Prevention of mental disorders with Data Science

Alumni of the HarvardX Data Science Professional Certificate Program

2023-09-25

1. Introduction

In the last century, prevention has become an increasingly important part of modern healthcare. For example, the decline in infectious diseases is due to preventive strategies such as immunization and hygiene than to Alexander Fleming's discovery of penicillin. Lifestyle changes and prophylactic medication have reduced coronary heart disease in recent decades. Now is the time to pay more attention to prevention of mental disorders.

Many studies have been conducted showing a relationship between psychosocial work characteristics and mental health outcomes, ranging from symptoms and psychological distress to diagnosed psychiatric disorders. However, many of these studies have been cross-sectional and thus make the causal direction between job stressors and mental health uncertain, especially in light of the demonstrated reciprocal relationship between job characteristics and mental health. Moreover, the results have not been consistent across studies, especially in the case of the job stress model. These inconsistencies could be resolved, and the causal direction clarified, by a data science-based analysis that provides causal associations between various worker characteristic factors and common mental disorders.

The present work provides a novel solution to the problem described, providing data science techniques to help identify high-risk individuals and provide interventions to prevent and treat mental illness. Although published research on data science applied to neuropsychiatry is quite limited, there are increasingly successful examples of its use in other healthcare fields such as oncology, radiology and dermatology.

1.1 Dataset and variables

This project makes use of the Open Sourcing Mental Illess (OSMI) Mental Health in Tech Survey 2016. OSMI has an ongoing survey from 2016, which "aims to measure attitudes towards mental health in the tech workplace, and examine the frequency of mental health disorders among tech workers." The survey is conducted online at the OSMI website and the OSMI team intends to use these data to help drive awareness and improve conditions for individuals with mental illness in the IT workplace.

1.2 Objective

This project is an assignment of the "Project Overview: Choose Your Own!" of "Data Science: Capstone" (HarvardX PH125.9x) course.

The main objective of this work is to demonstrate that data science can be a valuable tool in Occupational Medicine to predict employees at risk of suffering mental illness caused by their working conditions. This will be achieved by means of the following specific objectives:

- Key factors

- to identify patterns and fundamental factors that lead to mental illness in the work environment

- Prediction model
  - to design a model based on data science capable of making predictions about employees at risk of suffering mental illnesses caused by their working conditions

1.3 Key steps

The project consists of the following steps:

- Cleaning and preprocessing of input data
  - The database used has been generated with manual entries, so errors and inconsistencies may exist and should be checked and cleaned
  - Data gaps search and management
- Relevant variables
  - Search and selection of relevant characteristics to be used as explanatory variables in the prediction by our model
- Model generation and evaluation
  - A model will be generated with training dataset by means of a detailed search of algorithm hyperparameters
  - The performance of the final model will be tested in the testing dataset, which has not been used for the generation of the prediction model.

2. Methods and Analysis

In this Section, the process and techniques used are explained, including data source, cleaning, exploration and visualization, aolng with the insights gained. Finally, 2 alternative modeling approaches are presented.

2.1 Data download and cleaning

In this subsection we download the original database, and split it into the following subsets:

- training_dataset
  - used to develop our algorithm
- testing_dataset
  - used for a final test of our final algorithm

Install and load packages

```
if(!require(dplyr)) install.packages("dplyr", repos = "http://cran.us.r-project.org")
if(!require(ggplot2)) install.packages("ggplot2", repos = "http://cran.us.r-project.org")
if(!require(corrplot)) install.packages("corrplot", repos = "http://cran.us.r-project.org")
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(ranger)) install.packages("ranger", repos = "http://cran.us.r-project.org")

library(dplyr); library(ggplot2); library(corrplot); library(tidyverse); library(caret);library(ranger)
```

The following options sets as 120 s the amount of time to wait for a response from the remote name server before retrying the query via a different one.

```
options(timeout = 120)
```

Download and unzip the original dataset, *mental-heath-in-tech-2016_20161114.csv*

```
# The original file from Kaggle can be found here
# https://www.kaggle.com/datasets/osmi/mental-health-in-tech-2016/download?datasetVersionNumber=2

dl <- "mental-heath-in-tech-2016_20161114.csv"
if(!file.exists(dl))
  download.file(
    "https://github.com/cpperuch/mental_tech_survey/blob/c2bc59a8580025e1e4838f6e587e693e644c5ed2/mental

mental_tech_survey <- read_csv(dl)
```

The dataset is composed by 63 different columns, with 1433 rows. Columns correspond to the different questions answered by workers. The detailed description of the variables can be found here: https://osmi.typeform.com/report/Ao6BTw/U76z.

```
print(paste("Number of rows and columns: ", dim(mental_tech_survey), sep=""))
```

```
## [1] "Number of rows and columns: 1433" "Number of rows and columns: 63"
```

```
colnames(mental_tech_survey)
```

```
##  [1] "Are you self-employed?"
##  [2] "How many employees does your company or organization have?"
##  [3] "Is your employer primarily a tech company/organization?"
##  [4] "Is your primary role within your company related to tech/IT?"
##  [5] "Does your employer provide mental health benefits as part of healthcare coverage?"
##  [6] "Do you know the options for mental health care available under your employer-provided coverage
##  [7] "Has your employer ever formally discussed mental health (for example, as part of a wellness ca
##  [8] "Does your employer offer resources to learn more about mental health concerns and options for 
##  [9] "Is your anonymity protected if you choose to take advantage of mental health or substance abus
## [10] "If a mental health issue prompted you to request a medical leave from work, asking for that le
## [11] "Do you think that discussing a mental health disorder with your employer would have negative c
## [12] "Do you think that discussing a physical health issue with your employer would have negative co
## [13] "Would you feel comfortable discussing a mental health disorder with your coworkers?"
## [14] "Would you feel comfortable discussing a mental health disorder with your direct supervisor(s)?
## [15] "Do you feel that your employer takes mental health as seriously as physical health?"
## [16] "Have you heard of or observed negative consequences for co-workers who have been open about me
## [17] "Do you have medical coverage (private insurance or state-provided) which includes treatment of
## [18] "Do you know local or online resources to seek help for a mental health disorder?"
## [19] "If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to
## [20] "If you have revealed a mental health issue to a client or business contact, do you believe thi
## [21] "If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to
## [22] "If you have revealed a mental health issue to a coworker or employee, do you believe this has 
## [23] "Do you believe your productivity is ever affected by a mental health issue?"
## [24] "If yes, what percentage of your work time (time performing primary or secondary job functions)
## [25] "Do you have previous employers?"
```

```
## [26] "Have your previous employers provided mental health benefits?"
## [27] "Were you aware of the options for mental health care provided by your previous employers?"
## [28] "Did your previous employers ever formally discuss mental health (as part of a wellness campaig
## [29] "Did your previous employers provide resources to learn more about mental health issues and how
## [30] "Was your anonymity protected if you chose to take advantage of mental health or substance abus
## [31] "Do you think that discussing a mental health disorder with previous employers would have negat
## [32] "Do you think that discussing a physical health issue with previous employers would have negati
## [33] "Would you have been willing to discuss a mental health issue with your previous co-workers?"
## [34] "Would you have been willing to discuss a mental health issue with your direct supervisor(s)?"
## [35] "Did you feel that your previous employers took mental health as seriously as physical health?"
## [36] "Did you hear of or observe negative consequences for co-workers with mental health issues in y
## [37] "Would you be willing to bring up a physical health issue with a potential employer in an inter
## [38] "Why or why not?...38"
## [39] "Would you bring up a mental health issue with a potential employer in an interview?"
## [40] "Why or why not?...40"
## [41] "Do you feel that being identified as a person with a mental health issue would hurt your caree
## [42] "Do you think that team members/co-workers would view you more negatively if they knew you suff
## [43] "How willing would you be to share with friends and family that you have a mental illness?"
## [44] "Have you observed or experienced an unsupportive or badly handled response to a mental health
## [45] "Have your observations of how another individual who discussed a mental health disorder made y
## [46] "Do you have a family history of mental illness?"
## [47] "Have you had a mental health disorder in the past?"
## [48] "Do you currently have a mental health disorder?"
## [49] "If yes, what condition(s) have you been diagnosed with?"
## [50] "If maybe, what condition(s) do you believe you have?"
## [51] "Have you been diagnosed with a mental health condition by a medical professional?"
## [52] "If so, what condition(s) were you diagnosed with?"
## [53] "Have you ever sought treatment for a mental health issue from a mental health professional?"
## [54] "If you have a mental health issue, do you feel that it interferes with your work when being tr
## [55] "If you have a mental health issue, do you feel that it interferes with your work when NOT bein
## [56] "What is your age?"
## [57] "What is your gender?"
## [58] "What country do you live in?"
## [59] "What US state or territory do you live in?"
## [60] "What country do you work in?"
## [61] "What US state or territory do you work in?"
## [62] "Which of the following best describes your work position?"
## [63] "Do you work remotely?"
```

