ST. CLAIR COLLEGE

PROJECT REPORT HEALTHCARE ANALYTICS

Analyze Trends in Mental Health Issues in the Technical Industry

The aim is to understand the trend of mental health in the tech industry using OSMI Data over the duration of 2014-2020.

Authors

Suyash Chaudhari (769626)

Adrain Dsilva (756525)

Shane Marotical (769174)

Professor

Mahmoud Artima

INTRODUCTION

The World Health Organization conducted a study in 2004 called “the global burden of disease “which was done to measure the impact that a disease has on a population. They found that the burden of mental disorders was the largest of all disorder categories in North America, larger than cardiovascular disease and larger than cancer. this is a quote from this study. "In all regions, neuropsychiatric conditions are the most important causes of disability, accounting for around 1/3 of years lost to disability among adults aged 15 years over." Mental health conditions are the second leading cause of workplace absenteeism. Yet the work culture doesn’t do enough to deal with the crisis.

On an average 5.6 hours a week is lost due to workers not reporting to work because of depression (excluding other mental illnesses). Workers without mental illness lose about 1.5 hours a week. That’s more than 3 times the number of hours lost. In terms of economic value, just one major category of mental health disorder can lead to a loss $43.7 billion and 200 million days. Taking this into consideration we are planning to do our analysis on mental tech survey data from 2014 to 2020, apply concepts and strategies learned in this course and come up with insights, conclusions, and report if the mental health situation has changed in the recent past and if we can deal with this crisis better in terms of work culture in the tech world. We would be creating a major impact and awareness if we can contribute and find out how much effort has been taken in the recent years to deal with this problem and help in lowering the economic damage caused.

We use the OSMI mental health survey [3] data from 2014 to 2020 as a dataset in this study, and we try to find out the trend for mental health over the period. OSMI (Open Sourcing Mental Illness Initiative) is a non-profit organization whose mission is to improve mental health awareness problems of health. They have been conducting surveys over the year and provide this data for analysis. They show the rest of the world how to do business. Workers' productivity is affected by mental illnesses. They help to make the workplace a safer and more pleasant place to work for the employees.

HYPOTHESIS

Considering the advancement in technology and ease of access to mental health resources, the mental health of employees in the tech Industry should be better than before.

RELATED WORK

Machine Learning Techniques for Stress Prediction in Working Employees

In this paper [1] the author has used the OSMI Mental Health in Tech 2017 survey as the dataset, using which they trained different machine learning models to analyze the patterns of stress and mental health disorders among tech professionals and to determine the most influential factors that contribute to the same. The dataset consisted of 70 columns, but to have a specific model, some of these attributes have been neglected and finally 14 out of the 70 parameters were taken into consideration based on their relevance to the research. Whether an employee has taken treatment for stress-related disorders in past or not is used as a reference to be predicted by our trained models.

