UNIVERSITÉ JEAN MONNET

Mental Health in Tech Survey

Done by

Mohammad Badia ALFATHI

Submitted to

Dr. Fabrice Muhlenbach

April 2022

Contents

Figures Table

Introduction

Mental disorders are one of the most serious problems that the world suffers from from the past until now, and they are constantly increasing, especially within the rapid technological revolution that is taking place in the world.

In most cases, it is difficult to determine the causes of these disorders, and this makes their diagnosis more difficult, in addition to the lack of data in the psychological field. The diagnosis is made based on the patient's behavior, feelings, or thoughts. Therefore, researchers in OPEN SOURCING MENTAL ILLNESS OSMH[1] identified some of the influencing factors that lead to the occurrence or exacerbation of mental illness, which were collected in a Mental Health in Tech Survey[2] implemented in 2014 that measures attitudes towards mental health and the frequency of mental health disorders in the technology workplace.

The survey contained 27 different workers and questions, which were applied to about 1,300 workers in a technology workplace.

Data Analysis

Data Description

The data we used is provided on Kaggle[3]. It contains the questions represents the 27 features of the data, shown as follow:

Personal Info.	Timestamp, 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 primarily a tech company/organization?
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?

[1] OPEN SOURCING MENTAL ILLNESS OSMH.

[2] Mental Health in Tech Survey.

[3] Mental Health in Tech Survey Dataset.

seek_help:	Does your employer provide resources to learn more about mental health issues 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?
mentalhealthconsequence:	Do you think that discussing a mental health issue with your employer would have negative consequences?
physhealthconsequence:	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)?
mentalhealthinterview:	Would you bring up a mental health issue with a potential employer in an interview?
physhealthinterview:	Would you bring up a physical health issue with a potential employer in an interview?
mentalvsphysical:	Do you feel that your employer takes mental 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?
comments:	Any additional notes or comments

Data Preprocessing

First, after importing the data I started the work by understanding it and checking the Nulls values using str(), summary(), and is.na() functions. Then, I have used the ggplot() function to plot the histogram (distribution) of each feature of the dataset. In the below plots we can see that the data Age, Gender, self_employed, work_interfere attributes are not clean, we need to process them.

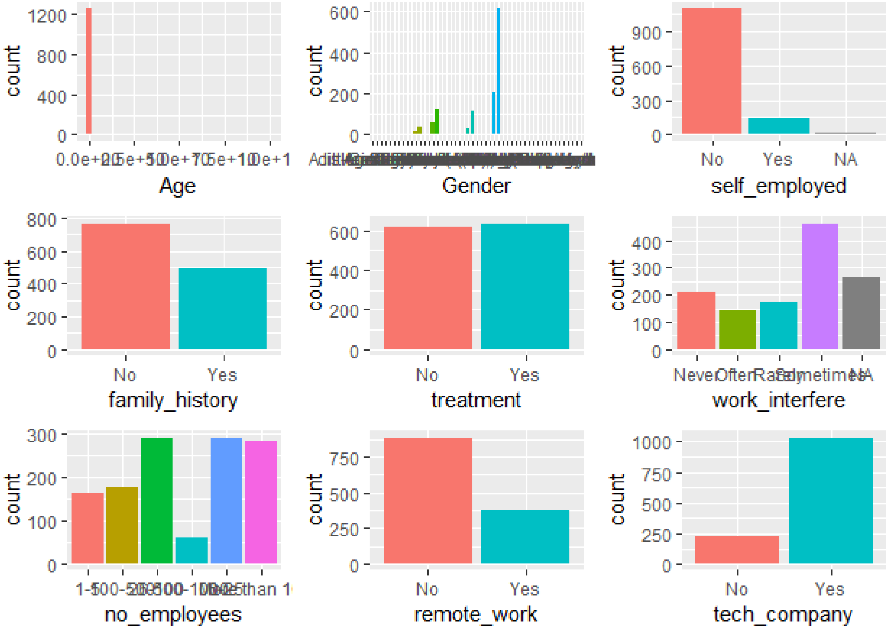


Figure 1: data distribution before processing

- I started by the **age** distribution that I split it into 2 categories who they are younger than 30, or older.

