

# Mental Health Quantifier

Daksh Gupta, Aashay Markale, Rishabh Kulkarni



**Abstract:** *The definition of mental disorders describes them as “health conditions involving changes in emotion, thinking or behavior or a combination of these”. Contemporary societies of 2020 still fall short in recognizing some of the most common afflictions as actual problems in people. Some of those are depression, anxiety and stress disorders. This paper proposes a Machine Learning based approach wherein the analysis of the multiple-choice inputs along with a neatly curated questionnaire based on feature extraction will be done and then supervised classification algorithms will be used to generate a mental health score as well as a detailed report based on responses the user gives.*

**Keywords:** *Classification, Feature Extraction, Machine Learning, Mental Health, Psychology*

## I. INTRODUCTION

Depression and poor mental health are one of the biggest problems in the contemporary world. WHO estimates that on average 1 in every 4 people go through mental or neurological disorders<sup>[27]</sup>. In India, the National Mental Health Survey of 2015-16 reported that approx. 15% adults from India need support for at least one mental health issue and depression clutches one in 20 Indians<sup>[15]</sup>. The mental health facilities designed to deal with the proportionate exponential increase in mental health issues are inadequate. A system is required which, at the very least, is able to provide a realistic estimate into ‘at risk’ individuals (namely their proportion, severity and type) for illuminating the current scenario of people having mental health issues.

This paper aims to propose using Machine Learning to find out high-risk individuals. Individuals will be given a questionnaire and based on those answers and previous ones, the algorithm will try to find the people undergoing mental depression. Diagnostic services at psychiatric hospitals and community support centers and reaching conclusions by observation/survey is a work reserved for psychologists and psychiatrists. Most used techniques, not limited to interviews, are the mental status examinations where the patient detail their behaviors, feelings and symptoms they have felt. These tests are generally not very accessible and sometimes expensive. Our goal is to simulate them and make them more

accessible to the general public.

## II. METHODS

### A. Dataset Collection

We have taken a Survey on Mental Health in the Tech Workplace in 2014 By OSMI. Frequency and attitudes towards mental health were measured in the tech workplaces. This dataset has been created with 1200 responses from working professionals across the globe in 2014 and at that time was probably the largest Mental Health in Tech Survey done in the Tech industry.

It was a critical part of the project to understand the characteristics of the questions answered in the data. Hence analysis of the various attributes was to be carried out. For this purpose, we looked closely at the contents. The survey includes responses given for several questions some of which are general (Age, Gender, Country, City etc.), some pertaining to workplace (remote work allowance, benefits, care options, supervisor etc.) and some personal questions (family history of mental illness).

### B. Data Pre-Processing

There were certain features in the dataset which didn't hold relevance for our purposes such as: ‘Timestamp’, ‘State’, ‘Comments’. These features were first removed from the dataframe.

The dataset contained several string attributes such as ‘Gender’, ‘Country’, ‘self\_employed’ etc. which had non-standard values like integers, float, null values in multiple rows which were standardized by filling “NaN”. Some rows contained values for all features as “31”. These rows were dropped from the dataframe because they can hamper the accuracy of models.

The **gender** attribute contained an entire spectrum of genders which were standardized. For “male” the entries included “m”, “M”, “Male”, “male”, “MaLE”, “mail” and lots of other variations and spelling mistakes. Furthermore, there were a lot of gender groups who don't identify as male or female such as “trans”, “queer”, “fluid”, “non-binary”. They were standardized as “trans”. Hence, for the gender attribute values were standardized into 3 categories - male, female and trans.

The **age** attribute also had several absurd values such as 999999, -175. Such values were replaced by the median of age excluding those.

In ‘self\_employed’ attribute ‘NaN’ was replaced by “Not Self Employed”.

In ‘work\_interfere’ attribute ‘NaN’ was replaced by “Don't Know”.

Manuscript received on May 25, 2021.

Revised Manuscript received on May 27, 2021.

Manuscript published on June 30, 2021.

\* Correspondence Author

**Daksh Gupta\***, Department of Information Technology, College of Engineering, Pune (Maharashtra), India. Email: guptads17.it@coep.ac.in

**Aashay Markale**, Department of Information Technology, College of Engineering, Pune, (Maharashtra), India. Email: markaleaa17.it@coep.ac.in

**Rishabh Kulkarni**, Department of Information Technology, College of Engineering, Pune, (Maharashtra), India. Email: kulkarnira17.it@coep.ac.in

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

### C. Data Encoding

Machine learning models mostly require all input data to be in numeric form. However, the dataset contained several attributes where the data was in string format.