In order to simplify the column names, we will rename them: instead of the whole question, we will reduce it to a name using as separators the symbols "_" and "." as follows:

```
# create new variable names
new.names <- c("self.employed", "num.employees", "tech.company", "tech.role", "mental_health.coverage",

colnames(mental_tech_survey) <- new.names
```

2.1.1 Cleaning up the gender variable

In this Subsection, we explore the responses to the questions about the gender of the respondents. Let's first explore how many different genres there are in our dataset.

```r
print(paste("Total of ", length(unique(mental_tech_survey$gender)), " different genres", sep = ""))
```

```
## [1] "Total of 67 different genres"
```

```r
head(table(mental_tech_survey$gender))
```

```
##
##        AFAB     Agender Androgynous    Bigender   Cis-woman  Cis female
##           1           2           1           1           1           1
```

We see a wide variety of responses, with a total of 67 genres. As these are handwritten responses, multiple responses representing the same gender are to be expected and will predictably need to be corrected.

For example, This is confirmed by exploring how many different responses include the word *female*.

```r
unique(mental_tech_survey$gender[str_detect(mental_tech_survey$gender, "female")])
```

```
## [1] "female"
## [2] "I identify as female."
## [3] "Cis female"
## [4] "Genderfluid (born female)"
## [5] "female/woman"
## [6] "male 9:1 female, roughly"
## [7] "female-bodied; no feelings about gender"
## [8] NA
```

Checking in detail the different answers, we grouped them as follows.

```r
# Females
female_cases <- c("Female", "Female ", " Female", "female", "female ", "Woman", "woman", "f", "Cis femal

# Males
male_cases <- c("Male", "male", "MALE", "Man", "man", "m", "man ", "Dude", "mail", "M|", "Cis male", "Ma
                "Sex is male","I'm a man why didn't you make this a drop down question. You should of as

# gender queers (GQ)
gender_queers_cases <- c(
"Agender", "Androgynous", "Bigender", "Female or Multi-Gender Femme", "female-bodied; no feelings about
"male 9:1 female, roughly", "N/A", "Other", "Genderfluid", "N/A", "Enby", "Queer" )

# transgender (TG)
transgender_cases <- c("Male (trans, FtM)", "Transgender woman", "Transitioned, M2F", "mtf")
```

We merge these genres in the following 4: *F*, *M*, *GQ* and *TG*, and explore the other answers

```r
mental_tech_survey$gender[which(mental_tech_survey$gender %in% male_cases)] = "M"
mental_tech_survey$gender[which(mental_tech_survey$gender %in% female_cases)] = "F"
mental_tech_survey$gender[which(mental_tech_survey$gender %in% gender_queers_cases)] = "GQ"
mental_tech_survey$gender[which(mental_tech_survey$gender %in% transgender_cases)] = "TG"

table(mental_tech_survey$gender)
```

```
##
##                    F                   GQ                    M
##                  337                   33                 1057
## none of your business                   TG
##                    1                    4
```

```r
paste("NA gender values: ", print(length(which(is.na(mental_tech_survey$gender)))), sep = "")
```

```
## [1] 1

## [1] "NA gender values: 1"
```

Finally, only the selected categories are used

```r
mental_tech_survey <- mental_tech_survey %>% filter(gender %in% c("M", "F", "GQ", "TG"))
```

2.1.2 Cleaning up the age variable

Exploring age variable, we found suspicious values: 3, 99 and 323

```r
table(mental_tech_survey$age)
```

```
##
##   3  15  17  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
##   1   1   1   4   6  15  32  23  42  44  63  63  74  79  94  82  72  69  69  74
##  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55
##  50  59  54  55  36  24  29  30  31  27  22  14   9  13   9   7   7   3   7  12
##  56  57  58  59  61  62  63  65  66  70  74  99 323
##   5   4   1   2   2   1   4   1   1   1   1   1   1
```

These suspicious values which we are discarded for our analysis

```r
mental_tech_survey <- mental_tech_survey %>% filter(age %in% 15:74)
```

We observe a large number of gaps in many of the variables in our dataset.

```r
names(which(colSums(is.na(mental_tech_survey)) > 1000))
```

```
##  [1] "tech.role"
##  [2] "private.med.coverage"
##  [3] "resources"
##  [4] "reveal.diagnosis.clients.or.business"
##  [5] "revealed.negative.consequences.CB"
##  [6] "reveal.diagnosis.coworkers"
##  [7] "revealed.negative.consequences.CW"
##  [8] "productivity.effected"
##  [9] "percentage"
## [10] "if.maybe.what"
```

We will select some columns of interest, eliminating those with the highest number of gaps, and we will eliminate the gaps in the selected ones.

```
mental_tech_survey = data.frame(mental_tech_survey[,c(1:3,5:12,15:16, 39, 41:44, 46:47,48,51, 53:54,56:
mental_tech_survey <- mental_tech_survey %>% na.omit()
```

Finally, we will classify the ages in four different categories: under 25 years old, between 25 and 40 years old, between 40 and 55 years old, and finally over 55 years old. We also select the cases of "Yes" and "No" in the current disorder question, and factored the rest of the variables

```
mental_tech_survey <- mental_tech_survey %>%
  mutate(age = ifelse(age < 25, "<25", ifelse(age <= 40, "25-40",ifelse(age < 55, "40-55", ">55")))) %>%
  mutate(age = factor(age))

mental_tech_survey = mental_tech_survey[which(mental_tech_survey$currently.have.mental.disorder %in% c(

mental_tech_survey <- mental_tech_survey %>% mutate_if(is.character, as.factor)
```

2.2 Data exploration and visualization

In this Section, we will explore the dataset and the relations between its parameters.

First, we split our original dataset into training (*training_dataset*, 90% of data) and testing (*testing_dataset*, 10% of data) datasets for model generation and testing, respectively.

```
set.seed(1)
test_index <- createDataPartition(y = mental_tech_survey$currently.have.mental.disorder,
                                  times = 1, p = 0.1, list = FALSE)

training_dataset <- mental_tech_survey[-test_index,]
testing_dataset <- mental_tech_survey[test_index,]
```

2.2.1 Variable importance

We will first determine the relative importance of the different variables based on *Ranger*, a fast implementation of random forests or recursive partitioning, particularly suited for high dimensional data. We can see that there are 4 variables of great importance (> 32), with the rest having an importance < 13.

