The Predictive Model of Mental Illness using Decision Tree and Random Forest classification in Machine Learning

Prithvipal Singh
Computer Science
GNDU
Amritsar, India
Prithvipalsingh89@gmail.com

Gurvinder Singh
Computer Science
GNDU
Amritsar, India
gurvinder4371@gmail.com

Sarveshwar Bharti
Computer Science
GNDU
Amritsar, India
sarveshwar.csc@gndu.ac.in

2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) | 978-1-6654-3789-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACITE53722.2022.9823761

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.**

Keywords— Mental Illness, Machine Learning, Decision Tree

I. 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 individually to nurture. This kind of atmosphere is missing today. [3]

Study on the identification of initial symptoms of mental health illness shall hypothetically drop the serious. Moreover, prescribed treatment shall help the affected person to manage the disputes. Issues related to mental illness can be comprised of mood disorders, think pattern change and behavioral change. Few examples embrace anxiety disorders, depression and stress.

The comparison of normal and unexpected behavior, and the early sign of mental health problem is not that much easy. The problems faced by the human being are because of the illness in the mind. This makes the situation miserable. Even the minute change in the behavior, feelings, and thoughts is noticed first by the family, friends, and teachers. The individuals also can notice the change in them, not at the first change; preferably at the mid of the illness. [2]

A. Signs and Symptoms of Mental Illness [4]

- Frequent changes to mood
- Having trouble with your memory
- Having feelings of harming themselves
- Confused thinking or over thinking
- Pulling out from friends in social media
- 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 behavior. High pulse rate, Sweating can be the symptoms.

II. Related Works

R. A. Rahman et al applied ML algorithms in Online Social Networks (OSNs) to detect mental illness. Dictionary-based and ML methods were the two methods to evaluate the data from OSNs texts sent by the customers. [6]

Jakub Tomasik et al developed a pinpointing algorithm based on the questionnaire through online and biomarker data to diminish the not diagnosis of bipolar disorder (BP); MDD. The threshold set for the depressive symptom is greater than or equal to 5 out of 9 questionnaires. These online questionnaires were posted only to the age group of 18-45 years. Based on the threshold value, the persons were picked and provided blood spot samples for biomarker analysis, followed by a Diagnostic Interview via telephone. The algorithms used to train the data were Extreme Gradient Boosting and nested cross-validation which yields the outcome as the differentiation among BP and MDD. Higher mood, grandiosity, talkativeness, wildness and unsafe behavior were the core parameters included in the analysis. The authors developed a proof of concept to detect the bipolar disorder among the recently diagnosed MDD. [7]

Xiaohui Tao et al described Remote Patient Monitoring (RPM) which has much popularity to monitor the patients in the clinics. The authors built the RPM system with RFID (Radio Frequency Identification) technology to detect suicidal behavior in the early stage. The set of ML like Decision Tree, Linear Regression, RF and XGBoost were applied and compared which helped to define the optimum position of RFID reader-antennas in the hospital ward. The developed RPM system retrieved the patient's heart speed, pulse speed, respiration speed and subtle motions. The decision tree gave the top result compared to RF and XGBoost algorithms. [8]

Nader Salari et all motivated on the stress and anxiety among the people in the covid-19 pandemic time. The authors examined the papers where the health problems were related to stress and anxiety. The collected studies were performed with the meta-analysis, and a random effects model was used, later I_2 index was applied. CMA software was used to conduct the data analysis. CMA stands for Comprehensive-Meta-Analysis. The sample size considered for analysis was 9074 and the occurrence of stress was 29.6% and the occurrence of stress was 31.9% for anxiety for the sample size of 63,439. With the 44,531 as sample size, the occurrence of depression was 33.7%. [9]

Yang Liu et al used the Patient Health Questionnaire (PHQ -9) as the screening tool. The early detect system developed can be deployed for the outpatients' centers, seek feedback from physicians and the patients. The system is either an application or it can be integrated with the already existing system. The factors included in their research were family log of mental illness, stressful events were considered in the assessment.

