SML 310 Final Project: Mental Health in the Tech Industry

Joyce Luo

December 1, 2020

1 Introduction

Mental health problems are extremely prevalent throughout the world, as around 1-in-7 people globally have one or more mental or substance abuse disorders.[1] Within the United States, this proportion is even larger, with around 1-in-5 people experiencing a mental illness, such as depression, anxiety, or bipolar disorder.[2] In addition to many other factors, negative or stressful workplace conditions can play a role in causing or exacerbating mental health problems.[3] The technology (tech) industry tends to foster these types of stressful workplace environments, even though tech is a fast-paced, exciting area of work that many students intend to enter into after college. The culture within the industry can be highly competitive and demanding, and there is an underlying disregard for mental health problems that has not been fully addressed by certain tech employers. Within startups, founders have to quickly transform their ideas into businesses, which often leads to late nights and a lack of work-life balance. In larger tech companies as well, there is a pressure to rise up the corporate ladder and produce as much as possible in the shortest amount of time. These characteristics like inflexible working hours or low levels of support for employees are listed as risks to mental health by the World Health Organization.[4] It is vital for tech companies to take into account employee perspectives about mental health and provide mental health resources for employees who need it.

In order to get a wide range of perspectives about an issue, questionnaires or surveys are often used as methods of data collection. This study uses yearly employee survey data from Open Sourcing Mental Illness (OSMI) to comprehensively assess employee perspectives on the tech industry's treatment of mental health over time. In addition, this study analyzes what characteristics or perceptions of tech companies might be good predictors of mental health problems in employees. These features in tandem with some demographic features are then used to develop a model that predicts whether or not an employee potentially has a mental illness. The main parts of this study are a longitudinal analysis of employee perspectives over time and a classification analysis which determines the best models for identifying whether employees have a mental illness or not. Through these analyses, we can illustrate what tech companies should do better to support their employees as well as develop a model that could potentially help employees who might have a mental health condition get the treatment they need.

[1] Ritchie, Hannah, Global mental health: five key insights which emerge from the data, 2018, url: https://ourworldindata.org/global-mental-health#:~:text=Around%5C%201%5C%2Din%5C%2D7%5C%20people,4%5C%20percent%5C%20of%5C%20the%5C%20population. (visited on 12/01/2020).

[2] of Mental Health, National Institute, Mental Illness, 2020, url: https://www.nimh.nih.gov/health/statistics/mental-illness.shtml (visited on 12/01/2020).

[3] Organization, World Health, Mental health in the workplace, url: https://www.who.int/teams/mental-health-and-substance-use/mental-health-in-the-workplace (visited on 12/01/2020).

[4] Ibid.

2 Related Work

There have been many previous studies that have developed models to predict mental illness or suicidal tendency. A previous study conducted by Pandey et al. performed a classification analysis using a 2014 survey from OSMI that focused on the tech industry. They chose a combination of work-related and demographic-related factors to predict suicidal tendency among employees within tech companies.[5] They used several models, specifically logistic regression, decision tree modeling, and neural networks, to predict suicidal tendency. After comparing the models, the authors concluded that the logistic regression model was the best model for their prediction task.[6] The study gives some context around how to go about testing models and presents models that could potentially be used for this type of classification analysis. The current study expands on this previous work by using data across multiple years, as well as aiming to predict whether or not an employee has a current mental illness rather than predicting suicidal tendency.

Another study conducted by Tate et al. used machine learning techniques to predict future mental health problems in children. The study used characteristics of a child's upbringing, personality, and family history obtained from questionnaire and register data as features in their models. They also investigated several different models, specifically tree-based models and logistic regression, and compared the performance across models.[7] This study illustrates that the onset of general mental health problems can be predicted to a reasonable degree from survey data, which is similar to what our current study aims to do as well. They concluded that their random forest model performed the best, although the accuracy was not significantly different from the other models.[8] Although these types of models are not necessarily viable in clinical settings as of yet, other factors could be added to the models that could be better predictors.

Another study conducted by Kessler et al. used predictive modeling to identify veterans who were "high-risk" for suicide based on data from the National Death Index.[9] This study compared their models to a very similar study conducted by McCarthy et al., which also used modeling on the same dataset.[10] The Kessler et al. study tried to improve on the logistic regression model that was used in McCarthy et al. by regularizing the model using ElasticNet and using less features. They succeeded in producing a model with comparable performance and less predictors.[11] This study illustrates the uses of regularization in logistic regression models to reduce dimensionality. Kessler et al. also used many different types of machine learning techniques to do their prediction task and concluded that Bayesian averaging of regression trees (BART) was advantageous over the other algorithms.[12] These studies illustrate that it is necessary to assess performance across models to determine which model yields the best predictions for a specific task. In addition, these studies give insight into which models might be the best for the survey-based prediction tasks that we undertake in our study.

[5] Pandey, Shambhavi / Khan, Mohammad B. / Thakkar, Sagar: Factors Affecting Mental Health in Employees and Their Relation to Suicide Rates at a Workplace, in: (2019).

[6] Ibid.

[7] Tate, Ashley E.: Predicting mental health problems in adolescence using machine learning techniques, in: PLOS ONE 4 (2020), 1–13.

[8] Ibid.

[9] Kessler, Ronald C.: Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration, in: International Journal of Methods in Psychiatric Research 3 (2017), e1575. (e1575 IJMPR-May-2017-0053.R1)

[10] McCarthy, J. F.: Predictive Modeling and Concentration of the Risk of Suicide: Implications for Preventive Interventions in the US Department of Veterans Affairs, in: Am J Public Health 9 (2015), 1935–1942.

[11] Kessler, R. C.: Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration (2017).

[12] Ibid.

3 Methods

3.1 Relevant Data

The survey datasets used for this analysis were obtained from the OSMI website (OSMI Research). OSMI is a non-profit organization whose main mission is to raise awareness about mental health in the tech industry. They release a mental health-related survey each year (since 2014) which is open to anyone working in the tech industry or in a tech role. These yearly surveys (2014, 2016-2019) are open source and available as separate *csv* files which contain numerous questions related to employee attitudes towards how mental health is addressed in the tech industry and the corresponding responses from employees. Each survey contains age and gender information as well as characteristics of the tech company that each employee works at which are related to mental health. The 2014 OSMI survey had 1298 employee responses, the 2016 survey had 1434 responses, the 2017 survey had 757 responses, the 2018 survey had 418 responses, and the 2019 survey had 353 responses. Although some of the survey questions overlapped across the yearly surveys, the 2016-2019 surveys contained significantly more questions than the 2014 survey in varying formats. In order to maintain relative consistency across all datasets, only certain fields were pulled from the 2016-2019 datasets. The 2016-2019 surveys also included certain fields that were not as relevant for this analysis, such as characteristics of previous employers, so those fields were omitted. Only questions that had categorical responses (like ”Yes”, ”No”, ”I don’t know”, etc.) were included in the dataset because few respondents answered the free-form questions. Table 1 shows the variables that are consistent across all of the datasets.

Variable Name	**Question**
age	What is your age?
gender	What is your gender?
country	What country do you work in?
state	If you live in the United States, which state or territory do you live in?
self_employed	Are you self-employed?
family_history	Do you have a family history of mental illness?
treatment	Have you sought treatment for a mental health condition?
no_employees	How many employees does your company or organization have?
tech_company	Is your employer primarily a tech company/organization?
benefits	Does your employer provide mental health benefits?
care_options	Do you know the options for mental health care your employer provides?
wellness_program	Has your employer ever discussed mental health as part of an employee wellness program?
seek_help	Does your employer provide resources to learn more about mental health issues and how to seek help?
anonymity	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
leave	How easy is it for you to take medical leave for a mental health condition?
coworkers	Would you be willing to discuss a mental health issue with your coworkers?
supervisor	Would you be willing to discuss a mental health issue with your direct supervisor(s)?
mental_health_interview	Would you bring up a mental health issue with a potential employer in an interview?
phys_health_interview	Would you bring up a physical health issue with a potential employer in an interview?

Table 1: Applicable survey questions and their corresponding variable name.

There were also some additional fields that were included in the 2016-2019 datasets (but not the 2014 dataset) that assisted with analysis, which are shown in Table 2.

Variable Name	**Question**
current_mental_disorder	Do you currently have a mental disorder?
diagnosed	Have you ever been diagnosed with a mental health disorder?

Table 2: Extra fields included in 2016-2019 surveys.

The 2017-2019 datasets included some additional fields that were used for analysis as well, and these are shown in Table 3.

Variable Name	**Question**
importance_physical	Overall, how much importance does your employer place on physical health?
importance_mental	Overall, how much importance does your employer place on mental health?
overall_support_rating	Overall, how well do you think the tech industry supports employees with mental health issues?