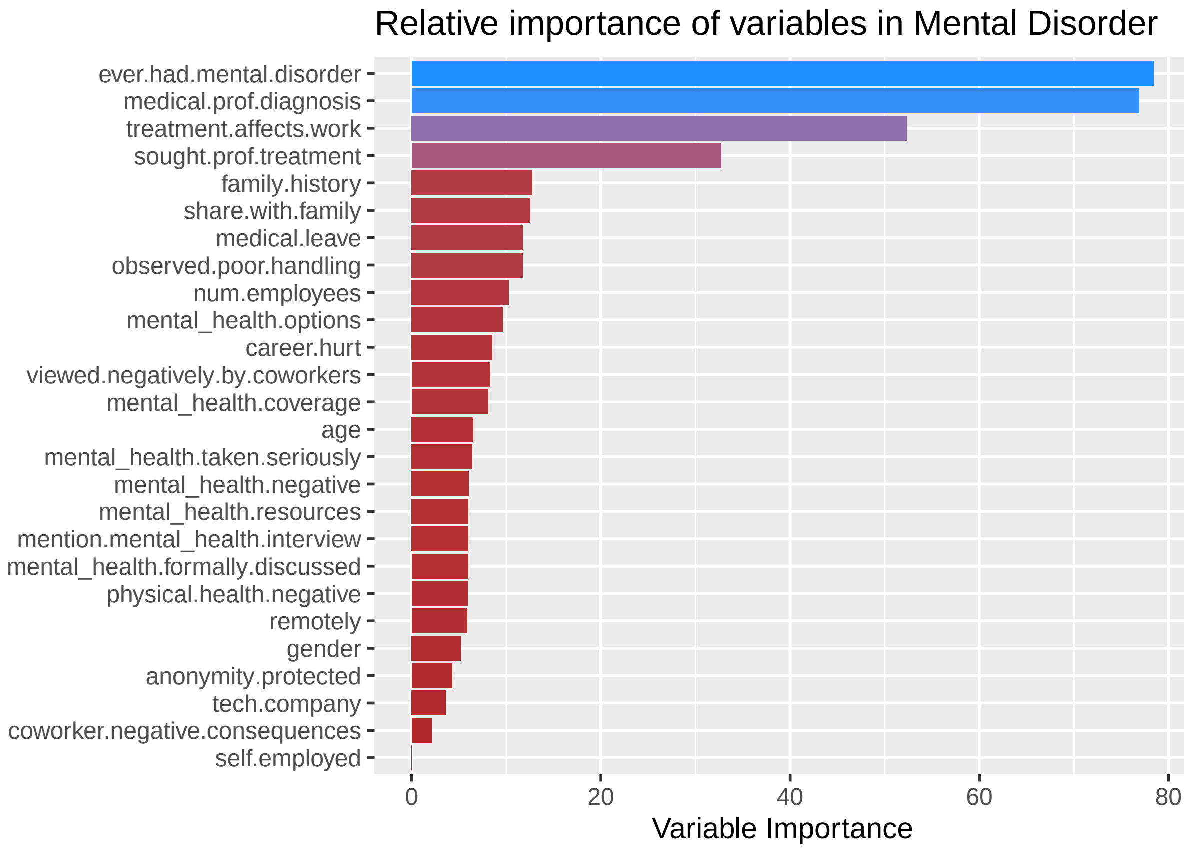
```
training_model <- ranger(currently.have.mental.disorder ~ .,
                   data = training_dataset,
                   importance = "impurity")

sort(training_model$variable.importance, decreasing=T)
```

```
##        ever.had.mental.disorder          medical.prof.diagnosis
##                       78.414913                       76.880040
##           treatment.affects.work            sought.prof.treatment
##                       52.318417                       32.696525
##                  family.history                share.with.family
##                       12.763831                       12.555905
##                   medical.leave          observed.poor.handling
##                       11.745890                       11.736536
##                   num.employees           mental_health.options
##                       10.280322                        9.634797
```

```
##                        career.hurt   viewed.negatively.by.coworkers
##                           8.526042                         8.279488
##            mental_health.coverage                              age
##                           8.071049                         6.530423
##     mental_health.taken.seriously           mental_health.negative
##                           6.401131                         6.054968
##           mental_health.resources  mention.mental_health.interview
##                           5.976474                         5.971280
## mental_health.formally.discussed        physical.health.negative
##                           5.968212                         5.899660
##                           remotely                           gender
##                           5.860141                         5.203602
##               anonymity.protected                     tech.company
##                           4.271030                         3.610031
##    coworker.negative.consequences                    self.employed
##                           2.100271                         0.000000
```

Let us graphically represent the variables ordered according to their relative importance.

```
ggplot(
  enframe(
    training_model$variable.importance,
    name = "variable",
    value = "importance"),
  aes(
    x = reorder(variable, importance),
    y = importance,
    fill = importance))+
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  ylab("Variable Importance") +
  xlab("") +
  ggtitle("Relative importance of variables in Mental Disorder") +
  guides(fill = "none") +
  scale_fill_gradient(low = "firebrick", high = "dodgerblue")
```



These are the variables of greatest weight in the Mental Disorder, as well as the questions to which they correspond in the survey:

- **ever.had.mental.disorder**: "Have you had a mental health disorder in the past?"
- **medical.prof.diagnosis**: "Have you been diagnosed with a mental health condition by a medical professional?"
- **treatment.affects.work**: "If you have a mental health issue, do you feel that it interferes with your work when being treated effectively?"
- **sought.prof.treatment**: "Have you ever sought treatment for a mental health issue from a mental health professional?"

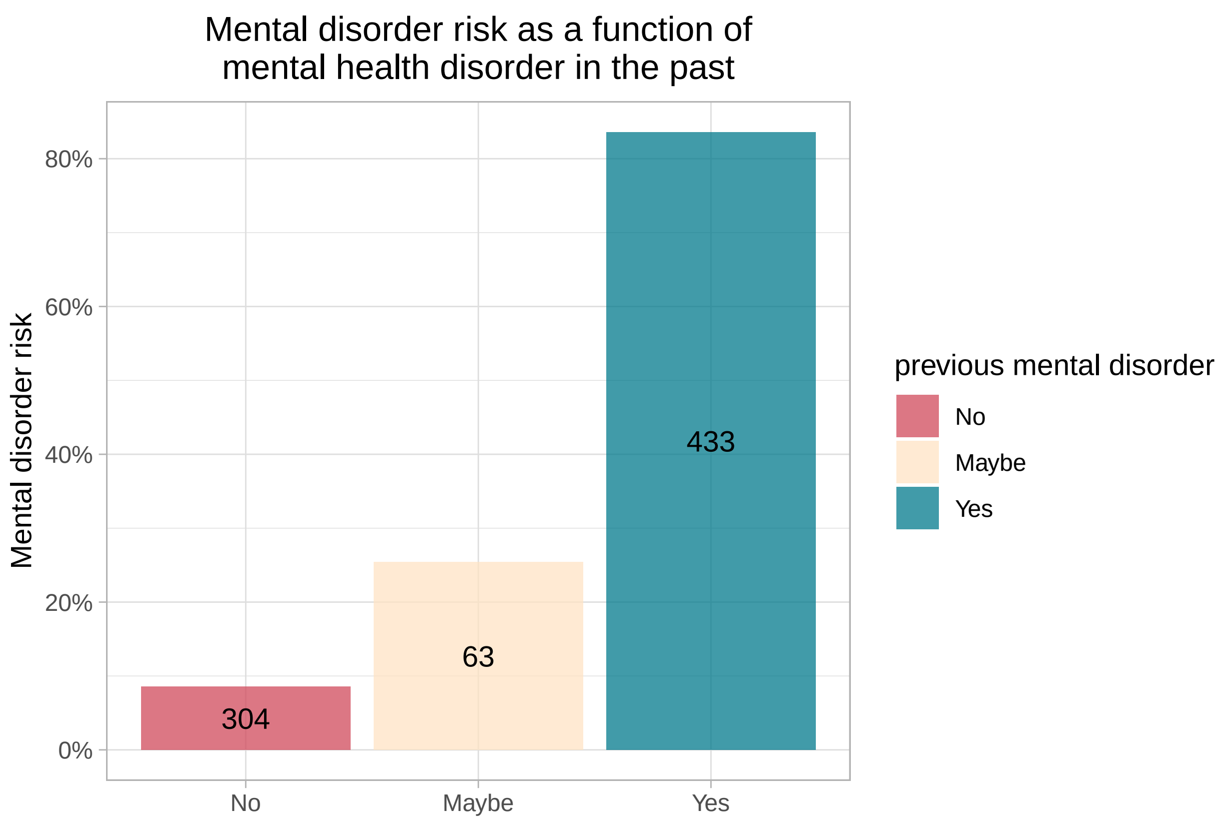
Let's explore in detail the influence of these 4 variables on Mental Disorder

2.2.1.a Ever had previous disorder

The following graph shows the percentage of cases with current mental disorder as a function of the previous mental disorders, the label being the total number of cases in each category.

