Data Science Project Report

Project Topic: Mental Health in Tech Industry

Group Members:

Noor Ahmed (2112282)

Keertan Kumar (2012161)

Submitted to:

Dr Imran Amin

Sir Muhammad Ahsan

Table of Contents

1. Abstract

Technology industries play a huge role in our everyday life and they are a backbone of other industries as well. Since tech industries are focused on being innovative and have growth to provide better services, the employees of organizations are an essential part behind their growth. Just like physical health a person's mental health also changes throughout their lives which can cause a person's health to become poor. Bad mental health conditions are increasing worldwide, especially in workplaces, if employees do not perform well organizations also suffer from this problem. In view of today's evolving workplaces, this project has been developed to gain insights on mental health problems in the tech industry by using OSMI (Open Sourcing Mental Illness) dataset to visualize the data and applying machine learning techniques to predict if an employee has ever sought treatment. Three machine learning models were used and they were evaluated using appropriate metrics. We performed further analysis using a Power BI dashboard which provides visualization and options to interact with the data.

2. Introduction

Mental health encompasses a wide range of psychological, emotional, and social aspects. These days, it leaves a lasting influence on both the prosperity of a company and the well-being of its employees. It plays an integral part in one's life. It basically affects the way we think, feel and act towards others. It helps in coping with stress and make better choices in life. Therefore, mental health is important at every stage of life. Likewise, in job culture a person's mental health matters a lot. In fact, having a good mental health positively affects job performance [1]. It is a right of employees to have a safe and healthy work environment. It is suggested that employees with good mental health show a positive working state and complete their tasks with more enthusiasm [2].

3. Literature Review

There have been various studies done on mental health. Mental health in workplace is a critical aspect, its importance should be highlighted in organizations worldwide. One study reviewed international guidelines and provided valuable insight for mental health in workplaces [3].

While relating to machine learning in mental health diagnosis, one study has expressed positive outcomes of using machine learning for the diagnosis and detection of mental health [4]. Using machine learning techniques, a study was made to predict mental health disorders it gave insight that gender and company type also play a role in mental health [5].

4. Problem Statement

Since, mental health illness is a growing concern worldwide, an employee cannot work efficiently with a bad mental health in the workplace. Technical workplaces too suffer as well if their workforce is not performing well.

This project aims to gain insights and look at factors that contribute to mental health problems in the tech industry. This objective is possible to achieve by using a mental health dataset provided by OSMI (Open Sourcing Mental Illness). OSMI is a non-profit organization that raises awareness and provides services to support mental health in tech communities. The dataset is utilized by using machine learning techniques to predict the treatment needed for an employee using various features of the dataset. An interactive Power BI dashboard has also been the goal of this project to further interact with the data.

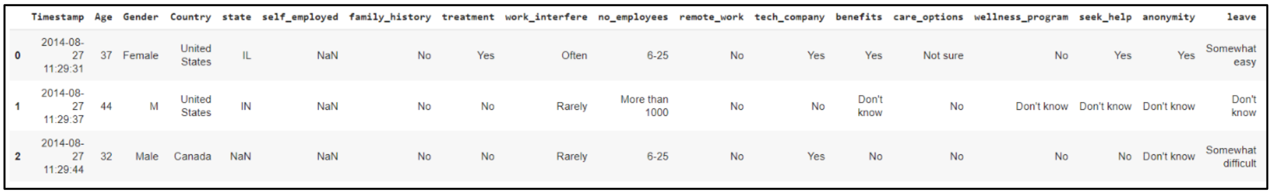
5. Methodology

5.1 Dataset

We have used a dataset from Kaggle named “Mental Health in Tech Industry” which is a survey dataset, conducted by OSMI (Open Sourcing Mental Illness) which provides insights towards mental health in the tech industry.

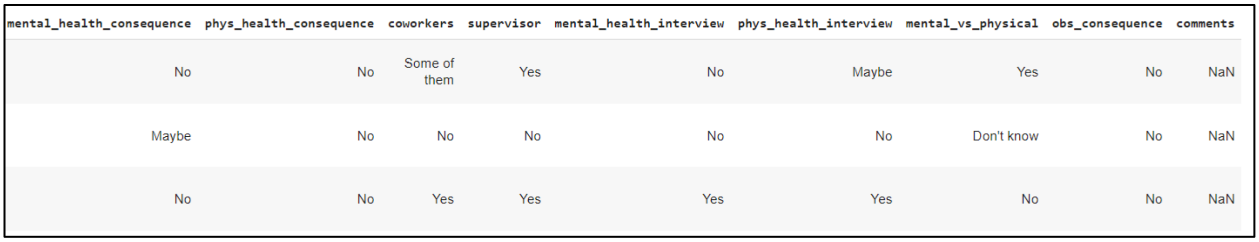
The dataset includes 1259 rows and 27 columns/variables. The variables are basically questioning of the survey which includes a person’s age, gender, country, treatment, family history, leave etc. Since, the columns are questions. Description regarding necessary columns has been provided below:

- **Timestamp:** Provides date time information
- **Age:** Age of the person
- **Country:** Country of the person
- **state:** Provides information if person lives in United States
- **self_employed:** Information if the person is self-employed or not
- **family_history:** Information of the person if they have family history of mental health problem
- **treatment:** If the person has ever sought treatment
- **work_interfere:** Provides information if work interferes with person’s mental health
- **no_employees:** Provides information about the number of employees in person’s company
- **remote_work**: Provides information if the person works remotely at least 50% of the time
- **tech_company**: Provides information if the person works in a tech company
- **benefits:** Provides information if the employer provides mental health benefits
- **care_options:** Provides information if the person knows about the care options provided by employer
- **wellness_program:** Provides information of employer giving mental health resources as part of wellness program
- **seek_help:** Provides information of the employer provides resources to help seek help
- **anonymity:** Provides information if the person’s anonymity protected when taking advantage of mental health condition
- **leave:** Provides information if it is easy to take leave



	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees	remote_work	tech_company	benefits	care_options	wellness_program	seek_help	anonymity	leave
0	2014-08-27 11:29:31	37	Female	United States	IL	NaN	No	Yes	Often	6-25	No	Yes	Yes	Not sure	No	Yes	Yes	Somewhat easy
1	2014-08-27 11:29:37	44	M	United States	IN	NaN	No	No	Rarely	More than 1000	No	No	Don't know	No	Don't know	Don't know	Don't know	Don't know
2	2014-08-27 11:29:44	32	Male	Canada	NaN	NaN	No	No	Rarely	6-25	No	Yes	No	No	No	No	Don't know	Somewhat difficult

Figure 1: Dataset Overview



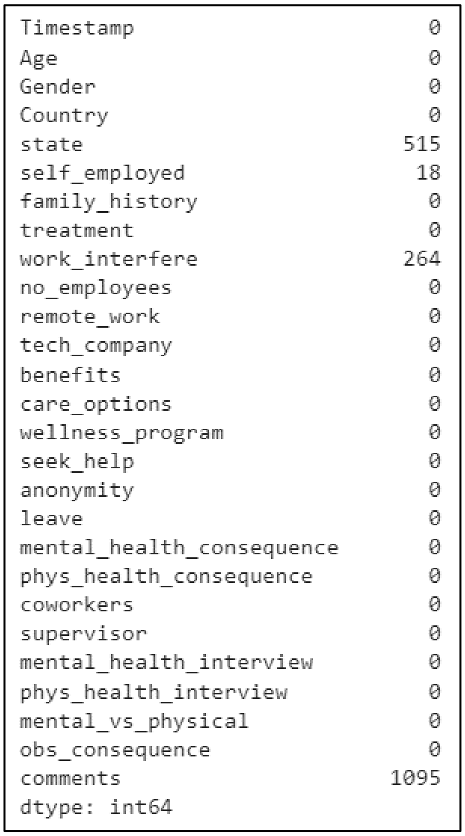
mental_health_consequence	phys_health_consequence	coworkers	supervisor	mental_health_interview	phys_health_interview	mental_vs_physical	obs_consequence	comments
No	No	Some of them	Yes	No	Maybe	Yes	No	NaN
Maybe	No	No	No	No	No	Don't know	No	NaN
No	No	Yes	Yes	Yes	Yes	No	No	NaN

Figure 2: Dataset Overview (cont.)

5.2 Data Cleaning

Since a dataset may contain anomalies or unnecessary data which need to be cleaned first before further processing or training using that data.

