Prediction of Mental Health in Tech Survey

1st Akhila Vemmana
Department of Computer Science (of Aff.)
University of North Carolina at Charlotte (of Aff.)
Charlotte, India
avemana@uncc.edu

2nd Jaya Sindhura Sadam
Department of Computer Science (of Aff.)
University of North Carolina at Charlotte (of Aff.)
Charlotte, India
jsadam@uncc.edu

3rd Neela Ayshwarya Alagappan
Department of Computer Science (of Aff.)
University of North Carolina at Charlotte (of Aff.)
Charlotte, India
nalagapp@uncc.edu

4th Srinath Muralinathan
Department of Computer Science (of Aff.)
University of North Carolina at Charlotte (of Aff.)
Charlotte, India
smuralin@uncc.edu

Abstract: The well-being of a person is the measure of mental health. The increasing use of technology will lead to a lifestyle of less physical work. Also, the constant pressure on an employee in any industry will make more vulnerable to mental disorder. These vulnerabilities consist of peer pressure, anxiety attack, depression, and many more. Here we have taken the dataset of the questionnaires which were asked to an IT industry employee. Based on their answers the result is derived. Here output will be that the person needs an attention or not. Different machine learning techniques like logistic Regression, Decision Tree classifier and Random Forest are used to get the results. The idea behind this project is to analyze mental health of technology employees to categorize deadly disease. Collect data across globe relating to mental health and perform exploratory data analysis and create visualizations. This prediction also tells us that it is very important for an IT employee to get the regular mental health checkup to tract their health. The employers should have a medical service provided in their company and they should also give benefits for the affected employees. Index Terms— mental health in tech, missing values, exploring dataset, logistic regression, decision tree classifier, random forest, confusion matrix.

I. INTRODUCTION

Mental health is the aggregation of emotional, social and psychological well-being of a person. Its effects on the person's thinking, acting and feeling capability. Mental health is a measure of handling stress and decision making with every step-in life. Mental Health is very important factor in every stage in life whether it be childhood or an adult. Mostly mental health is something which never discussed publicly, and no proper awareness is there in society. People would generally not talk about it in public. Mental health could affect one's thinking and behavior. Some common reasons of instable mental health could be: Past life experiences, such as ragging or bullying, Biological factors, such as genes and Hereditary problem from family [1]. People with mental disorder often face other anxiety disorder which eventually develops into depression. Hence authors are interested in online communities for data. They have crawled data from 247 online communities of 80,000 users. Then they have extracted the psycho-linguistic posts based on topics, which served as input to model. Machine learning techniques are applied to generate joint model for identifying mental health related features. At last they performed empirical validation of model on dataset where model performs best in recent techniques. Observe insights from visualization created using ML tools and techniques to draw conclusion on deadly mental diseases around the world. What are the strongest predictors of mental health illness or certain attitudes towards mental health in the workplace?

II. RELATED WORK

Machine learning and artificial intelligence can be applied to betterment of the society. The authors have proposed many machine learning techniques like support vector machine, navies Bayes, decision tree, K-nearest neighbor and logistic regression. Here the targeted group of people are students and working professionals [2]. There is a questionnaire which is asked to these people. Then unsupervised techniques were applied, and mean opinion score is calculated. Symptoms of mental disorder could be easily observed on social media platform. These are being automatically able to locate by methods. Here the

