Investigating Data Mining Techniques on Mental Health Issues

Long Nguyen
Goergen Institute of Data Science
University of Rochester
Rochester, NY
ngkhlong189@gmail.com

***Abstract*—Mental health includes our emotional, psychological, and social well-being. It affects how we think, feel, and act. It also helps determine how we handle stress, relate to others, and make healthy choices. Positive mental wellness allows one to reach his/her full potential in work productivity and life fulfillment. As a result, mental health remains a vital part of our life.**

***Index Terms*—mental health, treatment, work interference, association rules, clustering, and decision trees.**

I. Introduction

In recent years since the COVID-19 pandemic, mental health has become one of the most controversial issues with regard to one's well beings, particularly in the workplace. Statistics show that 60% of employees have been negatively affected by stress accumulated at work, and 87% of people reported actions taken by employers and the company positively impacted their mental well-being [1]. I suspect that this problem has been in existence prior to the outbreak of the pandemic. The prevalence of the viruses simulated such an emotional response to the peak of mental health issues. One of the main takeaways from the pandemic was the possibility of working from home [2]. The fact that this trend allowed employees to spend more time with family and friends, yet also burdened them with domestic and work-related responsibilities along with the concerns of job security. heightened feelings of guilt, insomnia, irritability, sadness, and emotional exhaustion are some of the common symptoms of mental health [3].

In this study, I want to explore what variables are directly correlated to the tendency of developing mental health issues or seeking treatments with a focus on the workforce. Exploring the data, association ruling, and building prediction models such as Decision Trees, Random Forest, Support Vector Machine (SVM), and Neural Networks to investigate the relationship between predictors and whether an employee sought mental health treatment or not.

II. Related work

My study took some of the inspirations of research on text mining and building classification models from ([2], [4]) on forming common patterns between people's interaction with COVID-19 that showed insights on mental health tendencies. Different from the papers' approach to conducting sentimental analysis, I utilized my data mining experience to perform association rulings between 2 predictors that most commonly result in seeking treatments. Additionally, based on previous research discoveries, I decided to design and evaluate specific classification models to obtain results that could complement the studies I referred to.

III. Dataset Overview

A. Data Collection

In this study, I used the data "Mental Health in Tech Survey" collected from Kaggle. It was a 2014 survey that measures attitudes toward mental health and the frequency of mental health disorders in the tech workplace. There are 1259 observations and 27 attributes, and in which I decided "treatment" was my target label.

B. Exploratory Data Analysis

My first approach to mining unique patterns was to explore the data. Here were some of the discoveries I found:

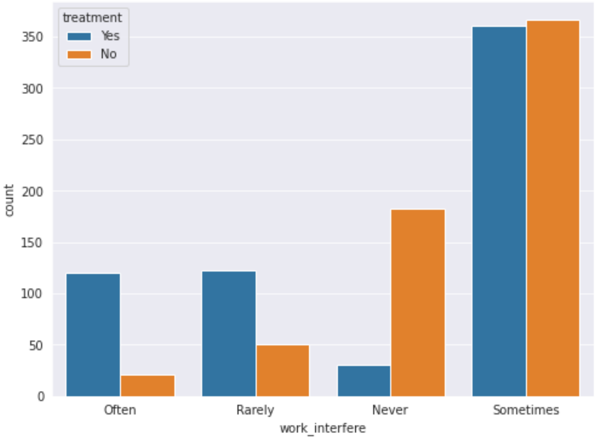


Fig. 1. Histogram of opinions on mental health interferes with work

It was implied that most people who sought treatments experienced a decrease in work productivity. I also investigated whether family history contributed to one's mental health or not:

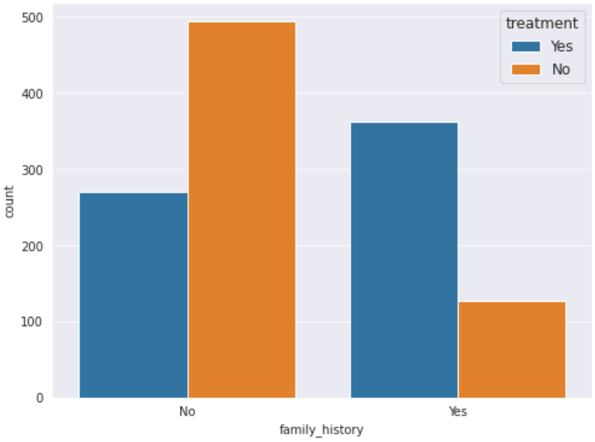


Fig. 2. Histogram of family history with regard to mental health

It was evidenced that people who sought treatment were most likely to have a family history of illness. This could be significant in conducting my research taking into account the theory that mental health was a genetic factor.

