Mental Health Prediction using Data Mining: A Systematic Review

Vidit Laijawala
K J Somaiya Institute of Engineering & Information Technology
University of Mumbai, India
vidit.l@somaiya.edu

Aadesh Aachaliya
K J Somaiya Institute of Engineering & Information Technology
University of Mumbai, India
aadesh.a@somaiya.edu

Hardik Jatta
K J Somaiya Institute of Engineering & Information Technology
University of Mumbai, India
hardik.jatta@somaiya.edu

Vijaya Pinjarkar
K J Somaiya Institute of Engineering & Information Technology
University of Mumbai, India
vkhirodkar@somaiya.edu

***Abstract*: The emotional, psychological and social welfare of a person is revealed by their mental health. It influences how an individual will think, feel or handle a situation. Positive mental health helps an individual to work productively and achieve their full potential. At each point in life, mental health is vital, from childhood to adulthood. Numerous factors contribute to mental health issues which lead to mental illness like stress, social anxiety, depression, obsessive compulsive disorder, drug addiction, workplace issues and personality disorders. The onset of mental illness should be determined without flaws for maintaining an appropriate life balance.**

We have collected data from online available datasets. The data has been label encoded for better prediction. The data is being subject to various machine learning techniques to obtain labels. These classified labels will then be used to build a model to predict the mental health of an individual. Our target population is in the working class i.e people above the age of 18. Once the model is built, it will be integrated to a website so that it can predict the outcome as per the details provided by the user.

Keywords: Random Forest, Decision Tree, Randomization, health.

I. INTRODUCTION

Mental wellness of an individual is the state of mind of that person and also provides an indication of his/her general nature. Mental illness is an outcome of imbalances in brain chemistry. The evaluation of mental wellness is extremely critical to understand and suggest therapies for patients with a deviated mental behavior. Most individuals are prone to stress while some are affected by depression due to various reasons. An administrative panel of World Health Organization (WHO) assessed in 2011 that, by 2030, depression will be the chief source of worldwide disease burden [1]. There is a fundamental change to incorporate the mental fitness outline of an affected individual by healthcare providers and it will be made obligatory in the approaching years to deliver improved medication and also promote speedy recoveries.

Since similar aspects and indications can point to multiple mental health issues, the diagnosis is a complex task and misdiagnosis can happen sometimes. The patient must co-operate for effectively identifying an issue. Detecting mental problem is a difficult task as a misdiagnosis can cause serious issues. Hence, appropriate care must be taken to recognize and treat the mental health issue precisely.

For our project, we have collected data from online available dataset, provided by an OSMI (Open Sourcing Mental Illness) survey. The dataset mainly consists of data of working individuals. It will predominantly be beneficial for employers and employees by creating greater awareness about work related mental illness [2]. We have applied machine learning algorithm to create a model. It shall be implemented on a website for users to get knowledge about their mental illness.

II. LITERATURE REVIEW

In [3] U. S. Reddy et. al. have applied various algorithms to find the most accurate one and compared the relationship between various parameters in the dataset.

In [4] M. P. Dooshima et. al. have used demographic, biological, psychological and environmental factors for prediction. Different mental health experts were consulted to validate the obtained parameters.

In [5] M. Srividya et. al. have used a questionnaire to obtain values for different attributes that can be helpful for prediction of mental health. The motive of this paper was to analyze different algorithms and predict the most accurate one. Various classification algorithms such as Decision Tree, Naïve Bayes as well as SVM were used in this paper. The labels form the data collected were used to compute a MOS. The above algorithms were then applied to find the most accurate one. The paper concluded that Support Vector Machine, K-Nearest Neighbour and Random Forest are the most accurate algorithms with similar accuracy results.

In [6] M. Srividya et. al. have used Neural Networks to predict the psychological conditions of humans such a depression, PTSD, anxiety etc. They also studied the effect of concussion or injuries on sportspersons.

In [7] S. G. Alonso et. al. have conducted extensive review of different algorithms used for mental health prediction. Different techniques such as Association Rule Mining and Randomization were studied and their predictions were noted for our project. This paper also reviewed other algorithms such as SVM, Decision tree, KNN, ANN, Naïve Bayes.