authors studied about social media activity to detect depression and other mental illness. Mentally ill people are distinguished by their membership in online forums, online screening or community distribution analysis on Twitter. These users are detected by their regularity in online presence and their use of languages. Many methods could be applied to detect mentally ill people on social media. Currently over 1300 responses, the ongoing 2016 survey aims to measure attitudes towards mental health in the tech workplace and examine the frequency of mental health disorders among tech workers. The dataset is taken from Kaggle, https://www.kaggle.com/osmi/mental-health-in-tech-survey. The dataset contains 27 fields like Timestamp Age, Gender, Country, family history, treatment, work interference, seek help, benefits, care options,mental health consequence, obs_consequence. The project is implemented using python jupyter notebook and visualization of analysis is carried out in tableau. Here the dataset used is a survey taken among the IT professionals from different regions. It mainly has information like age, gender, location of work, type of work, is he/she self-employed? and many more. Then now for machine learning to apply, we first need to remove unnecessary fields like comments and timestamp. This step is called as data cleaning. Data cleaning is the process by which data gets rid of from not required data, making it appropriate for further analysis. Incorrect format, errors while capturing, missing data acts as garbage data. Many of the attributes have empty values as input so default values will be assigned to it. For Integer it is 0, float is 0.0 and string is Nan. Now for gender attribute we have make it all in standard format by replacing any unknown inputs to standard input. The next step is encoding the data. Now after data is again checked whether nay data is missing or not. Then the dataset is scaled and fitted.

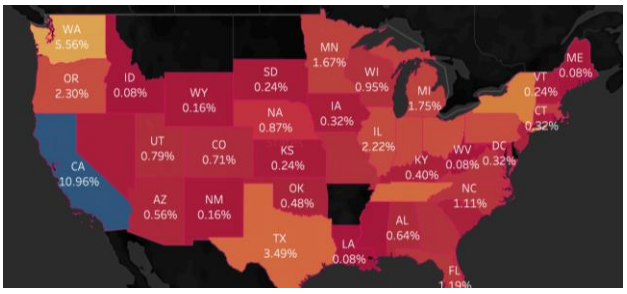

Fig 1: Distributing of mental health cases in the USA

Upon Exploring the dataset, it was observed that people between 25-30 years have highest frequency of mental issues. [3] [4] **2 out of 5** respondents said they have a current mental health problem Only **31%** feel that an employer takes mental health as seriously as physical health **68%** are unsure if their anonymity was protected on taking advantage of mental health treatment resources. Most prevalent conditions diagnosed were Anxiety Disorder and Mood Disorder.

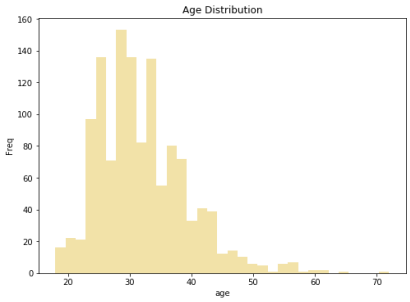

Fig 2: Age distribution

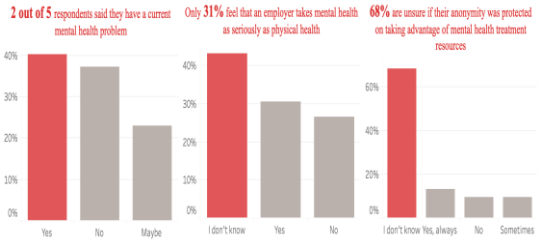

Fig 3: Further analysis

Fig 4: Most prevalent conditions

Now Label encoding is done to convert categorical text data into model-understandable numerical data. Import used from sklearn.preprocessing import Label Encoder. feature scaling is applied on the dataset. It basically helps to normalize the data within a particular range. from sklearn.preprocessing import MinMaxScale. Features Scaling age, because is extremely different from the others, scaler= MinMaxScaler(),train_df['age']=scaler.fit_transform(train_df[['age']]),train_df.head(). Machine learning techniques are applied and compared that which suits the dataset best. Since the dataset had categorical variables, supervised learning method was employed. [4] Three different modelling techniques were applied to the cleaned and preprocessed dataset. The first algorithm is logistic regression. Logistic regression is

a sigmoid function having an S-shaped curve that takes any real value and maps to value from 0 to 1. The equation is y = e^ (b0 + b1*x) / (1 + e^ (b0 + b1*x)). The second algorithm is Decision Tree classifier. In similar fashion, training was done using decision tree modelling technique on the split training data set. Accuracy score was calculated after prediction of testing set.