We observed that having had a previous mental disorder significantly increases the risk of having a current mental disorder (83.6%). This risk is reduced by an order of magnitude with the absence of a previous mental disorder (8.6%).

```
training_dataset %>%
  mutate(Ment_dis_binary = ifelse(currently.have.mental.disorder == "Yes", 1, 0)) %>%
  mutate(ever.had.mental.disorder = factor(ever.had.mental.disorder, levels = c("No", "Maybe", "Yes")))
  group_by(ever.had.mental.disorder) %>%
  summarise(Mental_disorder_ratio = mean(Ment_dis_binary), count = n()) %>%

  ggplot(aes(fill=ever.had.mental.disorder,
             label = count,
             y=Mental_disorder_ratio,
             x= ever.had.mental.disorder )) +
  geom_bar(position="stack", stat="identity", alpha = 0.75) +
  #scale_fill_manual(values = c("#00798c","#d1495b")) +
  geom_text(position = position_stack(vjust = 0.5)) +
  theme_light() +
  theme(#legend.title = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  guides(fill=guide_legend(title="previous mental disorder"))+
  xlab("") +
  ylab("Mental disorder risk") +
  ggtitle(paste("Mental disorder risk as a function of",
                "\nmental health disorder in the past"  , sep = "")) +
  scale_fill_manual(values = c("#d1495b", "bisque" , "#00798c")) + # "#8d96a3"
  scale_y_continuous(labels = scales::percent) #+ facet_wrap(~fulltime_parttime_description)
```



The following table details the numerical values of the above graph.

```r
averaged_ever.had.mental.disorder <-
training_dataset %>%
  mutate(Ment_dis_binary = ifelse(currently.have.mental.disorder == "Yes", 1, 0)) %>%
  mutate(ever.had.mental.disorder = factor(ever.had.mental.disorder, levels = c("No", "Maybe", "Yes")))
  mutate(medical.prof.diagnosis = ifelse(medical.prof.diagnosis == "Yes",
                                         "Prev. mental health cond. diagnosis",
                                         "No prev. mental health cond. diagnosis")) %>%
  group_by(ever.had.mental.disorder) %>%
  summarise(Mental_disorder_cases = sum(Ment_dis_binary), total = n(),
            Mental_disorder_ratio = round(100*mean(Ment_dis_binary), 1)) %>%
  arrange(desc(Mental_disorder_ratio))

averaged_ever.had.mental.disorder
```

```
## # A tibble: 3 x 4
##   ever.had.mental.disorder Mental_disorder_cases total Mental_disorder_ratio
##   <fct>                                    <dbl> <int>                 <dbl>
## 1 Yes                                        362   433                  83.6
## 2 Maybe                                       16    63                  25.4
## 3 No                                          26   304                   8.6
```

We then explore the statistical significance of the differences found. As a result, the three categories considered show statistically significant differences (p-value < 0.05).

```r
Yes_No_test <- prop.test(x = averaged_ever.had.mental.disorder$Mental_disorder_cases[c(1,3)],
                         n = averaged_ever.had.mental.disorder$total[c(1,3)],
                         alternative = "greater")

Yes_Maybe_test <- prop.test(x = averaged_ever.had.mental.disorder$Mental_disorder_cases[c(1,2)],
                         n = averaged_ever.had.mental.disorder$total[c(1,2)],
                         alternative = "greater")

paste("Mental health disorder in the past increases in a statistically significant way the risk on curr
```

```
## [1] "Mental health disorder in the past increases in a statistically significant way the risk on cur
```

```r
paste("No vs Yes: ", averaged_ever.had.mental.disorder$Mental_disorder_ratio[3],
      "% vs ", averaged_ever.had.mental.disorder$Mental_disorder_ratio[1],
      "% (p value = ",  signif(Yes_No_test$p.value,1), ")", sep = "")
```

```
## [1] "No vs Yes: 8.6% vs 83.6% (p value = 2e-89)"
```

```r
paste("Maybe vs Yes: ",
      averaged_ever.had.mental.disorder$Mental_disorder_ratio[2],
      "% vs ", averaged_ever.had.mental.disorder$Mental_disorder_ratio[1],
      "% (p value = ",  signif(Yes_Maybe_test$p.value,1), ")", sep = "")
```

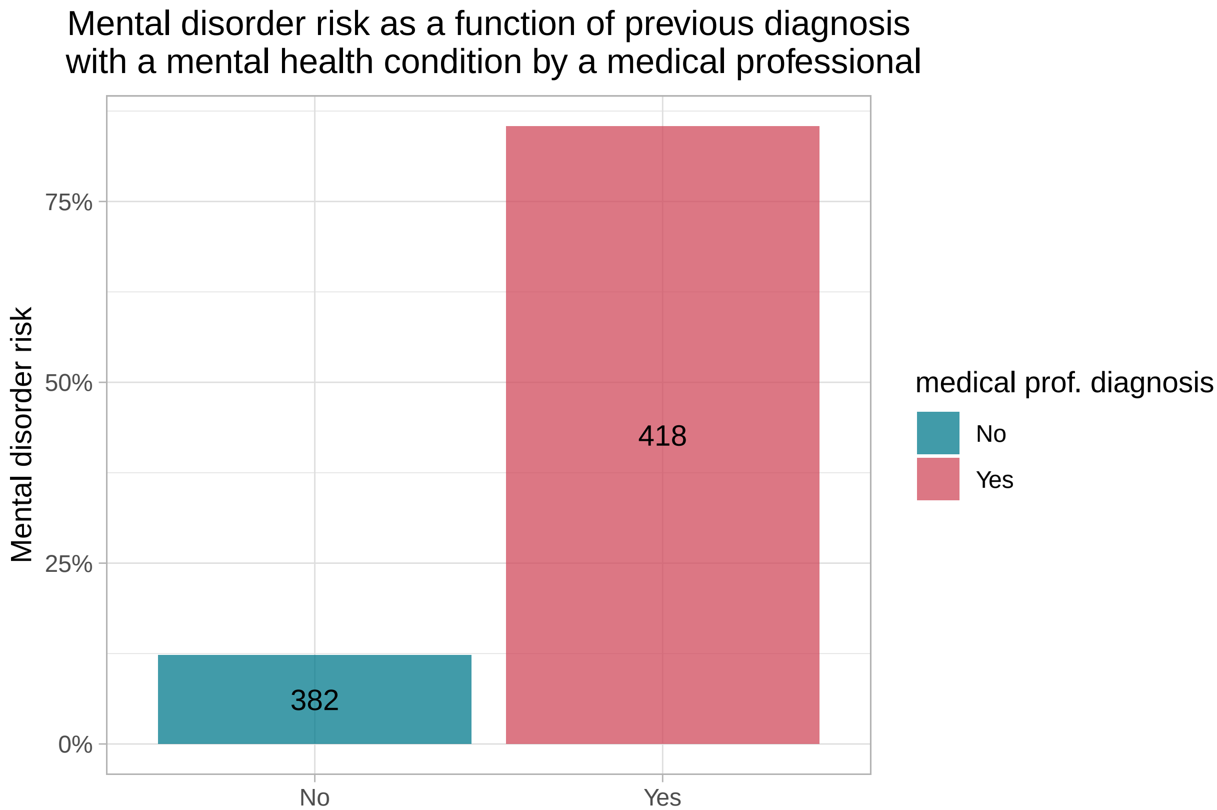
```
## [1] "Maybe vs Yes: 25.4% vs 83.6% (p value = 9e-24)"
```

2.2.1.b Medical proof diagnosis

This variable represents previous diagnosis of a mental health condition by a medical professional. The following graph shows the percentage of cases with current mental disorder as a function of this variable, the label being the total number of cases in each category.