There are different kinds of systems that currently exist. Most of them use different methodologies to predict mental illness. Some of the current systems include an online survey which predicts whether the user has mental illness or not. These surveys are illness specific i.e. a different survey for depression, a different one for stress and so on. All these surveys are available online and anyone can fill them.

There are a few systems which use chatbots to predict mental illness by questioning a user and then analyzing their response. Some systems also use Image Processing to monitor the facial expressions of users, analyze their behavior to a particular question to help in more accurate prediction of mental illness. Most of these surveys include the behavioral and physical aspects of an individual but not any work-related issues. Hence very minimal amount of research has been done on work related mental illness. [8]

Most of these systems are aimed are the general aspects of mental illness. They include the most common parameters to evaluate the result. Systems that focus on the employees and work-related mental illness are not available on a greater scale. [9]

Privacy-Preserving Data Mining algorithms:

- Randomization:

Here the records are shuffled vertically in the way that the semantic meaning or the record in the attribute is not distorted just vertical position of the record is changed hiding the correct identity. We can also add noise to the data so that the behavior of the individual records is masked. Nevertheless, the aggregate behavior of the data distribution can be reconstructed by subtracting out the noise from the data. It is one of the most used and simplest way to protect the data. [10]

- k-anonymity:

Data Anonymization is a technique to perform modification activity on the original data in such a manner that any information that is directly related to any of the individuals that could be deleted. In the k-anonymity method, the granularity of data representation is reduced by using suppression and generalisation techniques. [11]

- l-diversity:

The shortcomings of k-anonymity model are handled by l-diversity. Since protecting identities to the level of k-individuals is not the same as protecting the corresponding sensitive values, especially when there is homogeneity of sensitive values within a group. The anonymisation scheme promotes intra-group diversity of sensitive features. [11, 12]

Table 1. Comparison of existing systems

#	Paper	Algorithms used	Scope
1	U. S. Reddy et. al.	Logistical regression, Decision Tree, Random Forest.	Finds only the accuracy of algorithm. Not implemented.
2	M. P. Dooshima et. al.	Decision Tree Naïve Bayes.	Used WEKA Data of only 30 patients Does not focus on work predominantly.
3	M. Srividya et. al.	Naïve Bayes KNN Decision Tree Logistical regression Random Forest	Focuses mainly on depression, stress and anxiety Some aspects of work related stress considered.
4	D. Filip et. al.	Decision Tree SVM, Random Forest	Considers Stress, PTSD, Traumatic brain injuries related to sports.

III. PROPOSED SYSTEM

Based on the above survey, we propose a system with the primary goal of developing a website where users can enter values in a form and get results about potential or current mental illness based on their input. Firstly, we have collected a dataset which is available online. The data gathered is analysed and pre-processed. The data contains different labels such as age, gender, distance of workplace from home, previous mental illness, family history etc. We have label encoded the data for better prediction. We have planned to apply Random Forest or Decision Tree classification algorithm for classification of the data.

According to our goal, we will design a website where a user shall login and fill up a form which has questions based on the dataset gathered. The user will answer the questions and a result about his/her mental condition will be provided on the website as per the inputs provided. The website makes use of the model that we will build using the machine learning algorithms to provide the output. Since this project makes use of a dataset related to workplace mental illness, it will help raise awareness among employees and employers to provide greater attention to work related stress, depression and proper benefits can be provided to employees suffering from a mental illness.