Their results integrated the screening measures and family history in order to enhance the accuracy in prediction. 13% improvement was in specificity over PHQ-9 questions. The developed tool yields 72.0% accuracy with 74.2% sensitivity and 69.8% in specificity for MDD symptoms-only model. [10]

Sandip Roy et al proposed the algorithms like Random Forest, K-Nearest Neighbor and Judgment Analysis for the prediction of anxiety, stress and depression. 96% accuracy is achieved with K nearest neighbor algorithm. Random Forest achieves 93% of accuracy. The authors proposed the diagnosis model where the prediction of the abnormality with their prescriptions was mentioned. [1]

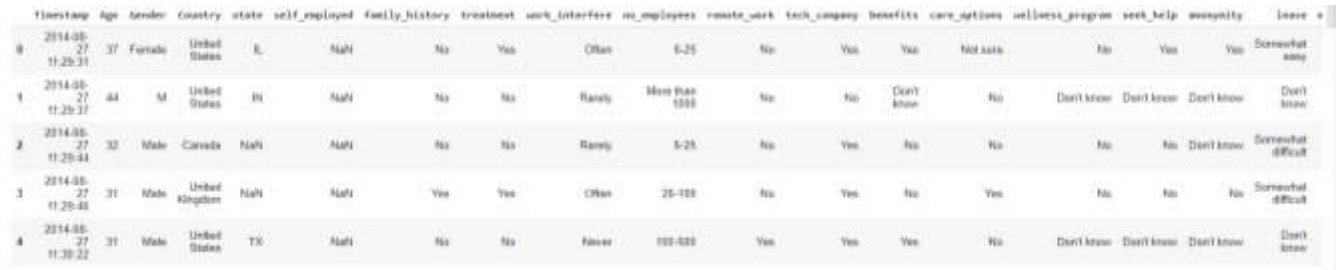
III. Workflow of Proposed Paper

Fig. 1. Workflow of the proposed system

IV. 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.

The abnormal can be split has anxiety, depression, stress. The parameter of the second level root is denoted in figure3.



	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

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:

- 0 Not applicable
- 1 Applicable to somewhat
- 2 Applicable for few degrees
- 3 Applicable to most of the time

The questions filled by the persons are tabulated below,

TABLE I. QUESTIONNAIRES ON ADS

Question No.	Anxiety	Depression	Stress
1	Dehydration i,e mouth dryness	Couldn't know the positive vibration	Feel hard to settle down after tension
2	Breathing Trouble	Problematic to start the things	Go over the top for the situations
3	Experience unsteadies at work or anything	Felt nothing to work upon	Uneasy energy
4	Self-Fear and fooling themselves	Felt sad and dejected	Getting Restless
5	Near to panic	Not able to an enthusiastic person	Hard to Calm
6	Heart rate increase when doing physical activities	Felt not worthy to alive	Unable to tolerant the actions of themselves
7	Scary always without any reason	Thought of meaningless life	Oversensitive

Table 1 represents 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 labeled according to severity levels – i.e nominal, low, reasonable, serious, and very serious. This is displayed in the table 2.

TABLE II. LEVELS OF SEVERITY

	A	D	S
Nominal	0-8	0-10	0-15
Low	9-10	11-14	16-19
Reasonable	11-15	15-21	20-26
Serious	16-20	22-28	27-34
Very Serious	21+	29+	35+

*A-Anxiety, D-Depression, S-Stress

V. DATA PREPROCESSING

Preprocessing is the essential phase in the case of handling the datasets with Machine Learning. This preprocessing 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 preprocessing of the data is done by filling the NAN values in the dataset with Mean of the particular field.