We observed that having had a previous diagnosis of a mental health condition by a medical professional increases the risk of having a current mental disorder (85.4%). This risk is reduced by 7 with the absence of a previous diagnosis (12.3%).

```r
training_dataset %>%
  mutate(Ment_dis_binary = ifelse(currently.have.mental.disorder == "Yes", 1, 0)) %>%
  group_by(medical.prof.diagnosis) %>%
  summarise(Mental_disorder_ratio = mean(Ment_dis_binary), count = n()) %>%

  ggplot(aes(fill=medical.prof.diagnosis,
             label = count,
             y=Mental_disorder_ratio,
             x= medical.prof.diagnosis )) +
  geom_bar(position="stack", stat="identity", alpha = 0.75) +
  #scale_fill_manual(values = c("#00798c","#d1495b")) +
  geom_text(position = position_stack(vjust = 0.5)) +
  theme_light() +
  theme(#legend.title = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  guides(fill=guide_legend(title="medical prof. diagnosis"))+
  xlab("") +
  ylab("Mental disorder risk") +
  ggtitle(paste("Mental disorder risk as a function of previous diagnosis",
              "\n with a mental health condition by a medical professional"  , sep = "")) +
  scale_fill_manual(values = c("#00798c", "#d1495b")) +
  scale_y_continuous(labels = scales::percent) #+ facet_wrap(~fulltime_parttime_description)
```



The following table details the numerical values of the above graph.

```
averaged_medical.prof.diagnosis <-
training_dataset %>%
  mutate(Ment_dis_binary = ifelse(currently.have.mental.disorder == "Yes", 1, 0)) %>%
  mutate(ever.had.mental.disorder = factor(ever.had.mental.disorder, levels = c("No", "Maybe", "Yes")))
  mutate(medical.prof.diagnosis = ifelse(medical.prof.diagnosis == "Yes",
                                         "Prev. mental health cond. diagnosis",
                                         "No prev. mental health cond. diagnosis")) %>%
  group_by(medical.prof.diagnosis) %>%
  summarise(Mental_disorder_cases = sum(Ment_dis_binary), total = n(),
            Mental_disorder_ratio = round(100*mean(Ment_dis_binary), 1)) %>%
  arrange(desc(Mental_disorder_ratio))

averaged_medical.prof.diagnosis
```

```
## # A tibble: 2 x 4
##   medical.prof.diagnosis       Mental_disorder_cases total Mental_disorder_ratio
##   <chr>                                        <dbl> <int>                 <dbl>
## 1 Prev. mental health cond. d~                   357   418                  85.4
## 2 No prev. mental health cond~                    47   382                  12.3
```

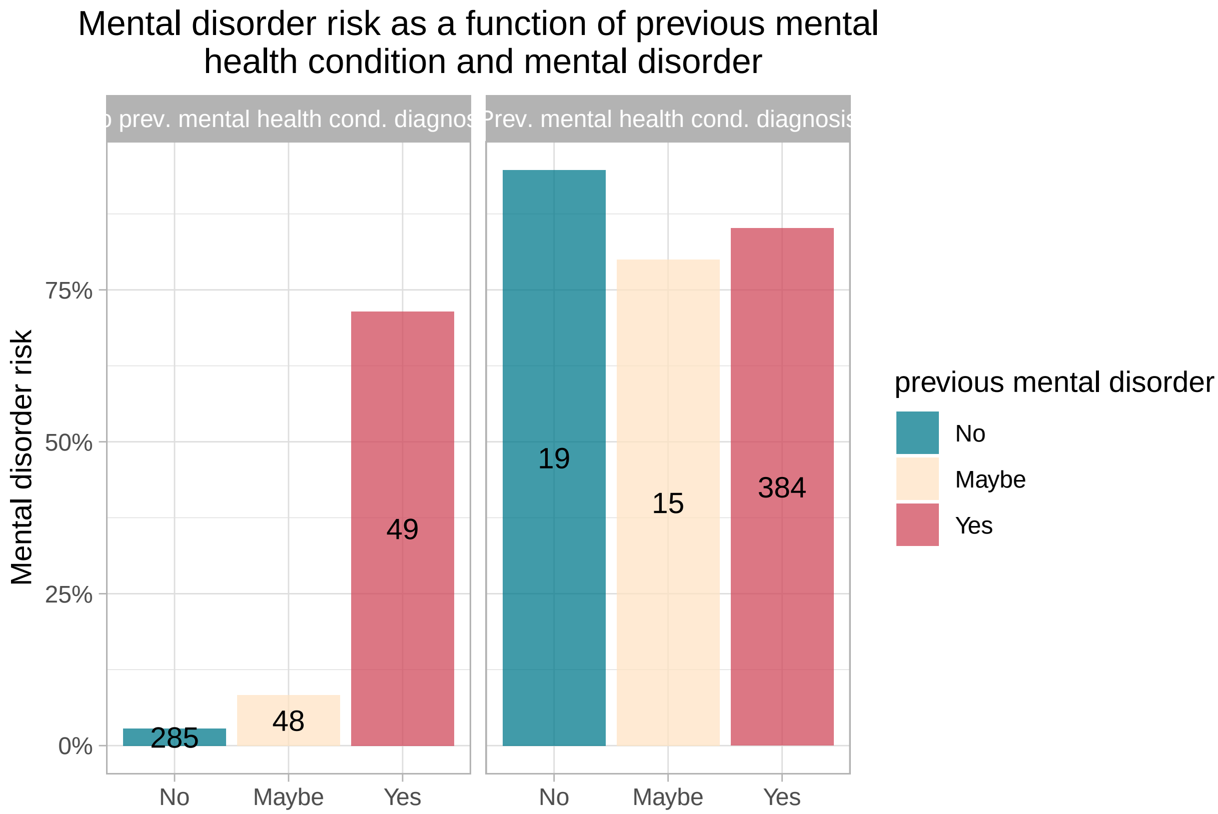
We then explore the statistical significance of the differences found. As a result, the two categories considered show statistically significant differences (p-value < 0.05).

```r
med.proof_test <- prop.test(x = averaged_medical.prof.diagnosis$Mental_disorder_cases,
                            n = averaged_medical.prof.diagnosis$total,
                            alternative = "greater")

paste("Previous diagnosis with a mental health condition by a medical professional increases in a stati
      "% vs ", averaged_medical.prof.diagnosis$Mental_disorder_ratio[1], "% (p value = ",
      signif(med.proof_test$p.value,1), ")", sep = "")
```

```
## [1] "Previous diagnosis with a mental health condition by a medical professional increases in a stati
```

The next graph shows the combined effect of these two variables. The highest-risk cases (>= 80%) are those in which there are previous diagnosis with a mental health condition by a medical professional. At the other extreme, the lowest risk case (2.8%) is the one in which there is no such diagnosis and there is no previous mental disorder.

```r
training_dataset %>%
  mutate(Ment_dis_binary = ifelse(currently.have.mental.disorder == "Yes", 1, 0)) %>%
  mutate(ever.had.mental.disorder = factor(ever.had.mental.disorder, levels = c("No", "Maybe", "Yes")))
  mutate(medical.prof.diagnosis = ifelse(medical.prof.diagnosis == "Yes",
                                         "Prev. mental health cond. diagnosis",
                                         "No prev. mental health cond. diagnosis")) %>%
  group_by(ever.had.mental.disorder, medical.prof.diagnosis) %>%
  summarise(Mental_disorder_ratio = mean(Ment_dis_binary), count = n()) %>%
  ggplot(aes(fill=ever.had.mental.disorder,
             label = count,
             y=Mental_disorder_ratio,
             x= ever.had.mental.disorder )) +
  geom_bar(position="stack", stat="identity", alpha = 0.75) +
  facet_wrap(~medical.prof.diagnosis) +
  #scale_fill_manual(values = c("#00798c","#d1495b")) +
  geom_text(position = position_stack(vjust = 0.5)) +
  theme_light() +
  theme(#legend.title = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  guides(fill=guide_legend(title="previous mental disorder"))+
  xlab("") +
  ylab("Mental disorder risk") +
  ggtitle(paste("Mental disorder risk as a function of previous mental",
                "\n health condition and mental disorder"  , sep = "")) +
  scale_fill_manual(values = c("#00798c", "bisque", "#d1495b")) +
  scale_y_continuous(labels = scales::percent) #+ facet_wrap(~fulltime_parttime_description)
```



