Mental Health Treatment Prediction using Machine Learning Techniques

Parmar Nirav, Patel Deep, Patel Dhruv and Patel Nihar

Dept. of Computer Science and Engineering,
Institude of Technology
Nirma University, Ahmedabad

Abstract. Excessive stress, lengthy workdays, pressure to succeed, the need to establish one's reputation, and an unbalanced work schedule between personal and professional life are all traits of mental health illnesses. Employees' mental problems are growing more prevalent as a result of workplace stress. A few more problems that can be brought on by mental illness are personality disorders, anxiety disorders, phobias, depression, mood disturbances, and eating disorders. The causes of mental health disorders among employees were examined in this paper. In this paper, we utilize the data from the mental health survey 2014 that contains data on mental health in tech workspace, which gathered information from a wide range of people around the world. This dataset included 27 features in it. The missing values data preprocessing approach is applied to the dataset to provide cleaned data with all feature values filled. Here, we used machine learning algorithms to analyze the severity of mental disorders for working employees based on a variety of factors or attributes, such as self-employment, mental health history in the employee's family, company offering positive health effects, whether the employee is receiving treatment for mental illness, and much more. To identify the model with the highest degree of accuracy, we employ a variety of machine learning techniques, including Logistic Regression, KNN, Naive Bayes, Decision Tree, SVM, Random Forest and Voting Classifier. By using Decision Trees, prominent features that influence stress were identified as work interface, family history, benefit and care option in the workplace. With these results, industries can now narrow down their approach to reduce stress and create a much comfortable workplace for their employees.

Keywords: Stress prediction, Mental Health, Healthcare, Machine Learning, KNN, Multinomial Naive Bayes, SVM, decision tree, Random Forest

1 INTRODUCTION

To achieve global development goals, mental health knowledge is crucial, according to WHO. World Health Organization (WHO) supports and recognizes World Mental Health Day, which is observed on October 10 each year, to increase global awareness of mental health concerns and address them as effectively as possible.

Like physical health, mental health is a subcategory of health that deals with the delicate balancing act between emotional and mental health and illness. One of the most prevalent issues with mental health is depression. Suicide is the second leading cause of death among people aged 15 to 29. People who suffer from mental illnesses may avoid getting the right support or medical attention because they feel ashamed. For this reason, research on mental health is necessary to raise public awareness of the issue, particularly in the fields of technology where many businesses are actively hiring.

Mental illness is a health problem that affects our mental health, affecting our ability to feel, think, communicate and act, also known as a mental disorder. Many people experience mental health problems from time to time, but when persistent signs and symptoms often cause stress, the problem becomes a mental illness. It can cause problems in offices, schools, and workplaces. There are various types of mental health problems. It includes Anxiety Disorders, Bipolar Disorders, Depressive Disorders, Obsessive- Compulsive Disorders.

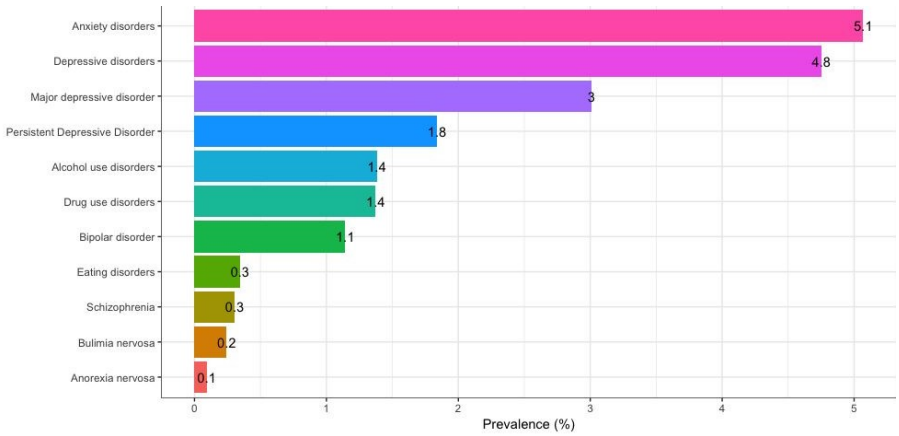

Fig. 1: UK Prevalance of Common Mental Health Disorder(2016)

