INTERNATIONAL JOURNAL OF RESEARCH IN MEDICAL SCIENCES & TECHNOLOGY

e-ISSN:2455-5134; p-ISSN: 2455-9059

DEVELOPING A SMART INTEGRATED MACHINE LEARNING BASED PREDICTIVE MODEL IN THE EARLY DIAGNOSIS OF MENTAL ILLNESS LEVERAGING THE DECISION TREE AND RANDON FOREST CLASSIFICATION

Bahisht Samar
Amity University, Noida, India

Paper Received: 13th November, 2021; **Paper Accepted:** 26th December, 2021;
Paper Published: 29th December, 2021

How to cite the article:

Bahisht Samar, Developing a Smart Integrated Machine Learning Based Predictive Model in the Early Diagnosis of Mental Illness Leveraging the Decision Tree and Random Forest Classification, IJRMST, July-December 2021, Vol 12, 279-286, DOI: http://doi.org/10.37648/ijrmst.v11i02.025

ABSTRACT

Ministry of HFW, Government of India ordered the NIMNS - National Institute of Mental Health and Neuro Sciences, Bengaluru, in alliance with 15 institutions from across India and made a survey on mental health issues. This commission covered 12 states, one among that is Punjab from Northern region. As per the report, 15% of the adults in India need treatment for mental disorder. Machine Learning is one of the most substantial proportions of Artificial Intelligence. Machine Learning is widely used in many fields like online fraud detection, speech recognition, and social media. It plays a vital role in healthcare sector. This boosts the interest on the detection of the mental illness using machine learning algorithm. The big challenge is to predict the state of mind. Psychologists impose assessment and therapy to their patients by one-to-one physical interactions. There are multiple causes to put the person into critical situation like depression, pressure etc. Hence, this research paper proposes an ideal solution to identify the sickness in the person by checking with the recorded dataset. The most preferred Supervised Machine Learning algorithm, Decision Tree Classifier is used for this purpose. The initial goal of the Decision Tress is to create training ideal which is used to forecast the target variable class. The parameters considered here are anxiety disorder, depression disorder and the stress. Random Forest algorithm is applied to predict the illness in the people. The result obtained is to have accurate prediction level compared to the existing model.

INTRODUCTION

Mental Illness is coined from the phrase called Mental Health Disorder. It states that the disorder may affects the mood, behavior, thinking of the person.

Even though the degrees of identifying the mental illness have better improvement over the past few decades, many cases stay undetected. The symptoms in association with mental illness are seen on social media like Twitter, FB, forums.

The person could be fine today, but the next day, the physician can tell something which is unfair to the previous day. These happen to people in day -by-day. This is the case for body or physical structure of the person. What if happens to the mind? Sympathy will not work out for the illness to the mind rather in the case of physical structure. Mental illness distorts the peaceful and happiness within themselves and in the surrounding. It is much painful for the affected people and even more aching for people around them. A soul or human requires a certain level of psychological, emotional, and space

x Confused thinking or over thinking

x Pulling out from friends in social media

x Dramatic changes in eating or sleeping pattern

B. Common Mental Health Issues

The most common heath issues listed by the health and medical news website WebMD were,

- Anxiety disorders: People who responds to certain objects or situations with fear is the sign of anxiety or panic. Those people have abrupt change in heartbeat.
- Mood Disorders: These involve the fluctuations in the mood. It may be either extreme happiness or extreme sadness.
- Psychotic disorders: Hallucinations is the major symptom of psychotic disorder.
- Stress: In this era, stress is the common problem which lives with many people as like the behaviour. High pulse rate, Sweating can be the symptoms.

WORKFLOW OF PROPOSED PAPER

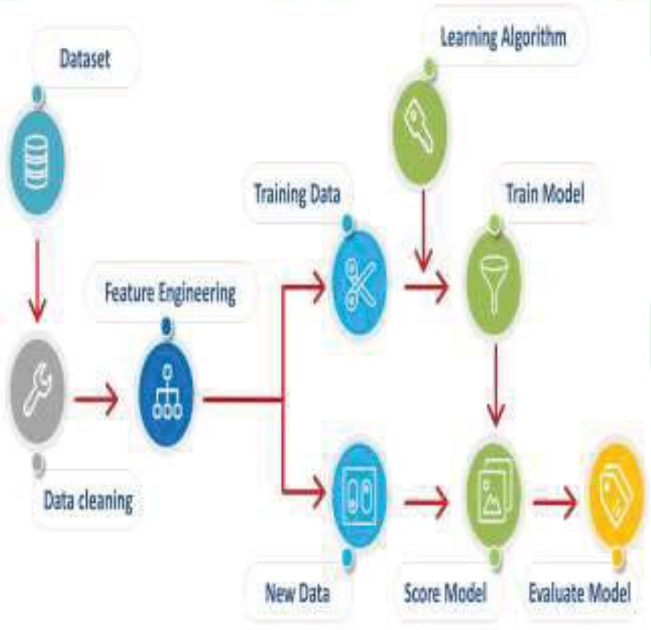
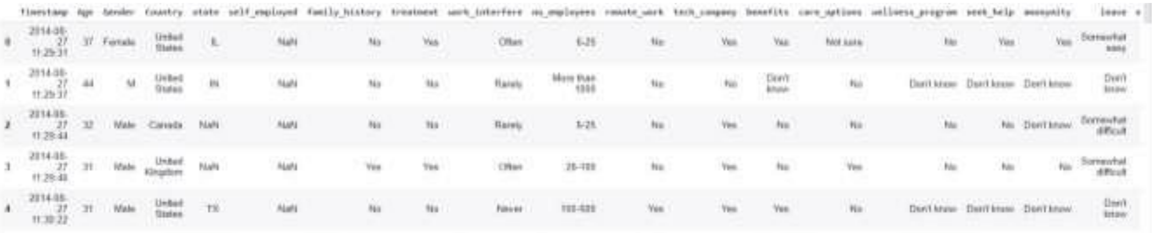