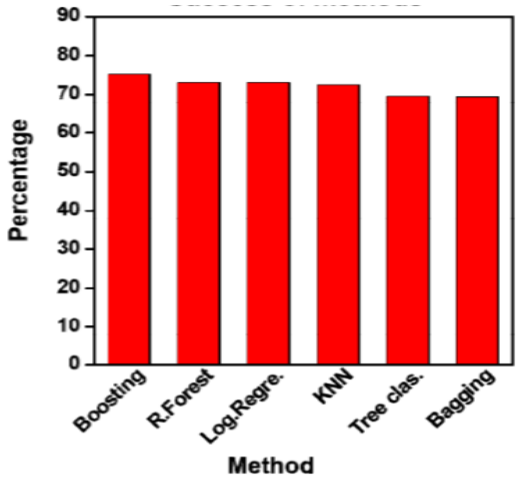

Fig 1.1 Accuracy comparison of different classifiers

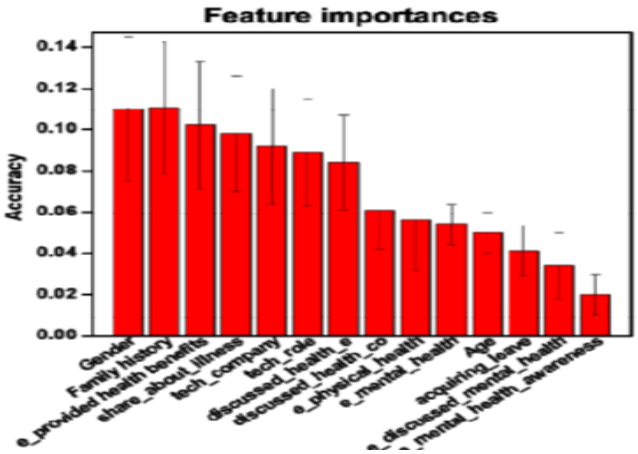

Fig 1.2 Feature importance of various attributes

From Figure 1.1, it can be stated that all the trained models performed well in classification with Boosting achieving highest accuracy of 75.13 and bagging achieved the least accuracy of 69.43.

The feature relevance of the 14 attributes analyzed is depicted in Figure 1.2 using a decision tree. Gender has the greatest impact on stress and anxiety, as shown in the graph.

Predicting Mental health disorders using Machine Learning for employees in technical and non-technical companies

In this paper [2] the author has used the data for the year 2019 from OSMI which has over 70 features. After selecting a few important features, they have used various ML algorithms to not only predict about the mental health but also find the important of each feature as how much they contribute to any mental health issue.

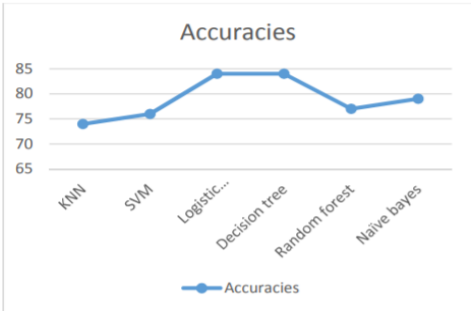

Fig 1.3 Accuracy comparison of different classifiers

The author has found out that Decision tree has the best accuracy for this. Also, Feature importance of the selected features showed that a history of mental health disorder contributes most during disorder prediction followed by family history. They have found that rest of the features contributes bare minimum to the

prediction with gender as their top rest of the features which includes mental health benefits or care provided by the employer, age and discussing mental health status with the employer barely makes any contribution to the prediction of mental health disorder.