For example:

- gender attribute has values: ['female', 'male', 'trans']
- self\_employed has values: ['No', 'Yes']
- work\_interfere has values: ['Don't know', 'Never', 'Often', 'Rarely', 'Sometimes']

All such string attributes were encoded using label encoder object provided by scikit-learn pre-processing class. Each unique label was mapped to an integer value.

### D. Feature Selection

Feature Selection can be done in various ways but is broadly classified into 3 categories – filter method, wrapper method, backward elimination method.

For this part, filter method was chosen since it's computationally less expensive and provided accuracy similar to wrapper method. Filter method does feature selection using correlation.

First correlation matrix was plotted for the dataset wherein correlation between each attribute was calculated with the formula mentioned above. Thereafter, attributes having absolute value of correlation with treatment greater than a threshold value  $\lambda$  were selected.  $\lambda$  was taken to be **0.1**.

In the selected features if any pair of attributes shows a high correlation then one of them can be dropped.

Finally, the independent features selected were = {'Gender', 'Anonymity', 'benefits', 'care\_options', 'family\_history', 'obs\_consequence', 'work\_interfere', 'Age'}.

### E. Applying Models

The dataset was split into training and testing datasets using a 70-30 split. In this case, we had input variables (**X: set of independent attributes**) and an output variable (**y: treatment**). The goal was to train the data using machine learning models and approximate the mapping function so that when we have new input data, the output value can be predicted. Since we needed a mapping function  $y = f(X)$ , supervised learning algorithms needed to be used. Therefore, the following machine learning models were applied on the dataset - *Logistic Regression, KNN, Decision Tree, Random Forest, Naive Bayes, Gradient Boosting*.

From this a binary score indicating requirement of treatment was obtained along with predicted probability value which will be used later.

### F. Sapien MHQ Integration

What has been achieved so far is the determination of the likelihood that a given individual working in the IT industry having a certain set of specific attributes would be in requirement of the assistance of mental health professionals. However mental health cannot be standardized for all people in any sector no matter how specialized to that sector. It ultimately depends on the individual's self-reflection of their mental state. To capture this angle of mental health, we resorted to the question format that has been used by Sapiens Labs. These set of questions are designed in accordance with the DSM-5 mental health criteria and cover the diagnosis of over 10 diseases officially documented in the DSM-5.

Furthermore, the questions stem from the condensed and prioritized version of the collection of over 120 standard mental health assessment tools (such as PHQ-9, Goldberg test etc).

The consolidated set of a total of 47 questions deal with mental health through two main sections: i) Spectrum and ii) Problem. The questions involved in the spectrum section serve the purpose of evaluating certain critical aspects of the functioning human mind. An important point to note here is that these questions do not necessarily deal with an issue that is almost always a potential hinderance to the mental well-being of the individual but could possibly be an asset as well depending on the response given by the individual. The questions have the rating scale from a 1 (which implies that the function does not work as intended and as a result, causes significant problems in daily situations) to a 9 (implying that the function is extremely effective and is an asset to the individual). Some examples of spectrum questions include 'Energy level' and 'Adaptability to change'.

Meanwhile the questions included in the Problem section deal with issues that can rarely be considered as an asset to the individual. These questions deal with discovering the intensity of a problem that the person might be suffering from based on their own self-evaluation.

### G. Sapien MHQ Scoring Method

The sapiens MHQ scoring algorithm is simple yet elegant since it gets effective results without complicating the internal process involved. Furthermore, it was taken into account that not all question scores received can be assigned the same weighting. Some scores have to be prioritized with respect to the seriousness of the issue faced. For example, Suicidal ideation would be weighed higher than anxiety since it calls for more immediate action. To put this idea into practice, the scores acquired are rescaled with the help of a Threshold value 'N'. N takes a value between 2 to 6 as is advised by the Sapiens Labs paper. When it comes to rescaling of Spectrum scores, the rescaled value is taken to be [received score] - N whereas for Problem scores the rescaled value comes from N - [received score]. When it comes to displaying the score output to the user of the MHQ, a summation is taken of all the scores received post the application of the above-mentioned operations on them. This 'intermediate score' can take either a positive or negative value depending upon the individual's mental health. Next, the positive scores are normalized to get a value between 0 and 200 whereas the negative scores are normalized to get values between -1 and -100. The negative scale was deliberately made smaller so as to avoid the score review process being perceived as a stressful activity.