Fig. 1. Workflow of the proposed system

DATA COLLECTION

The dataset has (1259, 21) rows*columns. The columns name act as the identifier to set as the root node for the tree from the top to the bottom. The Person can be split as Normal or Abnormal.



	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees	remote_work	tech_company	benefits	care_options	wellness_program	seek_help	anonymity	leave
0	2014-08-27 11:29:31	37	Female	United States	IL	NaN	No	Yes	Often	6-25	No	Yes	Yes	Not sure	No	Yes	Yes	Somewhat easy
1	2014-08-27 11:29:37	44	M	United States	IN	NaN	No	No	Rarely	More than 1000	No	No	Don't know	No	Don't know	Don't know	Don't know	Don't know
2	2014-08-27 11:29:44	32	Male	Canada	NaN	NaN	No	No	Rarely	6-25	No	Yes	No	No	No	No	Don't know	Somewhat difficult
3	2014-08-27 11:29:46	31	Male	United Kingdom	NaN	NaN	Yes	Yes	Often	26-100	No	Yes	No	Yes	No	No	No	Somewhat difficult
4	2014-08-27 11:30:22	31	Male	United States	TX	NaN	No	No	Never	100-500	Yes	Yes	Yes	No	Don't know	Don't know	Don't know	Don't know

Fig. 2. Dataset

The abnormal can be split has anxiety, depression, stress.

The parameter of the second level root is denoted in figure3.

Anxiety, Depression and Stress Scale questionnaire (ADSS 21) consists of 21 questions related to identify the abnormalities like stress, depression and anxiety. [11]

The rating scale is as follows:

x 0 Not applicable

x 1 Applicable to somewhat

x 2 Applicable for few degrees

x 3 Applicable to most of the time

The questions filled by the persons are tabulated below, The questionnaires from ADSS-21 related to anxiety, stress and depression. The dataset was encrypted with the values from zero (0) to three (3), and the levels were then manipulated by summing the values connected with the question with the given formula:

Value = Summation of class rating points*2

The calculated values were labelled according to severity levels – i.e. nominal, low, reasonable, serious, and very serious.

DATA PREPROCESSING

Pre-processing is the essential phase in the case of handling the datasets with Machine Learning. This pre-processing phase focus on cleaning the data, that is removing the unfilled rows from the dataset. This technique is not correct one to deal with this particular dataset. Another technique is to calculate the mean of the particular column or field to fill the NAN cases. There are other techniques to fill the missing value. Those are median and mode. Here mean is applied over the dataset. The pre-processing of the data is done by filling the NAN values in the dataset with Mean of the particular field.

EVALUATION METHODS

The pre-processed data is taken, and the features are extracted. The dataset is divided into 80:20 ratio which represents the 80% of the data is used to train the algorithm and the remaining 20% of the data is for testing the data. The classification algorithm is applied over the dataset to classify the data.

A. Decision Tree Classifier

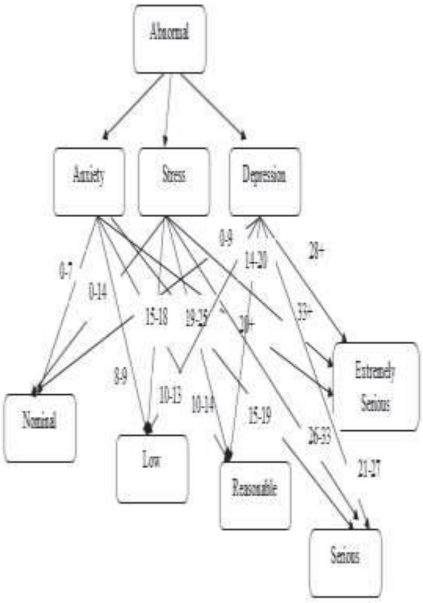


Fig. 3. Decision Tree Classifier

The decision tree strategy of machine learning is suitable for predictive problems. Decision tree is for both classification and regression.