Experiencing occasional anxiety attacks is a normal part of life. However, people with anxiety disorders often have intense, excessive, and persistent worries and fears in everyday situations. Some symptoms include rapid heart rate, rapid breathing, sweating, shivering, and feeling weak or tired. Depressive disorders are conditions that affect your body, mood, and thoughts. It reduces motivation and interferes with the normal functioning of daily life. This causes continuous sadness. Some Symptoms include sadness, tears, emptiness, anger, irritability, and frustration, over small matters. Bipolar disorder is a mental illness that causes abnormal mood swings, energy, concentration, and ability to perform daily tasks. Episodes of mood swings may occur infrequently or several times

a year. Some symptoms include Change in appetite, Agitation, Poor concentration and judgment, Rapid speech, Sadness, and crying. Obsessive Compulsive Disorder is a common, chronic, long-lasting disorder in which a person has uncontrollable repetitive thoughts and behaviors that a person wants to repeat over and over again. Some symptoms include Washing hands, cleaning, Checking, Counting, and Adhering to a strict routine.

For increased productivity and the well-being of the workers, maintaining a stress-free workplace must be given priority. We aim to simplify this process by using machine learning techniques to build a model that predicts the likelihood that stress will be experienced and whether a person needs treatment by taking some of his or her professional and personal factors as parameters collected in the form of carefully crafted surveys. We can also, perform early prediction if a person requires treatment for his mental health or not.

2 LITERATURE SURVEY

The XG Boost method performs better than GLMNet and other recently discovered methods for re-admitting individuals with mental health concerns. For the group of patients who utilized a conventional identification procedure [1], it shows a higher bias and is more accurate than the widely recognized forecasting model. In paper [2], the author provided guidelines for predicting anxiety disorders based on factors like a person's working environment and various other personal factors, and the prediction was carried out using logistic model trees. This offers more accuracy and is a hybrid model built on decision trees and logistic regression. Boosting and other ensemble techniques, followed by the random forest algorithm, helped to improve recognition accuracy and reliability overall. Family medical history of mental illnesses is more valuable than other factors when forecasting how a person may develop mental disease difficulties [3]. In [4], the author employed a smartphone-based sensor system to track or identify changes in the states of patients with bipolar disorder. Additionally, provide an early warning system with 97% accuracy and recall. According to a review of a few research papers, The SVM model has been the most commonly used algorithm, utilized to analyze stress with a 71% to 97% accuracy range.[5] In [6], the author employed ML algorithms to identify factors that lead to mental stress in working employees. The study indicated that random forest has the highest precision and accuracy (75.13%).

3 DATA DESCRIPTION

Open Sourcing Mental Illness (OSMI) is a non-profit organization that promotes awareness towards mental illness, disorders in the workspace and fights to eradicate the stigma surrounding them. They also help workplaces to identify the best

resources to help their employees in this aspect. As mentioned, OSMI Mental Health in Tech 2014 survey was taken as the dataset, using which we trained different machine learning models to analyze the patterns of stress and mental health disorders among tech professionals and to determine the most influential factors that contribute to the same.

OSMI Mental Health in Tech 2014 survey containing 1259 responses from various employees working in a wide range of tech divisions was used. These responses include both professional and personal factors of the individual and hence will give a complete view of the environment faced by professionals.In healthcare the right dataset is essential to obtain the right model and their performance. Therefore it is important for data analytics to have a good understanding in the dataset and also to extract the relevant information from the dataset.[10]