VI. EVALUATION METHODS

The preprocessed 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

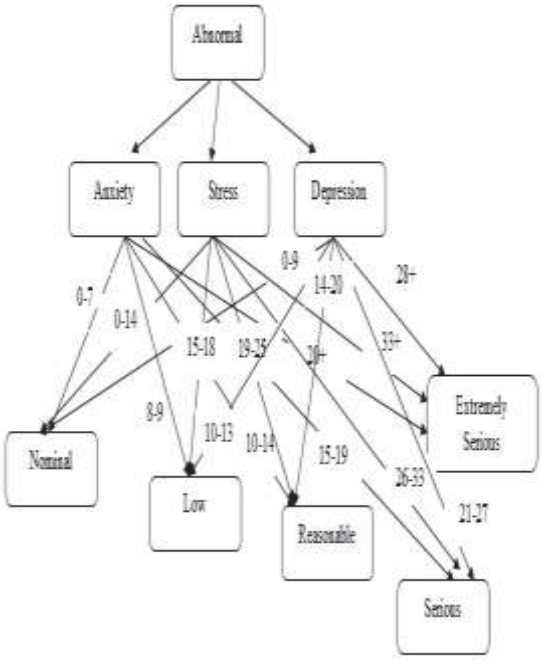


Fig. 3. Decision Tree Classifier

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

B. Random Forest Algorithm

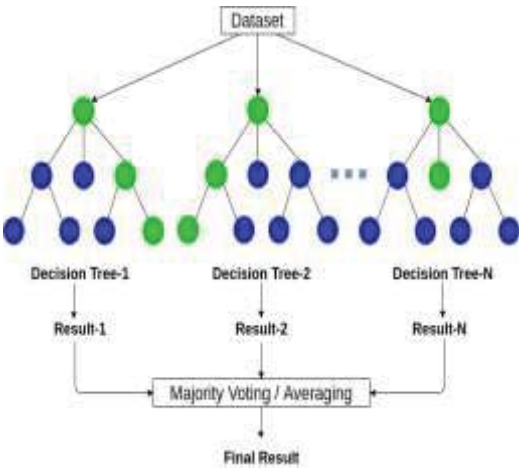


Fig. 4. Random Forest

VII. Result and Discussion

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

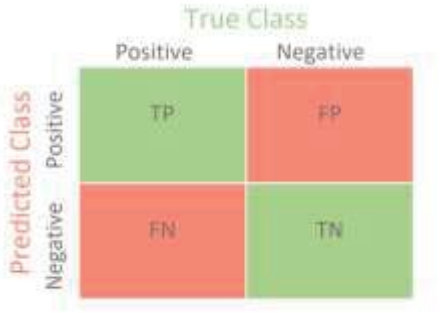


Fig. 5. Confusion Matrix

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.

$$\text{Αχχυραχψ Ρατε (AP)} = \frac{Summazation of Diagonals(TP)}{Total count} \quad (1)$$

$$\text{Ερρορ Ρατε (EP)} = 1 - \text{AP} \quad (2)$$

$$\text{Πρεχισιον} = \frac{True Positive}{True Positive + False Positive} \quad (3)$$

$$\text{Ρεχαλλ} = \frac{True Positive}{True Positive + False Negative} \quad (4)$$

$$\text{Σπεχιφιχιτψ} = \frac{True Negative}{True Negative + False Positive} \quad (5)$$

$$\Phi 1\ \text{Σχορε} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad (6)$$

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

TABLE III. CM - Confusion Matrix

	Anxiety	Stress	Depression

Decision Tee		1	2	3	4	5		1	2	3	4	5		1	2	3	4	5		
	1	22	0	5	0	0	1	32	5	3	0	0	1	32	5	3	0	0		
	2	3	0	4	0	0	2	5	8	2	0	0	2	5	8	2	0	0		
	3	6	0	31	0	0	3	0	0	25	0	0	3	0	0	25	2	0		
	4	0	0	10	0	3	4	0	0	6	6	2	4	0	0	6	6	2		
	5	0	0	1	0	27	5	0	0	0	6	15	5	0	0	0	6	17		
Random Forest		1	2	3	4	5		1	2	3	4	5		1	2	3	4	5		
	1	20	0	8	0	0	1	36	0	2	0	0	1	39	5	2	0	0		
	2	7	0	8	0	0	2	10	2	3	0	0	2	4	14	4	0	0		
	3	5	0	30	0	0	3	2	0	25	0	0	3	2	4	13	8	0		
	4	0	0	8	0	2	4	0	0	4	5	2	4	0	0	6	10	0		
	5	0	0	1	0	27	5	0	0	1	0	18	5	0	0	0	1	3		

TABLE IV. DIFFERENT MEASURES VALUES

Machine Learning algorithm	**Illness**	**Acc**	**Er**	**Pre$_{cision}$**	**Re$_{call}$**	**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

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.

VIII. 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