Mythri Partha

Mental Health in the Tech Industry: Final Report

I. Problem

Who goes to therapy in the technology industry? The technology industry is a vastly growing field and behind the growth are several innovative minds. Those who work in the technology industry work long hours trying to create the best products and software they can. Along with the territory of hard work, comes a lot of stress, some of it could be attributed to underlying mental illnesses or the stress could lead to more severe mental illness. My goal is to use a mental health survey conducted to see what leads to a person seeking therapy. I assume that more women face work interference and as a result they seek therapy more.

II. Data Wrangling

I used a mental health in tech survey that was conducted in 2014, it was found on Kaggle in this link: https://www.kaggle.com/osmi/mental-health-in-tech-survey

In the beginning, my dataset started off with about 26 features, most of which were qualitative, and it contained about 1259 rows. The next stage was data wrangling and cleaning the dataset, there were a lot of spelling errors in the gender section and there were a lot of other gender identities outside of the binary in which there was not enough data to create valid prediction models. After condensing that down, I noticed there were a disproportionate number of males in the study compared to females. In order to change that, I resampled the data by using a random selection process, this decreased my sample size to about 500 people where there was close to an even number of males and females in the study. After that, I made sure to drop the null values and for some columns, I just chose to ignore those completely and use other features

in my prediction models. The shape of the data set came out to be 509 rows and 37 rows, by the time I created dummy variables for the qualitative variables I most wanted to analyze, these were gender, whether or not the participant felt that they would face consequences if they told their employer that they were facing mental health issues, whether or not the participant sought out treatment, whether or not the participant faced work interference due to their mental health, and whether or not the participant had a family history of mental illness. Age of the participant was the only quantitative value that was important to this study, there were some values that were way too large and some that were way too small or negative. I converted those extraneous values to NaN values and then filled those in with the mean of the ages of participants of the study.

III. Exploratory Data Analysis

After cleaning up the data, I could finally observe the data visually. I started by observing trends in the participants' ages and their mental health. This modeling was best done by using swarmplots as shown below