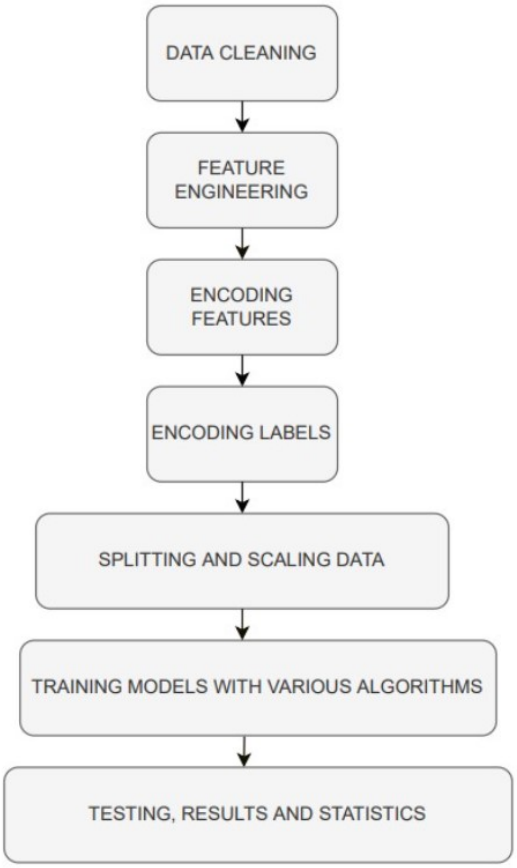

Fig. 2: Flow chart of Implemention

Data Cleaning: The original dataset contains 1259 response from different individuals and 27 attributes spanning both their personal and work life. The data has been cleaned using various standard methods that check for data consistency and validity of the survey responses [7]. Data cleaning is the process of eliminating the data that is redundant, lacking, unnecessary, duplicated, or formatted incorrectly and making it more suitable for our model for prediction. Main advantage of this method is that the maximum relevance criteria along with minimum redundancy criteria is used to choose features that are maximally relevant to the criteria and minimally redundant with respect to the criteria. We found the features having significant amount of null values/empty values out of the Dataframe and dropped them.

Feature Engineering: The preprocessing stage in machine learning known as feature engineering is used to turn raw data into features that may be used to build a prediction model using either machine learning or statistical modeling. Machine learning feature engineering aims to improve model performance. The gender column was filled haphazardly so we decided to engineer it using the first letter of the input and the rest in the category of others. We categorized the features into three categories -binary, ordinal, and multinomial features and encoded them into a numeric format.

Encoding Features: It is more convenient to use numerical data in machine learning algorithms hence it is important to convert the categorical values of the relevant attributes into numerical ones because of this. This method is referred to as feature encoding. We encoded the features having values yes and no as 1 and 0 respectively, ordering index in ordinal features and prefixes in multinomial features.

Encoding Labels: Label encoding is the process of transforming labels into a numeric form so that they may be machine-readable. The operation of those labels can then be better determined by machine learning algorithms.

Feature Scaling: Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. If feature scaling is not done, the machine learning algorithm tends to evaluate smaller values as the lower values regardless of the unit of the data and weigh greater values higher.

70 percent of the responses were used for training the model while the remaining 30 percent was utilized for testing. The data that we took almost had a similar number of candidates for both types of category (i.e. 1 - person requires treatment and 0 - does not need the treatment).

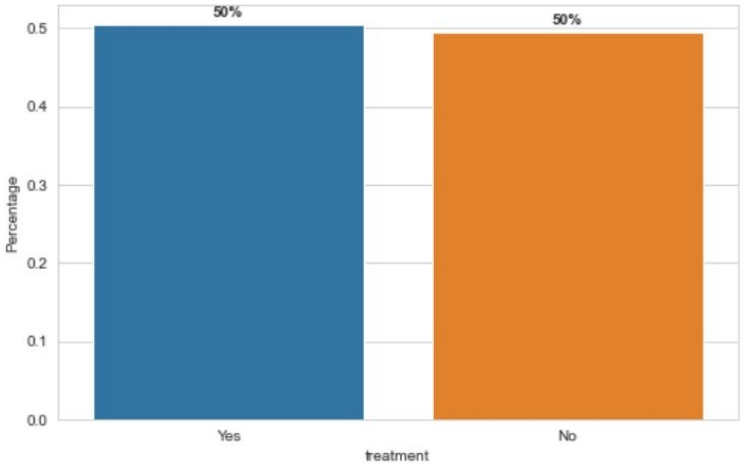

Fig. 3: Get Treatment of Survey Respondents

4 MACHINE LEARNING TECHNIQUES USED

Machine learning is a subset of artificial intelligence that allows computers and computing systems to learn and improve on their own without being explicitly programmed by humans. Machine learning is based on the creation of computing programs capable of retrieving data and learning for themselves. This is extremely useful in healthcare because there is an enormous amount of data, and if this data is properly fed to an intelligent system and trained appropriately, the resulting prediction model will be unparalleled, free of human errors, and reduce the time required for diagnostics. As a result, the OSMI 2014 dataset responses were used to train the following ML models, which have previously been tested in healthcare-based classification problems. We used Pandas for data analysis, NumPy for handling data matrices computation, and SKlearn for predefined ML algorithms.