B. Random Forest Algorithm

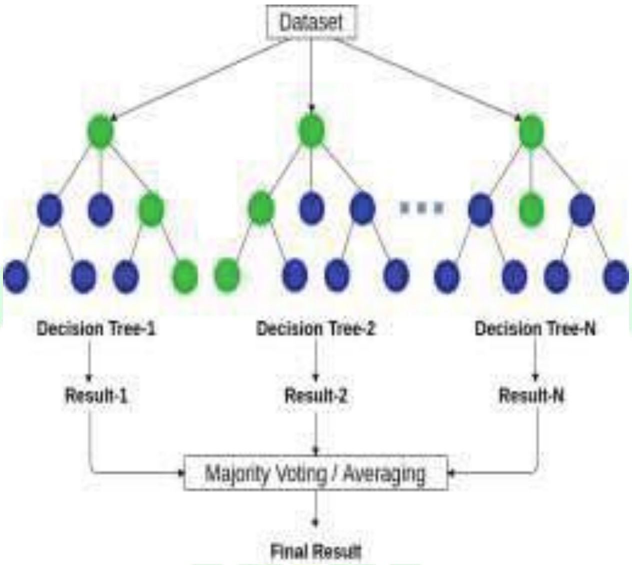


Fig. 4. Random Forest

RESULT AND DISCUSSION

TABLE I. DIFFERENT MEASURES VALUES

Machine Learning algorithm	Illness	Acc	Er	Precision	Recall	Specificity	F1 Value
Decision Tree	A	0.766	0.289	0.478	0.556	0.945	0.501
	D	0.799	0.245	0.767	0.745	0.923	0.745
	S	0.656	0.389	0.620	0.598	0.903	0.612
Random Forest	A	0.745	0.299	0.456	0.540	0.921	0.490
	D	0.822	0.230	0.899	0.698	0.913	0.789
	S	0.757	0.301	0.767	0.712	0.934	0.745

Decision tree and Random Forest algorithms are used to detect the stress, anxiety and depression. Based on this, confusion matrix is generated.

Equations 1 to 6 are used to manipulate the values of accuracy, precision, recall, error rates, and specificity which yield the CM - confusion matrix.

1, 2, 3, 4, 5 represents the severity levels. i.e normal, low, reasonable, serious, very serious.in the below table.

Whereas, True positive = Matrix diagonals

False Negative = Leaving TP of that class, sum of stable row for class

False Positive = Excludes TP of that class, summation of Equivalent class.

True Negative = Excludes of the class, sum of complete row and column A – Anxiety, D-Depression, S-Stress

The accuracy value is good for anxiety in decision tree classification. The accuracy of depression and stress while applying random forest are worthy. The error rate of depression is low when applying random forest algorithm. Depression and stress precision is fair enough in random forest than decision tree. F1 score of anxiety is notable in decision tree than random forest.

CONCLUSION AND FUTURE WORK

In this paper, to define the severity levels of ADS – Anxiety, Depression and Stress, machine leaning algorithms like Decision Tree and Random Forest algorithms were used. The dataset contains the general and basic information of the people along with the questionnaires mentioned by ADSS-21. The accuracy of Random Forest algorithm was discovered as worthy than with decision tree. F1 score was taken to identify which is the best suitable model for the prediction of the mental illness. Based on the F1 score, the best method is Random Forest.

The dataset with the anxiety, stress and depression can be applied with K nearest Neighbor, Support Vector Machine (SVM).

REFERENCES

[1] Roy, S., Aithal, P. S., & Bose, D. (2021). Judging Mental Health Disorders Using Decision Tree Models. International Journal of Health Sciences and Pharmacy (IJHSP), 5(1), 11-22.

[2] https://vertavahealth.com/addiction-resources/identifying-mental-health-issues

[3] https://isha.sadhguru.org/in/

[4] https://www.amazonswatchmagazine.com/health-wellbeing/mental-illness-is-nothing-to-be-ashamed-of/

[5] https://time.com/5727535/artificial-intelligence-psychiatry/

[6] Abd Rahman, R., Omar, K., Noah, S. A. M., Danuri, M. S. N. M., & Al-Garadi, M. A. (2020). Application of machine learning methods in mental health detection: a systematic review. IEEE Access, 8, 183952-183964.

[7] Tomasik, J., Han, S. Y. S., Barton-Owen, G., Mirea, D. M., Martin-Key, N. A., Rustogi, N., ... & Bahn, S. (2021). A machine learning algorithm to differentiate bipolar disorder from major depressive disorder using an online mental health questionnaire and blood biomarker data. Translational psychiatry, 11(1), 1-12.

[8] Tao, X., Shaik, T. B., Higgins, N., Gururajan, R., & Zhou, X. (2021). Remote patient monitoring using radio frequency identification (RFID) technology and machine learning for early detection of suicidal behaviour in mental health facilities. Sensors, 21(3), 776.

[9] Salari, N., Hosseinian-Far, A., Jalali, R., Vaisi-Raygani, A., Rasoulpoor, S., Mohammadi, M., ... & Khaledi-Paveh, B. (2020). Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis. Globalization and health, 16(1), 1-11.

[10] Liu, Y., Hankey, J., Cao, B., & Chokka, P. (2021). Screening for major depressive disorder in a tertiary mental health centre using EarlyDetect: A machine learning-based pilot study. Journal of affective disorders reports, 3, 100062.

[11] Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Computer Science, 167, 1258-1267