III. RESULT & EVALUATION

A **confusion matrix** is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. Used to evaluate quality of the output of a classifier on the data set. True positive (*TP),* True negative (*TN),* False negative (*FN), False positive (FP)***Accuracy metric: I**t's the ratio of the correctly labeled subjects to the whole pool of subjects. ***Accuracy = (TP+TN)/(TP+FP+FN+TN)***. In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values.

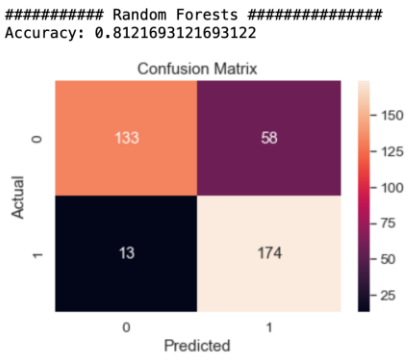

Fig 5: Random Forest

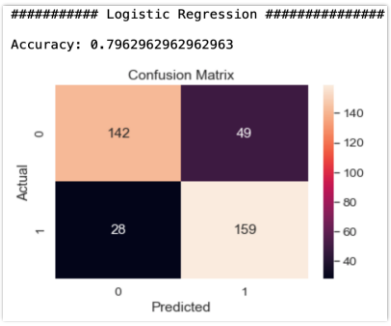

Fig 6: Logistic Regression

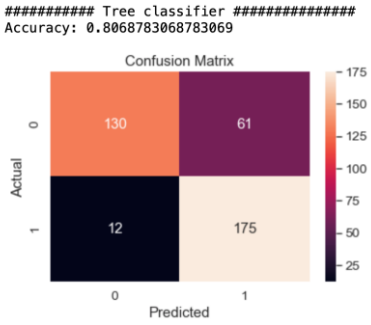

Fig 7: Tree Classifier

IV. MENTAL HEALTH AMERICA ANALYSIS

Online depression screening analysis: Screening improves the chances of getting treatment. Primary care physicians providing usual care miss 30% to 50% of depressed patients and likely fail to recognize many common mental health disorders. However, when results from a positive screening are included in the chart, doctors were over 3 times more likely to recognize the symptoms of mental illness and to plan to follow-up with people about their mental health disorders. [5][6] The screening most often taken by users online has been the depression screen (the Patient Health Questionnaire-9 or PHQ-9). Today, an average of 2,700 individuals come online to take a screen per day, and about 50 percent of those screens are depression screens. The following information includes analysis of our state level data from our depression screens from May 2014 through December 2016) and demographic data analysis from 2016. The PHQ-9 asks the questions below. For each question, individuals check among the following options: Not at all, Several Days, More than half the days, and Nearly Every day. Over the last 2 weeks, how often have you been bothered by any of the following problems? Little interest or pleasure in doing things, Feeling down, depressed, or hopeless, Trouble falling or staying asleep or sleeping too much, Feeling tired or having little energy, Poor appetite or overeating. Results are categorized based on scores. 1-4 = Minimal depression, 5-9 =Mild depression, 10-14 = Moderate depression, 15-19 = Moderately severe depression, 20-27 = Severe depression. Demographic data analysis: In 2016, 1,036,543 individuals visited MHA's website to take a screening test. This section breaks down the results from the Depression screenings by demographics. Over 44 million, or 18% of people will experience a mental health condition every year. Within this larger population, variation exists among sub- populations. Using an intersectional framework allows for an in-depth analysis of mental health trends. The intersection of sex, age, income, and

sexual orientation is a factor that should be considered when assessing prevalence rates and identifying potential barriers to treatment.

