Mind Matters - Mental Health in Tech Industry

Zainab Hasnain A20516879 | Aditya Sharma A20516668 | Muhammad Ans Qadri A20521573 | Abhishek Jaiswal A20380004

Project Proposal

1.1 Project Description/Research Goal

The project analyzes the responses from an annual survey that measures attitudes towards mental health and the frequency of mental health disorders in the tech workplace. The dataset contains different attributes of an employee working in a tech company, such as age, gender, family history, location, etc. Moreover, it has certain columns related to the work culture of the organization each employee is working in, for example, the size of a company, whether it is remote work, benefits the company offers, etc. Our goal here is to predict how inclined each employee is to seek for mental health treatment based on the above-mentioned features. The project will be a combination of predictive and inferential analysis with the use of various statistical methods studied in this class.

The set of questions that the project seeks to address

i. Is there a relationship between an employee's working conditions and his inclination to seek for help/treatment related to mental health?

ii. To what extent do work conditions impact mental health?

iii. Are there other factors, such as age and family history, which could help explain the correlation between workplace stress and mental health?

iv. What is the correlation between responses to questions about remote work, leave policies, and anonymity with whether individuals have sought treatment for a mental health condition?

v. Which of the personal features of a person, such as age, gender, location, affect an employee's mental health the most?

vi. Are there any interaction terms, meaning, is there a combination of features which affect an employee's mental health? How can we incorporate these complex relationships in our model?

vii. Can we determine the strongest predictors of mental health illness in the workplace using our statistical analysis and inferential summary?

viii. What is our observation of how the frequency of mental health illness and attitudes towards mental health vary by geographic location, company size, remote work, family history, age, gender, and other features?

ix. How accurate is a predictive model that can tell whether an employee will seek mental health treatment?

1.2 Proposed Methodology

I. **Data Augmentation**

- We have responses of the OSMI survey conducted in 2014. We will use those responses for this project as the baseline dataset.

- We also have access to the survey responses conducted in subsequent years. We aim to perform the same analysis on data for each year and compare our results.
- Finally, we can create new datasets (pre-covid and post-covid) by combining the yearly datasets. This will help us understand the pattern of mental health implications of workers in the tech industry.

II. **Data Preprocessing and Cleaning**

- Ensuring that each data point within the dataset has a recognizable data type.
- Handling null values within the dataset which can be replaced with a fixed value or depending on the data, by mean, mode, or median.
- Imputation: replacing the missing values with a similar approach for null values.
- Omitting any duplicate and invalid responses.
- Removing any redundant columns to reduce collinearity/dependency in the model.
- Merging and quantifying the data in categorical columns to create a more understandable and simpler dataset.
- Converting categorical variables to numerical variables that can be used in the analysis.
- Hot encoding a few variables for better prediction.
- Outlier detection and removal.

III. **EDA/Data Visualization**

- Using correlation plots to determine the relationship between various factors like country, benefits, remote work, etc.
- Drawing colored histograms by label (seek treatment) with the categorical/ variables to observe any trend in class pertaining to a certain group of people.
- Visualizing our EDA results using plotting libraries like ggplot to interpret findings.
- Using box plots/histograms/heatmaps/bar charts wherever appropriate to map relations between predictors (country, remote work, working in tech, etc) with the response (seek treatment or not).

IV. **Predictive Analysis**

- *Feature Selection*
  To avoid the curse of dimensionality, we will perform feature selection on **27** features in our dataset, selecting the most relevant features that contribute the most to the target. We will use methods like Chi-Squared Test, PCA, Lasso (or Ridge) Regression, etc to establish whether a feature is statistically significant.

  Since we are dealing with a categorical response variable (seek treatment (yes/no)), for our predictive analysis, we will rely on classification models primarily.
- *Classification Analysis*
  Use predictive modeling on selected features and compare their performance. Since we have a categorical target variable, we will focus on classification methods such as **Logistic Regression, LDA, Lasso, Ridge, Decision Trees,** and over time, other models such as ensemble methods, etc.