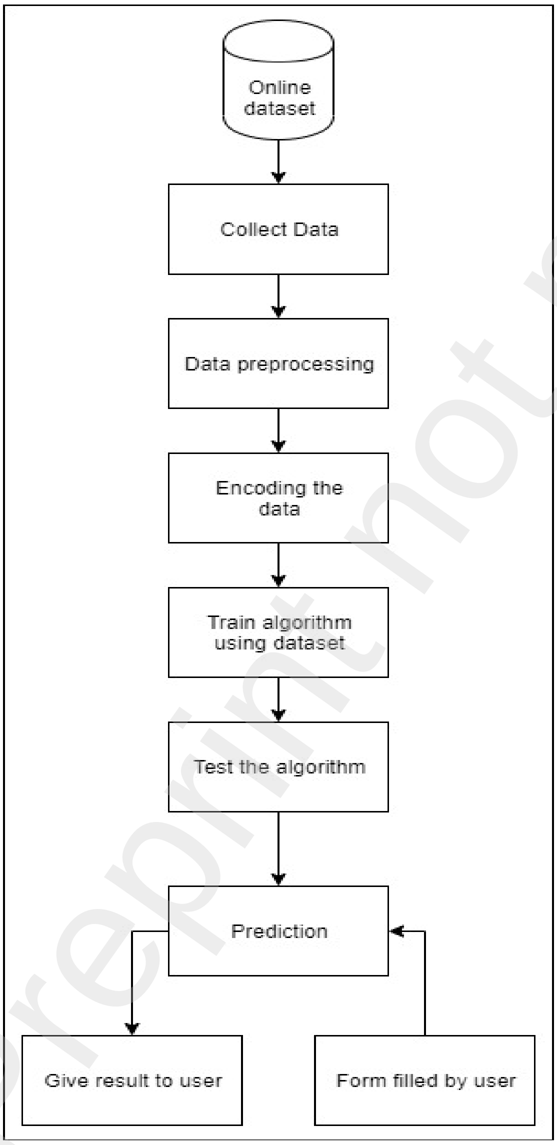

Figure 1: System Block diagram

IV. IMPLEMENTATION AND RESULT

Table 2. Data provided as input

Age	Gen	self_emp	fam_hist	work_int	past	diag	Treat
37	F	N	N	Often	Y	Y	Y
44	M	N	N	Rarely	Y	Y	N
32	M	N	N	Rarely	Myb	N	N
31	M	N	Y	Often	Y	Y	Y
31	M	N	N	Never	Y	Y	N
33	M	N	Y	Sometimes	N	N	N
35	F	N	Y	Sometimes	N	N	Y
39	M	N	N	Never	Y	Y	N
42	F	N	Y	Sometimes	Y	Y	Y
23	M	N	N	Never	Y	Y	N

We have collected data from online available dataset, provided by an OSMI (Open Sourcing Mental Illness) survey. The dataset mainly consists of data of working individuals. The data consists of string attributes which we later encoded to numeric attributes for better prediction. It consists of 26 attributes for prediction and 1 predicting label.

As represented in Table 2, our dataset consists of various parameters to predict mental illness of an employee. Some of the parameters are age, gender, work interference, family history, seeking help, remote work, past mental illness history, past diagnosis, anonymity, consequences etc. and the predicting label is treatment. The data mainly consists of values in terms of Yes and No, which means whether an individual should seek treatment or not.

Table 3. Comparison of Machine Learning algorithms

Algorithm	Accuracy %	Mean Absolute Error	Precision	Time Taken
Decision Tree	82.2	0.256	0.827	0.3 sec
Random Forest	79.3	0.316	0.793	0.6 sec
Logistical Regression	81.4	0.231	0.826	0.5 sec
OneR	82.1	0.178	0.834	0.3 sec
Naïve Bayes	78.7	0.24	0.787	0.5 sec
NNge	75.8	0.242	0.758	0.3 sec

From the above comparison of algorithms, we have found that the most optimal algorithm is Decision Tree because of its low execution time and high accuracy. The above analysis was done using WEKA.

The Decision Tree algorithm uses this dataset and selects the splitting nodes as per the entropies. The node with the lowest entropy is chosen as the root node. As per execution in WEKA, we found that work_int is the root node of the tree. The tree is further split by family history and gender. We found that people whose work is affected due to stress or depression need to consult a mental health professional. However, if the work life of an employee is unaffected, then he/she does not suffer from mental illness. Furthermore, employees whose work is sometimes affected and having a history of mental illness in the family, should seek treatment.

