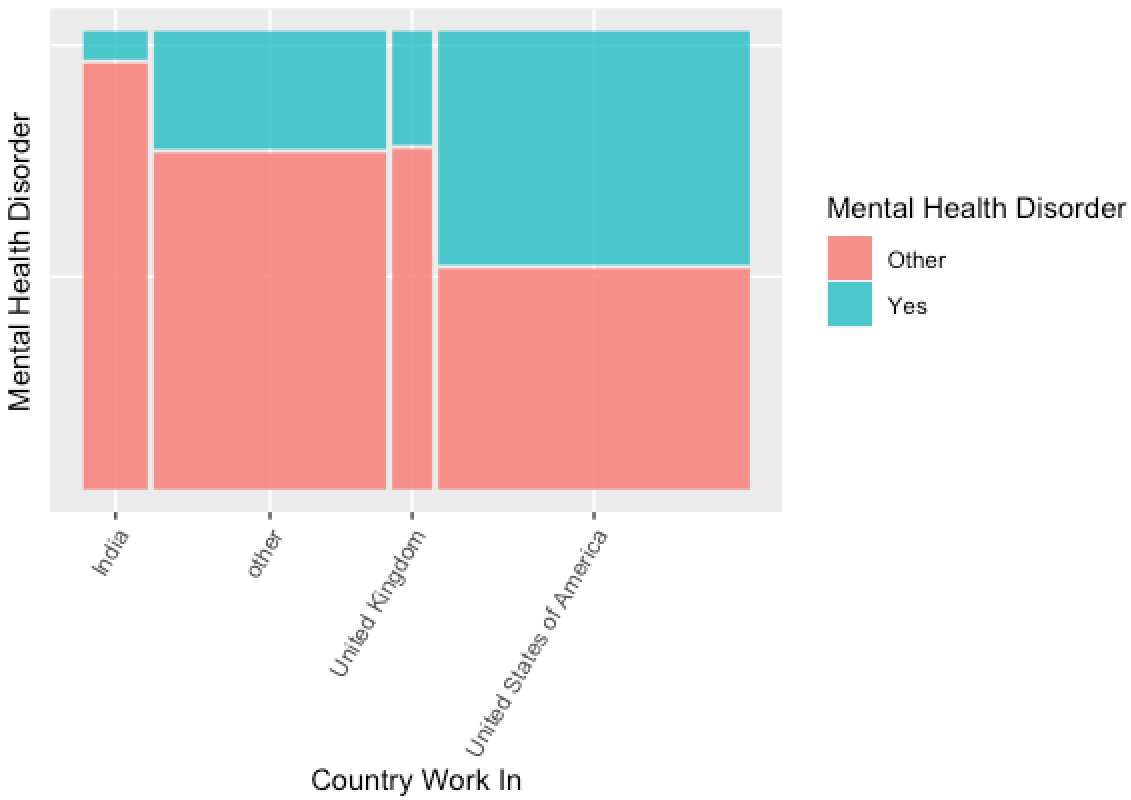
We ran the chi-squared test and found that we failed to reject the null hypothesis that there is no relationship between gender and mental health disorder as the p-value is >0.05

5. Country that people live in vs. Presence of Mental Health disorders



```
Chi_func(dat$CountryLiveIn, dat$CurrentlyMentalHealthDisorder)
# [1] 1.323527e-15
```