1.3 Performance Metrics

To validate the outcomes as we work to create a predictive classifier model, we can use the following metrics:

1. Confusion Matrix.

2. Precision, Recall, and F1-Score.

3. The area under the curve and ROC.

4. For feature selection, we can rely on a p-test and a 95% confidence Interval to remove irrelevant columns. Moreover, we can do a correlation analysis to remove dependent/redundant columns.

5. To validate our results, we will perform cross validation and select the best model.

Project Outline

2.1 Literature Review and Related Work

Research has shown that mental health issues are prevalent in the tech industry. They utilized an online survey to collect data from tech workers. The survey consisted of 27 questions that assessed respondents' experiences with mental health issues, workplace attitudes towards mental health, and experiences with seeking treatment. It received 1,135 responses from tech workers across the world and found that 56.4% of tech workers have experienced symptoms of a mental health disorder. The survey also found that only 17% of respondents were comfortable discussing their mental health issues with their employer.

Efforts have been made to address mental health issues in the tech industry. A survey by Quantic-Data, which consisted of 24 questions that assessed respondents' experiences with mental health issues, as well as their perceptions of mental health support in the workplace. The survey received over 5,000 responses from tech workers across the world and found that over 70% of tech workers believed that their employers were taking steps to support mental health in the workplace. Companies such as Google and Microsoft have also implemented mental health programs for their employees.

The International Journal of Industrial Engineering and Operations Management study (2022) utilized a survey to collect data from tech workers. The survey consisted of questions that assessed respondents' mental health status, work-related stress, productivity levels, and absenteeism rates. The survey received responses from 191 tech workers and found that employees with poor mental health had lower productivity levels and higher absenteeism rates. The study also found that mental health issues were a leading cause of turnover in the tech industry.

A study by BairesDev found that tech workers are particularly susceptible to burnout, which can have a negative impact on mental health. The study found that a significant number of tech workers experience high levels of stress due to tight deadlines, long working hours, and job insecurity. Several studies have highlighted the impact of mental health on employee productivity in the tech industry. A study by the International Journal of Industrial Engineering and Operations Management found that employees with poor mental health had lower productivity levels and higher absenteeism rates. The study also found that mental health issues were a leading cause of turnover in the tech industry.

However, despite these efforts, there is still much work to be done to address mental health issues in the tech industry. A study by LinkedIn found that only 30% of tech workers believed that their employers were doing enough to support mental health in the workplace.

The following works have been done relevant to our project:

"Analysis of the Impact of Mental Health on Employee Productivity and Work Attitudes in the Technology Industry" by Nadia U. Syed and Khadija U. Syed: This research paper uses the "Mental Health in Tech Survey" dataset to analyze the impact of mental health issues on employee productivity and work attitudes in the technology industry. The paper includes descriptive statistics, correlation analysis, and regression analysis.

"A Comparative Analysis of Machine Learning Algorithms for Predicting Mental Health Disorders in the Tech Industry" by Adarsh Mohata and Pratiksha Sutar: This research paper uses the "Mental Health in Tech Survey" dataset to compare the performance of different machine learning algorithms for predicting mental health disorders in the tech industry. The paper includes data cleaning and preparation, feature selection, and model training and evaluation.

"Mental Health in Tech: Exploring the Relationship Between Work Stress, Job Satisfaction, and Burnout" by Rhea Saini and Nirupma Singh: This research paper uses the "Mental Health in Tech Survey" dataset to explore the relationship between work stress, job satisfaction, and burnout in the tech industry. The paper includes descriptive statistics, correlation analysis, and regression analysis.

"Mental Health and Job Performance in the IT Industry: A Comparative Study of India and the United States" by Anuradha V. Daptardar, Mahesh S. Raisinghani, and Ajit K. Dasgupta: This research paper uses the "Mental Health in Tech Survey" dataset to compare mental health and job performance in the IT industry in India and the United States. The paper includes descriptive statistics, correlation analysis, and regression analysis.