Machine Learning algorithms:

- Decision tree:

Decision tree algorithm is a supervised learning algorithm. It can be used to solve both regression and classification problems. It uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. [13]

- Random Forest:

Random forest algorithm, like its name implies, consists of a large number of individual decision trees that operate together. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. [14]

Tool used:

- WEKA:

Named after a flightless New Zealand bird, Weka is a set of machine learning algorithms that can be applied to a data set directly, or called from your own Java code. Weka comprises of tools for data classification, data pre-processing, association rules, regression, data visualisation and clustering. [15]

V. CONCLUSION AND FUTURE SCOPE

Mental health is an extremely sensitive and important topic currently. It is integral for living a healthy and balanced life. Mental health impacts one's thoughts, behaviour and emotions. It can affect the productivity and effectiveness of an individual. As per the study by WHO, depression will be a major cause of mental illness in the world and people need to take more care about their mental well-being for a balanced social and professional life. People who are hesitant to approach humans for diagnosis can make use of online predictors for results. To do the prediction, we have encoded the data first. We will then implement the most accurate algorithm and later apply Random Forest or Decision Tree to the data for training. Once trained, this algorithm will be tested and model will be designed. It would then be deployed on a webpage where users can fill the forms and get the result accordingly.

VI. REFERENCES

[1]. DEPRESSION: A Global Crisis, WHO, https://www.who.int/mental_health/management/depression/wfmh_paper_depression_wmhd_2012.pdf March '12

[2]. Dataset: https://osmihelp.org/research: 2014 Dataset

[3]. U. S. Reddy, A. V. Thota and A. Dharun, "Machine Learning Techniques for Stress Prediction in Working Employees," *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Madurai, India, 2018, pp. 1-4.

[4]. M. P. Dooshima, E. N. Chidozie, B. J. Ademola, O. O. Sekoni, I. P. Adebayo, A Predictive Model for the Risk of Mental Illness in Nigeria Using Data Mining, *International Journal of Immunology*. Vol. 6, No. 1, 2018, pp. 5-16.

[5]. M. Srividya, M. Subramaniam and B. Natarajan, "Behavioral Modeling for Mental Health using Machine Learning Algorithms" "Journal of Medical Systems" Vol. 42(5):88 May 2018.

[6]. D.Filip & C. Jesus. (2015). A Neural Network Based Model for Predicting Psychological Conditions International Conference on Brain Informatics and Health 252-261.

[7]. S. G. Alonso, I. Torre-Díez, S. Hamrioui, M.l López-Coronado, D. C. Barreno, L. M. Nozaleda, and M. Franco. Data Mining Algorithms and Techniques in Mental Health: A Systematic Review. J. Med. Syst. Vol. 42, 9 (September 2018), 1–15

[8]. Sandhya P, M. Kantesaria "Prediction of Mental Disorder for employees in IT Industry", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6S, April 2019.

[9]. M. A. Haziq Megat S'adan, A. Pampouchidou and F. Meriaudeau, "Deep Learning Techniques for Depression Assessment," *2018 International Conference on Intelligent and Advanced System (ICIAS)*, Kuala Lumpur, 2018, pp. 1-5.

[10]. A. Kaur, "A hybrid approach of privacy preserving data mining using suppression and perturbation techniques," *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Bangalore, 2017, pp. 306-311.

[11]. A. Kiran and D. Vasumathi, "A Comprehensive Survey on Privacy Preservation Algorithms in Data Mining," 2017 *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Coimbatore, 2017, pp. 1-7.

[12]. A. Evfimievski. 2002. Randomization in privacy preserving data mining. SIGKDD Explor. Newsl. 4, 2 (December 2002), 43–48.

[13]. Bhakta, I and Sau, A. (2016). Prediction of Depression among Senior Citizens using Machine Learning Classifiers. International Journal of Computer Applications Vol. 144 No. 7 pp.11–16.

[14]. Deziel, M., Olawo, D., Truchon, L., &Golab, L. Analyzing the Mental Health of Engineering Students using Classification and Regression. *EDM (2013)*.

[15]. WEKA: https://www.cs.waikato.ac.nz/ml/weka/

[16]. A. 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.

[17]. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition.

[18]. P. N. Tan, M. Steinbach, Vipin Kumar, "Introduction to Data Mining", Pearson Education.

[19]. Charu C. Aggarwal, Philip S. Yu, "Privacy-Preserving Data Mining: Models and Algorithms", Springer.