A. Logistic Regression: Like all regression methods, logistic regression is a predictive analysis. It is used in scenarios where one binary variable is dependent on one or more independent variables. Here, we take the 14 relevant attributes to be independent variables and the possibility of an employee having stress and needing treatment as the dependent variable which is to be predicted by the trained model.

B. KNN Classifier: K-Nearest Neighbor (KNN) classifier is a supervised learning algorithm that can be implemented on labeled data. It was used here for predicting if a person needs treatment or not. KNN classifies the dependent variable based on how similar its independent variables are to a similar instance from the already known data.

C. Decision Trees: A decision tree can be used to model multiple choices or if-else statements/decisions in a tree-like fashion. Here, decision trees are used to find the most contributing factors among the 13 features that are used. This is highly helpful, as now more attention can be given to these areas and necessary steps are taken along those lines.

D. SVM: SVM takes data points as input and gives the output as a hyperplane. It divides the classes using a plane(hyperplane) also known as the decision boundary[9]. Where the decision boundary must maximize the distance of the nearest element of each class. Decision boundary separates the points into different classes.

E. Random Forest Classifier: Random Forests are a cluster of decision trees working together with each other and it has proved to be more effective than a single decision tree. Random Forest is a flexible, easy-to-use ML algorithm that produces a good results persistently, even without hyper tuning.

F. Voting Classifier: A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output. It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creating separate dedicated models and finding the accuracy for each them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

5 RESULT

All the discussed models were implemented in Python using Scikit-learn [10] to test the prediction if a person needs treatment or not. The results are visualized as follows:

	Model	Precision Score	Recall Score	Accuracy Score	f1-score
0	LogisticRegression	0.791878	0.821053	0.801061	0.806202
1	KNN	0.767773	0.852632	0.795756	0.807980
2	Multivariate Bernoulli	0.750000	0.757895	0.750663	0.753927
3	Multinomial	0.769231	0.842105	0.793103	0.804020
4	DecisionTree	0.757447	0.936842	0.816976	0.837647
5	SVC	0.768519	0.873684	0.803714	0.817734
6	RandomForest	0.766234	0.931579	0.822281	0.840855

Fig. 4: Performance Metrics of all implemented models

6 CONCLUSION

After analyzing, we found that Random forest has the best performance. As it has the best accuracy and precision with accuracy 82% and recall 93% .then other models like Decision Tree with 81% accuracy and 93% recall SVC with 80% accuracy and 87% recall Logistic Regression with 80% accuracy and 92% recall KNN with 79% accuracy and 85% recall Multinomial with 79% accuracy and 84% recall Multivariate Bernoulli with 75% accuracy and 75% recall. Also, Feature importance of the selected features showed that a Work interference has the largest contribution . Whether the employee's mental health issues interfering with the work is the thing that the company should ask for its employees. Family history and care options(programs and benefits) provided by company is also influential in employees who want to get treatment. For all the remaining features,there has been a little contribution. noticing/knowing some of these features beforehand can even help support an individual who may be experiencing a mental health issues and and connect them with the appropriate employee resources.