The following table details the numerical values of the above graph.

```
averaged_medical.prof.prev.dis <-
training_dataset %>%
  mutate(Ment_dis_binary = ifelse(currently.have.mental.disorder == "Yes", 1, 0)) %>%
  mutate(ever.had.mental.disorder = factor(ever.had.mental.disorder, levels = c("No", "Maybe", "Yes")))
  mutate(medical.prof.diagnosis = ifelse(medical.prof.diagnosis == "Yes",
                                         "Prev. mental health cond. diagnosis",
                                         "No prev. mental health cond. diagnosis")) %>%
  group_by(ever.had.mental.disorder, medical.prof.diagnosis) %>%
  summarise(Mental_disorder_cases = sum(Ment_dis_binary), total = n(),
            Mental_disorder_ratio = round(100*mean(Ment_dis_binary), 1)) %>%
  arrange(desc(Mental_disorder_ratio))

averaged_medical.prof.prev.dis
```

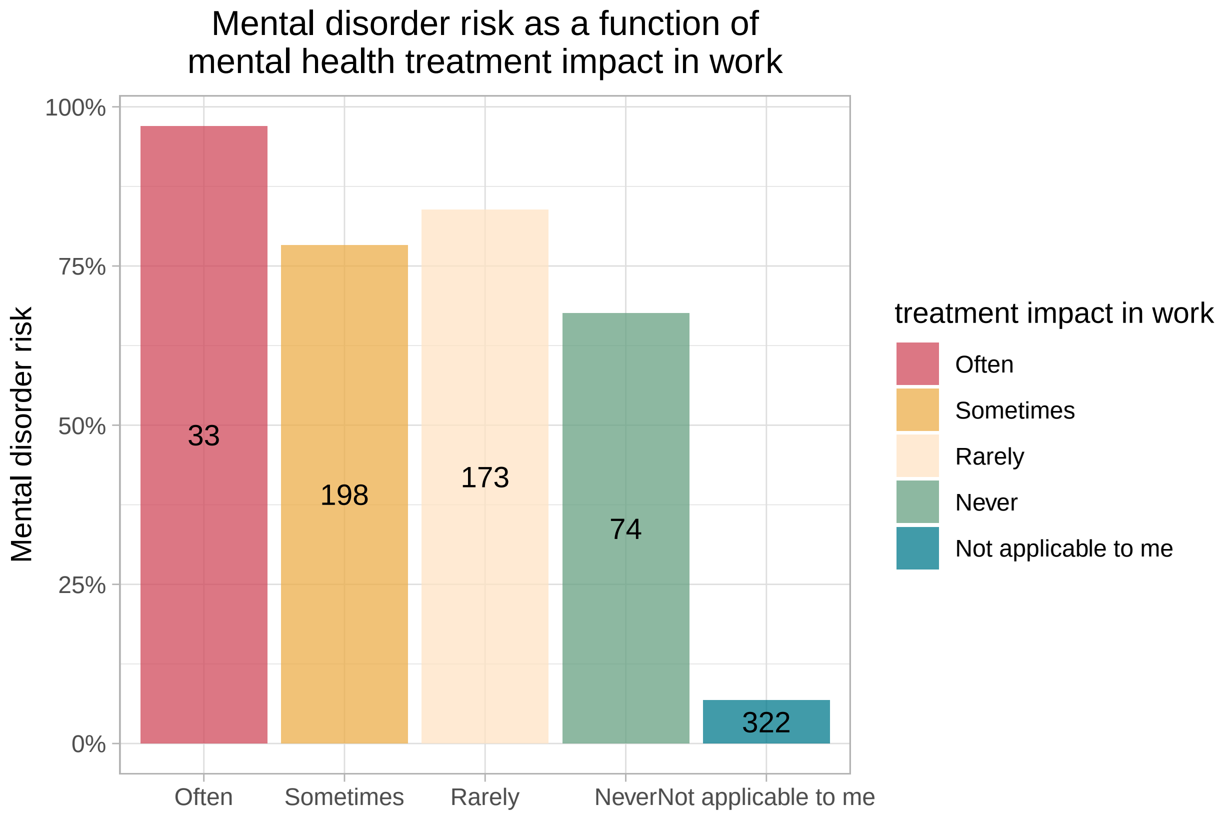
```
## # A tibble: 6 x 5
## # Groups:   ever.had.mental.disorder [3]
##   ever.had.mental.disorder medical.prof.diagnosis    Mental_disorder_cases total
##   <fct>                    <chr>                                     <dbl> <int>
## 1 No                       Prev. mental health cond~                    18    19
## 2 Yes                      Prev. mental health cond~                   327   384
## 3 Maybe                    Prev. mental health cond~                    12    15
## 4 Yes                      No prev. mental health c~                    35    49
## 5 Maybe                    No prev. mental health c~                     4    48
## 6 No                       No prev. mental health c~                     8   285
## # i 1 more variable: Mental_disorder_ratio <dbl>
```

2.2.1.c Treatment affects work

This variable represents the degree of interference with the work when being treated for a mental health issue. The following graph shows the percentage of cases with current mental disorder as a function of this variable, the label being the total number of cases in each category.

We observed that there is one case with a risk of one there is one case are an order of magnitude lower risk than the rest of the cases: *Not applicable to me* (6.8%), which is followed by *Never* option (67.6%).

```
training_dataset %>%
  mutate(Ment_dis_binary = ifelse(currently.have.mental.disorder == "Yes", 1, 0)) %>%
  mutate(treatment.affects.work = factor(treatment.affects.work, levels = c(
    "Often", "Sometimes", "Rarely", "Never", "Not applicable to me"))) %>%
  group_by(treatment.affects.work) %>%
  summarise(Mental_disorder_ratio = mean(Ment_dis_binary), count = n()) %>%
  mutate(label_val = count) %>%
  ggplot(aes(fill=treatment.affects.work,
             label = label_val,
             y=Mental_disorder_ratio,
             x= treatment.affects.work )) +
  geom_bar(position="stack", stat="identity", alpha = 0.75) +
  #scale_fill_manual(values = c("#00798c","#d1495b")) +
  geom_text(position = position_stack(vjust = 0.5)) +
  theme_light() +
  theme(#legend.title = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  guides(fill=guide_legend(title="treatment impact in work"))+
  xlab("") +
  ylab("Mental disorder risk") +
  ggtitle(paste("Mental disorder risk as a function of",
               "\nmental health treatment impact in work"  , sep = "")) +
  scale_fill_manual(values = c("#d1495b", "#edae49",  "bisque", "#66a182",  "#00798c")) +
  scale_y_continuous(labels = scales::percent) #+ facet_wrap(~fulltime_parttime_description)
```



The following table details the numerical values of the above graph.

```
averaged_impact_work <- training_dataset %>%
  mutate(Ment_dis_binary = ifelse(currently.have.mental.disorder == "Yes", 1, 0)) %>%
  mutate(treatment.affects.work = factor(treatment.affects.work, levels = c(
    "Often", "Sometimes", "Rarely", "Never", "Not applicable to me"))) %>%
  group_by(treatment.affects.work) %>%
  summarise(Mental_disorder_cases = sum(Ment_dis_binary), total = n(),
            Mental_disorder_ratio = round(100*mean(Ment_dis_binary), 1))

averaged_impact_work
```

```
## # A tibble: 5 x 4
##   treatment.affects.work Mental_disorder_cases total Mental_disorder_ratio
##   <fct>                                  <dbl> <int>                 <dbl>
## 1 Often                                     32    33                  97
## 2 Sometimes                                155   198                  78.3
## 3 Rarely                                   145   173                  83.8
## 4 Never                                     50    74                  67.6
## 5 Not applicable to me                      22   322                   6.8
```

We then explore the statistical significance of the differences found. As a result, the lowest risk category is statistical significantly lower than the others (p-value < 0.05).

```
Ofter_Never_test <- prop.test(x = c(averaged_impact_work$Mental_disorder_cases[1],averaged_impact_work$M
                              n = c(averaged_impact_work$total[1], averaged_impact_work$total[4]), alte
Never_NA <- prop.test(x = c(averaged_impact_work$Mental_disorder_cases[4],
                            averaged_impact_work$Mental_disorder_cases[5]),
                      n = c(averaged_impact_work$total[4], averaged_impact_work$total[5]), alternative =

paste("Health treatment impact in work: a not applicable case is statistically significant lower than an
      "% vs ", averaged_impact_work$Mental_disorder_ratio[4], "% (p value = ",
      signif(Never_NA$p.value,1), ")", sep = "")
```

```
## [1] "Health treatment impact in work: a not applicable case is statistically significant lower than a
```