Four columns in the dataset contained missing values, the "state" column contained 515 missing values, "self_employed" column contained 18 missing values, "work interfere" column contained 264 missing values and the "comments" column contained 1095 missing values.



```
Timestamp                      0
Age                            0
Gender                         0
Country                        0
state                        515
self_employed                 18
family_history                 0
treatment                      0
work_interfere               264
no_employees                   0
remote_work                    0
tech_company                   0
benefits                       0
care_options                   0
wellness_program               0
seek_help                      0
anonymity                      0
leave                          0
mental_health_consequence      0
phys_health_consequence        0
coworkers                      0
supervisor                     0
mental_health_interview        0
phys_health_interview          0
mental_vs_physical             0
obs_consequence                0
comments                    1095
dtype: int64
```

Figure 3: Columns with missing values

Since the columns "Timestamp", "state", "comments" were not significant. they were dropped from the data frame. The "self_employed" column's missing values (1.4%) were replaced with the mode which was "No". The "work_interfere" column's missing values (20%) were replaced with "Unknown". The "Age" column contained some outliers. In a tech industry, it is not possible that there are people with age less than 18 or greater than 100. So, the values less than 18 were replaced with 18 and the values greater than 75 were replaced with 75.

It seems the gender section of the survey was a text field because the “Gender” column contained values like “male”, “M”, “Make”, “femail”, “All” etc. So, the genders similar to the male gender were replaced with “Male”, the gender similar to female were replaced with “Female” and rest of the genders were replaced with “Other”.

5.3 Data Visualization

Using Python, we can visualize datasets to gain insights or visualize the data to find patterns or answers. The data visualization was done using Matplotlib and Seaborn libraries.

In data visualization we have made plots for the relevant variables to gather insights and understand the factors which may be affecting mental health of an employee. The plots of variables are also combined with the treatment column to check the relation of the variables affecting the employee to seek treatment.

In the first plot, we can observe that number of responses in the treatment column are almost close. “Yes” (51%) and “No” (49%). The treatment column basically states if an employee has ever sought any treatment.

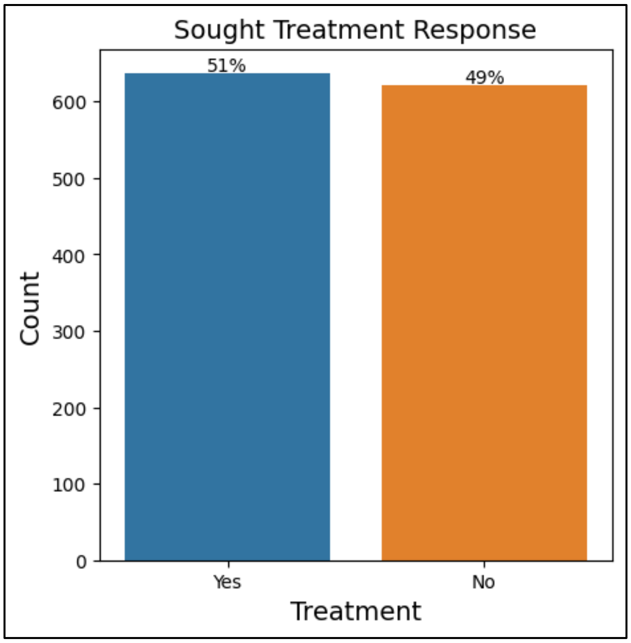


Figure 4: Showing response of employees if they have sought treatment

The gender distribution in the dataset is shown below. It shows the number of males in the dataset are 79% and females are 20% and other genders make up 2% of the dataset.

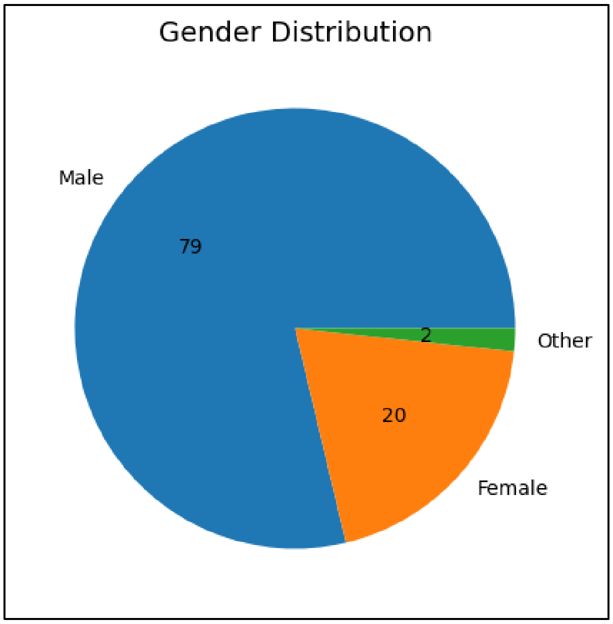


Figure 5: Pie chart showing gender distribution

Here, employees were asked if they were self-employed. 88% said no and 12% said yes which means majority of the employees belong to working class. In the right plot we can observe that treatment percentage is almost similar when relating to self-employment.

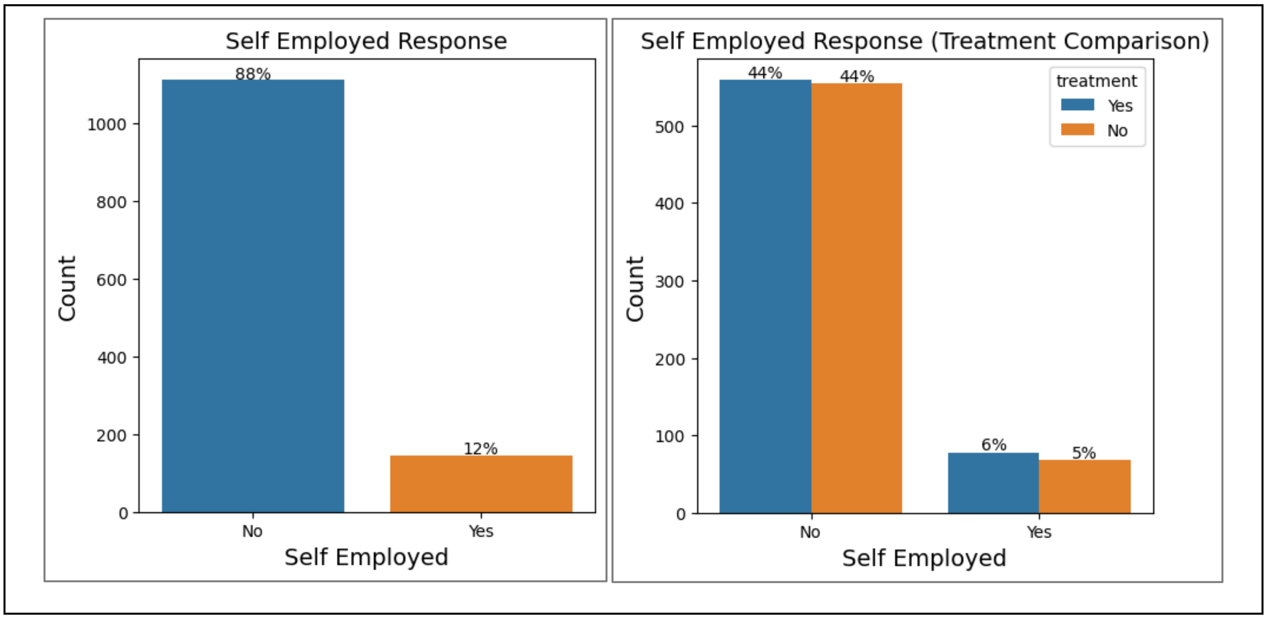


Figure 6: Left plot shows self-employment response and right plot shows self-employment response with seeking treatment

The employees were asked whether they have a family history of mental illness 61% said no and 39% said yes.
In the bottom plot we can observe that 29% of the employees have sought treatment who have a family history of mental illness while 10% have not.