METHODS

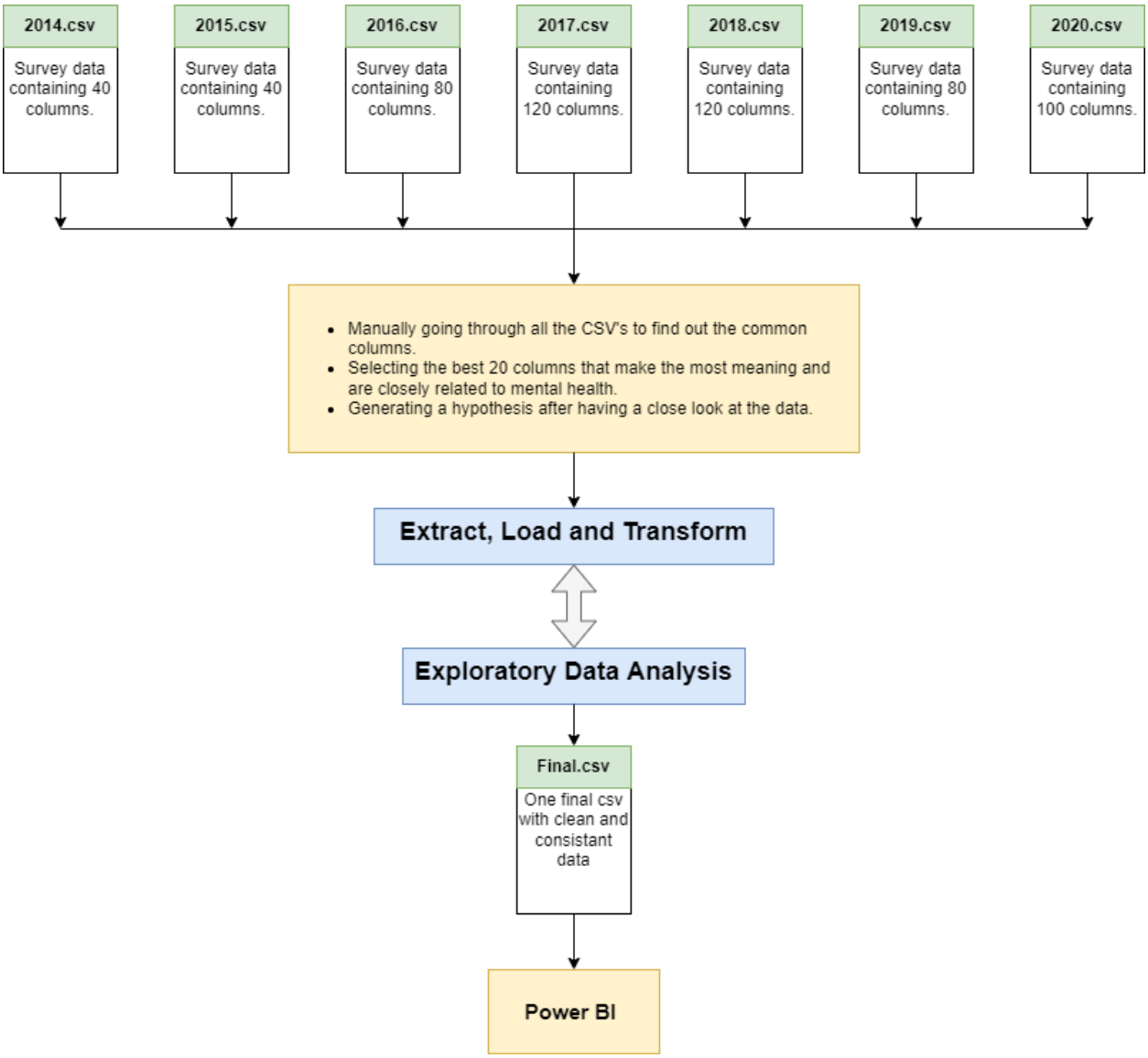

Fig 2.1 Process flow diagram

As we had proposed that we plan to find a trend in the data over time. This was a challenge as no one had done it earlier. In the citations above we can see that analysis on individual years has been done vastly.

The biggest task was to find all the common data from all the csv files, as all files had a different number of columns. As the data was divided into 3 major sections namely Personal information, individual's opinion about the company, and individual's opinion about themselves. This was a rigorous task and had to be done manually.

As the data was very inconsistent and had a lot of null values it had to undergo multiple rounds of EDA and ETL. Below is an example of some data inconsistency and how it was dealt with.

	How many employees does your company or organization have?
0	100-500
0	100-500
0	Jun-25
0	More than 1000
1	
0	100-500
0	Jun-25
0	26-100
0	100-500
1	
0	100-500
0	100-500
0	More than 1000
0	More than 1000
0	26-100
0	100-500
0	26-100
1	
0	Jun-25
0	More than 1000
0	Jun-25
0	100-500
0	26-100
0	01-May
0	100-500
0	Jun-25
0	More than 1000

Fig 2.2 organization size data inconsistency

In this column, we see that the header is the survey question. As this is not a suitable name for any analysis it was changed to company_size. Similar changes have been made for all the columns.

The next task here is that we see the value 'Jun-25' in this column. Which cannot be the case, as the number of employees cannot be a date. The value should have been '6-25'.

To achieve this, python replace function comes in handy.

```
# clean company size
data["company_size"] = data["company_size"].str.replace("25-Jun", "6 to 25")
data["company_size"] = data["company_size"].str.replace("Jun-25", "6 to 25")
data["company_size"] = data["company_size"].str.replace("01-May", "1 to 5")
data["company_size"] = data["company_size"].str.replace("05-Jan", "1 to 5")
data['is_mh_benefit_provided'].value_counts()
```

Fig 2.3 Code to fix inconsistency in company_size

Using this we can see that inconsistent data values like “25-Jun”, “Jun-25”, “01-May” and “05-Jan” have been replaced with their original value making the column data useful for analysis.