However, the final normalized scores are not the final result. It would be quite useful for the questionnaire to provide an indepth review of certain key functioning components of the individual's mental health. Hence the final result is also displayed in the form of 6 dimensions, namely 'Core cognition', 'Mood and Outlook', 'Complex cognition', 'Mind and Body', 'Social self' and 'Drive and Motivation'.

Each of these dimensions are scored independently from each other. Every question asked in the MHQ is evaluated w.r.t relevance against these 6 dimensions.

If found to be an integral part, then a coefficient of 1 is multiplied to the score for the dimension. Else, partial relevance awards a coefficient of 0.5 to the score added for that particular dimension. Similar to the overall range, the scores for each category are normalized to take values between -50 to 100.

### III. RESULTS

After applying the aforementioned supervised models, training and testing accuracy was calculated for each model. It was observed that gradient boosting gave the best test accuracy.

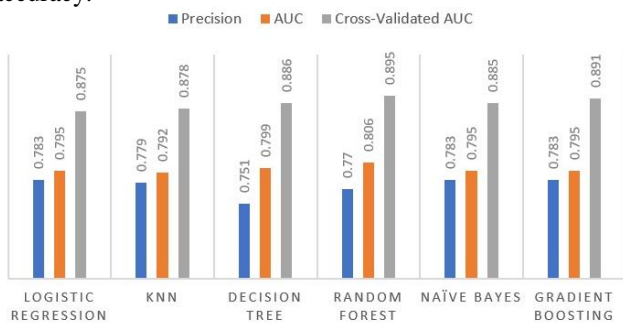


Fig. 1. Evaluation Metrics For Models

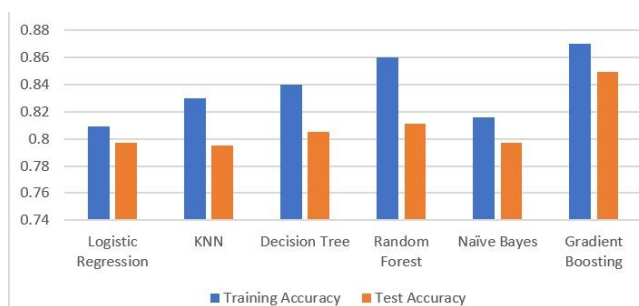


Fig. 2. Performance of different models on OSMI dataset

#### A. Intermediate Result

After receiving the responses from the user against the extracted OSMI questions, the supervised ML algorithms give a binary score of 1 or 0 indicating requirement of treatment or might not have to seek treatment along with the predicted probability score generated for the same.

#### B. Final Result

To get a better overall picture, users are encouraged to further take the Sapien MHQ which takes the previously calculated probability into account along with a host of questions asked related to mental health spectrum and other related problems.

After answering all the questions, the user gets a score between -100 to 200 which classifies them into one of the either categories - Clinical (-100 to -50), At-Risk (-50 to 0), Enduring (0 to 50), Managing (50 to 100), Succeeding (100 to 150), Thriving (150 to 200).

Along with this the algorithm also calculates sub-scores across 6 dimensions (Core Cognition, Complex Cognition, Drive Motivation, Mood Outlook, Social Self, and Mind Body Connection) each scored between -50 to 100 which help user give a better understanding of the specific areas he needs to work with.

### IV. CONCLUSION

In this project we have applied supervised machine learning algorithms to classify whether a person should go for mental health treatment based on OSMI Tech Survey 2014 dataset. Feature selection was performed using the filter method and Gradient Boosting model gave the highest test accuracy of 84.92%.

We further performed hyperparameter optimization with Random Search and Grid Search to see if accuracy can be increased even further. Additional methods like Naive Bayes and Gradient Boosting were tested with respect to accuracy and efficiency. The result obtained by this model was used to determine whether or not the user should seek treatment based on workplace factors.