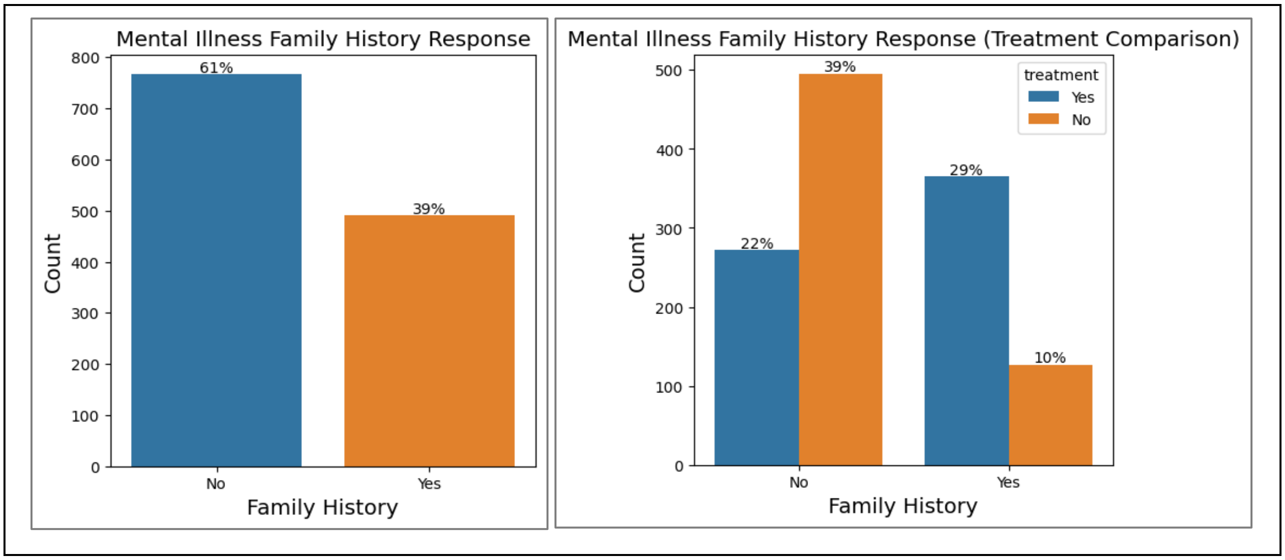


Figure 7: Left plot shows response to family history of mental illness and right plot shows family history with seeking treatment

Employees were asked if they have a mental health condition does it interfere with their work. 37% of the employees said “Sometimes” which is not a clear answer but looking at the work interference with seeking treatment category it shows that employees who said “Sometimes” have the highest amount of seeking treatment i.e., 28%.

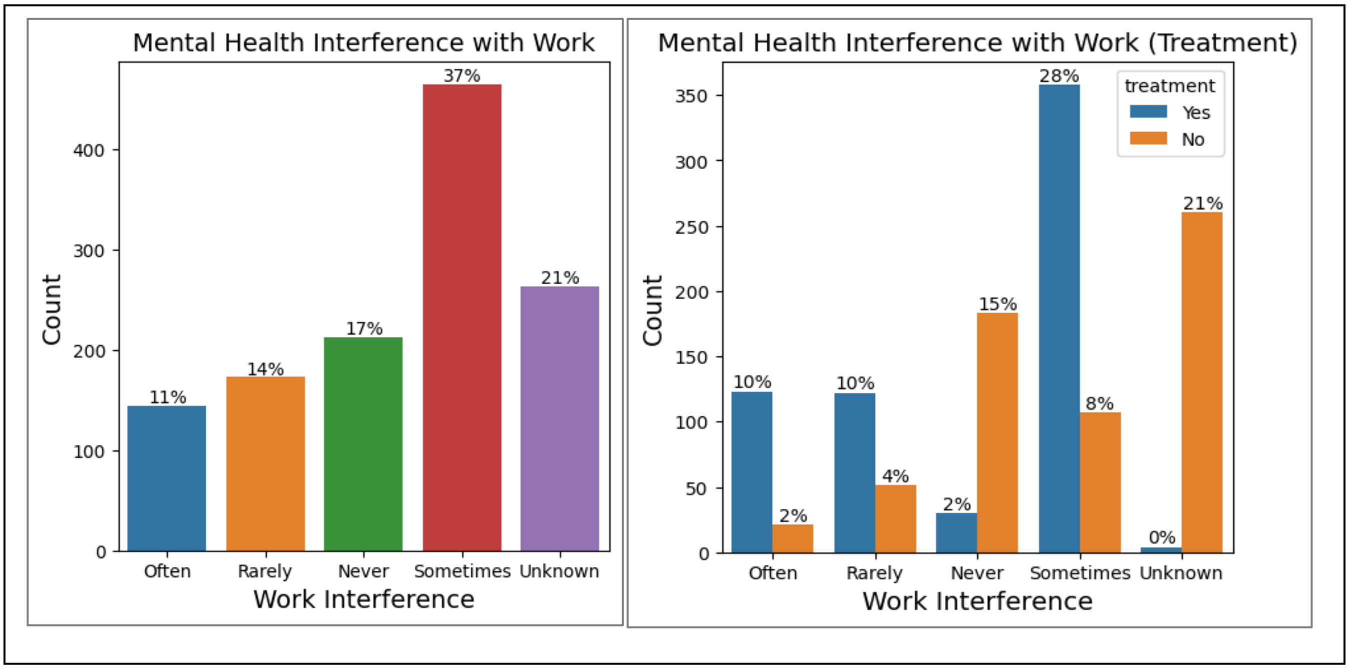


Figure 8: Left plot shows response to if mental health condition interferes with work and right plot shows work interference with seeking treatment

Here, employees were asked if they work remotely 50% of the time. 70% employees said no. When looking at remote work response with seeking treatment, there is not much insight as both categories seem close.

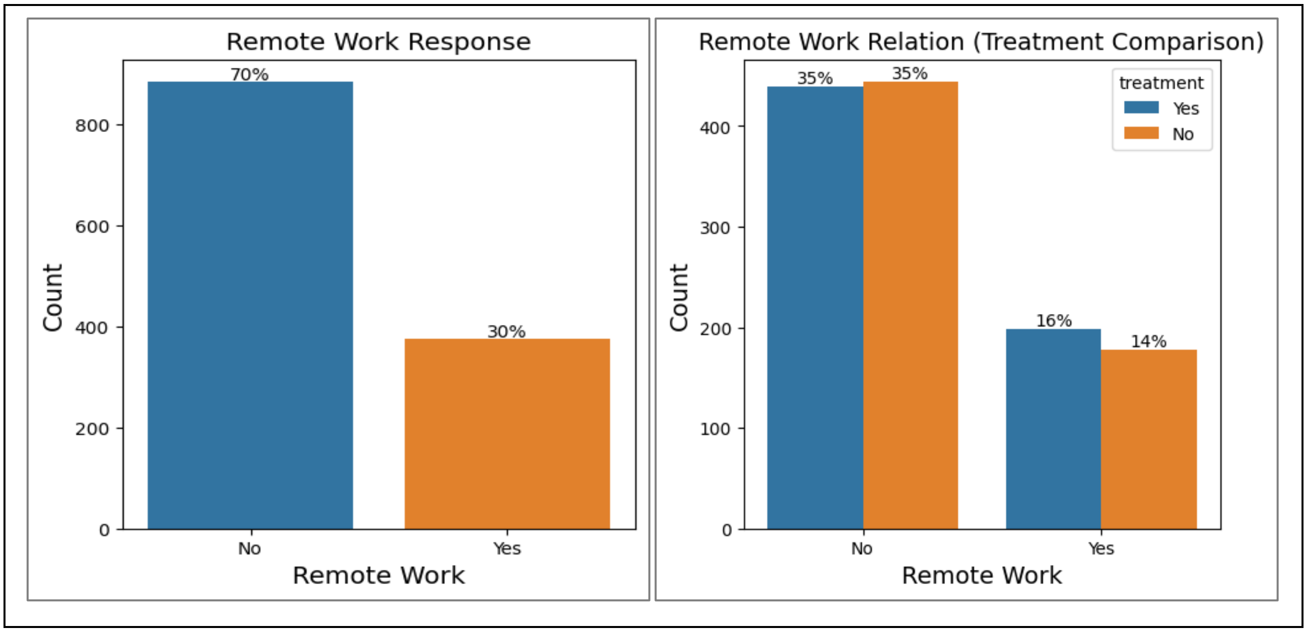


Figure 9: Left plot shows response of employee if they work remotely and right plot shows their response with seeking treatment

Employees were asked if their company provides them mental health benefits. 38% employees said yes, if we look at benefits provided along with seeking treatment, we can observe that the majority has said yes to seeking treatment which means that the employees who have sought treatment also get benefits from their company.