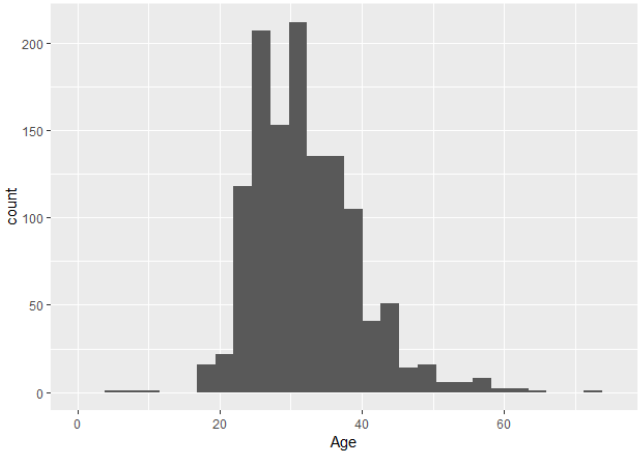


Figure 2: Age distribution after processing

- In the **gender** attribute, there were multiple categories like female, F, f, woman, etc. So, I categorized them into "male", "female", and "undecided" for those not understandable.

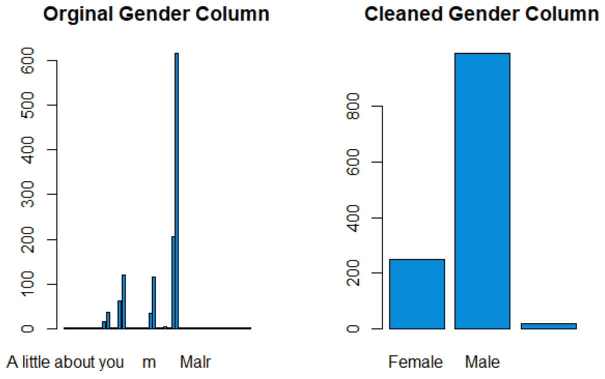
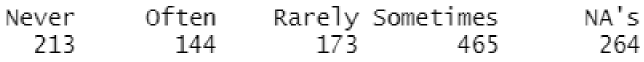


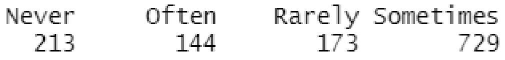
Figure 3: Gender attribute comparison

- Finally, **self_employed, work_interfere** columns have some Nulls data that I replaced them by the most frequency value in each column (mode).



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

Figure 4: work_interfere before cleaning



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

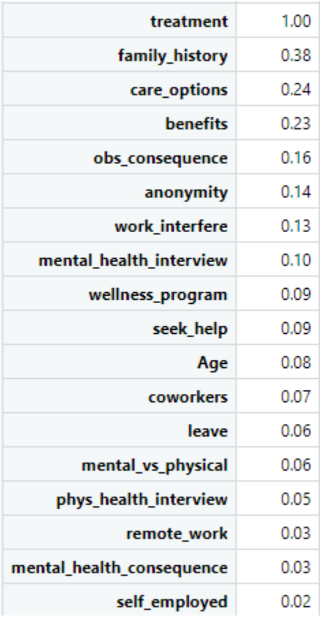
Figure 5: work_interfere after cleaning

Drop Columns

In the **comment** column most of the values are Nulls, so I decided to drop the column where there is no information can be used there. Also, I see that Timestamp, country, and state attributes can not provide us important data that can be used to improve the model. So, I dropped them.

Correlation Analysis

Numerical data is required in order to analyze the correlation between different attributes. Since we just have the Age as a numeric column and the others as factors, we need to convert them into numerical data. Then, we show the correlation between the target label **treatment** with all other attributes. We can see that the highest correlation found with family_history, care options, and the mental health benefits that the employer provides. On the other hand, the lowest correlation shown in the Gender of employee, and the number of the employees in the company.



treatment	1.00
family_history	0.38
care_options	0.24
benefits	0.23
obs_consequence	0.16
anonymity	0.14
work_interfere	0.13
mental_health_interview	0.10
wellness_program	0.09
seek_help	0.09
Age	0.08
coworkers	0.07
leave	0.06
mental_vs_physical	0.06
phys_health_interview	0.05
remote_work	0.03
mental_health_consequence	0.03
self_employed	0.02