6. Country that people work in vs. Presence of Mental Health disorders

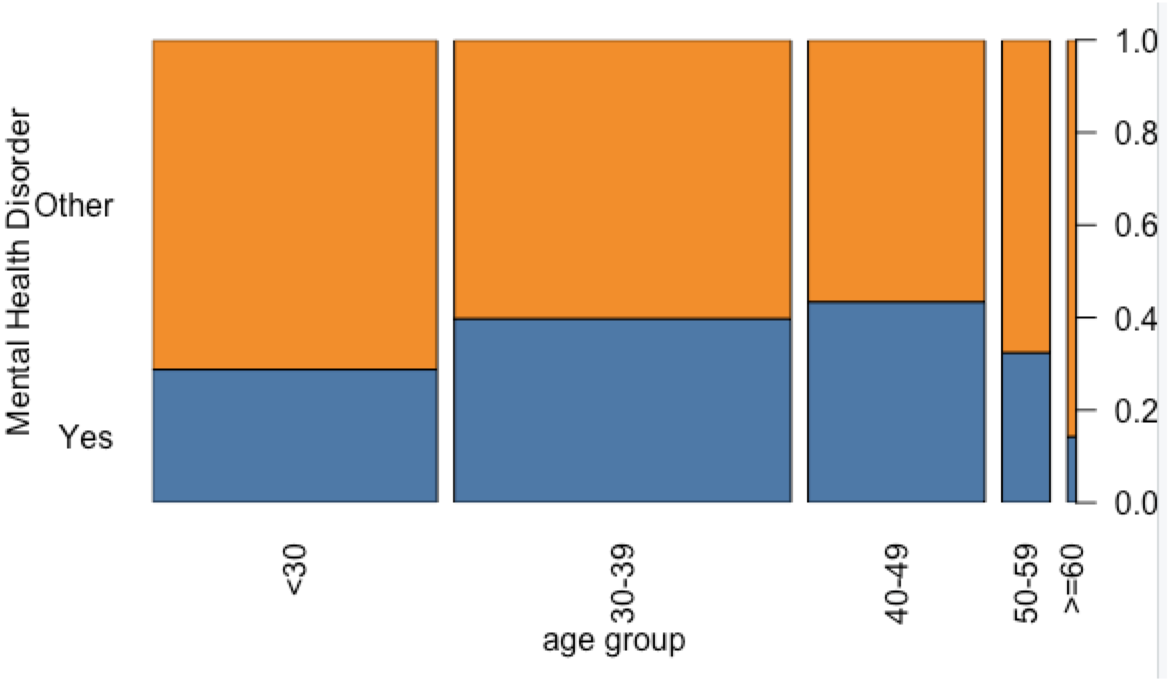


```
# CountryWorkIn ***
Chi_func(dat$CountryWorkIn, dat$CurrentlyMentalHealthDisorder)
# [1] 3.592461e-15
```

It is observed that the people who live in or work in the USA have a higher likelihood of developing mental health disorders. However, this could also be due to the small amount of dataset available from people in the other countries.

We ran the chi-squared test for both factors and found that we can reject the null hypothesis that there is no relationship between mental health disorder and CountryWorkIn and CountryLiveIn as the p value is <0.05.

7. Age vs. Presence of Mental Health Disorders



For age, we decided to split into 5 different age groups so that we can identify the age group that is most likely to have a mental health disorder.