2.2 Data Source and Reference Data

We are using the OSMI survey responses of 2014 in a csv file format. The survey contains 27 questions, each of which is a column in our dataset. The number of responses is the total number of records in our dataset. We have data for the following years:

Year	**Number of responses**
2014	**1260**
2016	**1570**
2017	**756**
2018	**417**
2019	**352**
2020	**180**
2021	**131**

This dataset contains survey responses from individuals in the tech industry about their mental health, including questions about treatment, workplace resources, and attitudes towards discussing mental health in the workplace. By analyzing this dataset, we can better understand how prevalent mental health issues are among those who work in the tech sector.
This dataset tracks key measures such as age, gender, and country to determine overall prevalence, along with responses surrounding employee access to care options; whether mental health or physical illness are being taken as seriously by employers; whether anonymity is protected with regards to seeking help; and how coworkers may perceive those struggling with mental illness issues such as depression or anxiety.

Data Columns

Other than age, most of the columns of our data contain qualitative (categorical) responses. Moreover, our target variable, Treatment, contains a binary response (yes, no).

- Timestamp: Date and time of survey submission. (DateTime)
- Age: Age of the respondent. (Numeric)
- Gender: Gender of the respondent. (Categorical)
- Country: Country of residence of the respondent. (Categorical)
- state: US state of residence of the respondent. (Categorical)
- `self_employed`: Whether the respondent is self-employed. (Categorical)

- `family_history`: Whether the respondent has a family history of mental health issues. (Categorical)
- `work_interfere`: How much work is interfered with due to mental health issues. (Categorical)
- `no_employees`: Number of employees at the respondent's workplace. (Numeric)
- `remote_work`: Whether the respondent works remotely. (Categorical)
- `tech_company`: Whether the respondent works in a tech company. (Categorical)
- benefits: Whether the respondent's workplace offers mental health benefits. (Categorical)
- `care_options`: Whether the respondent's workplace offers mental health care options. (Categorical)
- `wellness_program`: Whether the respondent's workplace offers a wellness program. (Categorical)
- `seek_help`: Whether the respondent has sought help for a mental health issue. (Categorical)
- anonymity: Whether the respondent's workplace allows for anonymity when seeking help for a mental health issue. (Categorical)
- leave: Whether the respondent's workplace offers leave for mental health issues. (Categorical)
- `mental_health_consequence`: Whether the respondent has experienced negative consequences due to discussing mental health issues at work. (Categorical)
- `phys_health_consequence`: Whether the respondent has experienced negative consequences due to discussing physical health issues at work. (Categorical)
- coworkers: Whether the respondent has discussed mental health issues with coworkers. (Categorical)
- supervisor: Whether the respondent has discussed mental health issues with their supervisor. (Categorical)
- **treatment: Whether the respondent has sought treatment for a mental health issue. (Categorical Target variable)**

2.3 Data processing and pipeline

As discussed previously, we will be using the data for each year and perform analysis on it. That means, we will have to pre-process and clean each year's data separately, according to the level of cleanliness and type of responses. Here are the steps we will take to reprocess the data, divided into the following sections:

I. **Data Augmentation**

- We have responses to the OSMI survey conducted in 2014. We will use those responses for this project as the baseline dataset.
- We also have access to the survey responses conducted in subsequent years. We aim to perform the same analysis on data for each year and compare our results.
- Finally, we can create new datasets (pre-covid and post-covid) by combining the yearly datasets. This will help us understand the pattern of mental health implications of workers in the tech industry.

II. **Data Preprocessing and Cleaning**

- Ensuring that each data point within the dataset has a recognizable data type.

- Handling null values within the dataset which can be replaced with a fixed value or depending on the data, by mean, mode, or median.
- Imputation: replacing the missing values with a similar approach for null values
- Omitting any duplicate and invalid responses.
- Removing any redundant columns to reduce collinearity/dependency in the model.
- Merging and quantifying the data in categorical columns to create a more understandable and simpler dataset.
- Converting categorical variables to numerical variables that can be used in the analysis.
- Hot encoding a few variables for better prediction.
- Outlier detection and removal.

2.4 Data Stylized Facts

Distributional Analysis:

Distributional analysis is a statistical method used to describe and analyze the distribution of a variable or set of variables. We may use it in our project since we need to examine the shape, central tendency, and variability of the data, and identify any significant pattern(s) or outliers in the distribution.