Table 3: Extra fields included in 2017-2019 surveys.

These extra fields were included in the datasets because they seemed to be important factors that would be interesting to visualize over time. Certain factors such as *current_mental_disorder* or *diagnosed* had potential to be used as targets in the classification analysis.

3.2 Data Cleaning

Significant data cleaning was required for each year's survey dataset because the surveys were not standardized across years. Each question was converted to a variable name that was the same across all datasets, and all of the responses from individuals who were self-employed ($self\_employed = 1$) were dropped. This was because self-employed individuals did not respond to any questions that were related to a tech workplace, so their responses would not be helpful for analysis. Null values were either removed or replaced with "N/A", as this was usually one of the possible responses for the survey questions. Spelling or formatting mistakes were resolved by parsing through the unique values of each field and correcting or categorizing the unique values. The *age* variable, which was initially a continuous numerical field, was converted into *age_range*, which is a categorical variable for different age ranges. *gender*, which was a free-form field on the survey, was also converted into a categorical variable with three categories: "M", "F", and "Other". The "Other" category was intended to be very general because it includes those who identify as non-binary, transgender, genderqueer, etc. A new variable called *country_label* was also created, which was equal to "Non-USA" if the individual works in a country other than the United States, and "USA" if the individual works in the United States. Responses were standardized across datasets (i.e., a "Maybe" response was equated with a "Possibly" response in different datasets), and all variables other than *country* and *state* were label encoded for ease of manipulation.

3.3 Exploratory Data Analysis

3.3.1 General EDA

To get more of a general idea about the individuals in the survey datasets, exploratory data analysis was performed on a merged version of all 5 datasets. In order to get a clearer view of the demographics and location of respondents, we first had to look at the distributions of location, age, and gender. After counting unique location instances, we discovered that the majority of respondents work in the United States (2268), followed by the United Kingdom (375), Canada (144), and Germany (103). Overall, 2268 respondents work in the United States and 1175 work in countries outside the United States. Within the United States, we can see from Figure 1 that most respondents are from California (355), which makes sense because numerous tech companies have their headquarters in California. Slightly less respondents are from Illinois (184), followed by Washington (149).

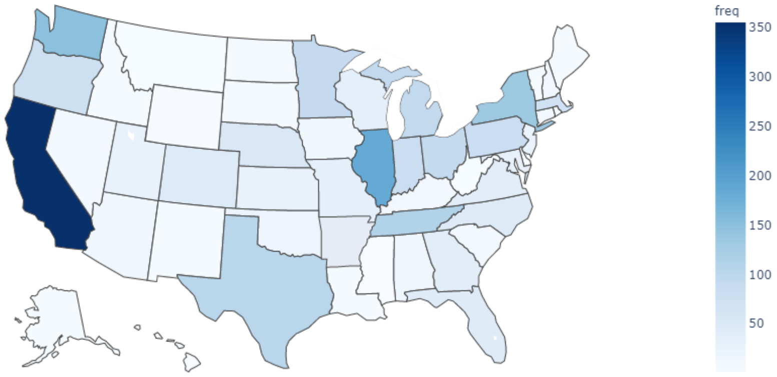

Figure 1. Choropleth map of the number of respondents from each state.

We can also see that the majority of individuals who filled out the survey are male (71.9%), followed by female (24.9%), and Other (3.11%) (Appendix Figure 1). The overwhelming percentage of males who filled out these surveys seems to be fairly indicative of the demographics of the tech industry currently. According to a report by the U.S Equal Employment Opportunity Commission, there is only about a 30% participation rate for women in Silicon Valley tech firms.[13] In addition, the distribution of age among the respondents is highly concentrated between 20 and 40 years of age (Appendix Figure 2). The average age of an employee in the tech industry is 29 years old,[14] so this fits with the distribution of our data.

In order to get a measure of how the tech industry supports employees with mental health issues overall, we looked at the survey question represented by *overall_support_rating*, which is a rating on the scale of 1-5 (1 being the worst and 5 being the best). From Figure 2, it seems like the majority of employees do not think that the tech industry is doing a good job of supporting their employees with mental health issues. The percentage of

[13] Commission, U.S. Equal Employment Opportunity, Diversity in High Tech, URL: https://www.eeoc.gov/special-report/diversity-high-tech?renderforprint=1#:~:text=Compared%5C%20to%5C%20overall%5C%20private%5C%20industry,percent%5C%20to%5C%208%5C%20percent)%5C%2C%5C%20and (visited on 12/01/2020).

[14] Training, Recruiting Innovation Online, Diversity in Tech: We've Got a Long Way to Go, URL: https://recruitinginnovation.com/blog/diversity-in-tech/ (visited on 12/01/2020).

employees who gave their tech employers a 1, 2, or 3, far exceeds those who gave them a 4 or 5. This will be explored further in the longitudinal analysis.

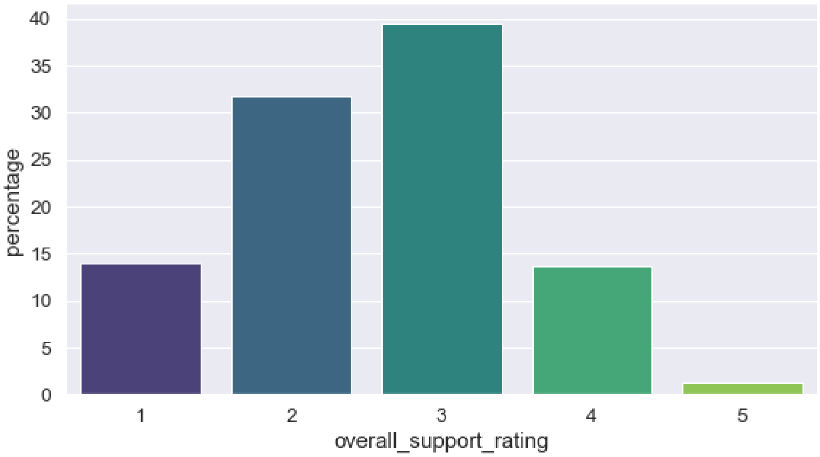

Figure 2. Overall support rating percentages from 2017-2019 surveys.

We also wanted to see whether current employees who have mental illnesses are actually getting the treatment that they need, since we wanted to create models to help improve this statistic. We got a measure of this by calculating the percentage of employees that have been treated, given that they have a mental health disorder. The following variables were important for this: *treatment* and *current_mental_disorder*. From Table 4, we can see that the majority of people who currently have a mental disorder have sought treatment for the disorder (93.24%). However, there are still around 7% of people who have not been treated even though they have a current mental disorder. For those who might have a mental disorder, 54.65% have sought treatment, but 45.35% have not. Since there are discrepancies among the two variables, it would be useful to predict both.

Current Mental Disorder				
		Yes	Maybe	No
Treatment	Yes	93.24%	54.65%	25.74%
	No	6.76%	45.35%	74.26%

Table 4

We then wanted find out which variables were most correlated with the targets *treatment* and *current_mental_disorder*. We plotted correlation heatmaps with the variables that had the highest correlations with *treatment* or *current_mental_disorder*. The variables that are most correlated with *treatment* are *family_history*, *benefits*, *care_options*, and *gender* (Figure 3). The variables that are most correlated with *current_mental_disorder* are *gender*, *benefits*, *care_options*, *family_history*, and *no_employees* (Figure 4). A higher correlation is usually indicative that the variable will be a better predictor of the target. Therefore, we can expect that the most correlated variables are more important for the prediction of each of the targets.

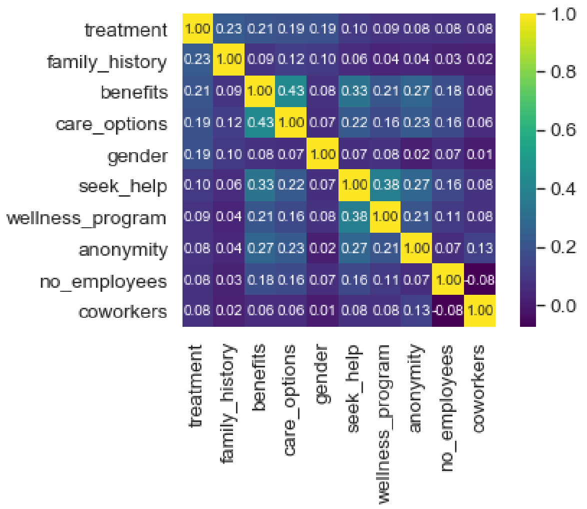

Figure 3. Heatmap of variables that are most correlated with *treatment*.