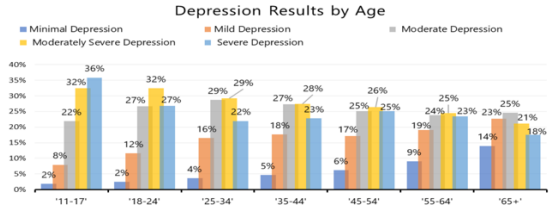

Fig 8: Depression Results by Age

Youth is at great risk. Sixty-two percent of Female youth scored Moderately Severe Depression or Severe Depression. This was the case for 52% of Male youth. The 2018 State of Mental in America Report continues to show a negligent response to youth who require treatment for Severe Depression. [2][3] On average, it takes 10 years between the onset of symptoms and when individuals receive treatment. Given that this population is more likely to engage in risky behavior, it is important that mental health services and treatments be made available and accessible. A timely response to the mental health needs of youth, can prevent them from entering adulthood in crisis.

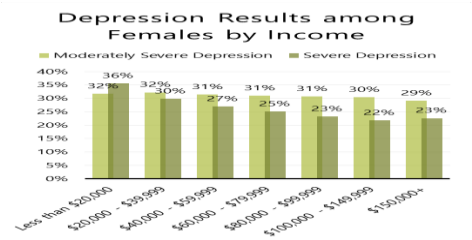

Fig 9: Depression among females by income

Females and males reporting household income of less than $20,000 a year scored highest rates of Severe Depression. Outreach and awareness are critical among members of special populations.

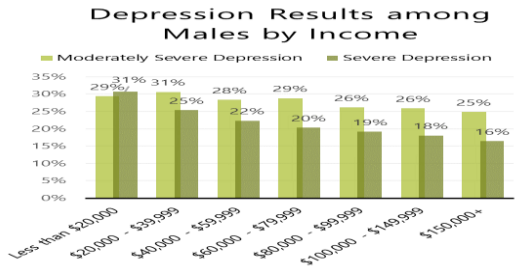

Fig 10: Depression among males by income

Those with a greater need for treatment, cannot afford it. Depressive symptoms are common among individuals who are afflicted with chronic financial stress. The highest percentage of men and women who scored severely depression earned a household income of less than $20,000. Next steps: Although most screeners report that they would take NO action following their results, even among screeners who score with Severe Depression, another third report that they will discuss the results with a family member, a friend or a professional. Young screeners (11-17) were least likely to seek treatment, and most likely to take no action. Screeners aged 35- 44 were more likely to find treatment or discuss the results with someone. Screeners aged 55-64 were most likely to want to conduct additional research online and those 65+ were most likely to monitor their health. This is a particularly vulnerable population that often must rely on the actions of adults to address mental health concerns. [7] This may explain the increase in youth ages 18-24 who were more likely discuss their results with someone and seek treatment.

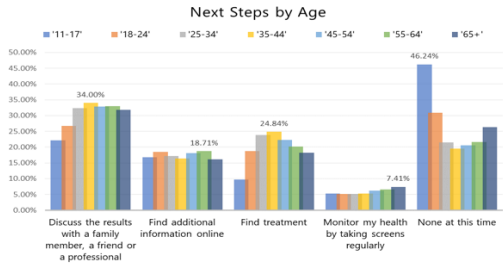

Fig 11: Next steps by age

Males were more likely than females to report that they would do nothing after screening.

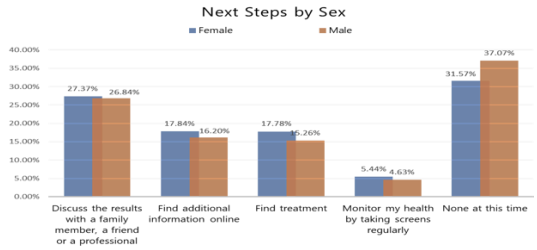

Fig 12: Next steps by sex

Individuals making $100,000 – $149,000 annual income reported most likely to discuss results with someone. Individuals making more than $150,000 annual income are most likely to do nothing. Individuals making less than $40,000 annual income are most likely to want to find treatment.