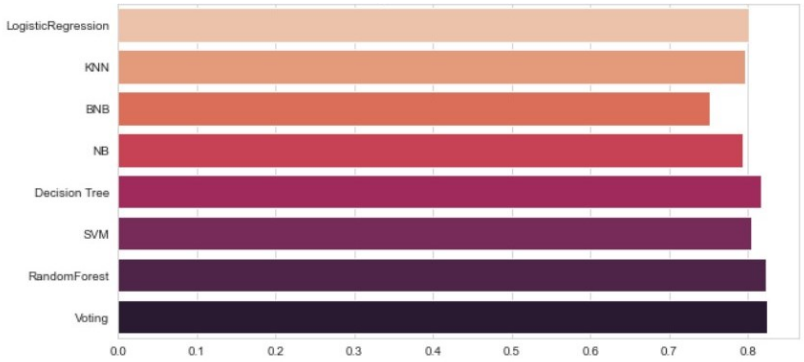

Fig. 5: Ploting All model accuracies

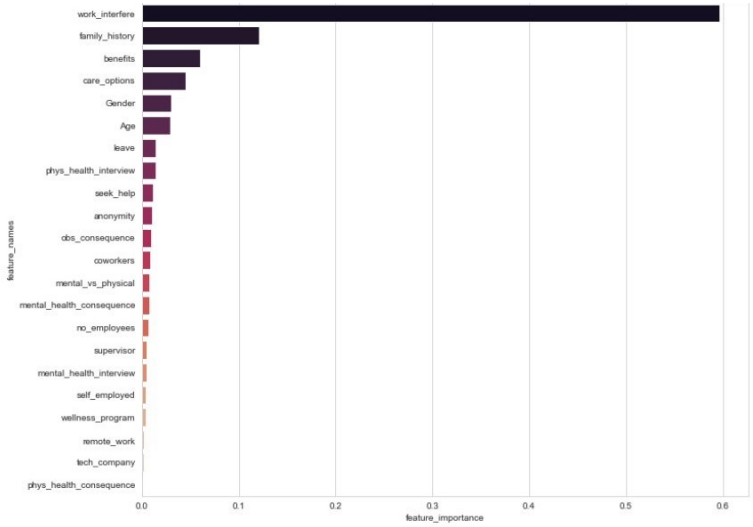

Fig. 6: Key Feature in the dataset

References

1. Didier Morel, Kalvin C. Yu, Ann Liu-Ferrara, Ambiorix J. CaceresSuriel, Stephan G. Kurtz, Ying P. Tabak, "Predicting hospital readmission in patients with mental or substance use disorders: A machine learning approach", International Journal of Medical Informatics, Vol.139, July 2020, pp. 1-11
2. S. Dmonte, G. Tuscano, L. Raut, and S. Sherkhane, "Rule generation and prediction of Anxiety Disorder using Logistic Model Trees," 2018 Int. Conf. Smart City Emerg. Technol. ICSCET 2018, 2018, doi: 10.1109/ICSCET.2018.8537258.
3. A Srinivasulu Reddy, Aditya Vivek Thota, A Dharun, "Machine Learning Techniques for Stress Prediction in Working Employees", IEEE International Conference on Computational Intelligence and Computing Research, 2018, pp.1-4.
4. A. Grünerbl et al., "Smartphone-based recognition of states and state changes in bipolar disorder patients," IEEE J. Biomed. Heal. Informatics, vol. 19, no. 1, pp. 140–148, 2015, doi: 10.1109/JBHI.2014.2343154.
5. Nor Safika Mohd Shafiee, Sofianita Mutalib, "Prediction of Mental Health Problems among Higher Education Students Using Machine Learning", International Journal of Education and Management Engineering (IJEME), pp.1-9, 2020.
6. U. S. Reddy, A. V. Thota, and A. Dharun, "Machine Learning Techniques for Stress Prediction in Working Employees," 2018 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2018, pp. 1–4, 2018, doi: 10.1109/ICCIC.2018.8782395.
7. Van den Broeck, J., Cunningham, S. A., Eeckels, R., Herbst, K.(2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS medicine, 2(10), e267.
8. Gaikwad Kiran P, Dr C M Sheela Rani ,"Regression Model with Modified Linear Discriminant Analysis Features for Bimodel Emotion Recognition", "International Journal of Scientific Technology Research", ISSN : 2277-8616, Volume 9, Issue 3, pp. 1355 to 1360, March 2020
9. A. Goel and S. Mahajan, "Comparison: KNN SVM Algorithm," Int. J. Res. Appl. Sci. Eng. Technol., vol. 887, no. Xii, pp. 2321–9653, 2017, [Online]. Available: www.ijraset.com.
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.