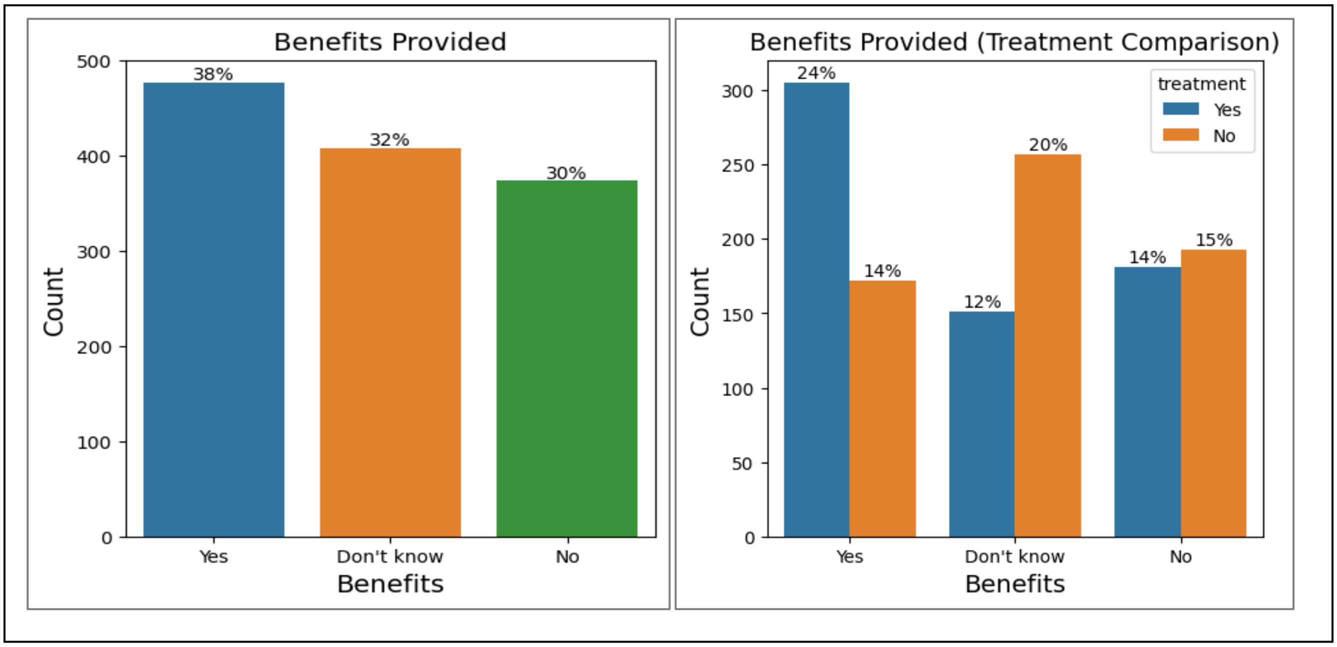


Figure 10: Left plot shows employees response if company provides mental health benefits and right plot shows benefits response along with seeking treatment

We can see that 3 categories stand out when we look at the employees' company size. 23% of employees belong to a company where their workforces are "6-25", "26-100" and 22% of the employees belong to companies where total workforce is "More than 1000".

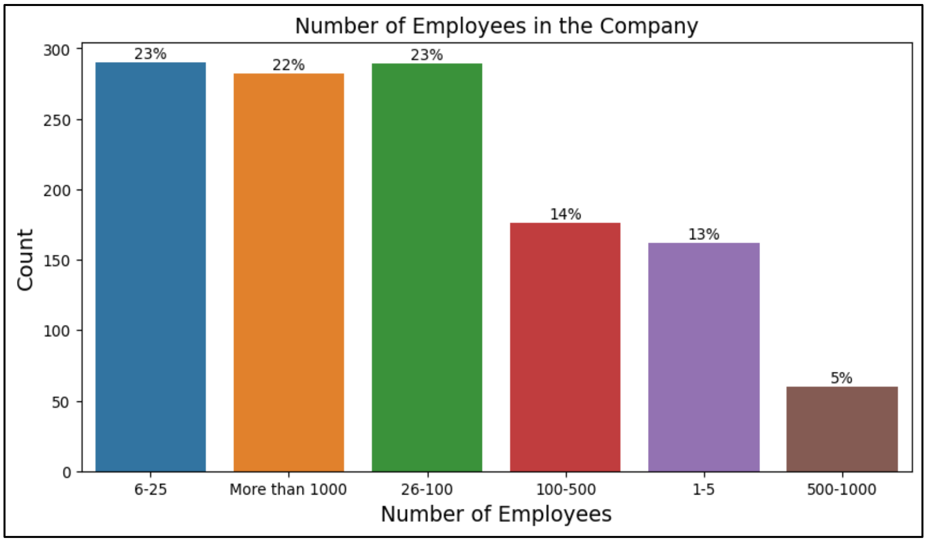


Figure 11: Plot shows number of employees in the company

Employees were asked if leave is easily provided for a medical condition or mental health condition. The majority, 45% have responded “Don’t know” and 21% have responded “Somewhat easy”. When we look at the treatment comparison, we can observe that people who responded with “Somewhat difficult” or “Very difficult” have high number of seeking treatment.

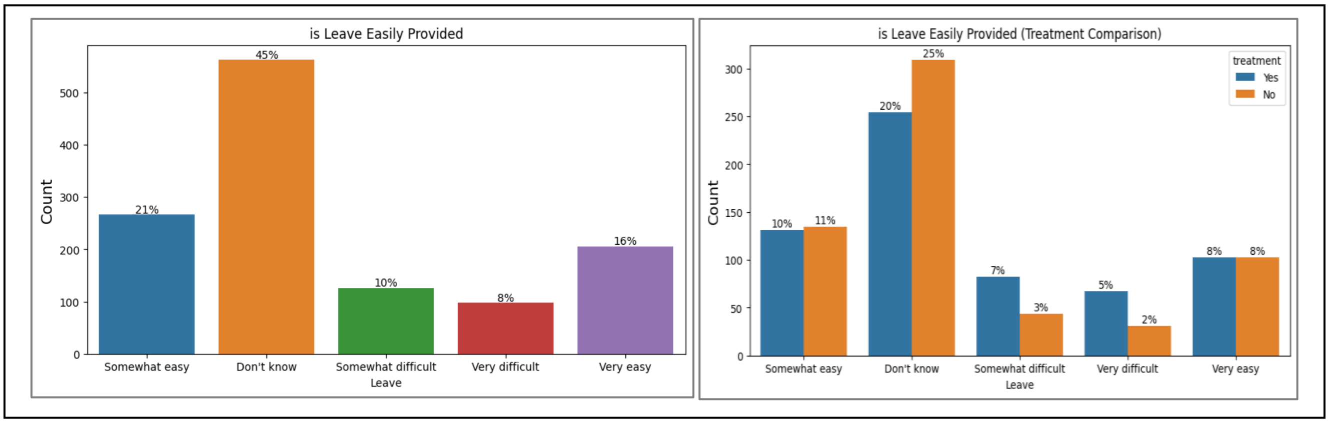


Figure 12: Left plot shows employee response to leave provided by company and right plot shows leave provided along with seeking treatment

Employees were asked if discussing mental health issues with employer has negative consequences. 39% said no and 38% said maybe. If we look in comparison with seeking treatment, we can see that out of the employees who said “Maybe” have high seeking treatment count.

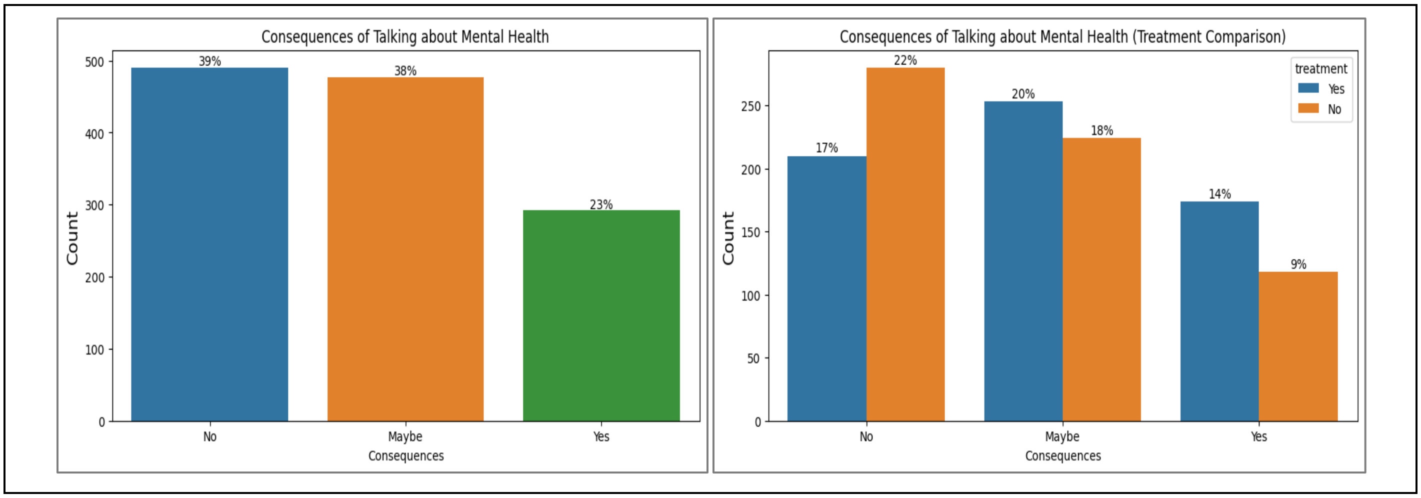


Figure 13: Left plot shows consequences of talking about mental health and right plot shows consequences along with seeking treatment

The employees were asked if they knew about any care options for mental health provided by their company. 40% said No and 35% said Yes but when looking at care options comparison with treatment variable we can see that companies that have care options have more employees seeking treatment and where company does not offer care options employees are also not getting treatment.