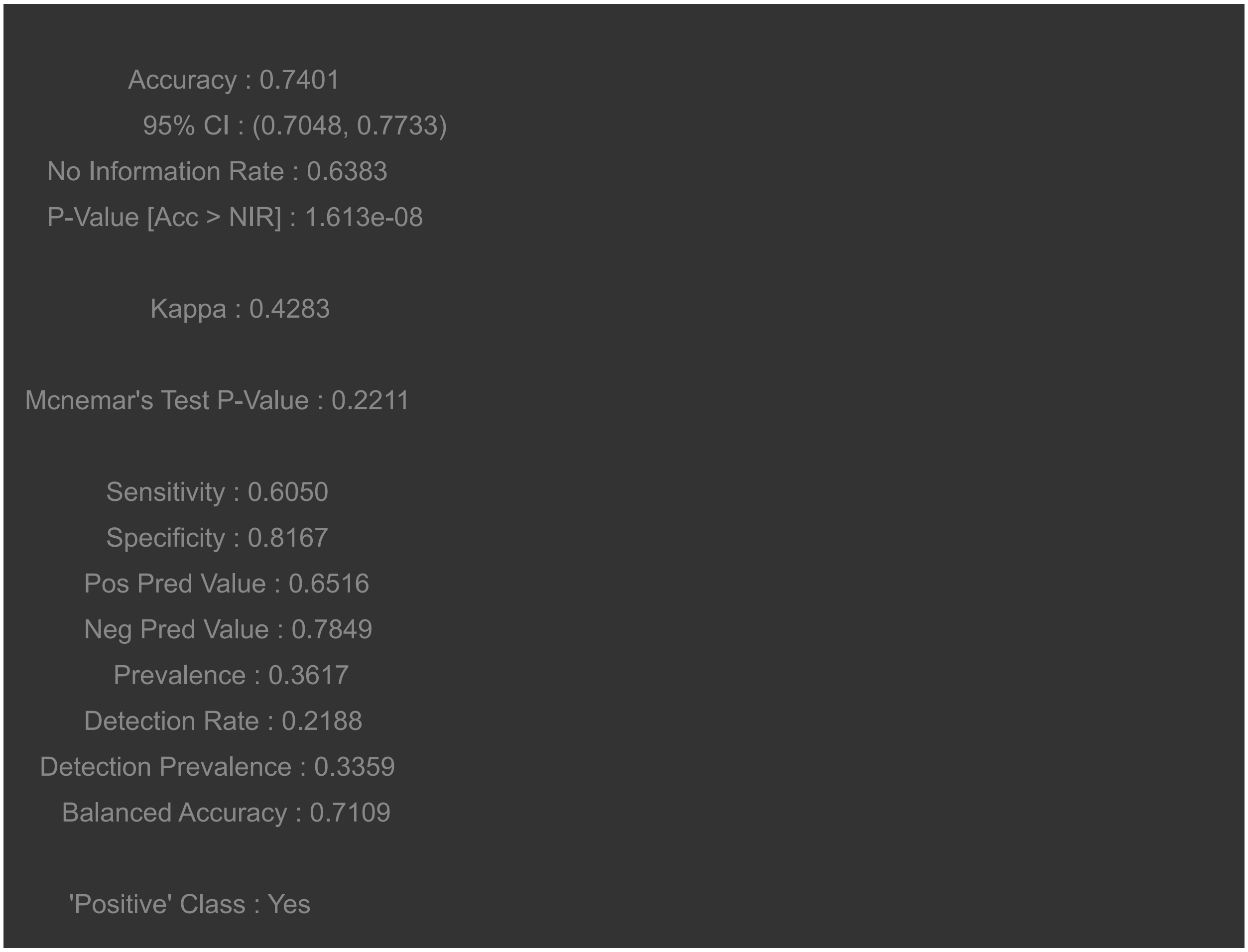
From this graph, we can see that the likelihood of one having a mental health disorder increases gradually from the age group < 30 and eventually peaks at 40-49. Beyond that, the likelihood of developing the mental health disorder decreases.

```
chisq.test(dat$AgeGroup, dat$CurrentlyMentalHealthDisorder, simulate.p.value = T
# Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)
# data:  dat$AgeGroup and dat$CurrentlyMentalHealthDisorder
# X-squared = 11.391, df = NA, p-value = 0.02149
```

We ran the chi-squared test and found that we can reject the null hypothesis that there is no relationship between age group and mental health disorder as the p-value is <0.05.

To find out which feature is the most important, we build a random forest model to predict ‘Do you currently have a mental health disorder’. The confusion matrix we have is listed below.

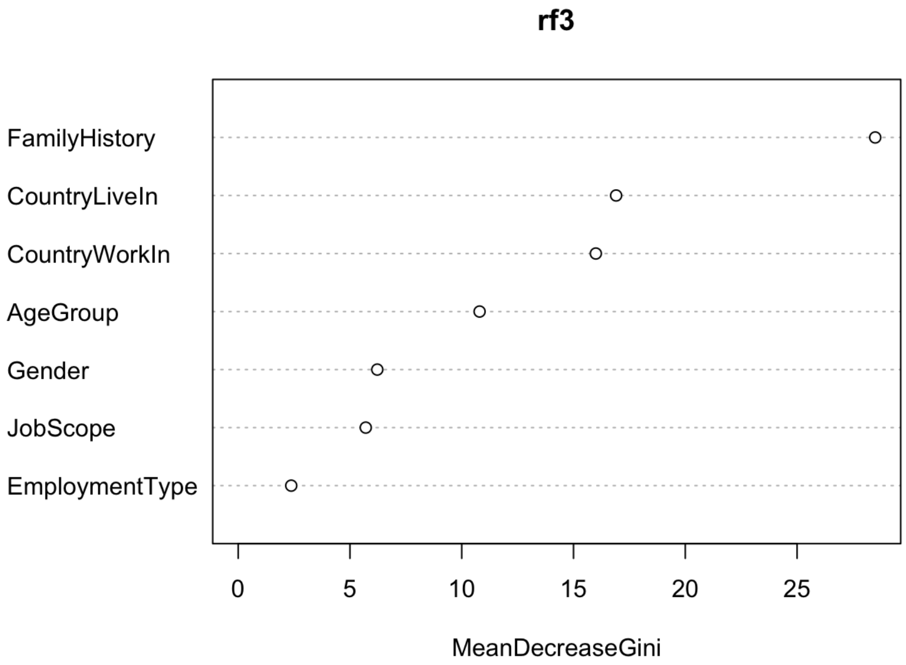
```
Confusion Matrix and Statistics


       Reference
Prediction Yes Other
    Yes   144    77
    Other  94   343
```