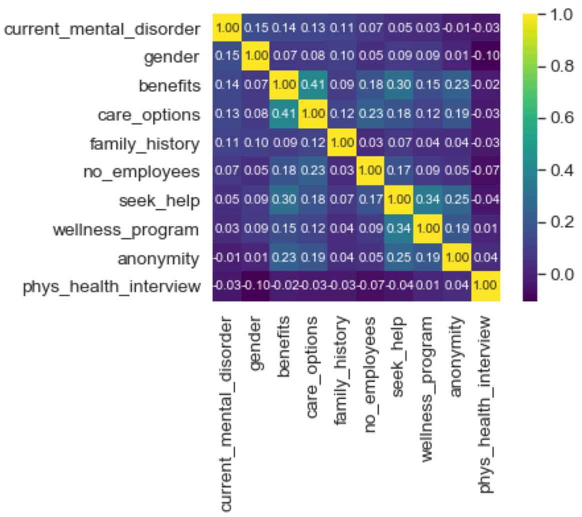

Figure 4. Heatmap of variables that are most correlated with *current_mental_disorder*.

3.3.2 Longitudinal Analysis

For the longitudinal analysis, we aimed to see whether there was any change in employee perceptions of the tech industry's handling of mental health over time. We also wanted to see whether tech companies were taking any more measures to support their employees' mental health in recent years. In order to do this, we had to partition out the necessary data from each year and plot the percentages for responses to specific questions over time to see if there had been a positive, negative, or no change in the percentages. Since only certain years had specific data fields, we had to make 3 merged datasets, one with data from the years 2014 and 2016-2019, another with data from the years 2016-2019, and another from the years 2017-2019.

3.4 Modelling

3.4.1 Data Preprocessing

In order to compile as much data as possible for prediction, the 2014 and 2016-2019 datasets were merged into one large dataset called *merged_all* and the 2016-2019 datasets were merged into their own separate dataset called *merged_no*2014. The *merged_all* dataset has data from 3443 individuals, and the *merged_no*2014 dataset has data from 2349 individuals. We performed two prediction tasks: the first using *treatment* as the target, and the second using *current_mental_disorder* as the target. *diagnosed*, which is a variable that equals 1 if the employee had ever been diagnosed with a mental health disorder and 0 otherwise, was also tested as a target. However, the models did not perform well at predicting this target, so this analysis was discontinued. *current_mental_disorder* only appears as a field in the 2016-2019 datasets, which is why we had to merge them separately for the *current_mental_disorder* prediction task. For both prediction tasks, the following variables were used as features: *age_range*, *gender*, *no_employees*, *family_history*, *benefits*, *care_options*, *wellness_program*, *seek_help*, *anonymity*, *supervisor*, *leave*, *coworkers*, *mental_health_interview*, and *phys_health_interview*. These features are a combination of employee perspectives about the tech company they are working at, demographic information, and characteristics of their employer.

Data manipulation was required in order to make the features suitable to input into classification models. All variables were label encoded during data cleaning, but some of the variables were not as conducive to label encoding since they didn't have a natural ordering. For example, the possible responses to *wellness_program* were "Yes", "No", or "I don't know". However, there is no natural numerical ordering that is reasonable for these categories. Therefore, these features were one-hot encoded, which created a new binary variable for each response. Although one hot encoding increases the dimensionality of the feature space, it also helps certain algorithms, especially logistic regression, do a better job in prediction. For each of the one-hot encoded variables, one of the binary variables that represented a response was preemptively dropped in order to avoid multicollinearity in the models. There were also certain variables like *no_employees* that had a natural ordering to the categories, so these variables were left label encoded. *no_employees* had responses that described how many employees worked at the tech company, and respondents could select from "1-5", "6-25", "26-100", etc. We can see that these categories have an ordering, so they were labeled numerically starting from 0. *age_range*, *no_employees*, and *leave* were label encoded, while the rest of the predictors were one-hot encoded.

To make the targets suitable for modeling, *treatment* was converted into a binary variable with 1 representing that the employee had been treated for a mental illness and 0 representing that they had not. *current_mental_disorder* initially had 4 possible responses: "I don't know", "Maybe", "No", and "Yes". The "I don't know" responses were removed from each dataset, because they only took up a small percentage of the responses and did not give any information for classification. The "Maybe" responses were all treated as "Yes" responses for the purposes of the classification. This is due to the fact that "Maybe" responses could be indicative of the employee being unsure about their mental health status, since they have not previously talked to a health professional about it. The purpose of our study is to identify those with possible mental health issues in order for them to get the help that they need, so it is better to over-identify rather than under-identify those who might have mental health problems. In addition, mental health problems can worsen if they are unattended to, so it is better to have a false positive rather than a false negative. This simplifies *current_mental_disorder* to a binary variable indicating 1 for "Yes" and 0 for "No", which also simplifies the prediction task to binary classification.

3.4.2 Models

A 70/30 train-test split was performed on the datasets *merged_all* and *merged_no*2014 for the purposes of cross-validation. The following models were trained on the *merged_all* training set to predict *treatment*: L1 Penalized Logistic Regression, Random Forest, Decision Tree, Polynomial Support Vector Machine (SVM), Bernoulli Naive Bayes, and K-Nearest Neighbors. The same models were trained on the *merged_no*2014 training set to predict *current_mental_disorder*. These models were selected based on prior recommendation from previous literature and modified to fit the prediction tasks for this study.[15][16][17] All models were generated from the Scikit-learn package in Python 3.[18]

For the Logistic Regression model, L1 regularization was used in order to help with feature

[15] Kessler, R. C.: Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration (2017).

[16] Tate, A. E.: Predicting mental health problems in adolescence using machine learning techniques (2020).

[17] Pandey, S. / Khan, M. B. / Thakkar, S.: Factors Affecting Mental Health in Employees and Their Relation to Suicide Rates at a Workplace (2019).

[18] Pedregosa, F.: Scikit-learn: Machine Learning in Python, in: Journal of Machine Learning Research (2011), 2825–2830.

selection in the model. This was necessary in order to reduce the higher dimensionality that was caused by one-hot encoding most of the features. L1 regularization adds a penalty term to the loss function of the logistic regression, which aims to shrink the absolute value of the magnitude of each coefficient. The SAGA solver was used because it supports L1 regularization, and it is an incremental gradient algorithm that is shown to quickly converge to a solution.[19] The Random Forest model was used with default hyper-parameters which are shown in the Scikit-Learn documentation (Random Forest). The Decision Tree model was used with a maximum tree depth of 3, because this depth produced the highest training and testing accuracies. A SVM model with a polynomial kernel of degree 3 was used for modeling because it produced the highest recall score and accuracy compared to other kernels. A polynomial kernel introduces a non-linear decision boundary and calculates the similarity between two vectors according to a d-degree polynomial.[20] It is used to account for interactions between features, and since there are correlations between features in our model, this kernel is more useful. The linear kernel was problematic because it was only using one feature for prediction and didn't take the model's complexity into account. Bernoulli Naive Bayes was used because most of the features were binary variables after one-hot encoding. K-Nearest Neighbors was used with K = 47 for predicting *treatment* and K = 51 for predicting *current_mental_disorder*. The parameter K was determined by using the rule of thumb, which is $K = \sqrt{n}$, where n is the sample size. It was also recommended to use an odd number K for binary classification.[21] The rule of thumb was used as a starting point and then several values of K near the rule of thumb value were tested to see which yielded the best accuracies and recall.

All models were evaluated on their respective testing sets. For each model, the training classification accuracy, testing classification accuracy, confusion matrix, receiver operating characteristic (ROC) curve, AUC score, and recall were calculated. Recall was chosen as an important metric to evaluate the performance of the models because there is a high cost associated with false negatives when trying to predict whether or not an individual has a mental illness. Classifying an individual as not having a mental illness when they actually have one is extremely bad, because they will not get the treatment that they need. Recall captures the percentage of employees that the model correctly predicts have a mental illness out of all the employees that actually have a mental illness (and likewise for *treatment*).

4 Results

4.1 Longitudinal Analysis

This analysis was carried out in order to visualize potential changes in employee perceptions related to mental health in their workplaces and to see if there was a shift in tech company policies surrounding mental health over time. There were several notable plots that illustrated changing trends in tech company policies over the past years. Figure 5 illustrates that the percentage of "Yes" responses for whether or not the employer has discussed mental health as part of an employee wellness program has a clear increasing trend over time. Correspondingly, the percentage of "No" responses is decreasing.

[19] Defazio, Aaron / Bach, Francis R. / Lacoste-Julien, Simon: SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives, in: CoRR (2014).

[20] Scikit-learn, Pairwise metrics, Affinities and Kernels, url: https://scikit-learn.org/stable/modules/metrics.html (visited on 12/01/2020).