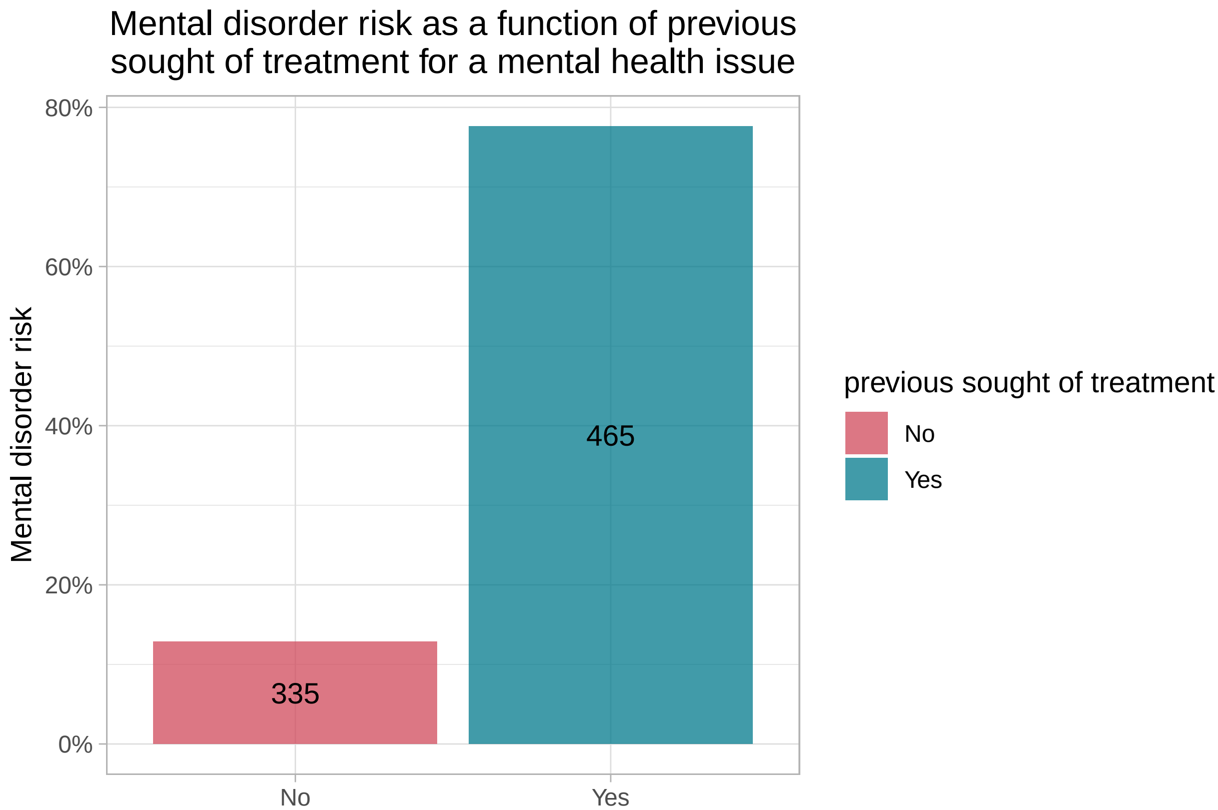
```
paste("Health treatment impact in work: an often impact is statistically significant higher than a never
      "% vs ", averaged_impact_work$Mental_disorder_ratio[4],
      "% (p value = ",  signif(Ofter_Never_test$p.value,1), ")", sep = "")
```

```
## [1] "Health treatment impact in work: an often impact is statistically significant higher than a neve
```

2.2.1.d Sought of proof treatment

The following graph shows the percentage of cases with current mental disorder as a function of previous sought treatment for a mental health issue from a mental health professional, the label being the total number of cases in each category. We observed that the previous sought of that professional treatment is correlated with high risk of having a current mental disorder (77.6%). This risk is reduced by 6 times with the absence of a previous sought (12.8%).

```
training_dataset %>%
  mutate(Ment_dis_binary = ifelse(currently.have.mental.disorder == "Yes", 1, 0)) %>%
  mutate(sought.prof.treatment = ifelse(sought.prof.treatment == 1, "Yes", "No")) %>%
  group_by(sought.prof.treatment) %>%
  summarise(Mental_disorder_ratio = mean(Ment_dis_binary), count = n()) %>%
  mutate(label_val = count) %>%
  ggplot(aes(fill=sought.prof.treatment,
             label = label_val,
             y=Mental_disorder_ratio,
             x= sought.prof.treatment )) +
  geom_bar(position="stack", stat="identity", alpha = 0.75) +
  #scale_fill_manual(values = c("#00798c","#d1495b")) +
  geom_text(position = position_stack(vjust = 0.5)) +
  theme_light() +
  theme(#legend.title = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  guides(fill=guide_legend(title="previous sought of treatment"))+
  xlab("") +
  ylab("Mental disorder risk") +
  ggtitle(paste("Mental disorder risk as a function of previous",
                "\nsought of treatment for a mental health issue"  , sep = "")) +
  scale_fill_manual(values = c("#d1495b","#00798c")) +
  scale_y_continuous(labels = scales::percent) #+ facet_wrap(~fulltime_parttime_description)
```



The following table details the numerical values of the above graph.

```
averaged_sought.prof.treatment <- training_dataset %>%
  mutate(Ment_dis_binary = ifelse(currently.have.mental.disorder == "Yes", 1, 0)) %>%
  mutate(sought.prof.treatment = ifelse(sought.prof.treatment == 1, "Yes", "No")) %>%
  group_by(sought.prof.treatment) %>%
  summarise(Mental_disorder_cases = sum(Ment_dis_binary), total = n(),
            Mental_disorder_ratio = round(100*mean(Ment_dis_binary), 1))

averaged_sought.prof.treatment
```

```
## # A tibble: 2 x 4
##   sought.prof.treatment Mental_disorder_cases total Mental_disorder_ratio
##   <chr>                                 <dbl> <int>                 <dbl>
## 1 No                                       43   335                  12.8
## 2 Yes                                     361   465                  77.6
```

We then explore the statistical significance of the differences found. As a result, the two cases analyzed in this Section are statistical significantly different (p-value < 0.05).

```
No_Yes_test <- prop.test(x = c(averaged_sought.prof.treatment$Mental_disorder_cases[2],
                               averaged_sought.prof.treatment$Mental_disorder_cases[1]),
                         n = c(averaged_sought.prof.treatment$total[2],
                               averaged_sought.prof.treatment$total[1]), alternative = "greater")
```

```r
paste("Risk of previous sought of treatment for a mental health issue is lower: ",
      averaged_sought.prof.treatment$Mental_disorder_ratio[1],
      "% vs ", averaged_sought.prof.treatment$Mental_disorder_ratio[2], "% (p value = ",
      signif(No_Yes_test$p.value,1), ")", sep = "")
```

```
## [1] "Risk of previous sought of treatment for a mental health issue is lower: 12.8% vs 77.6% (p valu
```

3. Results

3.1 Model approach design

In this subsection, the final model will be generated from the insights obtained in last section, and using the whole dataset reserved for generating the model, the *training_dataset*. This dataset will be used to test different models, and the selected one will be validated against the *testing_dataset*. The target metrics will be the F1 Score, the harmonic mean of precision and recall.