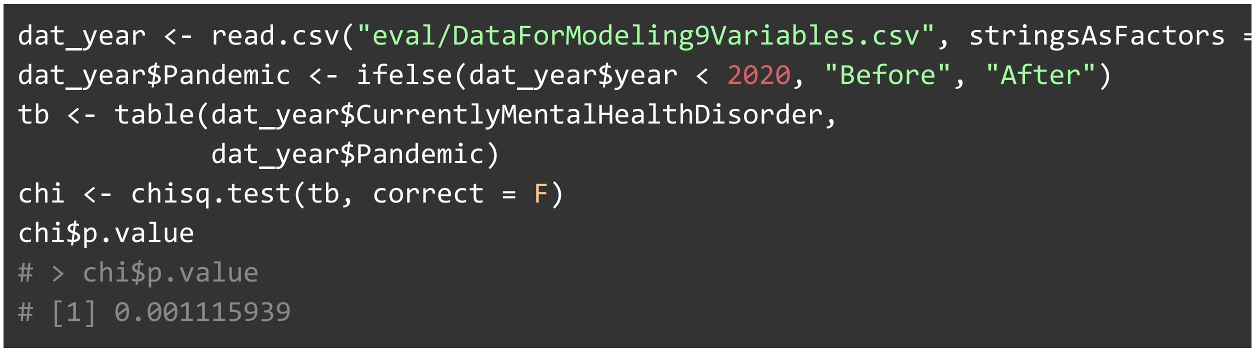
```
               Accuracy : 0.7401
                 95% CI : (0.7048, 0.7733)
    No Information Rate : 0.6383
    P-Value [Acc > NIR] : 1.613e-08

                  Kappa : 0.4283

 Mcnemar's Test P-Value : 0.2211

            Sensitivity : 0.6050
            Specificity : 0.8167
         Pos Pred Value : 0.6516
         Neg Pred Value : 0.7849
             Prevalence : 0.3617
         Detection Rate : 0.2188
   Detection Prevalence : 0.3359
      Balanced Accuracy : 0.7109

       'Positive' Class : Yes
```

The accuracy we managed to achieve is 0.74. Even though this may not be high, it was already a significant improvement compared to our original random forest, which has an accuracy rate of 0.47.

Using the Gini importance, the top 5 most important features in order of importance are as follows: family history, country that one lives in, country that one works in, age group and gender.



2. Does the pandemic have an effect on mental health disorders?



```
dat_year <- read.csv("eval/DataForModeling9Variables.csv", stringsAsFactors =
dat_year$Pandemic <- ifelse(dat_year$year < 2020, "Before", "After")
tb <- table(dat_year$CurrentlyMentalHealthDisorder,
            dat_year$Pandemic)
chi <- chisq.test(tb, correct = F)
chi$p.value
# > chi$p.value
# [1] 0.001115939
```

We decided to do the chi-squared test as both feature and outcome are categorical values. Since the p-value < 0.05, we can reject the null hypothesis. In other words, there is a relationship between the pandemic and the presence of mental health disorders.