Another common data inconsistency we came across was columns with the same means but different names. E.g., some CSV’s had ‘Gender’ column while others had ‘What is your Gender’. As this would have been a problem while merging all the files together, we had to manually check, change, and update the column names.

have_sought_mh_treatment	have_sought_mh_treatment
0	FALSE
1	FALSE
1	FALSE
1	FALSE
1	FALSE
1	TRUE
0	TRUE
1	TRUE
1	

Fig 2.3 Inconsistency in have_sought_mh_treatment

In the above picture we see that in some CSV’s the values are stored as 0 and 1 while in some it TRUE and FALSE. To have consistent data we have replaced all the 0’s and FALSE’s with ‘No’ and all the 1’s and TRUE’s with ‘Yes’ using the replace function.

```
# clean have_sought_mh_treatment
data["have_sought_mh_treatment"] = data["have_sought_mh_treatment"].str.replace("0", "No")
data["have_sought_mh_treatment"] = data["have_sought_mh_treatment"].str.replace("1", "Yes")
data["have_sought_mh_treatment"] = data["have_sought_mh_treatment"].str.replace("FALSE", "No")
data["have_sought_mh_treatment"] = data["have_sought_mh_treatment"].str.replace("TRUE", "Yes")
data['have_sought_mh_treatment'].value_counts()
```

```
Yes    2007
No     1289
Name: have_sought_mh_treatment, dtype: int64
```

Fig 2.4 Code Snippet to fix inconsistency in have_sought_mh_treatment

We have used this approach with a lot to columns which had inconsistent data. After an iterative EDA and ETL process we were able to generate 1 CSV which consisted of data from the years 2014 -2020. It included 20 selected columns with clean and consistent data.

As we used Microsoft Power BI to generate this newly generated CSV came in handy to generate visuals and find patterns.

Our goal was creating visuals depicting

1. Individual’s opinion about the company
2. Individual’s opinion about themselves

Based on the survey questions available in the dataset we created 5 visuals per section and tried find out if there is a positive to negative trend.

Finding the trend will allow us to see the hidden story in the data. As our hypothesis expects that there should be a positive trend in the data, i.e., the mental health of the employees is getting better.

RESULTS

To keep the trend consistent, we should note that an increase in the green bar towards the right (from 2014 -2019) is a positive trend. This indicate that the thing had got better over the years.

On the other had if we see that the red bars are increasing meaning there is a negative trend indicating things have got bad.