[21] Band, Amey, How to find the optimal value of K in KNN?, url: https://towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb (visited on 12/01/2020).

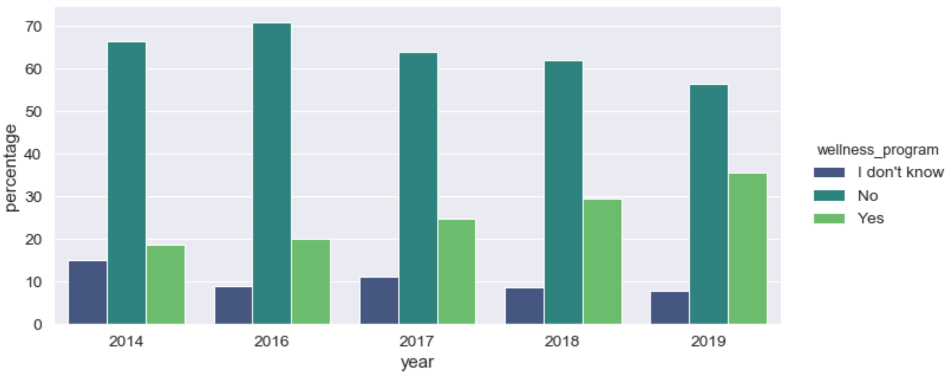

Figure 5. Trend in responses to *wellness_program* from 2014 to 2019.

In Figure 6, we can see that in regards to how easy it is for an employee to take medical leave for a mental health condition, the percentage of "Very difficult" responses decreases from 2016 to 2019 and the percentage of "Very easy" responses has gone up over time. This could be an indication that employers are being more reasonable about letting employees take medical leave due to mental illness.

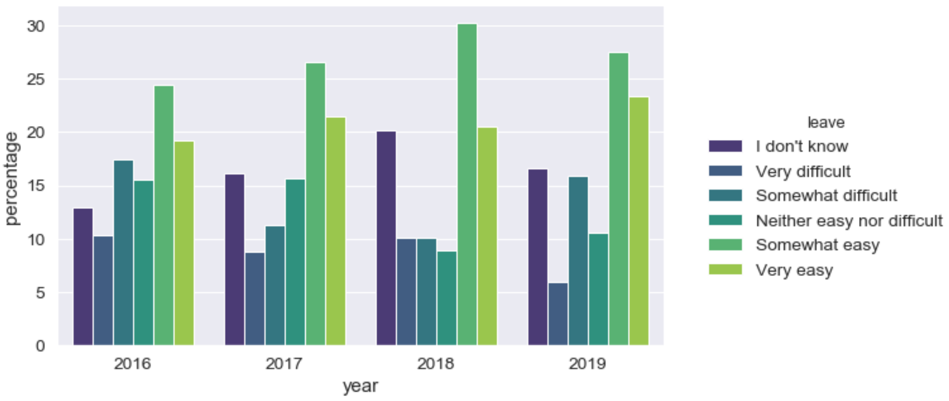

Figure 6. Trend in responses to *leave* from 2016 to 2019.

Figure 7 illustrates that a larger percentage of employers are providing resources to learn more about mental health issues and how to seek help in 2019 compared to 2014. There is an upward trend in the percentage of "Yes" responses and a downward trend in the percentage of "No" responses.

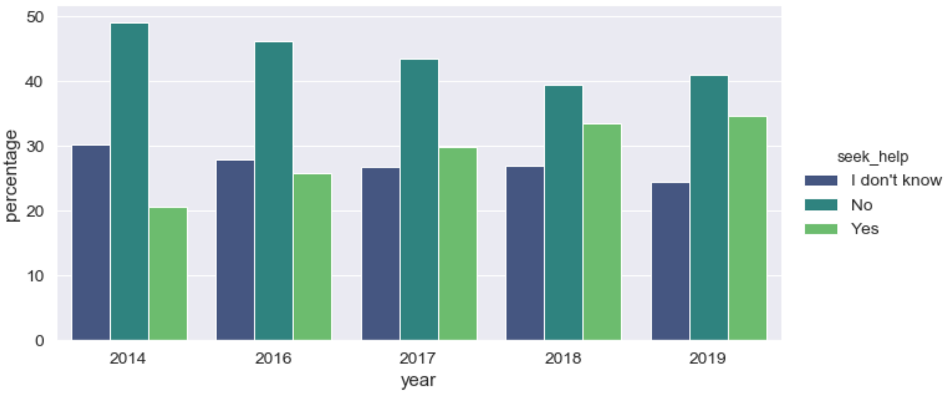

Figure 7. Trend in responses to *seek_help* from 2014 to 2019.

Nevertheless, there were quite a few features that did not have any changes in their trends from 2014 to 2019. For instance, *mental_health_interview*, *supervisor*, *anonymity*, *benefits*, and *overall_support_rating* had plots that had the same general distribution of responses in each year (Appendix Figures 3-8). The majority of the plots that did not have any changes over time were plots regarding employee perceptions about the tech industry. *mental_health_interview* asks whether the respondent would bring up a mental health issue with a potential employer in an interview, and the overwhelming majority of respondents indicated that they would not do so in every year. The *overall_support_rating* ratings distribution was also exactly the same each year, with only about 1% of employees giving the tech industry a rating of 5 out of 5 for how well they support their employees with mental health issues (Figure 2). *coworkers* had a slight change in the response distribution from 2016 to 2017, as after 2017, a greater percentage of respondents answered ”Yes” than ”No” to whether they would be willing to discuss a mental health issue with a coworker. However, there was no significant change in the trend of the distribution of responses after that shift (Appendix Figure 5). Variables like *importance_physical* and *importance_mental*, which asked about how much importance the respondent's employers placed on physical and mental health respectively also did not have any changes over time. The plots illustrated that throughout the years employees perceived that employers have consistently placed more importance on physical health rather than mental health (Appendix Figures 9-10).

4.2 Classification Analysis

4.2.1 Feature Importance

Feature importance measures are provided for certain classification models, which give an idea of what features contribute the most to predicting the target variable. For tree-based models like Random Forest or Decision Tree models, the feature importance is based on the Gini impurity, which is computed from the structure of each tree. The Gini impurity is the probability that you misclassify an instance from a set of data, if the instance were randomly classified according to labels from the dataset.[22] At each split in the tree, if a feature decreases the Gini impurity more, this indicates that it is a more important feature. For Logistic Regression models, we can get a sense of feature importance from the size of the model coefficients. These coefficients act as a ”weight” for each feature, so the larger the magnitude of the coefficient, the more important the feature is. In order to get a general idea of which features were the most important for our prediction tasks, we decided to mainly look at the feature importance for these three models.

Figures 8 and 9 illustrate the most important features in the Logistic Regression model and Decision Tree model, respectively, which were used for predicting *treatment*. Only the top 10 most important features are displayed in each plot. For the L1 regularized Logistic Regression model, features like *family_history_1*, *care_options_3*, *coworkers_2*, *gender_2*, and *family_history_2* were the most important for prediction (Figure 8). In Figure 9, we can see that the Decision Tree only uses 5 of the features for prediction: *family_history_1*, *care_options_3*, *gender_1*, *mental_health_interview_1*, and *family_history_2*. It is interesting that both of the models indicated that similar variables were important for prediction. *family_history*-, *care_options*-, and *gender*-related binary variables were important for both of the models. The Random Forest model feature importance plot is included in the Appendix, with the most important features

[22] Ambielli, Brian, Gini Impurity (With Examples), url: https://bambielli.com/til/2017-10-29-gini-impurity/#:~:text=Gini%5C%20Impurity%5C%20is%5C%20a%5C%20measurement,labels%5C%20from%5C%20the%5C%20data%5C%20set. (visited on 12/01/2020).

being $family\_history\_1$, $leave$, $no\_employees$, age_range, and $family\_history\_2$ (Appendix Figure 11).

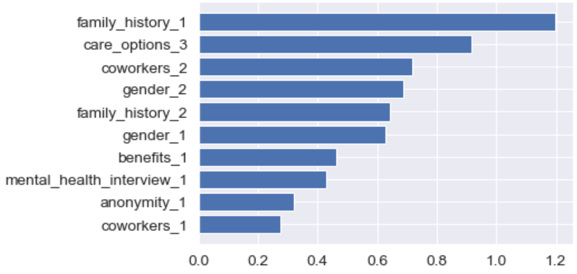

Figure 8. Most important features in Logistic Regression predicting $treatment$.

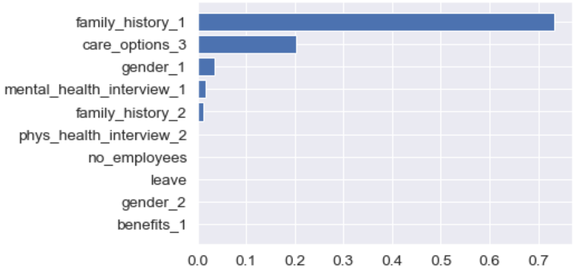

Figure 9. Most important features in Decision Tree model predicting $treatment$.