To get an even better picture the predicted probability score from previous model was combined with Sapien MHQ algorithm to give user a broad picture of his mental state as well as sub-scores calculated across 6 dimensions which can help him/her understand the areas he/she needs to work upon.

Since we intended the test to be easily accessible to anyone, the entire project was packaged as a Django web-app such that any person with browser and internet connection can give the test rather than keeping the same on command line which can be tricky for people not intimately familiar with technology.

### REFERENCES

- Holm, Liselotte Holm, Per Bech, J. (2001). Monitoring improvement using a patient-rated depression scale during treatment with anti-depressants in general practice. *Scandinavian Journal of Primary Health Care*, 19(4), 263–266. <https://doi.org/10.1080/02813430152706819>
- Kroenke, K., Spitzer, R. L., Williams, J. B. W. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606–613 <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>
- D. Goldberg, K. Bridges, P. Duncan-Jones, D. Grayson Detecting anxiety and depression in general medical settings <https://doi.org/10.1136/bmj.297.6653.897>
- Kroenke, K., Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, 32(9), 509–515 <https://psycnet.apa.org/doi/10.3928/0048-5713-20020901-06>
- Aron Halfin, MD Depression: The Benefits of Early and Appropriate Treatment November 1, 2007 Volume 13, Issue 4 Suppl <https://www.ajmc.com/view/nov07-2638ps092-s097>
- Víctor M. Prieto, Sérgio Matos, Manuel Alvarez, Fidel Cacha, and José Luí's Oliveira Twitter: A Good Place to Detect Health Conditions *PLoS One*. 2014; 9(1): e86191. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3906034/>
- Rumi Chunara, Jason R Andrews, John S Brownstein Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak January 2012 *The American journal of tropical medicine and hygiene* 86(1):39-45 [https://www.researchgate.net/publication/221735905\\_Social\\_and\\_News\\_Media\\_Enable\\_Estimation\\_of\\_Epidemiological\\_Patterns\\_Early\\_in\\_the\\_2010\\_Haitian\\_Cholera\\_Outbreak](https://www.researchgate.net/publication/221735905_Social_and_News_Media_Enable_Estimation_of_Epidemiological_Patterns_Early_in_the_2010_Haitian_Cholera_Outbreak)
- Ahmet Emre Aladag, Serra Muderrisoglu, Naz Berfu Akbas, Oguzhan Zahmacioglu Detecting Suicidal Ideation on Forums and Blogs: Proof-of-Concept Study June 2018 *JMR* 20(6):e215 [https://www.researchgate.net/publication/325043548\\_Detecting\\_Suicidal\\_Ideation\\_on\\_Forums\\_and\\_Blogs\\_Proof-of-Concept\\_Study](https://www.researchgate.net/publication/325043548_Detecting_Suicidal_Ideation_on_Forums_and_Blogs_Proof-of-Concept_Study)
- Alice Xue, Dr. Hannah Rohde, Adam Finkelstein An Acoustic Automated Lie Detector Department of Computer Science, Princeton University



[https://www.cs.princeton.edu/sites/default/files/alice\\_xue\\_spring\\_2019.pdf](https://www.cs.princeton.edu/sites/default/files/alice_xue_spring_2019.pdf)