1. Creating visuals for individual's opinion about the company

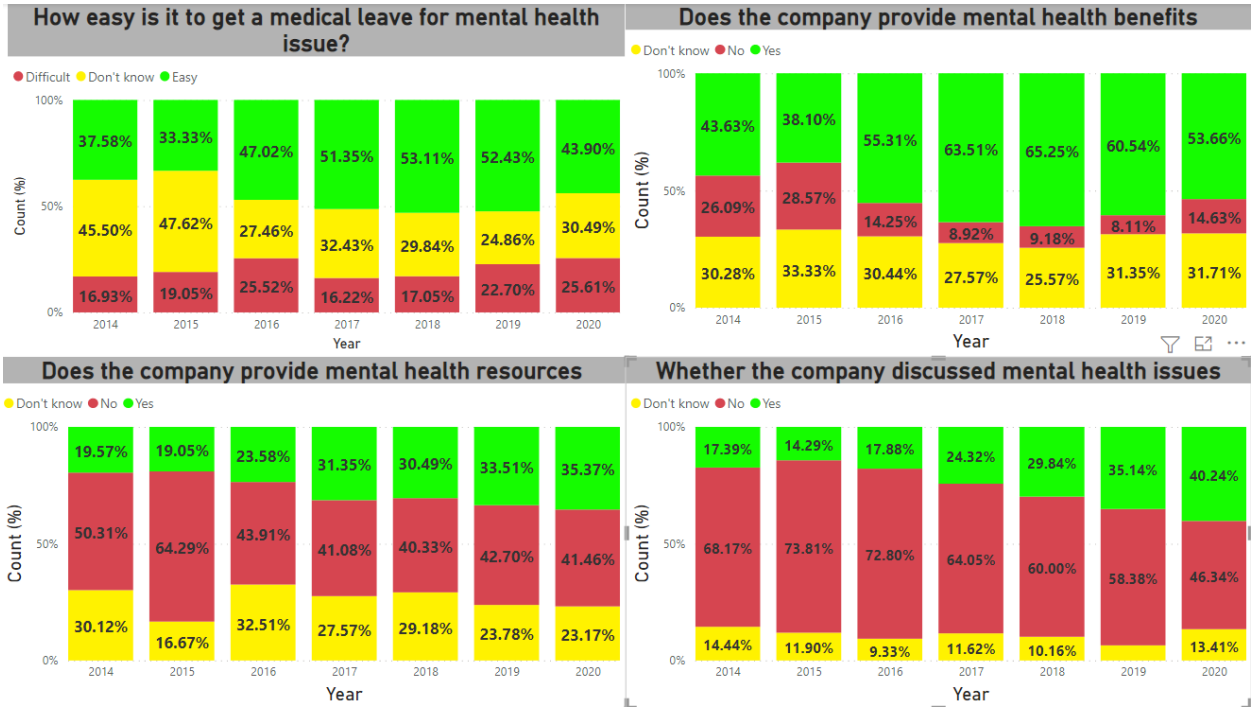

Fig 3.1 Individual's opinion about the company dashboard

2. Creating visuals for individual's opinion about themselves

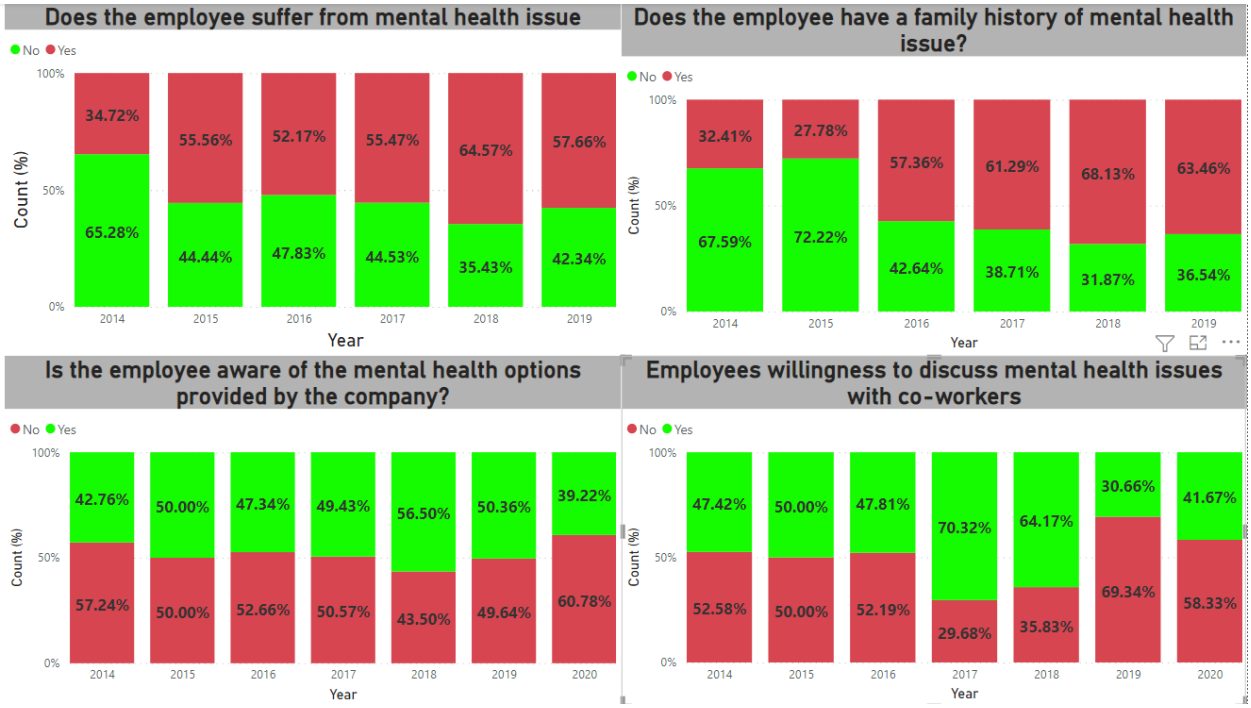

Fig 3.2 Individual's opinion about themselves dashboard

We created a matrix to summarize the trend for all the questions asked. Red indicates a negative trend and green indicates a positive trend. The figure below shows the matrix.