For the $current\_mental\_disorder$ prediction task, Figures 10 and 11 illustrate the most important features for prediction in the Logistic Regression model and Decision Tree model, respectively. For the L1 regularized Logistic Regression model, features like $family\_history\_1$, $care\_options\_3$, $gender\_2$ were still extremely important for prediction similar to the $treatment$ prediction task (Figure 10). However, other variables like $mental\_health\_interview\_1$ and $seek\_help\_2$ became more important for predicting whether or not an employee currently has a mental disorder. In Figure 11, we can see that the Decision Tree model indicated that $family\_history\_1$ and $care\_options\_3$ were the most important, but in comparison to the $treatment$ prediction task, $family\_history\_1$ became disproportionately important for prediction. The Random Forest model feature importance plot is included in the Appendix once again, and the most important features were the same for this prediction task as they were for the $treatment$ prediction task (Appendix Figure 12).

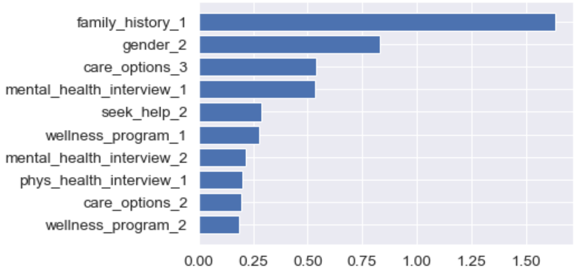

Figure 10. Most important features in Logistic Regression predicting $current\_mental\_disorder$.

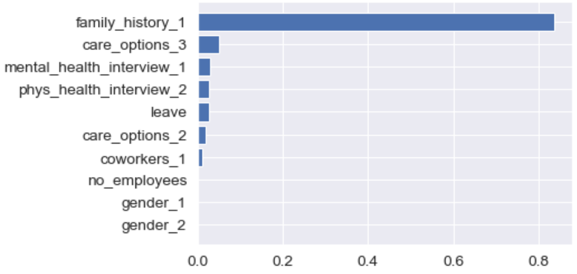

Figure 11. Most important features in Decision Tree model predicting $current\_mental\_disorder$.

4.2.2 Model Comparison

In order to establish a baseline for the classification analyses, the percentage of employees in $merged\_all$ who had $treatment = 1$ was calculated, as was the percentage of employees who had $current\_mental\_disorder = 1$ in $merged\_no2014$. If the models predicted all 1's for the respective targets, the accuracies would be around 57.86% for $treatment$ and 65.43% for $current\_mental\_disorder$. The models in the classification analysis should be doing better than this baseline to be successful.

The models from this study have relatively good accuracies that are better than their respective baselines. For the $treatment$ prediction task, we can see from Table 5 that the best-performing model in terms of training accuracy is the Random Forest model, with an accuracy of 99.54%. The model that performs the best on the testing data is the Polynomial SVM model, with an accuracy of 71.64%. The Decision Tree model also performs reasonably well on the testing set, with an accuracy of 71.25%. For the recall metric, the model that performs the best is the Decision Tree model, with a recall score of 84.79%. The Penalized Logistic Regression model has the highest AUC score out of all of the models (0.764).

Model	**Training Accuracy**	**Testing Accuracy**	**Recall**	**AUC Score**
Penalized Logistic Regression	73.28%	71.06%	78.18%	0.764
Random Forest	99.54%	70.58%	77.19%	0.747
Decision Tree	73.03%	71.25%	84.79%	0.749
Naive Bayes	71.29%	69.12%	74.55%	0.744
Poly SVM	75.44%	71.64%	82.31%	0.758
KNN	72.11%	69.12%	76.69%	0.738

Table 5: Predicting *treatment* using the 2014, 2016-2019 datasets.

Compared to the *treatment* prediction task, all of the models are better at predicting *current_mental_disorder* according to the performance metrics. We can see from Table 6 that the best-performing model for predicting *current_mental_disorder* in terms of training accuracy is still the Random Forest model, with an accuracy of 99.94%. The model that performs the best on the testing data is the Penalized Logistic Regression model, with an accuracy of 75.04%. The Logistic Regression model actually does better on the testing data than the training data for this prediction task, which seems to mean that the model is learning the pattern of the data and not overfitting. The Penalized Logistic Regression model once again has the highest AUC score out of all of the models (0.776). Representative ROC curves for each model predicting *current_mental_disorder* are included in the Appendix (Appendix Figures 13-18). ROC curves for predicting *treatment* are not included because they are very similar. For the recall metric, the model that performs the best is K-Nearest Neighbors, with a recall score of 94.36%. This is a surprisingly high recall compared to the accuracies for all the models which are in the low to mid-70s. In fact, all of the models have a high recall score in comparison to their accuracies.

Model	**Training Accuracy**	**Testing Accuracy**	**Recall**	**AUC Score**
Penalized Logistic Regression	73.97%	75.04%	86.12%	0.776
Random Forest	99.94%	72.77%	85.68%	0.754
Decision Tree	73.42%	72.06%	86.98%	0.742
Naive Bayes	72.26%	73.76%	81.78%	0.771
Poly SVM	73.42%	72.62%	90.02%	0.750
KNN	71.17%	71.21%	94.36%	0.741

Table 6: Predicting *current_mental_disorder* using the 2016-2019 datasets.

In order to see why it is the case that all the models have high recall scores, we can look at the confusion matrices for certain models. We can see that in Figure 12, the confusion matrix for the K-Nearest Neighbors model shows that the model tends to predict that $current\_mental\_disorder = 1$ the majority of the time. There are a large number of false positives (177) compared to true negatives (67), and there are also a large number of true positives (435) compared to false negatives (26). This illustrates that the model is not very good at predicting that $current\_mental\_disorder = 0$, which is why the testing accuracy is so low compared to the recall score. This model tends to err on the side of predicting that the individual currently has a mental disorder but sacrifices correctly classifying those who do not have a mental disorder. We can compare this to the Penalized Logistic Regression model's confusion matrix (Figure 13). The Logistic Regression model gives a slightly higher accuracy because it is not as biased as the K-Nearest Neighbors model towards predicting that $current\_mental\_disorder = 1$. The Logistic Regression model predicts more true negatives than false positives, which is good. However, this comes at a slight detriment to the recall score, which is 86.12% compared to 94.36% for the K-Nearest Neighbors model. If we look at the rest of the confusion matrices for the models, we can see a similar trend where models which have high recall scores tend to over-predict $current\_mental\_disorder = 1$ (Appendix Figures 19-22). Over-predicting the number of people who currently have a mental disorder is not necessarily a bad outcome for the models, depending on what is done with the prediction results.

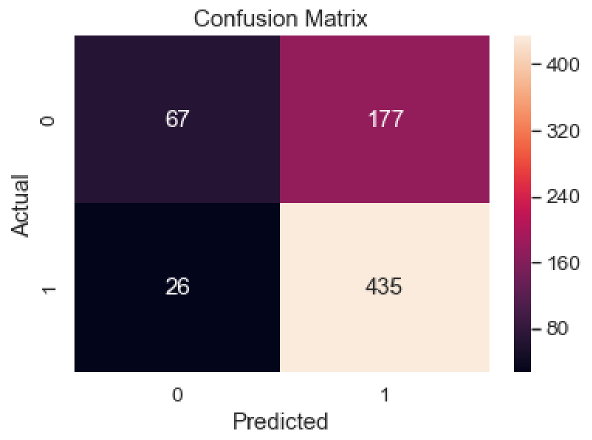

Figure 12. K-Nearest Neighbors confusion matrix for predicting $current\_mental\_disorder$.

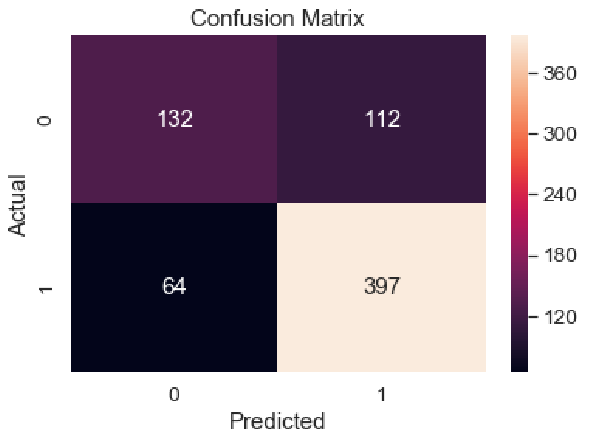

Figure 13. Logistic Regression confusion matrix for predicting $current\_mental\_disorder$.