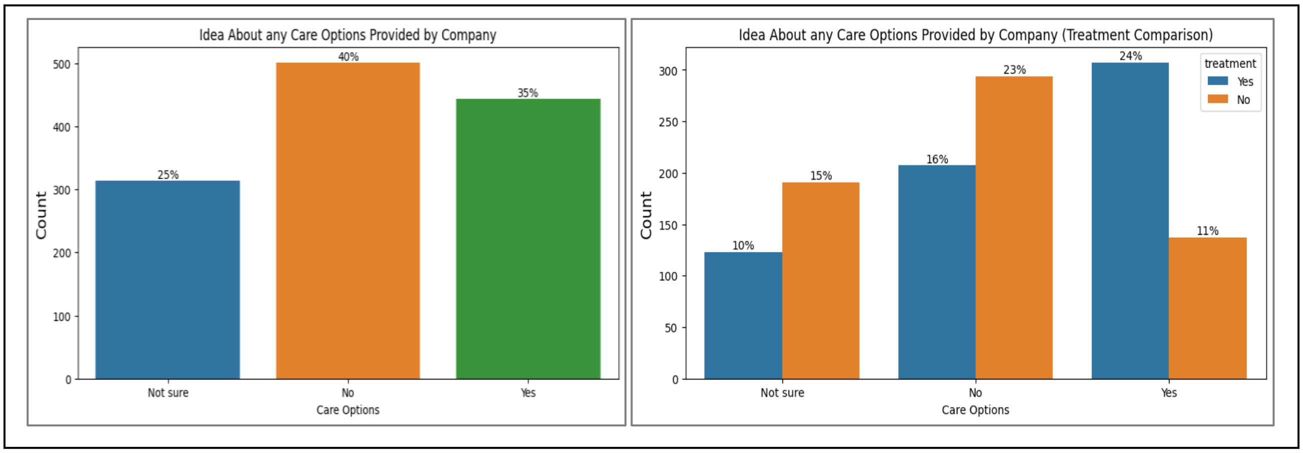


Figure 14: Left plot shows care options provided by company and right plot shows care options provided by company along with seeking treatment

5.4 Model Training & Prediction

In machine learning we have classification models which are basically trained on the target data that contains categorical values. Classification models are used for supervised machine learning problems, supervised machine learning is the method of machine learning which uses labeled datasets to predict outcomes. In the dataset majority of the data is categorical so we have used 3 classification models on the dataset where “treatment” is the target column for prediction. Prediction method is to predict whether an employee has sought treatment. The model’s performance is evaluated based on the classification metrics such as accuracy, precision, recall and f1-score. Before training the models the columns with object data type were encoded using

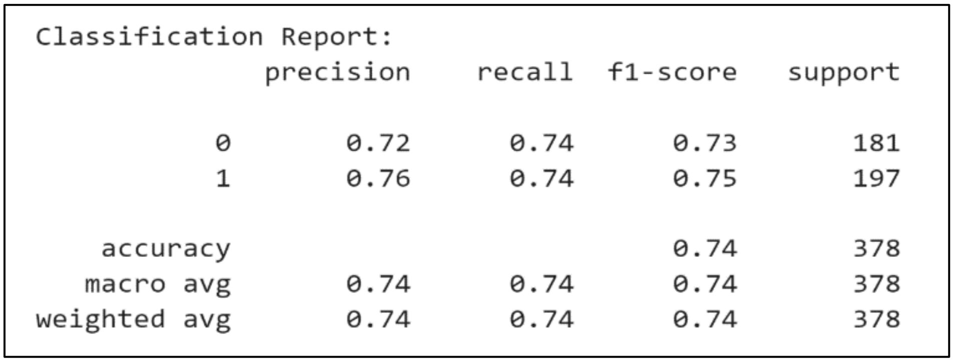
Label Encoder and “train_test_split” was used to split the dataset into training and testing batches, where the test size was kept 30% and 70% was reserved for training. The models used for training are:

- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boost Classifier

These models are imported using the scikit-learn library.

1. Decision Tree Classifier

Decision Tree Classifier was able to achieve 74% accuracy, precision for 0 class is 72% and 76% for 1 class. Recall for 0 and 1 class is 74%. The classification report is provided below:

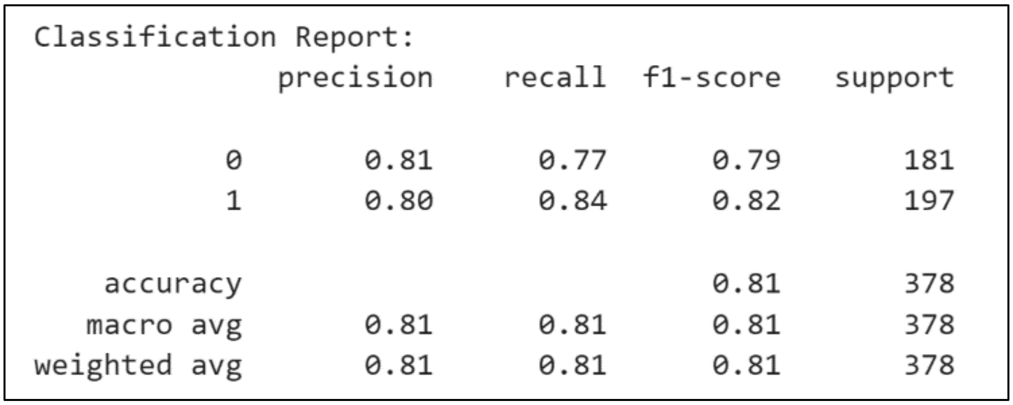


```
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.74      0.73       181
           1       0.76      0.74      0.75       197

    accuracy                           0.74       378
   macro avg       0.74      0.74      0.74       378
weighted avg       0.74      0.74      0.74       378
```

Figure 15: Classification Report of Decision Tree Classifier

2. Random Forest Classifier

Random Forest Classifier was able to achieve 81% accuracy. Precision for 0 class is 81% and 80% for 1 class. Recall for 0 class is 79% and 82% for 1 class. The classification report is provided below:

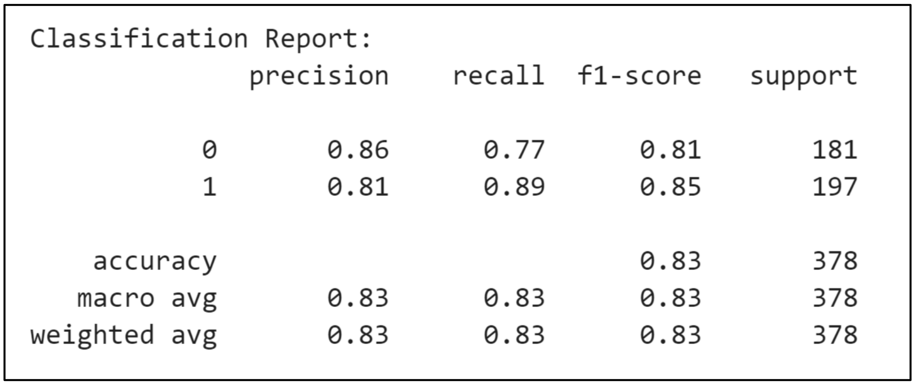


```
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.77      0.79       181
           1       0.80      0.84      0.82       197

    accuracy                           0.81       378
   macro avg       0.81      0.81      0.81       378
weighted avg       0.81      0.81      0.81       378
```

Figure 16: Classification Report of Random Forest Classifier

3. Gradient Boost Classifier

Gradient Boost Classifier was able to achieve the highest accuracy out of the two others i.e., 83%. Precision for 0 class is 86% and 81% for 1 class. Recall for 0 class is 77% and 89% for 1 class. The classification report is provided below:



```
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.77      0.81       181
           1       0.81      0.89      0.85       197

    accuracy                           0.83       378
   macro avg       0.83      0.83      0.83       378
weighted avg       0.83      0.83      0.83       378
```

Figure 17: Classification Report of Gradient Boost Classifier

5.5 Power BI Dashboard

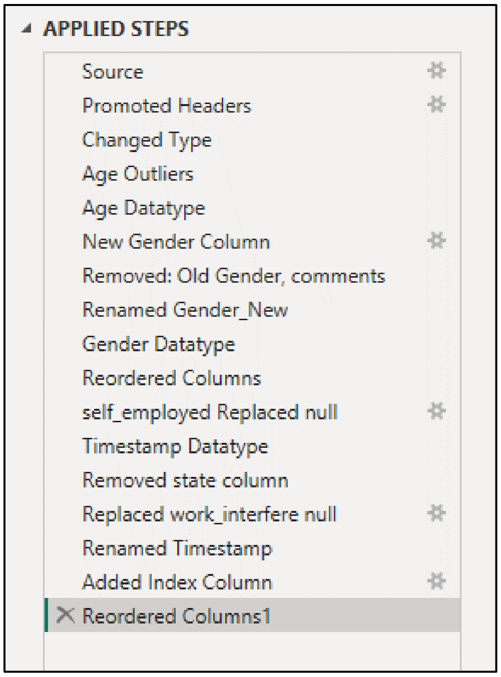
Power BI is an interactive data visualization tool developed by Microsoft to make report/dashboards using datasets.