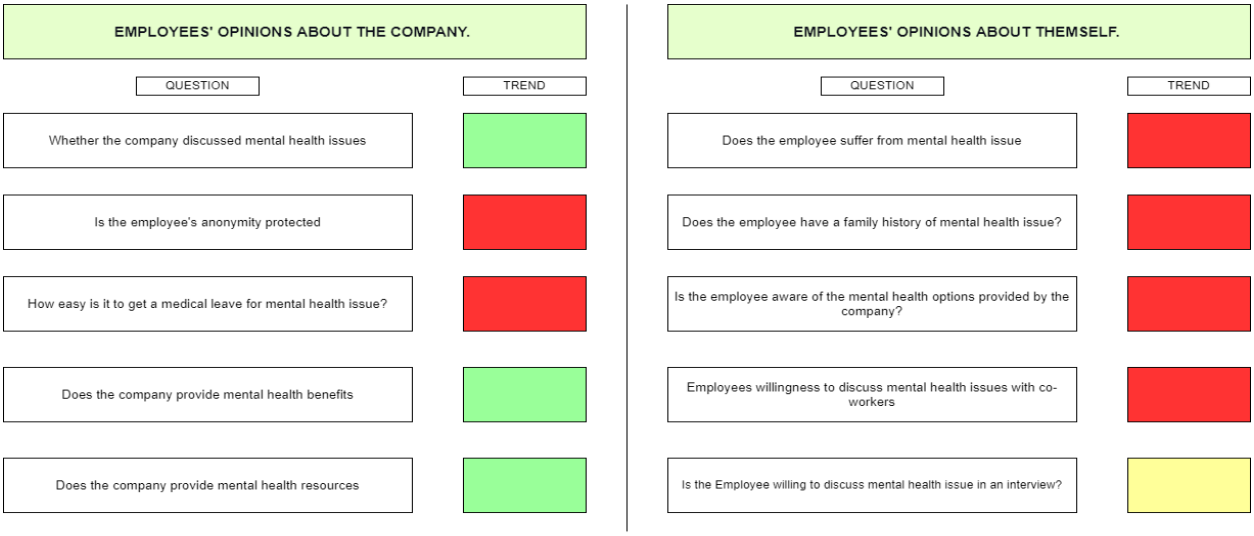

Fig 3.3 Summary matrix for all questions and their trends

DISCUSSIONS

Looking at the left side of the fig 3.3 we see that there are 2 negative trends and 3 positive trends in the employees' opinion about the company. While on the right there are 4 negative trends and 1 neutral trend in the employees' opinion about themselves.

The positive trends on the left side of fig 3.3 clearly indicates that the companies are trying to help their employees with the resources and benefits to deal with mental health issues but that's not the only factor which is responsible for the employees' mental wellbeing.

From the results, we can clearly say that we were able to come up with a conclusion for the hypothesis.

CONCLUSIONS

As mentioned in fig 3.3 we can say that majority of the trends/graphs are pointing towards the deterioration of mental health among the employees.

We can say that the mental health of the employees in the tech industry is getting worse over the years and hence we reject the Hypothesis.

Our future objectives is to deeply understand and analyze the reasons behind the employees opinion about the company work culture and what measures can be taken to bridge the gap between the employees and their employers.

CONTRIBUTIONS

Activity	Suyash	Shane	Adrian
Project Topic Research	✓	✓	✓
Project Proposal		✓	✓
Data source search			✓
Weekly minutes		✓	
Data description merge for different years	✓		✓
Data merging	✓	✓	
Data cleaning	✓		✓
Data Visualization	✓	✓	
Presentation		✓	
Presentation report	✓		✓
Final project report	✓	✓	✓

REFERENCES

[1] Reddy, U Srinivasulu; Thota, Aditya Vivek; Dharun, A (2018). [IEEE 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) - Madurai, India (2018.12.13-2018.12.15)] 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) - Machine Learning Techniques for Stress Prediction in Working Employees. , (), 1–4. doi:10.1109/ICCIC.2018.8782395

[2] Katarya, R., & Maan, S. (2020). Predicting Mental health disorders using Machine Learning for employees in technical and non-technical companies. 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE). doi:10.1109/icadee51157.2020.9368923

[3] Osmihelp.org. 2021. *OSMI Home: Open Sourcing Mental Illness - Changing how we talk about mental health in the tech community - Stronger Than Fear*. [online] Available at: <https://osmihelp.org/> [Accessed 9 December 2021].