5 Discussion

This study provides a comprehensive view about how employees have perceived the tech industry's treatment of mental health issues in the workplace over the past few years. In our longitudinal analysis, we showed that most employee perceptions of how tech companies are treating mental health have not changed over time. There is still a clear stigma about discussing mental health issues in the workplace with coworkers and supervisors that has not significantly improved since 2014. However, it seems like tech companies have started to address some of the aspects that are lacking in their treatment of mental health, as our analysis has shown that there has been an upward trend in the percentage of employees who indicate that their companies are offering resources to seek help, discussing mental health as a part of employee wellness programs, and making it easier for employees to take medical leave for a mental health issue.

Since many more people have been advocating for elevating the discussion around mental health in recent years,[23][24] we should expect to see improvement in most tech company

[23] Barkved, Kirsten, Let's Talk: It's Time to Get Serious About Mental Illness in Tech, URL: https://www.iqmetrix.com/blog/lets-talk-its-time-to-get-serious-about-mental-illness-in-tech#:~:text=According%5C%20to%5C%20OSMI%5C%20data%5C%2C%5C%2051,National%5C%20Alliance%5C%20on%5C%20Mental%5C%20Illness. (visited on 12/01/2020).

[24] Greenbaum, Zara: Catering to the needs of workers in big tech, in: Monitor on Psychology 2 (2019), 64.

policies surrounding mental health. However, we did not see any improvement in certain aspects, such as offering mental health benefits, educating employees about the options for mental health care that the employer provides, and protecting employee anonymity if they choose to take advantage of mental health resources. In addition, there was no significant positive shift in perceptions regarding discussing mental health in the workplace over the years. The change in the percentage of employees who said that they would be willing to discuss mental health issues with their coworkers or supervisors over time was insignificant. Employees also still indicated that tech employers were doing a less than mediocre job of supporting employees who had mental illnesses, as most respondents gave their tech employers a support rating of 1, 2, or 3 out of 5 in every year.

We also compared several models for predicting whether or not an employee has been treated for a mental disorder and also whether or not an employee currently has a mental disorder. The results from the prediction of *current_mental_disorder* seem to be more useful because they indicate whether or not a patient could currently have a mental illness. Being treated for a mental disorder or not does not dictate the time frame, so the individual could have gotten treatment earlier in their life and no longer need it now. In addition, model performance for predicting *current_mental_disorder* was better for every single model compared to predicting *treatment*. Nevertheless, if one wanted to use the *treatment* prediction models, the model performance results illustrate that Polynomial SVM should be used out of all the models, since it performs the best (with the exception of the Random Forest model) on the training set, the best on the testing set, has the highest AUC score, and has the second highest recall. If we wanted to put more weight on the recall, then the Decision Tree model would be the most ideal, since it has the highest recall. For predicting whether or not an employee currently has a mental disorder, the L1 Penalized Logistic Regression has the best overall performance. It has the second highest training accuracy, the highest testing accuracy, and the highest AUC score. Its recall score is also relatively high compared to the other models. This is consistent with one of the previous studies conducted by Pandey et al., which also concluded that the Logistic Regression model was the best for predicting mental health problems and suicide tendency.[25] However, the K-Nearest Neighbors model which yields an extremely high recall could be useful in certain cases. In cases where false negatives are intolerable but false positives are okay, this model would be extremely useful for predicting mental health problems. In the case of mental illness, false positives seem to be more tolerable, as the first treatment for a mental illness is going to see a therapist, rather than some life altering treatment. A medical professional could in theory use this model if they wanted to be sure about catching almost everyone who could potentially have a mental illness.

In terms of good predictors for current mental illness, the models seem to indicate that family history and gender are very important predictors of mental illness. Another variable that had a notably high feature importance had to do with whether employees knew about the mental health care options that their employer provides (*care_options*). The fact that family history and gender are the most important predictors in the models shows that we need more historical and demographic information to yield more accurate predictions. However, our models still show that tech workplace characteristics and perceptions can be important predictors of mental illness.

The K-Nearest Neighbors or L1 Penalized Logistic Regression models could potentially be used by mental health professionals who are working with tech companies that want

[25] Pandey, S. / Khan, M. B. / Thakkar, S.: Factors Affecting Mental Health in Employees and Their Relation to Suicide Rates at a Workplace (2019).

to focus more on mental health and provide employees with more targeted resources. A mental health professional could feasibly provide a similar survey to employees and use this data to predict which of the employees might possibly have a mental illness. In addition, both of these models are very explainable to those who might not have as much data science experience. The predicted probabilities from the models rather than a strict 1 or 0 cutoff between having a mental illness or not could potentially be useful to get a measure of the risk of an individual developing a mental illness. For instance, a probability that is closer to 1 would indicate that they have a higher potential risk of developing a mental illness. The threshold for when an intervention is necessary should then be determined by the mental health professional. Nevertheless, this could bring up ethical and privacy concerns if employees are not willing to reveal their mental health condition to their employers. It would be very necessary in this case to protect employee data and to make clear exactly what the predictions would be used for.

The models presented in this study are limited by their accuracy, although K-Nearest Neighbors has a very good recall score for predicting whether or not an employee currently has a mental illness. In the future, more demographic characteristics and specific information about the employees' family and clinical histories would likely be necessary to improve model accuracies. Additionally, the models could be further improved by specifically using methods to tune the hyperparameters of the models. There are various methods such as Exhaustive Grid Search and Randomized Parameter Optimization[26] that could efficiently tune the models to be as accurate as possible.

6 Conclusion

In this study, we were able to quantitatively measure the trends of how the tech industry has addressed mental health over time through OSMI yearly surveys. The fact that there hasn't been an improvement in certain factors like protecting employee anonymity if they take advantage of mental health resources is a cause for concern, especially since mental health is being discussed more freely in the media in recent years. There also seems to still be a stigma around discussing mental health with coworkers or in interviews with potential employers that has been the same since 2014. These factors that we measured over time were also used as predictors for developing models to predict mental illness in employees. It can be seen from the models that certain characteristics of tech employers (related to mental health) are good predictors of mental illness in employees, in addition to other factors like gender and family history. Certain models like K-Nearest Neighbors and L1 Penalized Logistic Regression seemed to perform relatively well and could potentially be used to predict mental illness based on the characteristics given in a survey like the OSMI surveys. These models could also be generalized to other workplaces, not necessarily tech companies, as none of the characteristics used in modeling were specific to the tech industry. Tech companies still need to do a better job of supporting their employees' mental health, and this study provides insights about what tech companies need to improve on and a model to help employees who potentially have a mental illness get the help they need.

[26] Scikit-learn, Tuning the hyper-parameters of an estimator, url: https://scikit-learn.org/stable/modules/grid_search.html (visited on 12/01/2020).

References

Ambielli, Brian, Gini Impurity (With Examples), url: https://bambielli.com/til/2017-10-29-gini-impurity/#:~:text=Gini%5C%20Impurity%5C%20is%5C%20a%5C%20measurement,labels%5C%20from%5C%20the%5C%20data%5C%20set. (visited on 12/01/2020).

Band, Amey, How to find the optimal value of K in KNN?, url: https://towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb (visited on 12/01/2020).

Barkved, Kirsten, Let's Talk: It's Time to Get Serious About Mental Illness in Tech, url: https://www.iqmetrix.com/blog/lets-talk-its-time-to-get-serious-about-mental-illness-in-tech#:~:text=According%5C%20to%5C%20OSMI%5C%20data%5C%2C%5C%2051,National%5C%20Alliance%5C%20on%5C%20Mental%5C%20Illness. (visited on 12/01/2020).

Commission, U.S. Equal Employment Opportunity, Diversity in High Tech, url: https://www.eeoc.gov/special-report/diversity-high-tech?renderforprint=1#:~:text=Compared%5C%20to%5C%20overall%5C%20private%5C%20industry,percent%5C%20to%5C%208%5C%20percent)%5C%2C%5C%20and (visited on 12/01/2020).

Defazio, Aaron / Bach, Francis R. / Lacoste-Julien, Simon: SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives, in: CoRR (2014) url: http://arxiv.org/abs/1407.0202.

Greenbaum, Zara: Catering to the needs of workers in big tech, in: Monitor on Psychology 2 (2019), 64 url: https://www.apa.org/monitor/2019/02/workers-big-tech.