Another finding that was very surprising in my research was that genders showed insights into the tendencies of treatment:

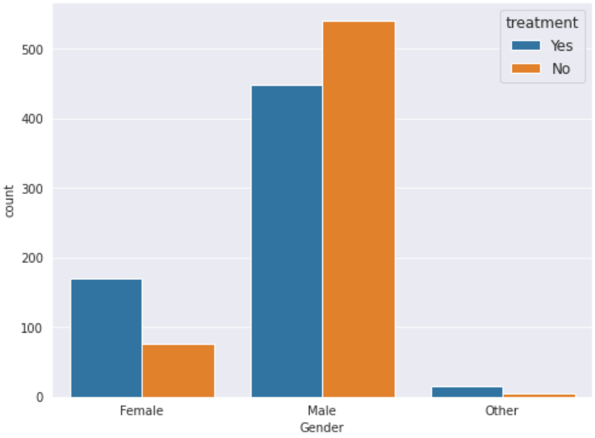


Fig. 3. Histogram of family history with regard to mental health

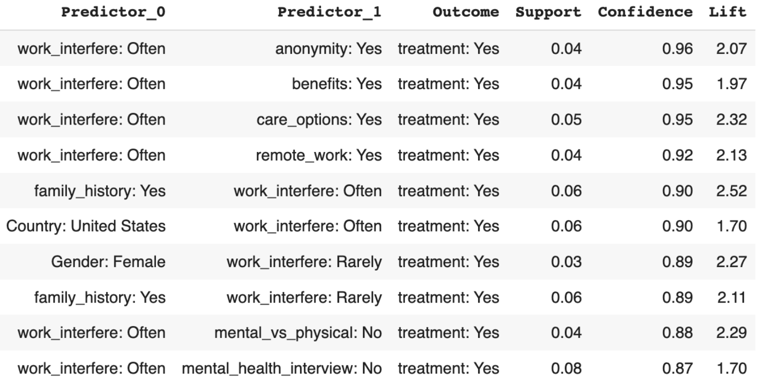
Male employees were more likely to have mental health issues and sought treatments compared to other genders.

IV. Methodology

A. Association Rules

After learning about some of the noticeable patterns of correlation in the data, I wanted to investigate the association rules. Here I decided to pick out the two most common predictors that contribute to whether a person seeking treatment (treatment == ’Yes’). The minimum support threshold is 3% of the data set since this is the most optimal percentage to produce a table that satisfies the conditions for support, confidence, and lift.

Conditions: min_sup = 0.3, min_conf > 0.5, min_lift > 1.



Predictor_0	Predictor_1	Outcome	Support	Confidence	Lift
work_interfere: Often	anonymity: Yes	treatment: Yes	0.04	0.96	2.07
work_interfere: Often	benefits: Yes	treatment: Yes	0.04	0.95	1.97
work_interfere: Often	care_options: Yes	treatment: Yes	0.05	0.95	2.32
work_interfere: Often	remote_work: Yes	treatment: Yes	0.04	0.92	2.13
family_history: Yes	work_interfere: Often	treatment: Yes	0.06	0.90	2.52
Country: United States	work_interfere: Often	treatment: Yes	0.06	0.90	1.70
Gender: Female	work_interfere: Rarely	treatment: Yes	0.03	0.89	2.27
family_history: Yes	work_interfere: Rarely	treatment: Yes	0.06	0.89	2.11
work_interfere: Often	mental_vs_physical: No	treatment: Yes	0.04	0.88	2.29
work_interfere: Often	mental_health_interview: No	treatment: Yes	0.08	0.87	1.70

Fig. 4. Table of association rulings

The most common patterns implied that employees who sought mental health treatment after experiencing interference with work and having a family history of illness. Moreover, mental health was most common in the United States compared to other countries. Working from home was suspected to have an impact on whether he/she has a mental illness. This conclusion supported the research from [2].

B. Classification Models

Given the features in the data set such as gender, family history of mental illness, the extent to which a person disclosed their mental illness with their coworkers, etc., I would like to predict whether a person would decide to seek mental health support. I had chosen 20 features from the data set as the independent variables for the label "treatment".

Since the "treatment" variable only contained two values - "Yes" and "No", it was clear that this was a binary classification problem. Thus, I had chosen some popular classification models to test on the data, which are Logistic Regression, Support Vector Machine (SVM), Decision Trees, Random Forest, and Multilayer Perceptrons (MLP). All the aforementioned machine learning models are called from the Python Scikit-learn package except for the MLP model which is constructed using the Keras library.

1) Logistic Regression: Logistic Regression is one of the most popular algorithms for binary classification problems. This algorithm computes the probability of an event happening given a set of features, and the probability is bounded between 0 and 1.

2) Support Vector Machine (SVM): Support vector machine is a model for both linear and non-linear classification. The goal of the model is to find decision boundary that maximizes the margin - the distance between the decision boundary and the closest data points. Some advantages of the SVM algorithm are that it produces results with high accuracy with less computational power, has a solid mathematical foundation, and yields efficient inference. In this research project, I would build an SVM model after tuning the parameters. I set the kernel to be the Radial Basis Function, the regularization parameter c to be 0.5, and the scaling factor gamma to be 0.1.