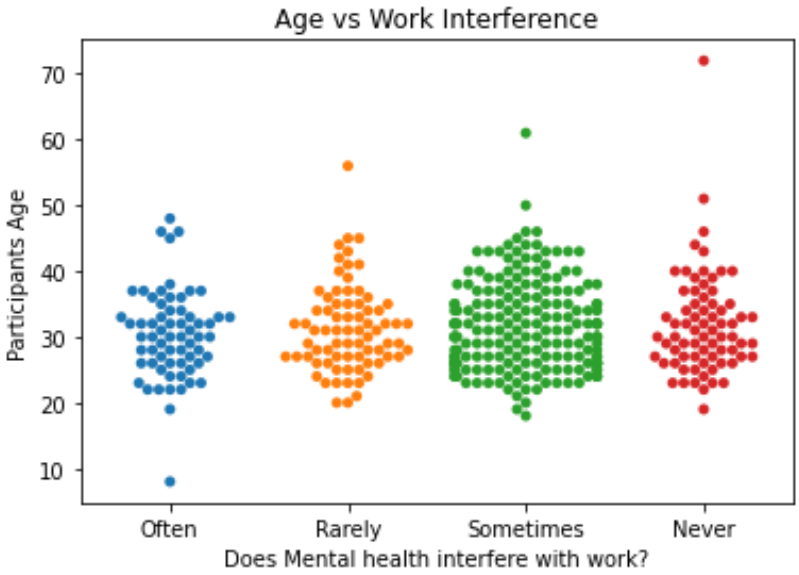

Figure 1: Swarmplot of Work Interference and Participant Age

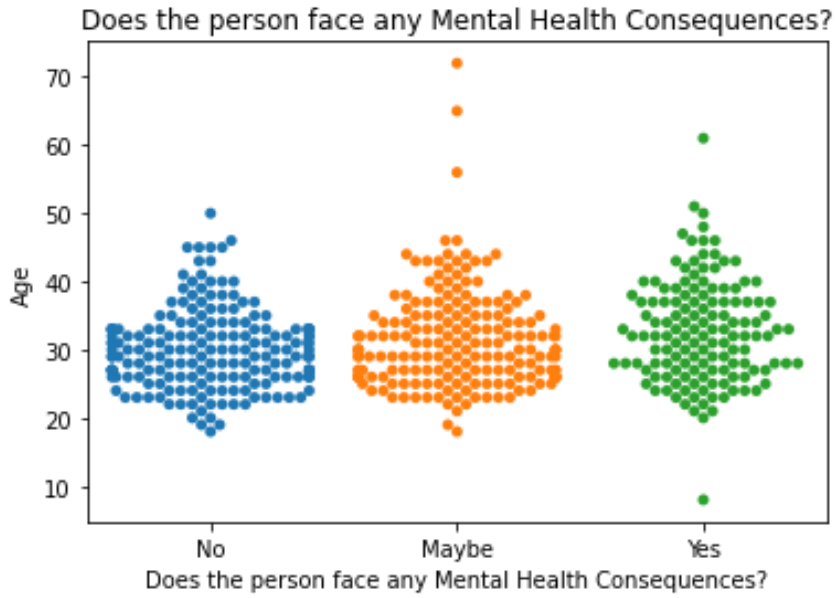

Figure 2: Does the participant fear talking to their employer about mental health by age

According to the figures above, we can see that a majority of the participants in their 20-40s have faced some type of work interference. Some of those in their 20s to early 50s have faced some fear that there would be some consequence of them coming out with their struggles to their employers, however, there still is a substantial amount of people in that age range who have not had that fear. The oldest participants feel like there might be consequences to discussing mental health.

I then observed trends in gender and mental health, since this is primarily qualitative, I had to create countplots on seaborn to observe the frequency of males and females who experience work interference, have fears of work consequences, and those who actually go to therapy. First, figure 3 below is the distribution of genders that participated in the study and figure 4 is a comparison of how men and women face work interference because of their mental health.