Kessler, Ronald C.: Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration, in: International Journal of Methods in Psychiatric Research 3 (2017), e1575 url: https://onlinelibrary.wiley.com/doi/abs/10.1002/mpr.1575. (e1575 IJMPR-May-2017-0053.R1)

McCarthy, J. F.: Predictive Modeling and Concentration of the Risk of Suicide: Implications for Preventive Interventions in the US Department of Veterans Affairs, in: Am J Public Health 9 (2015), 1935–1942 url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4539821/.

of Mental Health, National Institute, Mental Illness, 2020, url: https://www.nimh.nih.gov/health/statistics/mental-illness.shtml (visited on 12/01/2020).

Organization, World Health, Mental health in the workplace, url: https://www.who.int/teams/mental-health-and-substance-use/mental-health-in-the-workplace (visited on 12/01/2020).

Pandey, Shambhavi / Khan, Mohammad B. / Thakkar, Sagar: Factors Affecting Mental Health in Employees and Their Relation to Suicide Rates at a Workplace, in: (2019) url: https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3966-2019.pdf.

Pedregosa, F.: Scikit-learn: Machine Learning in Python, in: Journal of Machine Learning Research (2011), 2825–2830.

Ritchie, Hannah, Global mental health: five key insights which emerge from the data, 2018, url: https://ourworldindata.org/global-mental-health#:~:text=Around%5C%201%5C%2Din%5C%2D7%5C%20people,4%5C%20percent%5C%20of%5C%20the%5C%20population. (visited on 12/01/2020).

Scikit-learn, Pairwise metrics, Affinities and Kernels, url: https://scikit-learn.org/stable/modules/metrics.html (visited on 12/01/2020).

Scikit-learn, Tuning the hyper-parameters of an estimator, url: https://scikit-learn.org/stable/modules/grid_search.html (visited on 12/01/2020).

Tate, Ashley E.: Predicting mental health problems in adolescence using machine learning techniques, in: PLOS ONE 4 (2020), 1–13 url: https://doi.org/10.1371/journal.pone.0230389.

Training, Recruiting Innovation Online, Diversity in Tech: We've Got a Long Way to Go, url: https://recruitinginnovation.com/blog/diversity-in-tech/ (visited on 12/01/2020).

Appendix

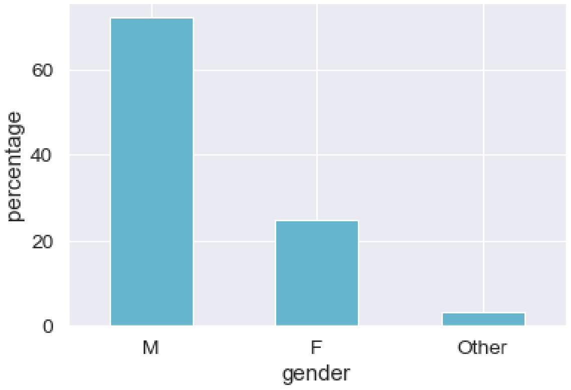

Figure A.1. Percentage of individuals in each gender category.

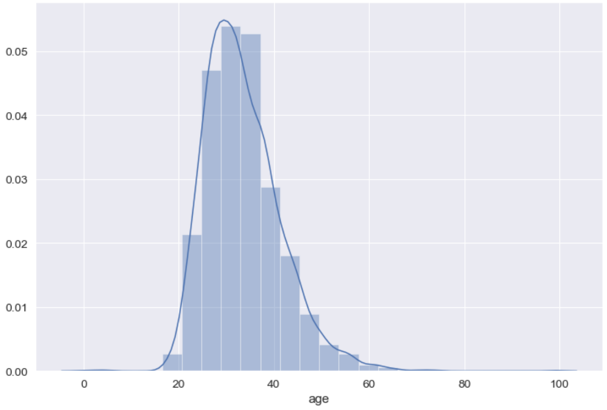

Figure A.2. Distribution of age.

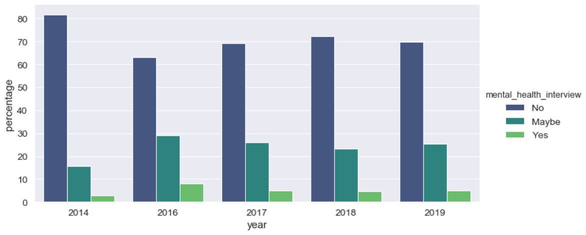

Figure A.3. *mental_health_interview* response distribution trend from 2014-2019.

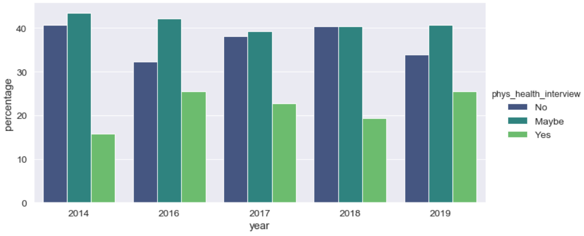

Figure A.4. *phys_health_interview* response distribution trend from 2014-2019.

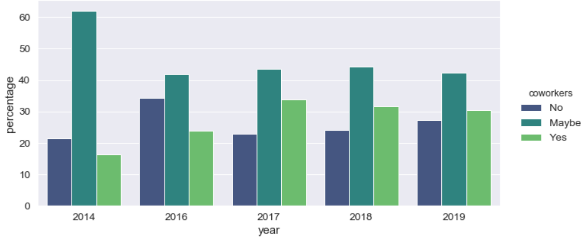

Figure A.5. *coworkers* response distribution trend from 2014-2019.

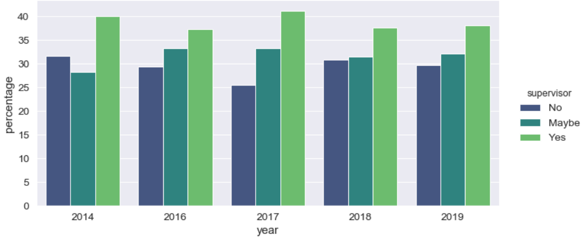

Figure A.6. *supervisor* response distribution trend from 2014-2019.

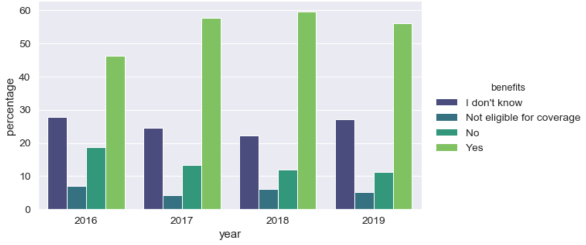

Figure A.7. *benefits* response distribution trend from 2016-2019.

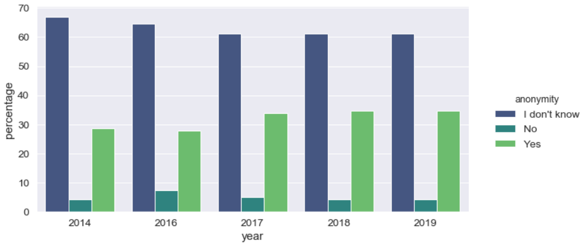

Figure A.8. *anonymity* response distribution trend from 2016-2019.

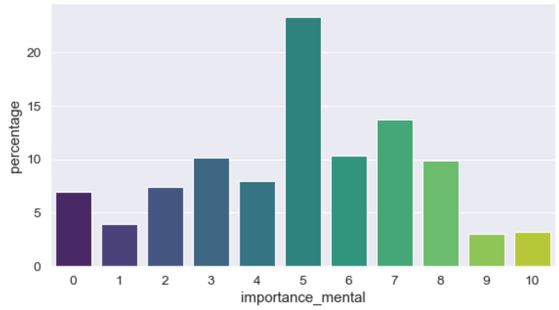

Figure A.9. *importance_mental* response distribution with data from 2017-2019.

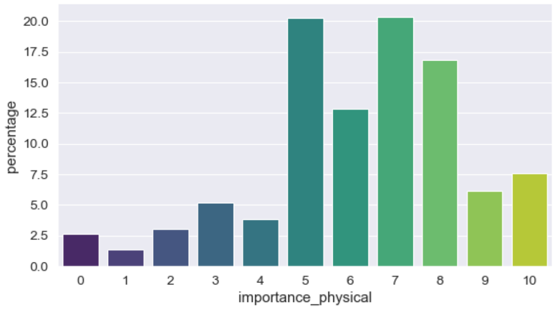

Figure A.10. *importance_physical* response distribution with data from 2017-2019.

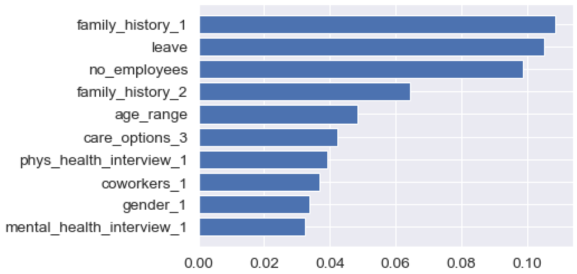

Figure A.11. Most important features in Random Forest predicting *treatment*.