We have made an interactive Power BI dashboard to further gain insights and visualize the dataset. Similar to the data cleaning before we have transformed the dataset and then made the visuals using the tools provided.

1. Data Transformation

Similar to the data cleaning done using Python we have transformed the dataset before visualizing except that we kept the “Timestamp” column to get the years data to visualize the data year wise as well. The dataset was first imported as usual and then transformation was done using Power BI transform tools. For “Age” and “Gender” columns power queries were used to replace the values in that column because “Age” contained outliers that required conditional replacement and “Gender column had to be replaced conditionally as well.

Applied steps are provided below:



APPLIED STEPS

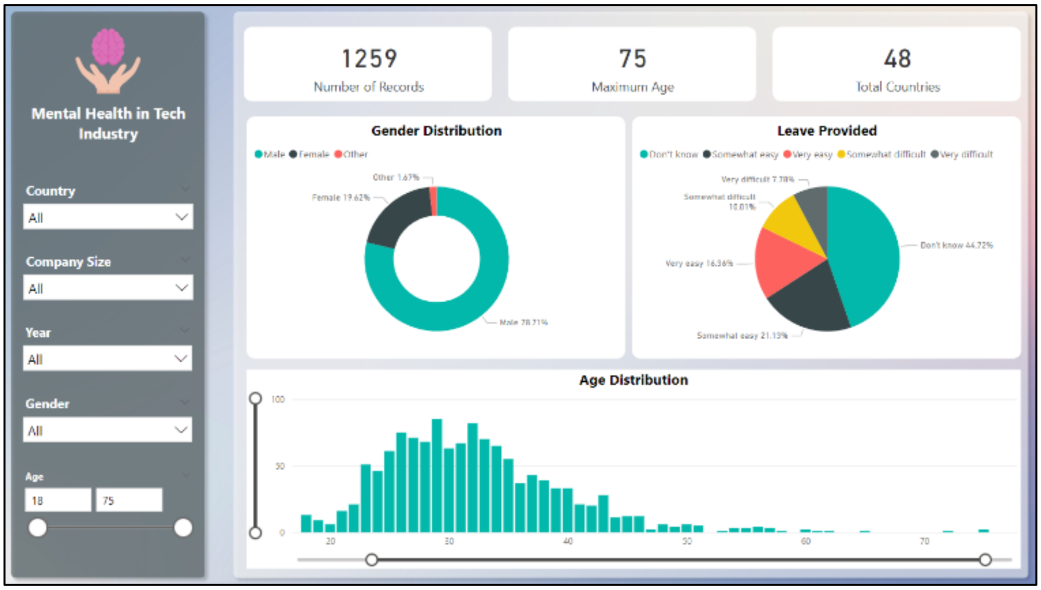
1. Source
2. Promoted Headers
3. Changed Type
4. Age Outliers
5. Age Datatype
6. New Gender Column
7. Removed: Old Gender, comments
8. Renamed Gender_New
9. Gender Datatype
10. Reordered Columns
11. self_employed Replaced null
12. Timestamp Datatype
13. Removed state column
14. Replaced work_interfere null
15. Renamed Timestamp
16. Added Index Column
17. Reordered Columns1

Figure 18: Steps applied to transform the dataset

2. Dashboard Design

In the dashboard design, left side has been made as a filtering option to filter using variables that are Country, Year, No of Employees (Company Size), Gender and Age. The right side provides charts for visualization. The dashboard contains a total of 6 pages.

The first page provides overview of the data, showing number of records, maximum age, total countries, gender distribution, age distribution and leave provided chart. The left menu can be used to filter the data.



Mental Health in Tech Industry

Country: All

Company Size: All

Year: All

Gender: All

Age: 18 – 75

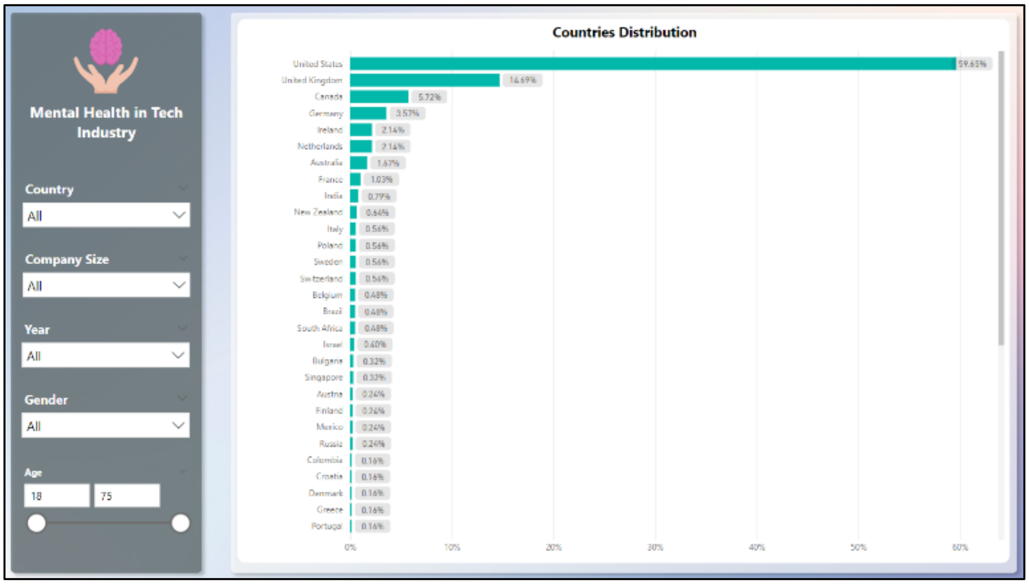
1259 Number of Records

75 Maximum Age

48 Total Countries

Figure 19: Power BI dashboard design (page 1)

The second page shows country distribution. From this bar chart we can see that majority of the survey respondents were from United States.



Mental Health in Tech Industry

Country: All

Company Size: All

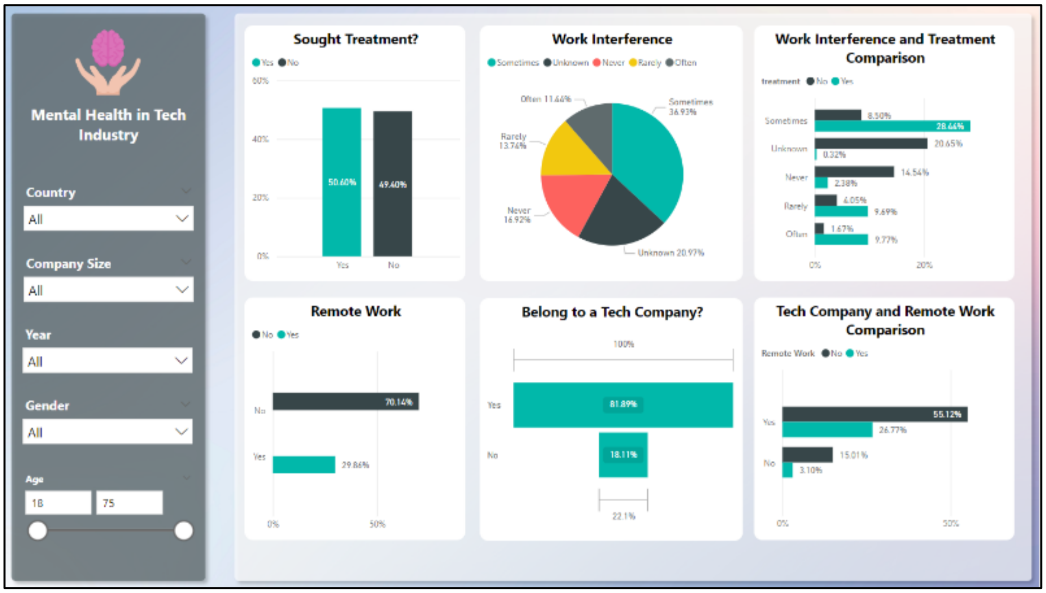
Year: All

Gender: All

Age: 18 – 75

Figure 20: Power BI dashboard design (page 2)

The third page shows treatment, work interference comparison and remote work, tech company comparison.