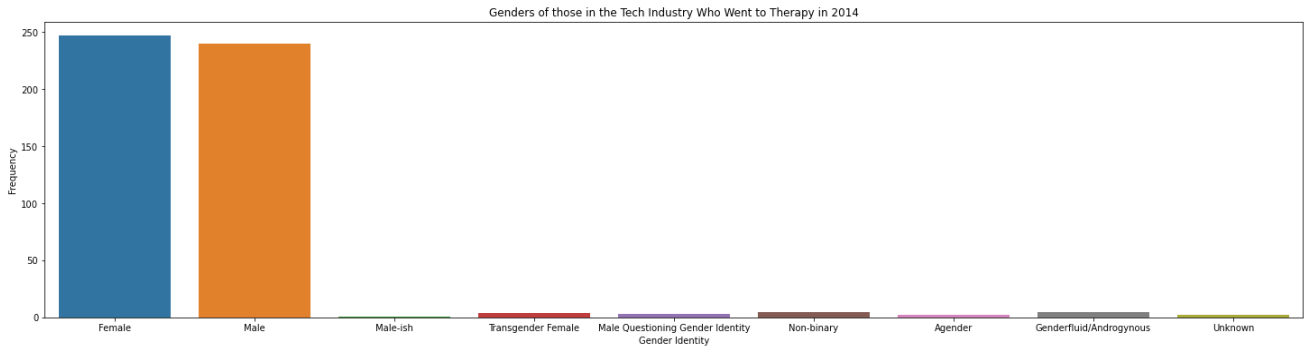

Figure 3: Gender distribution of study participants

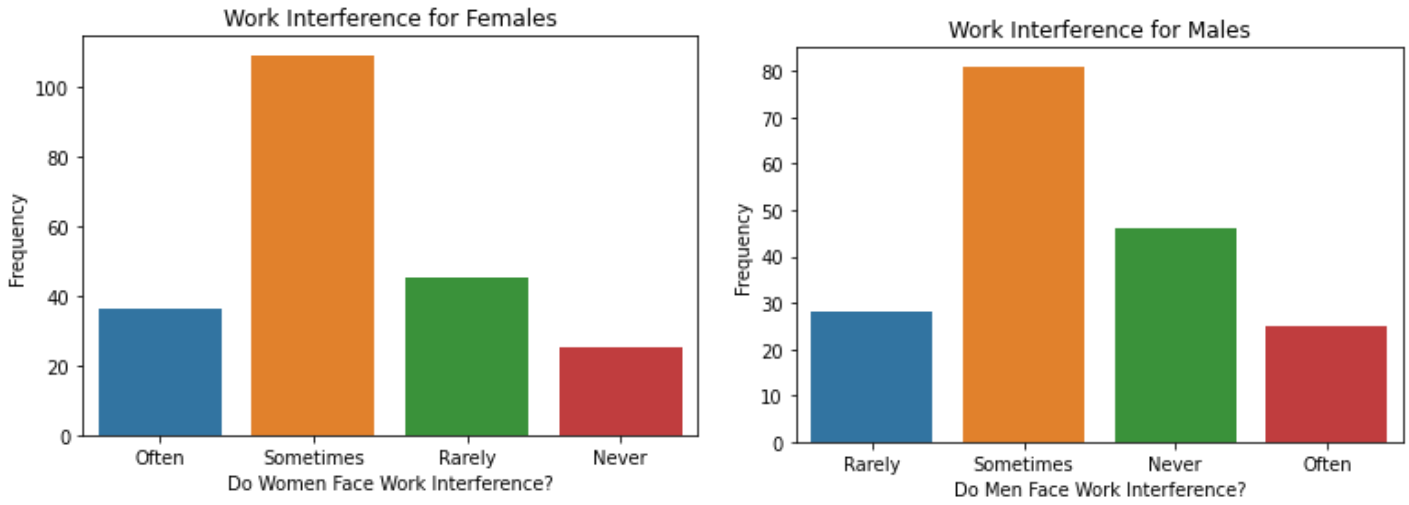

Figure 4: Work Interference for Females and Males

Figure 4 shows that, compared to males, females seem to find that their mental health interferes with their work as more said often or sometimes more than men did. There were nearly two times as many men saying they never face work interference than women saying they never face work interference.

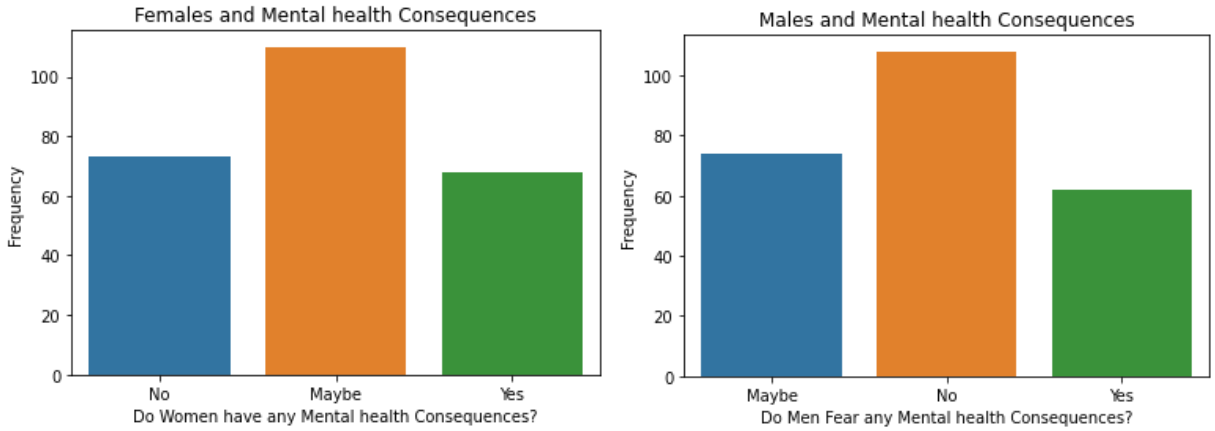

Figure 5: Does the participant fear talking to their employer about mental health by gender

When observing figure 5, we can see that men rarely are afraid to come out to their employer and co-workers since about half of males in this sample do not feel like discussing their mental health at work would have any consequences, some feel there may be consequences, and about a quarter of them do not feel there would be any consequences. When observing females, about half of them said they do feel like they would face consequences for discussing their mental health issues in the workplace, approximately a quarter of them said they would face consequences.

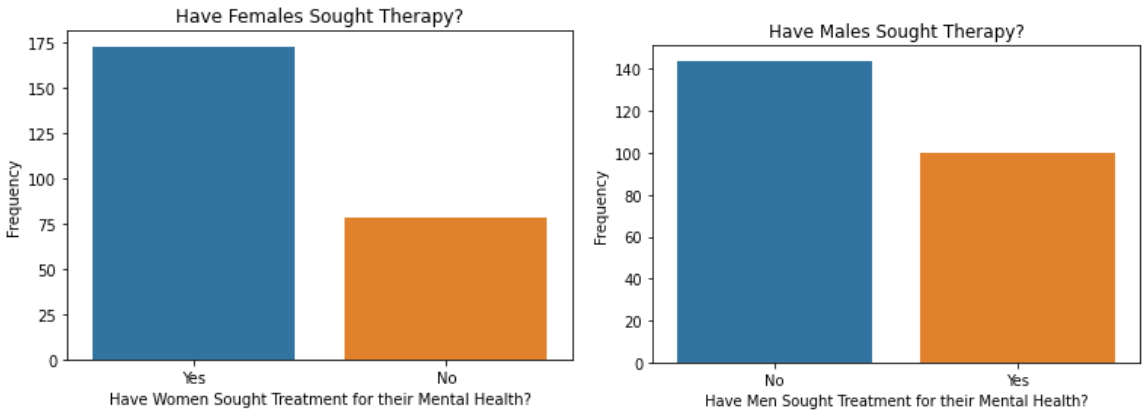

Figure 6: Therapy by Gender

After observing figure six, we can see that women are much more likely to actually go to therapy and seek help, while there are much less men who go to therapy than women, a significant portion of men still do seek help.