Figure 6: correlation with target label

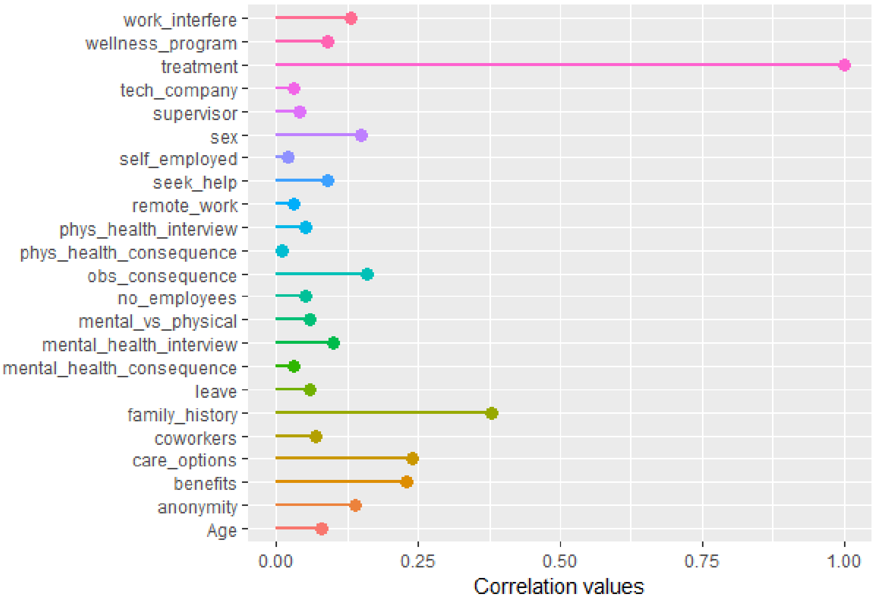


Figure 7: correlation map

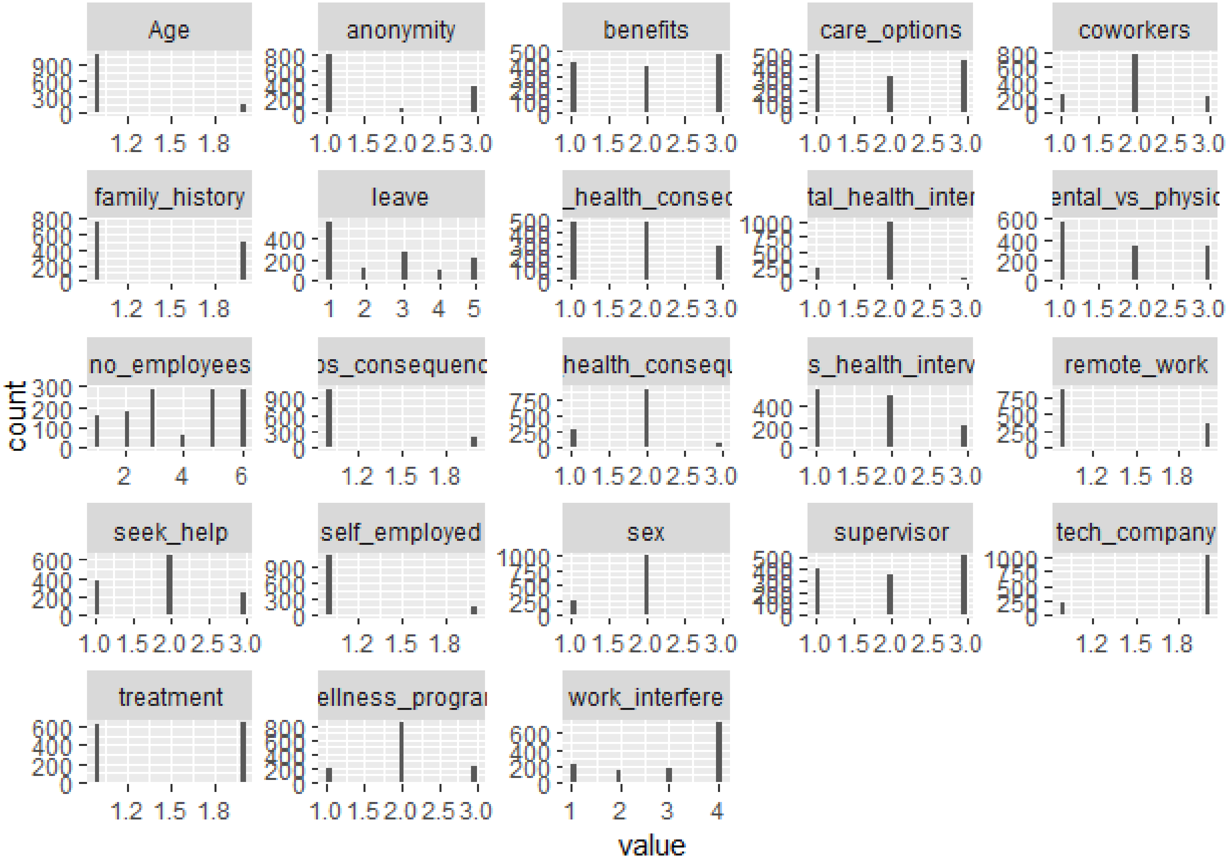


Figure 8: Final distribution of the data

Normalization

I tried to normalize the data, but it does not make any improving in the model because the data is discrete and categorical.