Mental Health in Tech Industry

Country: All

Company Size: All

Year: All

Gender: All

Age: 18 – 75

Figure 21: Power BI dashboard design (page 3)

The fourth page shows self-employment, work interference comparison and number of employees in company information.

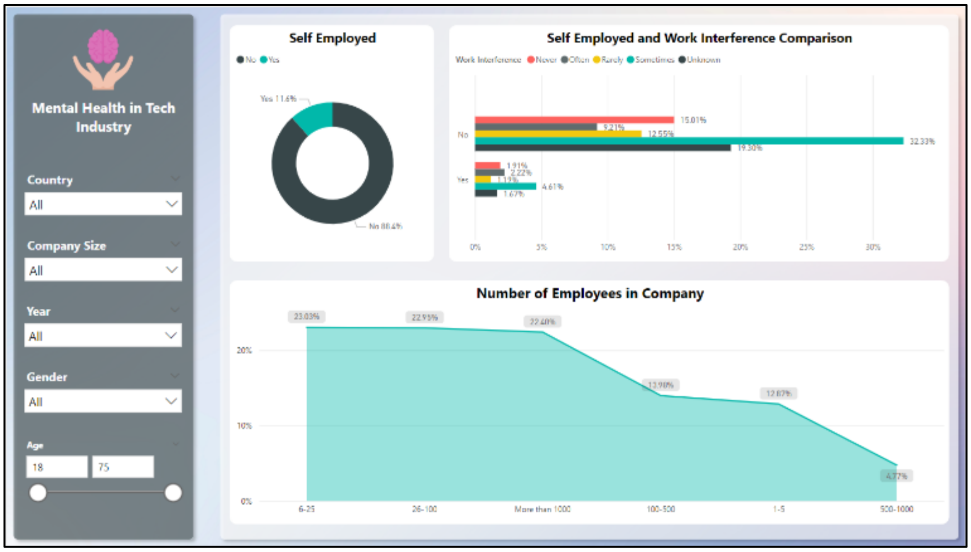


Figure 22: Power BI dashboard design (page 4)

The fifth page shows family history, mental health treatment comparison and discussing mental health with coworkers, treatment comparison information.

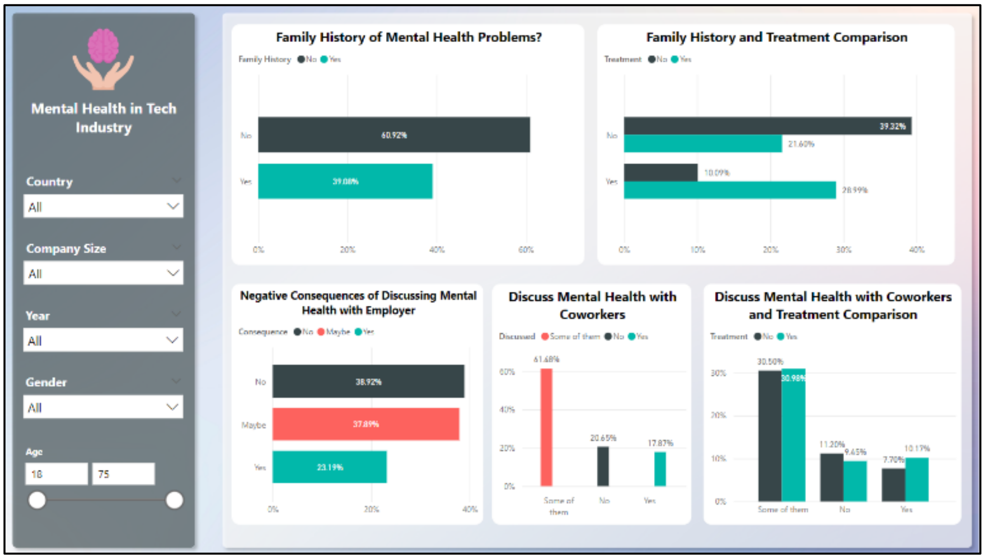
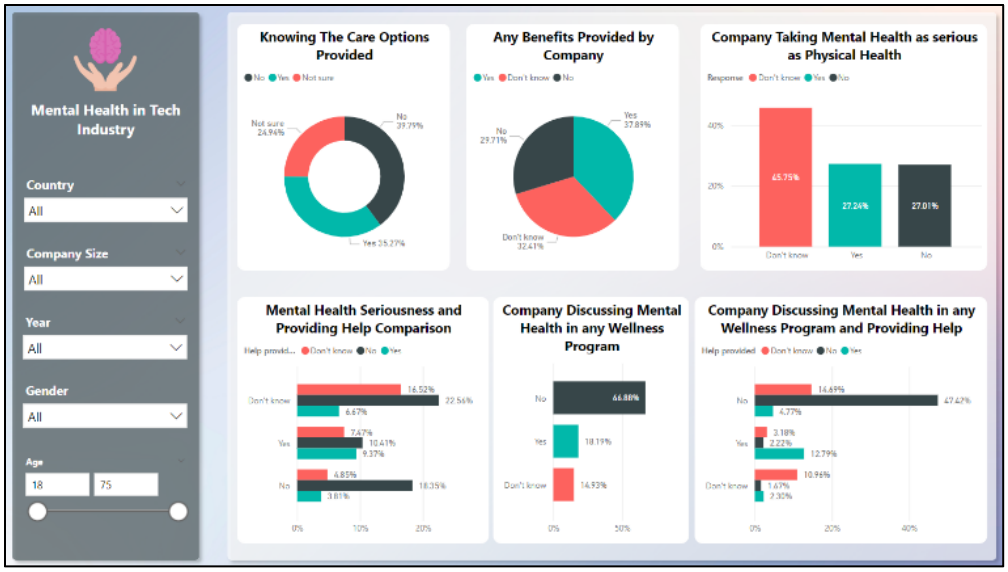

Mental Health in Tech Industry
Country
All
Company Size
All
Year
All
Gender
All
Age
18
75
Family History of Mental Health Problems?
Family History ● No ● Yes
No
Yes
60.92%
39.08%
0%
20%
40%
60%
Family History and Treatment Comparison
Treatment ● No ● Yes
No
Yes
39.32%
21.60%
10.09%
28.99%
0%
10%
20%
30%
40%
Negative Consequences of Discussing Mental Health with Employer
Consequence ● No ● Maybe ● Yes
No
Maybe
Yes
38.92%
37.89%
23.19%
0%
20%
40%
Discuss Mental Health with Coworkers
Discussed ● Some of them ● No ● Yes
61.68%
20.65%
17.87%
60%
40%
20%
0%
Some of them
No
Yes
Discuss Mental Health with Coworkers and Treatment Comparison
Treatment ● No ● Yes
30.50%
30.98%
11.20%
9.45%
7.70%
10.17%
30%
20%
10%
0%
Some of them
No
Yes

Figure 23: Power BI dashboard design (page 5)

The sixth page shows awareness of the company regarding mental health condition. The charts provided are insights for care options provided by company, benefits provided by company, mental health seriousness and providing help comparison.