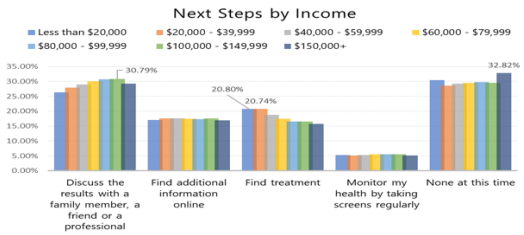

Fig 13: next steps by Income

Individuals making $100,000 – $149,000 annual income reported most likely to discuss results with someone. Individuals making more than $150,000 annual income are most likely to do nothing. Individuals making less than $40,000 annual income are most likely to want to find treatment. Low income screeners want to act. Low-income screeners showed a great interest for taking next steps. For this group, socio-economic barriers may prevent them from acquiring mental health services. Online resources and tools can bridge this gap, providing options for addressing their mental health concerns.[4][5] Overall, individuals that took the depression screening in 2016 had some mental health concern, with the majority screening at high risk for Moderately Severe Depression and Severe Depression. Increasing mental health coverage can increase the number of individuals that are diagnosed and, if necessary, treated before they encounter extreme consequences (including self-harm, substance abuse, incarceration, etc.). In addition to expanding Medicaid, the largest payer for mental health services, providers must negotiate higher reimbursement rates for services, incentivizing more mental health professionals to take private and/or public insurance. Lastly, investment in preventive services, as well as recovery services (e.g. peer services, supportive employment, and supportive housing), would provide support and opportunity for individuals with mental health conditions.

V. CONCLUSION

There are many suggestions that employers and employees could keep in mind. Employers need to keep track of number of their employees having mental disorder. Employers should allow flexible work environment with flexible work scheduling and break timings. They should allow employees to work from home or have flexible place of work. They should give day-to-day feedback and guidance for nurturing employees' health. This type of model could be used to detect metal health progress among employees and also could lead to policy changes. Employees could talk to colleagues and their managers about their problem freely. Hence upper management could help them to get correct aid with beneficiaries like work from home, flexible timings, more leaves, many more. Employees should know health benefits provided by their organization participate in any wellness programs. Proper feedback should be provided to employee when they resign from their job. This could help them to improve their health.

REFERENCES

[1] G. Azar, C. Gloster, N. El-Bathy, S. Yu, R. H. Neela and I. Alothman, "Intelligent data miningand machine learning for mental health diagnosis using genetic algorithm," 2015 IEEE International Conference on Electro/Information Technology (EIT), Dekalb, IL, 2015, pp. 201-206.

[2] B. Saha, T. Nguyen, D. Phung and S. Venkatesh, "A Framework for Classifying Online Mental Health-Related Communities with an Interest in Depression," in IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 4, pp. 1008-1015, July 2016.

[3] T. Simms, C. Ramstedt, M. Rich, M. Richards, T. Martinez and C. Giraud-Carrier, "Detecting Cognitive Distortions Through Machine Learning Text Analytics," 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, 2017, pp. 508-512.

[4] R. Subhani, W. Mumtaz, M. N. B. M. Saad, N. Kamel and A. S. Malik, "Machine Learning Framework for the Detection of Mental Stress at Multiple Levels," in IEEE Access, vol. 5, pp. 13545-13556, 2017.

[5] Christensen, K. S., Toft, T., Frostholm, L., Ørnbøl, E., Fink, P., & Olesen, F. (2005). Screening for common mental disorders: who will benefit? Results from a randomised clinical trial. Family practice, 22(4), 428-434.

[6] Pignone, M. P., Gaynes, B. N., Rushton, J. L., Burchell, C. M., Orleans, C. T., Mulrow, C. D., & Lohr, K. N. (2002). Screening for depression in adults: a summary of the evidence for the US Preventive Services Task Force. Annals of internal medicine, 136(10), 765-776.

[7] O'Connor, E. A., Whitlock, E. P., Beil, T. L., & Gaynes, B. N. (2009). Screening for depression in adult patients in primary care settings: a systematic evidence review. Annals of Internal Medicine,151(11),793-803.