10. Manish Munikar, Sushil Shakya, Aakash Shreshtha Fine-grained Sentiment Classification using BERT Department of Electronics and Computer Engineering Pulchowk Campus, Institute of Engineering, Tribhuvan University Lalitpur, Nepal  
<https://arxiv.org/pdf/1910.03474v1.pdf>
11. 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 (ICIC), Madurai, India, 2018, pp. 1-4, doi: 10.1109/ICIC.2018.8782395.
12. Islam, M. R., Miah, S. J., Kamal, A. R. M., Burmeister, O. (2019). A Design Construct of Developing Approaches to Measure Mental Health Conditions. Australasian Journal of Information Systems, 23. <https://doi.org/10.3127/ajis.v23i0.1829>
13. Anu Priya, Shruti Garg, Neha Prerna Tigga, Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms, Procedia Computer Science, Volume 167, 2020, Pages 1258-1267, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2020.03.442>. (<https://www.sciencedirect.com/science/article/pii/S1877050920309091>)
14. Coutanche, M., Hallion, L. (2020). Machine Learning for Clinical Psychology and Clinical Neuroscience. In A. Wright M. Hallquist (Eds.), *The Cambridge Handbook of Research Methods in Clinical Psychology* (Cambridge Handbooks in Psychology, pp. 467-482). Cambridge: Cambridge University Press. doi:10.1017/9781316995808.041
15. National Mental Health Survey 2015-16 <http://indianmhs.nimhans.ac.in/Docs/Summary.pdf>
16. NLTK Documentation <http://www.nltk.org/>
17. The Calculator Mental Illness Test <https://www.thecalculator.co/personalityMentalIllness-Test-368.html>
18. Sapiens MHQ Assessment <https://sapienlabs.org/mhq/>
19. Link for Spanish tweets dataset used in Prieto's research <https://doi.org/10.1371/journal.pone.0086191.t001>
20. Link for Spanish tweets dataset used in Prieto's research <https://doi.org/10.1371/journal.pone.0086191.t002>
21. Link for Dataset performance: dataset tested on Naïve Bayes, SVM, Decision Trees and kNN in Prieto's research <https://doi.org/10.1371/journal.pone.0086191.t005>
22. <https://1drv.ms/u/s!At-FuMThpgzUhmnaoUYncdd89-AF?e=Fpz3Ch>
23. Dataset for Ahmet Emre Aladag's research <https://www.jmir.org/2018/6/e215/table1>
24. Dataset for Alice Xue's research <https://1drv.ms/u/s!At-FuMThpgzUhmqLj093nHm27IsH?e=vviWdj>
25. Dataset for Manish Munikar's research <https://1drv.ms/u/s!At-FuMThpgzUhmuzwh5vyy-aAh7?e=VLY0ei>
26. Mental Health Taskforce NE. The Five Year Forward View for Mental Health. 2016 [cited 2017 May 23] <https://www.england.nhs.uk/wp-content/uploads/2016/02/MentalHealth-Taskforce-FYFVfinal.pdf>
27. <https://www.who.int/news/item/28-09-2001-the-world-health-report-2001-mental-disorders-affect-one-in-four-people>
28. NCRB Survey Report <https://timesofindia.indiatimes.com/india/15-people-ended-life-every-hour-in-India-during-2014-NCRB/articleshow/48135887.cms>
29. Jones, P. (2013). Adult mental health disorders and their age at onset. *British Journal of Psychiatry*, 202(S54), S5-S10. <https://doi.org/10.1192/bjp.bp.112.119164>
30. Afifi, Mustafa. (2007). Gender differences in mental health. *Singapore medical journal*. 48. 385-91. <https://pubmed.ncbi.nlm.nih.gov/17453094/>
31. McGrath, J. J., Wray, N. R., Pedersen, C. B., Mortensen, P. B., Greve, A. N., Petersen, L. (2014). The association between family history of mental disorders and general cognitive ability. *Translational psychiatry*, 4(7), e412. <https://doi.org/10.1038/tp.2014.60>
32. Brohan, E., Henderson, C., Wheat, K. et al. Systematic review of beliefs, behaviours and influencing factors associated with disclosure of a mental health problem in the workplace. *BMC Psychiatry* 12, 11 (2012). <https://doi.org/10.1186/1471-244X-12-11>

## AUTHORS PROFILE



**Daksh Gupta**, Department of Information Technology, College of Engineering Pune (COEP), an autonomous college affiliated to Savitribai Phule Pune University in Pune, Maharashtra, India. His primary interests are computer science and algorithms. [dakshgupta60000@gmail.com](mailto:dakshgupta60000@gmail.com)



**Aashay Markale**, Department of Information Technology, College of Engineering Pune (COEP), an autonomous college affiliated to Savitribai Phule Pune University in Pune, Maharashtra, India. His interests include machine learning, web development and psychology. [markaleaa17.it@coep.ac.in](mailto:markaleaa17.it@coep.ac.in)



**Rishabh Kulkarni**, Department of Information Technology, College of Engineering Pune (COEP), an autonomous college affiliated to Savitribai Phule Pune University in Pune, Maharashtra, India. Working towards set goals has always been his priority. He has had an interest in the workings of the human mind since a young age and as a result have been working towards the discovery and improvement of the multiple aspects of human mental health. [rishmeditates@gmail.com](mailto:rishmeditates@gmail.com)