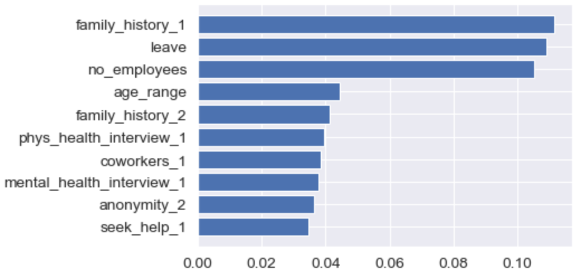

Figure A.12. Most important features in Random Forest predicting *current_mental_disorder*.

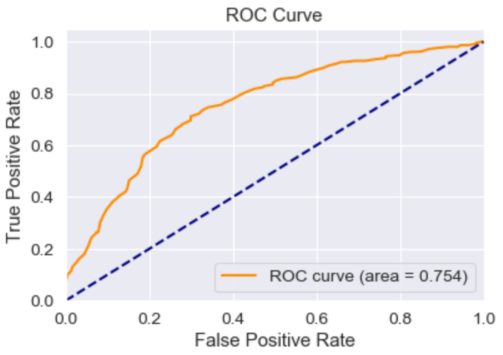

Figure A.13. Random Forest ROC curve for predicting *current_mental_disorder*.

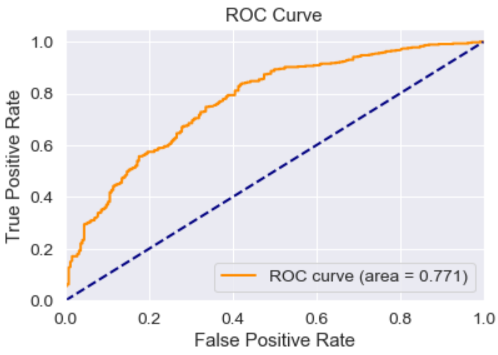

Figure A.14. Naive Bayes ROC curve for predicting *current_mental_disorder*.

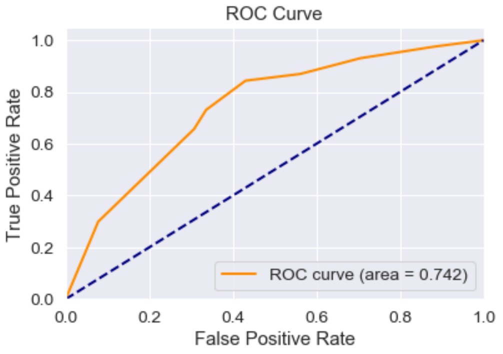

Figure A.15. Decision Tree ROC curve for predicting *current_mental_disorder*.

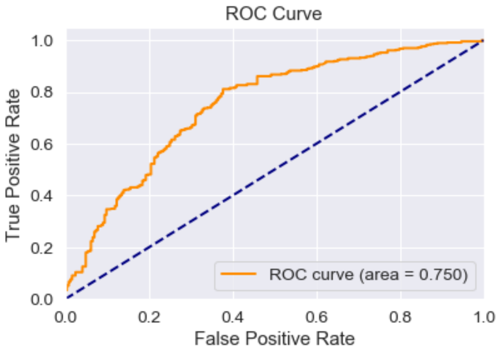

Figure A.16. Poly SVM ROC curve for predicting *current_mental_disorder*.

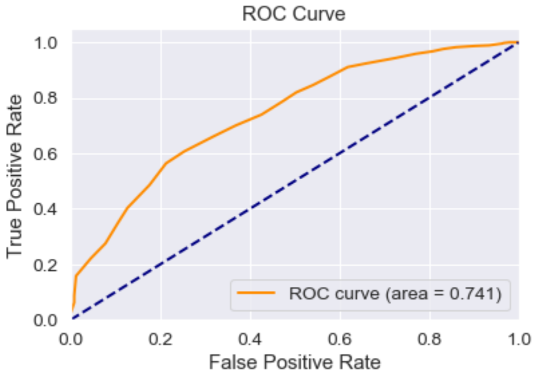

Figure A.17. KNN ROC curve for predicting *current_mental_disorder*.

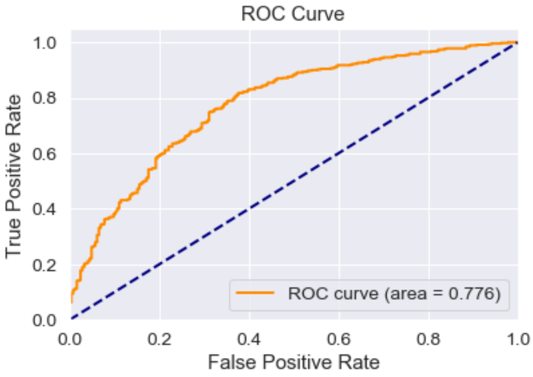

Figure A.18. Logistic Regression ROC curve for predicting *current_mental_disorder*.

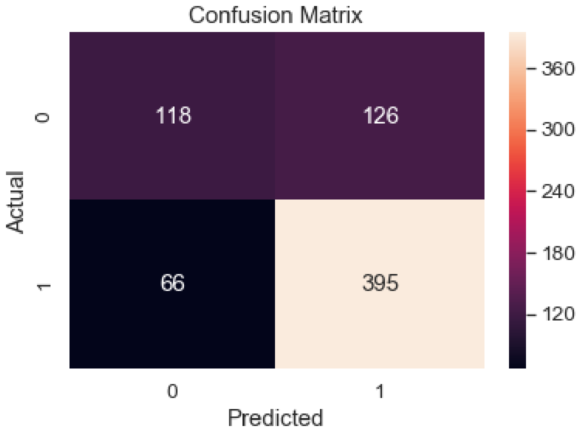

Figure A.19. Random Forest confusion matrix for predicting *current_mental_disorder*.

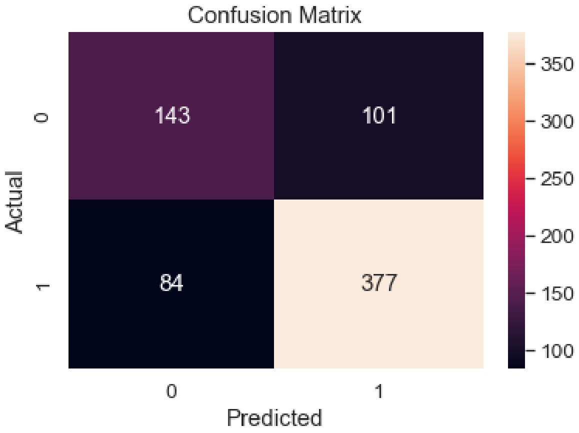

Figure A.20. Naive Bayes confusion matrix for predicting *current_mental_disorder*.

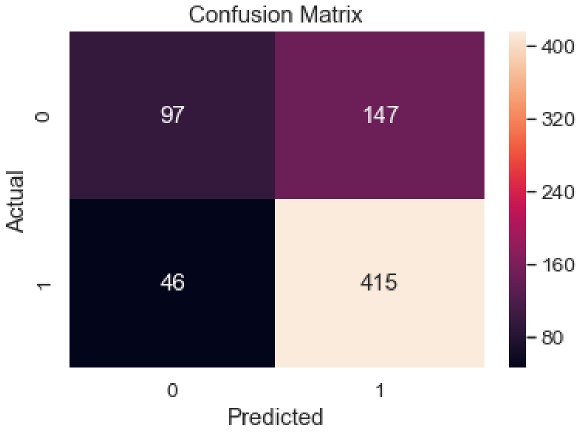

Figure A.21. Poly SVM confusion matrix for predicting *current_mental_disorder*.

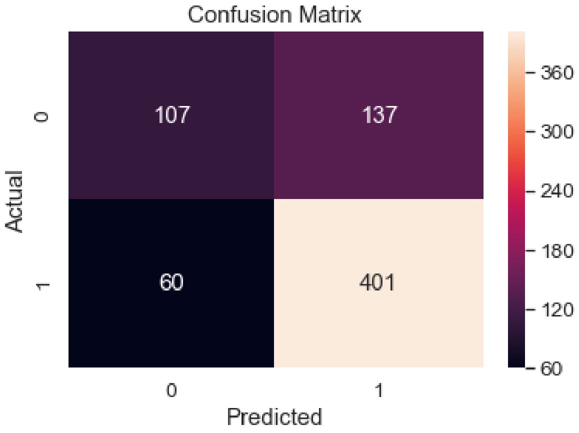

Figure A.22. Decision Tree confusion matrix for predicting *current_mental_disorder*.