3. Does the pandemic have an effect on the attitudes towards mental health disorders?

The features that are associated with this research question are as follows:

1. Would you feel comfortable discussing a mental health issue with your direct supervisor?
2. Would you feel comfortable discussing a mental health issue with your coworkers?

```
# 3-1 DiscussWithSupervisor

dat2019 <- read.csv("data/OSMI 2019 Mental Health in Tech Survey Results - OSMI Mental Hea
Tech Survey 2019.csv", stringsAsFactors = T) %>%
  select(DiscussWithSupervisor =
`Would.you.feel.comfortable.discussing.a.mental.health.issue.with.your.direct.supervisor.s..`) %
  mutate(year = 2019)
dat2020 <- read.csv("data/OSMI 2020 Mental Health in Tech Survey Results .csv", stringsAsFa
T) %>%
  select(DiscussWithSupervisor =
`Would.you.feel.comfortable.discussing.a.mental.health.issue.with.your.direct.supervisor.s..`) %
  mutate(year = 2020)
dat2021 <- read.csv("data/OSMI 2021 Mental Health in Tech Survey Results .csv", stringsAsFa
T) %>%
  select(DiscussWithSupervisor =
`Would.you.feel.comfortable.discussing.a.mental.health.issue.with.your.direct.supervisor.s..`) %
  mutate(year = 2021)
dat <- rbind(dat2019, dat2020) %>%
  rbind(dat2021)

dat$Pandemic <- ifelse(dat$year < 2020, "Before", "After")
tb <- table(dat$DiscussWithSupervisor,
            dat$Pandemic)
chi <- chisq.test(tb, correct = F)
chi$p.value
# [1] 0.04401081
```

```
# 3-2 DiscussWithCoworker

dat2019 <- read.csv("data/OSMI 2019 Mental Health in Tech Survey Results - OSMI Mental Hea
Tech Survey 2019.csv", stringsAsFactors = T) %>%
  select(DiscussWithCoworkers =
`Would.you.feel.comfortable.discussing.a.mental.health.issue.with.your.coworkers.`) %>%
  mutate(year = 2019)
dat2020 <- read.csv("data/OSMI 2020 Mental Health in Tech Survey Results .csv", stringsAsFa
T) %>%
  select(DiscussWithCoworkers =
`Would.you.feel.comfortable.discussing.a.mental.health.issue.with.your.coworkers.`) %>%
  mutate(year = 2020)
dat2021 <- read.csv("data/OSMI 2021 Mental Health in Tech Survey Results .csv", stringsAsFa
T) %>%
  select(DiscussWithCoworkers =
`Would.you.feel.comfortable.discussing.a.mental.health.issue.with.your.coworkers.`) %>%
  mutate(year = 2021)
dat <- rbind(dat2019, dat2020) %>%
  rbind(dat2021)

dat$Pandemic <- ifelse(dat$year < 2020, "Before", "After")
tb <- table(dat$DiscussWithCoworkers,
            dat$Pandemic)
chi <- chisq.test(tb, correct = F)
chi$p.value
# [1] 0.5038958
```

We use chi-squared tests for both questions as both feature and outcome are categorical data. We can reject the null hypothesis that the pandemic does not have a relationship with the feature “discussion of mental health disorders with direct supervisors” as the p-value < 0.05. In other words, the pandemic has an effect on the discussion of mental health disorders with direct supervisors

However, for co-workers, we found that we failed to reject the null hypothesis that the pandemic does not have a relationship with the feature “discussion of mental health disorders with coworkers” as the p-value > 0.05.

A possible reason for this phenomenon could be because people are more comfortable sharing their personal lives/ issues with their coworkers but not their direct supervisors. However, the pandemic might have increased the importance of mental health disorders and changed this perception.

Conclusion & Discussion

In conclusion, we can conclude that the factors that have a relationship with mental health disorders are family history, the country that people live or work in and age group. Employees who have a family history of mental health disorders, age between 40 and 49 and from the USA have a higher probability of having mental health disorders.

We also found that the pandemic has an effect on mental health disorders and the pandemic has an effect on the discussion of mental health disorders with direct supervisors but not with coworkers.

However, due to time constraints, we were not able to see if there is an increase or decrease in mental health disorders or find out if more or less people are willing to share their mental health disorders with their direct supervisors. Both of which are good research questions for future research.

We believe that the pandemic has caused an increase in the mental health disorders and changed the perception of mental health disorders. People now understand the severity of mental health disorders and the importance of having mental wellness. Therefore, they are now more willing to share their condition with their direct supervisors at work. Both of these information is supported in this article by WHO.

Another possible research idea would be trying to find other models that have a better accuracy rate. One other possible model would be a logistic regression model.