3) Decision Trees: A decision tree is a flowchart-like structure in which each internal node represents a test on a feature, the root nodes represent the class labels, and the branches represent the conjunctions of the features that result in the class labels. The tree is constructed through an algorithmic approach which identifies ways to split the data based on several conditions. Again, I would build a decision tree with my tuned parameters. I set the criterion to be "entropy", the maximum depth of the tree to be 5, and the minimum leaf samples to be 20.

4) Random Forest: A random forest is an ensemble of multiple decision trees in which each tree will produce a class prediction and the class with the most votes will be the outcome of the model. By utilizing multiple, low-correlated trees, a random forest can avoid over-fitting and thus typically outperforms a single decision tree. I would construct a random forest after fine-tuning the parameters - my forest would have 30 trees, each tree would have a maximum depth of 4.

5) Multilayer Perceptrons (MLP): Multilayer perceptron is a type of neural network that maps the inputs to outputs, is controlled by a set of parameters, and utilizes backpropagation and gradient descent to adjust the parameters to solve classification problems. MLP is highly flexible and is capable of learning and extracting complex patterns and relationships from various types of data. For my classification problem, an MLP model with one hidden layer was constructed. I used the tanh activation for the hidden layer and the sigmoid activation for the output layer. Firstly, I set the number of hidden units in the hidden layer to be 20. I also set the learning rate to 0.1 and the number of epochs to 1000.

C. Model Analysis

To evaluate the efficiency of each model, I decided to split the whole data into training data and testing data with a ratio of 85:15. The train-test-split process was stratified to preserve the proportion of samples in each class as observed in the data set. I would observe the scores on the training data and as well as the score on the testing data.

For my evaluation metrics, I decided to choose F1-score which is the harmonic mean of precision and recall of my classifier models. The formula for F1-score was the following:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{1}$$

I also decided to investigate the accuracy of each model which represented the number of correctly classified data instances over the total number of data instances. The formula for accuracy was the following:

$$accuracy = \frac{TP + TN}{TN + FP + FN + TP}, \tag{2}$$

where TP, TN, FP, FN indicated true positive, true negative, false positive, and false negative, respectively. The models' performances were recorded in the following table:

Name	F1-train	F1-test	Accuracy-train	Accuracy-test
Decision Tree	0.783760	0.743719	0.769953	0.730159
Random Forest	0.781505	0.746114	0.773709	0.740741
SVM	0.848263	0.750000	0.844131	0.746032
Log Regression	0.742210	0.739583	0.743662	0.735450
MLP	1	0.659686	1	0.656085

TABLE I
PERFORMANCE MEASURES OF MODELS

According to Table 1, SVM performed the best among the models with the highest F1-score and the highest accuracy score on the testing data. Surprisingly, the MLP model had the lowest scores when trained on the testing data. The random forest classifier performed a little better than the decision tree classifier.

D. Feature Importance

Another aspect I would like to learn from the models was how much impact each feature had on a person's decision to seek mental health support. After re-investigating the models, I found out that family history and interference in work would more likely lead the person to reach out for help than other features.

V. CONCLUSION AND FURTHER WORK

From my experiments with association rules and machine learning models (where MLP did not perform well, surprisingly), I found out that if a person had a family history of mental illness and has mental illness interfering with their work, then they were more likely to seek support. Furthermore, male and female employees both actively sought mental help, thus gender was not an important aspect when it comes to treatment. I had also found out that having to work remotely might affect an employee's mental, which was an important insight that companies should take notice of in order to support their employees.

More work still needs to be done for my project. Firstly, the data set size was relatively small (about 1250 observations), which affected my ability to explore and validate insights about mental health issues. In the future, I want to retrieve more data from other sources to gain more insights. Secondly, I would like to continue tuning model parameters to obtain higher scores on the data, especially with the MLP model parameters such as the activation function, the number of epochs, or batch size. Finally, I am interested in investigating how different communities in different cities, states, or regions deal with mental health problems and identifying the similarities and differences between their responses.

References

[1] Williams M. Taryn, “The Intersection of Work and Wellbeing, for All Workers” Mental Health, Disability Employment, U.S Department of Labor, March 30 2022.

[2] Xiong Z, Li P, Lyu H, Luo J. Social Media Opinions on Working From Home in the United States During the COVID-19 Pandemic: Observational Study. JMIR Med Inform. 2021 Jul 30;9(7):e29195. doi: 10.2196/29195. PMID: 34254941; PMCID: PMC8330633. Accessed Dec 09, 2022.

[3] Elflein J. Coronavirus Impact Mental Health Symptoms Workers 2020. Percentage of Workers Who Reported Select Mental Health Symptoms Since the Coronavirus Outbreak in 2020. Published December 15, 2020. Available at: https://www.statista.com/statistics/1169854/covid-related-mental-health-symptoms-in-workers/. Accessed Dec 09, 2022.

[4] V. Duong, J. Luo, P. Pham, T. Yang and Y. Wang, "The Ivory Tower Lost: How College Students Respond Differently than the General Public to the COVID-19 Pandemic," 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2020, pp. 126-130, doi: 10.1109/ASONAM49781.2020.9381379. Accessed Dec 10, 2022.