Once gender and age were looked at separately, I decided to observe gender and age together along with whether or not the participant goes to therapy. In order to do this, I created separate dataframes for men and women then created swarmplots to observe the relationships.

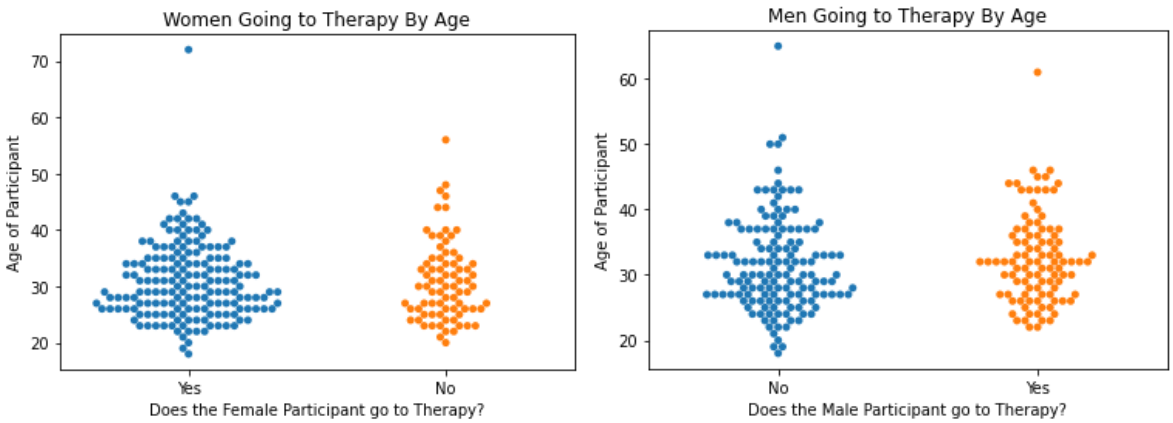

Figure 7: Females and Males going to therapy by age

Figure seven shows that younger males seem to go to therapy, but those who do not go still outweigh those who do. For females, majority of those who go to therapy are under fifty, with one outlier who is in their 70s. After putting all of these findings together, I assume that those who are most afraid of talking to their coworkers about mental health and those who face more work interference are more likely to seek out therapy. Now that I have some speculations, I can perform statistical analysis.

IV. Statistical Analysis

My goal with statistical analysis is to see if there is a normal distribution between males and females who go to therapy. I set the null hypothesis to say that there is a normal distribution

among the ages of people who go to therapy and the alternate hypothesis was that there was not a normal distribution of the ages of those who go to therapy. I first did a normal test on the full dataset on those who go to therapy, then I did two more normal tests (one for each gender data set). All three tests yielded p-values that were less than 0.05, which means that we can reject the null hypothesis that there is a normal distribution of ages of those who go to therapy. The p values were about $1.48*10^{-18}$, $3.43*10^{-17}$, and $3.01*10^{-5}$ for both genders going to therapy, females who go to therapy, and males who go to therapy respectively.