PCA

I also used prcomp() function to do the PCA, but it is also does not add any improvement on the predictions. So, it was enough to use the attributes that I kept and leave what I dropped.

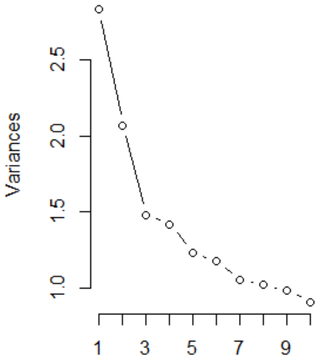


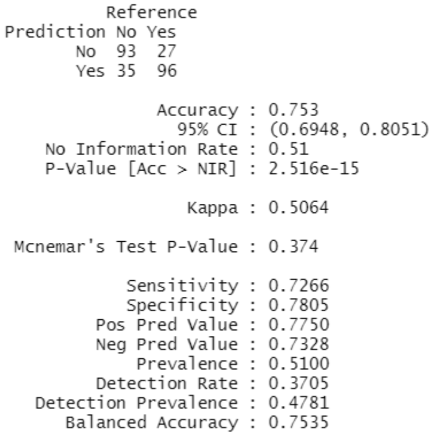
Figure 9: PCA

Model

I split the data into the ratio of 80% for training, 20% for testing using the createDataPartition() function. Then, I learned these model as follow:

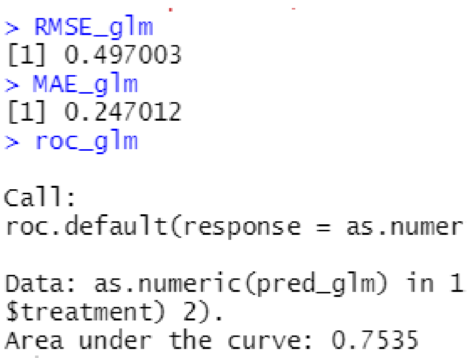
Logistic Regression

I started with Logistic Regression model where I get the Confusion matrix, RSME, MAE, and ROC as shown:



```
          Reference
Prediction No Yes
       No  93  27
       Yes 35  96

               Accuracy : 0.753
                 95% CI : (0.6948, 0.8051)
    No Information Rate : 0.51
    P-Value [Acc > NIR] : 2.516e-15

                  Kappa : 0.5064

 Mcnemar's Test P-Value : 0.374

            Sensitivity : 0.7266
            Specificity : 0.7805
         Pos Pred Value : 0.7750
         Neg Pred Value : 0.7328
             Prevalence : 0.5100
         Detection Rate : 0.3705
   Detection Prevalence : 0.4781
      Balanced Accuracy : 0.7535
```

Figure 10: Logistic Regression Confusion matrix

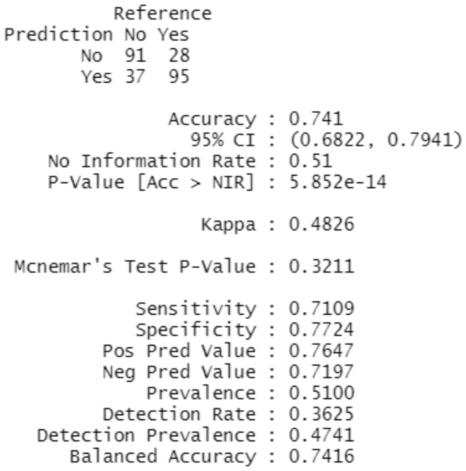


```
> RMSE_glm
[1] 0.497003
> MAE_glm
[1] 0.247012
> roc_glm

Call:
roc.default(response = as.numer

Data: as.numeric(pred_glm) in 1
$treatment) 2).
Area under the curve: 0.7535
```