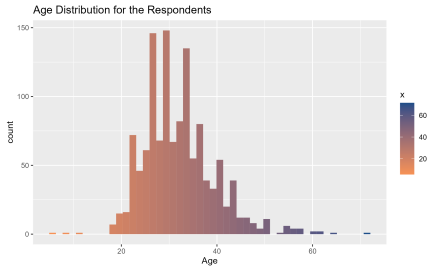

Figure 1: Age Distribution of the Participants

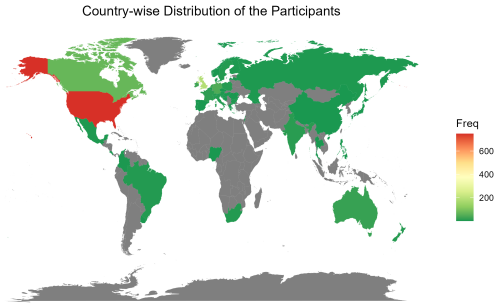

Figure 2: Country Wise Distribution of the Respondents

For example, we can see that the age of most respondents lies between 20-40 and the majority of the respondents are from North America, as compared to other parts of the world. We can further check for skewness in our data by using histograms, box plots, and bar graphs. Moreover, we can also plot colored (according to response variable) bar graphs and scatter plots to see if there is a particular age group/sex or any other group more closely linked to having mental health issues.

Dimensionality Reduction

Dimensionality reduction is a powerful technique that can be used to simplify complex data, improve the performance of machine learning models, and visualize high-dimensional data. As our data set is complex we are using this technique which in turn will help in increasing the performance of the learning models.

Since most of our columns are categorical variables, we will have to introduce dummy variables for our baseline logistic regression. This means, the number of columns will further increase. Therefore, it is imperative for us to remove redundant columns. We can do so by analyzing any dependent columns during EDA. For example, the column ***seek help*** is very similar to the target variable ***treatment***. We can omit or merge these two columns to reduce collinearity in our dataset.

2.5 Model Selection

Feature Selection

In order to reduce the effect of the dimensionality of the data we are doing feature scaling. In this process, we will remove irrelevant and redundant features among others. This in turn will help in improving the performance of machine learning models by reducing overfitting and improving generalization.

Classification Methods

Because we are concentrating on a quantitative response variable (yes/no), we would like to employ classification-based algorithms such as LDA, Logistic Regression, Decision Trees, Naive Bayes, and others for our predictive analysis and further evaluate which approach provides better performance metrics while also understanding the relationship between predictors and response. The algorithms we chose are divided into both 'generative' and 'discriminative' model approaches, having a more in-depth insight regarding which is more reliable.

Baseline Model

For benchmarking and helping us evaluate the performance of different classifiers, we have chosen the ***'Logistic Regression'*** as the baseline because of its simplicity and to attain informative evaluation for a model, and fine-tuning for other models.

2.6 Tools

- **Libraries/Packages:** tibble, DT, knitr, tm, ggplot2, dplyr, fitdistrplus, plotly, plyr.
- **Softwares:** RStudio, R Project Management
- **Source Control:** Kaggle Dataset/Website, GitHub

References

Syed, N. U., Syed, K. U. (2019). Analysis of the Impact of Mental Health on Employee Productivity and Work Attitudes in the Technology Industry. https://ieeexplore.ieee.org/document/8937389

Mohata, A., Sutar, P. (2021). A Comparative Analysis of Machine Learning Algorithms for Predicting Mental Health Disorders in the Tech Industry. https://dl.acm.org/doi/10.1145/3454548.3454582

Saini, R., Singh, N. (2020). Mental Health in Tech: Exploring the Relationship Between Work Stress, Job Satisfaction, and Burnout. https://dl.acm.org/doi/10.1145/3441876.3441895

Daptardar, A. V., Raisinghani, M. S., Dasgupta, A. K. (2018). Mental Health and Job Performance in the IT Industry: A Comparative Study of India and the United States.
https://www.tandfonline.com/doi/abs/10.1080/1097198X.2018.1438897

Supplemental Resources

- 2014 Survey Data: Data World
- Medium - Mental Health in the Tech Industry (OSMI Survey 2014)
- LinkedIn - Unraveling the State of Mental Health in Tech
- UoSD - Mental Health Analysis in Tech Workplace
- BairesDev - Supporting Mental Health in Tech Culture