We will split the *training_dataset* into a further training and testing datasets, so that ony original training data are used to test and select the models, as follows.

```r
sub_train_index <- createDataPartition(y = training_dataset$currently.have.mental.disorder,
                                      times = 1, p = 0.1, list = FALSE)

sub_training_dataset <- training_dataset[-sub_train_index,]
sub_testing_dataset <- training_dataset[sub_train_index,]
```

The following code define a 10-fold cross validation with 3 repeats, as well as *accuracy* as target variable (no need of F1 metric as target, as the target variable is balanced).

```r
trainControl <- trainControl(method="repeatedcv", number=10, repeats=3, verboseIter = F)
metric <- "Accuracy"
```

3.1.1 k-Nearest Neighbors (KNN) K-Nearest Neighbors (KNN) is a supervised machine learning model that can be used for both regression and classification tasks. The algorithm is non-parametric, which means that it doesn't make any assumption about the underlying distribution of the data. The KNN algorithm predicts the labels of the test dataset by looking at the labels of its closest neighbors in the feature space of the training dataset. The knn-based classification is applied by means of the caret package as follows.

```r
knn_model <- caret::train(currently.have.mental.disorder~.,
                          data = sub_training_dataset, method = "knn",
                          metric = metric,trControl = trainControl)
# Model predictions
pred_m_knn = predict(knn_model, sub_testing_dataset)

# Confusion matrix
conf_m_knn <- confusionMatrix(
pred_m_knn,
sub_testing_dataset$currently.have.mental.disorder, positive = "Yes",
mode = "everything")

conf_m_knn
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  32   5
##        Yes  8  36
##
##                Accuracy : 0.8395
##                  95% CI : (0.7412, 0.9117)
##     No Information Rate : 0.5062
##     P-Value [Acc > NIR] : 3.753e-10
##
##                   Kappa : 0.6787
##
##  Mcnemar's Test P-Value : 0.5791
##
##             Sensitivity : 0.8780
##             Specificity : 0.8000
##          Pos Pred Value : 0.8182
##          Neg Pred Value : 0.8649
##               Precision : 0.8182
##                  Recall : 0.8780
##                      F1 : 0.8471
##              Prevalence : 0.5062
##          Detection Rate : 0.4444
##    Detection Prevalence : 0.5432
##       Balanced Accuracy : 0.8390
##
##        'Positive' Class : Yes
##
```

3.1.2 Random forest Random forest is a machine learning technique that uses ensemble learning, a technique that combines many classifiers to provide solutions to complex problems, and is based on many decision trees. Random forests improve predictive accuracy by generating a large number of bootstrapped trees (based on random samples of variables), classifying a case using each tree in this new "forest", and deciding a final predicted outcome by combining the results across all of the trees (an average in regression, a majority vote in classification). The random forest-based classification is applied by means of the caret package as follows (it may take 2 min).

```r
rf_model <- caret::train(currently.have.mental.disorder~.,
                         data = sub_training_dataset, method = "rf",
                         metric = metric,trControl = trainControl)

# Model predictions
pred_m_rf = predict(rf_model, sub_testing_dataset)

# Confusion matrix
conf_m_rf <- confusionMatrix(
pred_m_rf,
sub_testing_dataset$currently.have.mental.disorder, positive = "Yes",
mode = "everything")

conf_m_rf
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  32   3
##        Yes  8  38
##
##                Accuracy : 0.8642
##                  95% CI : (0.77, 0.9302)
##     No Information Rate : 0.5062
##     P-Value [Acc > NIR] : 1.223e-11
##
##                   Kappa : 0.7279
##
##  Mcnemar's Test P-Value : 0.2278
##
##             Sensitivity : 0.9268
##             Specificity : 0.8000
##          Pos Pred Value : 0.8261
##          Neg Pred Value : 0.9143
##               Precision : 0.8261
##                  Recall : 0.9268
##                      F1 : 0.8736
##              Prevalence : 0.5062
##          Detection Rate : 0.4691
##    Detection Prevalence : 0.5679
##       Balanced Accuracy : 0.8634
##
##        'Positive' Class : Yes
##
```

3.1.3 Model selection The selected model, then, is the Random Forest one.

```
models_df <- data.frame(rbind(conf_m_knn$byClass,conf_m_rf$byClass))
rownames(models_df) <- c("knn", "random forest")
models_df[, c(5:7, 11)]
```

```
##               Precision    Recall        F1 Balanced.Accuracy
## knn           0.8181818 0.8780488 0.8470588         0.8390244
## random forest 0.8260870 0.9268293 0.8735632         0.8634146
```

3.2 Testing of the prediction model

In this subsection, the selected approach, Random Forest, is generated with *training_dataset* and tested with *testing_dataset*, not used in the model generation. To achieve this, we will first generate the predictions made by our model with the input variables of the testing dataset, and then we will compare these predictions with the actual values.

```
final_rf_model <- caret::train(currently.have.mental.disorder~.,
                               data = training_dataset, method = "rf",
                               metric = metric,trControl = trainControl)
```

```r
# Model predictions
 pred_m_rf = predict(final_rf_model, testing_dataset)

# Confusion matrix
conf_m_final_rf <- confusionMatrix(
pred_m_rf,
testing_dataset$currently.have.mental.disorder, positive = "Yes",
mode = "everything")

conf_m_final_rf
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  37   1
##        Yes  7  44
##
##                Accuracy : 0.9101
##                  95% CI : (0.8305, 0.9604)
##     No Information Rate : 0.5056
##     P-Value [Acc > NIR] : 2.858e-16
##
##                   Kappa : 0.8199
##
##  Mcnemar's Test P-Value : 0.0771
##
##             Sensitivity : 0.9778
##             Specificity : 0.8409
##          Pos Pred Value : 0.8627
##          Neg Pred Value : 0.9737
##               Precision : 0.8627
##                  Recall : 0.9778
##                      F1 : 0.9167
##              Prevalence : 0.5056
##          Detection Rate : 0.4944
##    Detection Prevalence : 0.5730
##       Balanced Accuracy : 0.9093
##
##        'Positive' Class : Yes
##
```

4. Conclusion

Mental health has been the great forgotten in the general ideology when we refer to what we mean by health, being even today the object of stigma and invisibilization. The World Health Organization (WHO) has updated its definition of Mental Health by focusing not so much on the absence of illness, but on a complete state of well-being, and in a recent report urged the reorganization of environments that influence mental health, such as workplaces, as scientific literature suggests that certain types of work may increase the risk of common mental disorders, although the exact nature of this relationship has been controversial. Notwithstanding, the prevention, detection and treatment of mental health problems in the workplace is not a simple task due to its multidimensional nature, involving personal, organizational and sociocultural factors. This complexity is compounded by the stigma attached to mental illness, which is responsible, among other factors, for the fact that less than one third of people with mental disorders (in the general population) receive health care. Addressing these aspects therefore requires a multidisciplinary perspective, with contributions from occupational medicine, family and community medicine, psychiatry, psychology, sociology, nursing and social work, among others.

In this context, this work has shown key variables that correlate with mental disease, as well as a model for predicting the risk of developing such disease. In particular, two variables are highlighted with very high importance ($> 70\%$): *medical.prof.diagnosis* (i.e., previously diagnosed with a mental health condition by a medical professional) and *ever.had.mental.disorder* (mental health disorder in the past), as well as two others of medium importance (30%-50%): *treatment.affects.work* (interference of mental health issue treatment with work) and *sought.prof.treatment* (sought treatment for a mental health issue in the past). The rest have relative importance $< 13\%$. The prediction model provides high balanced accuracy and f1 using a random forest approach (outperforming the knn-based approach by 0.03).

Future works may include detailing the method for different countries/sectors separately, as well as the inclusion of new input data regarding detailed past history of workers and their environment. The results of this work can also be linked in the future to attrition rates in certain companies (or jobs), since certain work environments may increase the risk of common mental disorders: According to data collected by Infojobs and Esade within the study State of the labor market during 2022, the main reason among those willing to leave their job, 27% of the total - 4% more than the previous year - was to protect their mental health (32%). Even ahead of the search for better financial conditions (27%) or a job with better work-life balance (24%). In this context, it is crucial to predict in advance risky work situations that may lead to mental disorders, in order to take actions aimed at preventing them, thus avoiding negative effects on the worker and the company.

5. References

Irizarry, R. A. (2019). Introduction to data science: Data analysis and prediction algorithms with R. CRC Press.

OSMI: Osmi Mental Health in Tech Survey (2016). https://osmi.typeform.com/report/Ao6BTw/U76z. Accessed 01 Sep 2023.

Breiman, L. (2001). Random forests. Mach Learn, 45:5-32. tools:::Rd_expr_doi("10.1023/A:1010933404324").

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

ESADE & Infojobs (2023). State of the labor market. Annual Report. https://nosotros.infojobs.net/wp-content/uploads/2023/03/Informe-Anual-InfoJobs-Esade-2022.pdf