Mental Health in Tech Industry
Knowing The Care Options Provided
● No ● Yes ● Not sure
Not sure 24.94%
No 39.79%
Yes 35.27%
Any Benefits Provided by Company
● Yes ● Don't know ● No
Yes 37.89%
No 29.71%
Don't know 32.41%
Company Taking Mental Health as serious as Physical Health
Response ● Don't know ● Yes ● No
40%
20%
0%
45.75%
27.24%
27.01%
Don't know
Yes
No
Mental Health Seriousness and Providing Help Comparison
Help provid... ● Don't know ● No ● Yes
Don't know
16.52%
22.56%
6.67%
Yes
7.47%
10.41%
9.37%
No
4.85%
18.35%
3.81%
0%
10%
20%
Company Discussing Mental Health in any Wellness Program
No
66.88%
Yes
18.19%
Don't know
14.93%
0%
50%
Company Discussing Mental Health in any Wellness Program and Providing Help
Help provided ● Don't know ● No ● Yes
No
14.69%
47.42%
4.77%
Yes
3.18%
2.22%
12.79%
Don't know
10.96%
1.67%
2.30%
0%
20%
40%

Country: All

Company Size: All

Year: All

Gender: All

Age: 18 – 75

Figure 24: Power BI dashboard design (page 6)

6. Results

In the data visualization we can see that family history, care options, benefits and work interference variables contributed more towards treatment variable. For machine learning we trained 3 classifier models on the dataset. Results for the models can be found below:

Model	Accuracy
Decision Tree Classifier	74%
Random Forest Classifier	81%
Gradient Boost Classifier	83%

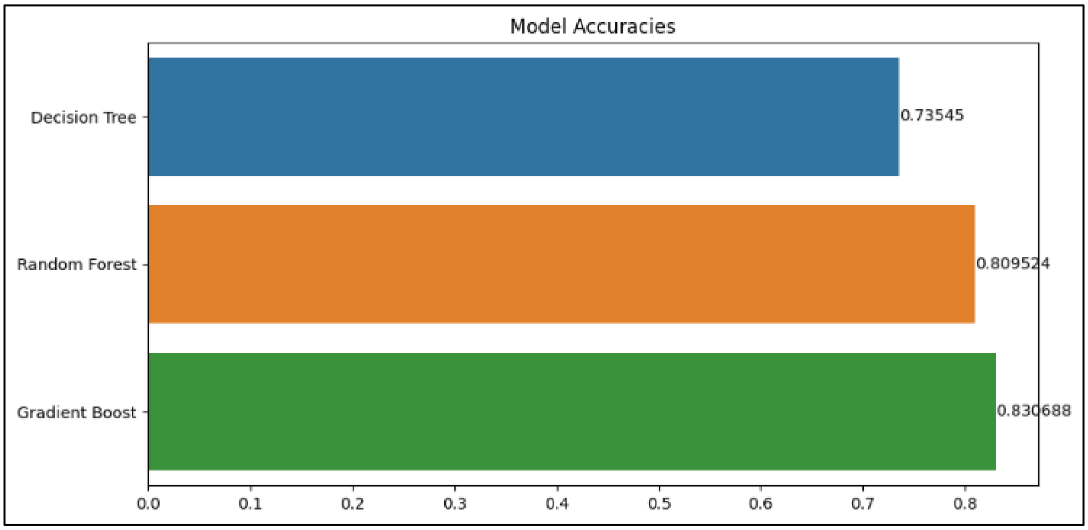


Figure 25: Accuracy of Models

Confusion matrix of Decision Tree Classifier provided following count information of 0 and 1 class:

- True Positive: 134
- False Positive: 47
- False Negative: 52
- True Negative: 145

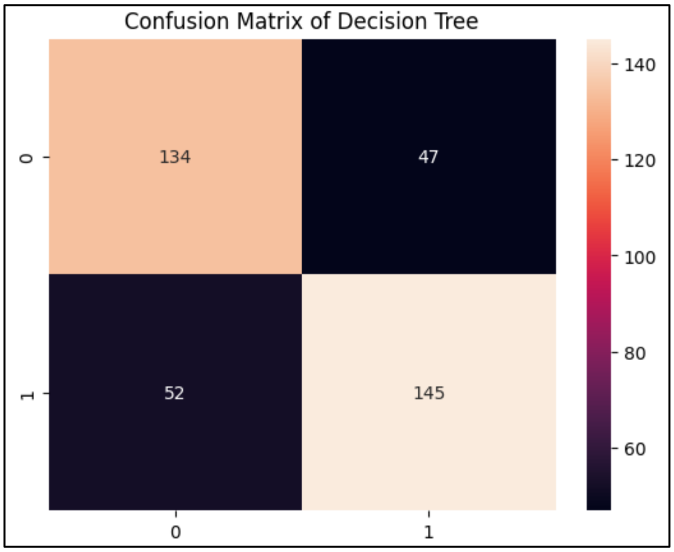


Figure 26: Confusion Matrix of Decision Tree Classifier

Confusion matrix of Random Forest Classifier provided following count information of 0 and 1 class:

- True Positive: 140
- False Positive: 41
- False Negative: 32
- True Negative: 165

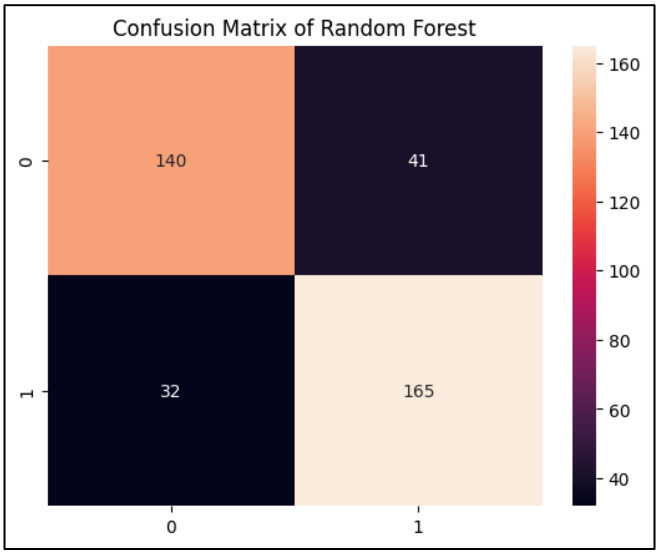


Figure 27: Confusion Matrix of Random Forest Classifier

Confusion matrix of Gradient Boost Classifier provided following count information of 0 and 1 class:

- True Positive: 139
- False Positive: 42
- False Negative: 22
- True Negative: 175

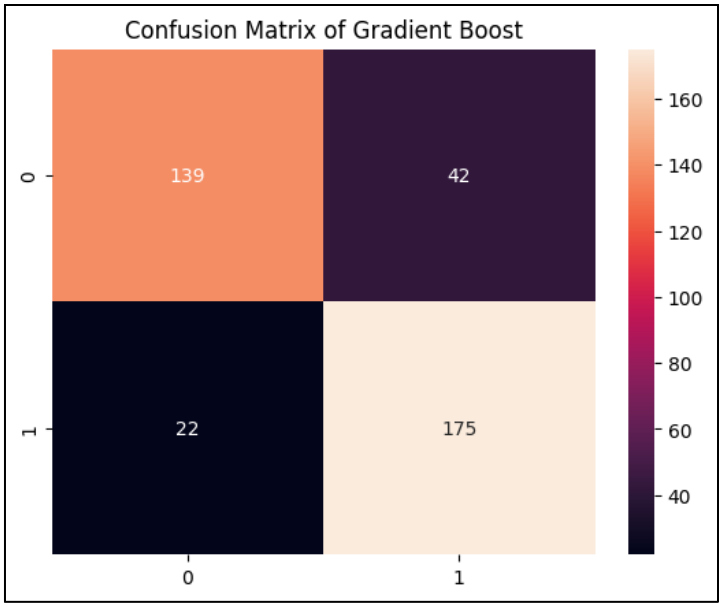


Figure 28: Confusion Matrix of Gradient Boost Classifier

7. Conclusion

Mental health is an important issue in the tech industry, and while awareness of mental health is increasing, the impact of these efforts may be limited as there is still a need to further support and provide resources, especially in the tech industry where employees face many challenges like long working hours, high pressure and layoffs. To address issues like these, it is necessary for tech industry companies to prioritize mental health issues and provide a supportive working environment. An employee's performance in their workplace and well-being are closely tied to their mental health. OSMI (Open Sourcing Mental Health), a non-profit organization provides help in the tech industry to improve mental wellness. By utilizing their dataset in this project, we found that more awareness should be made to address these issues and provide assistance to employees who are facing many challenges due to bad mental health.

8. References

[1] “Relationship between Employee Mental Health and Job Performance: Mediation Role of Innovative Behavior and Work Engagement - PMC.” Accessed: Dec. 24, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9180763/

[2] M. T. Ford, C. P. Cerasoli, J. A. Higgins, and A. L. Decesare, “Relationships between psychological, physical, and behavioural health and work performance: A review and meta-analysis,” Work & Stress, vol. 25, no. 3, pp. 185–204, Jul. 2011, doi: 10.1080/02678373.2011.609035.

[3] K. Memish, A. Martin, L. Bartlett, S. Dawkins, and K. Sanderson, “Workplace mental health: An international review of guidelines,” Preventive Medicine, vol. 101, pp. 213–222, Aug. 2017, doi: 10.1016/j.ypmed.2017.03.017.

[4] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, “Machine learning in mental health: a scoping review of methods and applications,” Psychological Medicine, vol. 49, no. 9, pp. 1426–1448, Jul. 2019, doi: 10.1017/S0033291719000151.

[5] R. Katarya and S. Maan, “Predicting Mental health disorders using Machine Learning for employees in technical and non-technical companies,” in 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE), Dec. 2020, pp. 1–5. doi: 10.1109/ICADEE51157.2020.9368923.