Figure 11: Logistic Regression Evaluation

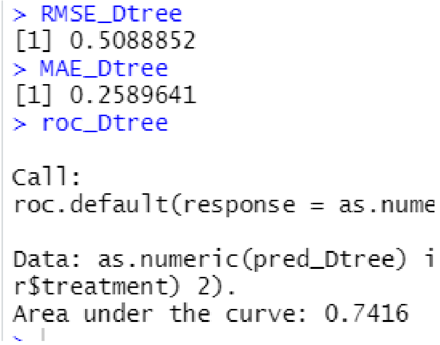
Decision Trees

Here I trained the Decision Trees model and here we can see the evaluation of the model:



```
          Reference
Prediction No Yes
       No  91  28
       Yes 37  95

               Accuracy : 0.741
                 95% CI : (0.6822, 0.7941)
    No Information Rate : 0.51
    P-Value [Acc > NIR] : 5.852e-14

                  Kappa : 0.4826

 Mcnemar's Test P-Value : 0.3211

            Sensitivity : 0.7109
            Specificity : 0.7724
         Pos Pred Value : 0.7647
         Neg Pred Value : 0.7197
             Prevalence : 0.5100
         Detection Rate : 0.3625
   Detection Prevalence : 0.4741
      Balanced Accuracy : 0.7416
```

Figure 12: Decision Trees Confusion matrix



```
> RMSE_Dtree
[1] 0.5088852
> MAE_Dtree
[1] 0.2589641
> roc_Dtree

Call:
roc.default(response = as.nume

Data: as.numeric(pred_Dtree) i
r$treatment) 2).
Area under the curve: 0.7416
```

Figure 13: Decision Trees Evaluation

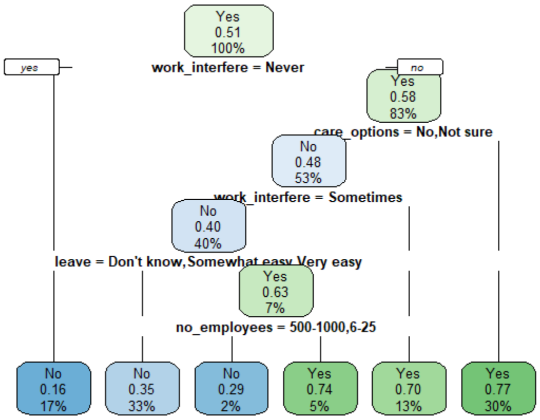


Figure 14:Decision Trees Representation

Cross Validation

I trained the Logistic regression, SVM, and Decision trees models using Cross validation. Here we can see the comparison between the models.

Logistic regression		SVM			Decision trees		
Accuracy	Kappa	C	Accuracy	Kappa	cp	Accuracy	Kappa
0.7399871	0.4799431	0.25	0.7331454	0.4663030	0.01982851	0.7076967	0.4147065
		0.50	0.7402758	0.4805364	0.06109325	0.6883378	0.3781594
		1.00	0.7408069	0.4816147	0.35852090	0.5703540	0.1358309

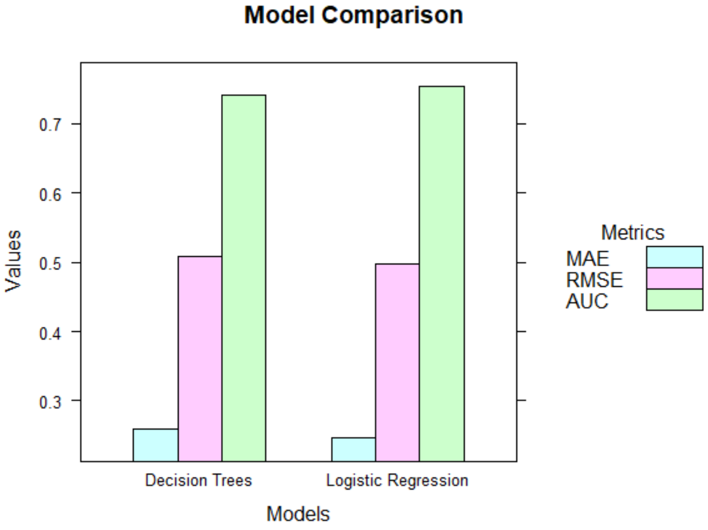


Figure 15: Models Comparison

Conclusion

In this project, we have analyzed the predictors of mental illness in the tech industry. We figured out the factors that may cause or exacerbates the disease. Employers have a significant impact on the psychological impact on employees. As we saw we could get some expected results in this project that I chose based on my big interest in the psychological field. The results were good for this size of data. I learned how to deal with categorical data and the data based on surveys. Finally, my passion in this field has increased, and I hope to work on the new surveys that I found during my work and research in this project.

REFERENCES

[1] OPEN SOURCING MENTAL ILLNESS OSMH.
[2] Mental Health in Tech Survey.
[3] Mental Health in Tech Survey Dataset.

[4] Notebook.