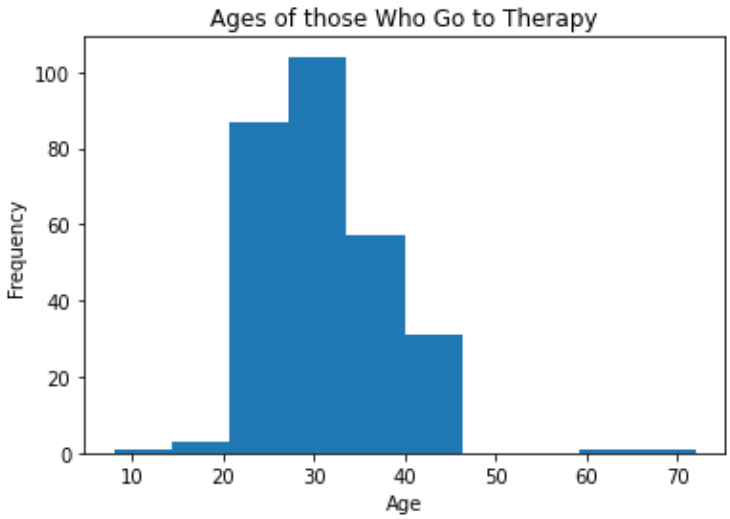

Figure 8: Distribution of those who go to therapy by age

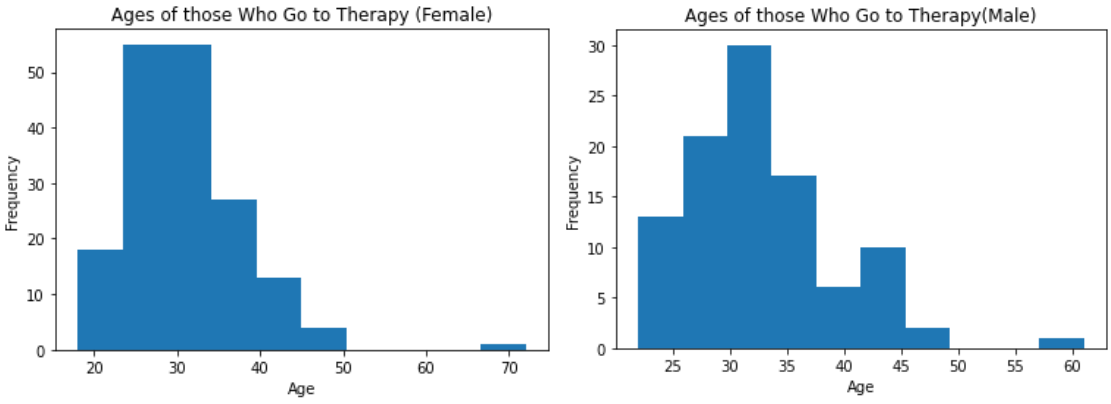

Figure 9: Distribution of those who go to therapy by gender

While observing the histograms of who goes to therapy, all the distributions are skewed to the left, which means that younger people are the ones who seek out therapy most. There are a couple people who seem to be outliers in their 60s to 70s.

V. Modeling/Machine Learning

After seeing the basic distributions, I went on to the modeling and machine learning stage. I first created dummy variables for my qualitative columns of my data so I could easily manipulate it and form accurate models. Once the dummy variables were made, I split my data into training and test sets based on whether or not a participant went to therapy. I performed seven different machine learning algorithms on the training and testing datasets. I did logistic regression, random forest, gradient boost, k-nearest neighbors, support vector machines (SVM), and ADABoost. The table below shows how the tests fared with regards to accuracy, precision, f1 score, recall, and ROC-AUC Score.

	Algorithm	Accuracy	Precision	Recall	F-Score	ROC-AUC Score
0	Logistic Regression	0.877551	0.880702	0.869048	0.873276	0.869048
1	Random Forest	0.836735	0.837719	0.827381	0.831034	0.827381
2	Gradient Boost	0.857143	0.856034	0.851190	0.853231	0.851190
3	K Nearest Neighbors	0.795918	0.794737	0.785714	0.788793	0.785714
4	Support Vector Machine (SVM)	0.877551	0.880702	0.869048	0.873276	0.869048
5	Naive Bayes	0.714286	0.710702	0.714286	0.711279	0.714286
6	ADA Boost	0.857143	0.856034	0.851190	0.853231	0.851190

Figure 10: Model Performance

Figure 10 shows us that the models with the best accuracy and precision are Logistic Regression and Support Vector Machine (SVM) as they both have the same scores of about 0.88 for both accuracy and precision.

After deciding that logistic regression and SVM were the best models, I performed hyperparameter tuning to see if there was any improvement, but for logistic regression the accuracy and precision rate went down to 0.85 and 0.84 respectively. For SVM, however, the accuracy and precision remained at 0.88.

After hyperparameter tuning, I also wanted to look into the feature importance for both logistic regression and SVM. When observing the feature importance, I noticed that mental health consequences, work interference, and family history were the most important features in deciding whether or not a person goes to therapy, contrary to my original belief that gender and age were main driving forces on whether or not a person goes to therapy.

VI. Conclusion and Future Direction

Now comes the most important part, choosing the final model! SVM seems to be the best before and after tuning the hyperparameters, the feature importance was the same, or at least, very close after tuning for both logistic regression and SVM, so both are valid models. The accuracy and precision values could definitely be improved a bit more by possibly using a different train-test split combination to see how each model could perform differently.