PAPER • **OPEN ACCESS**

Predicting Mental Health Illness using Machine Learning Algorithms

To cite this article: Konda Vaishnavi *et al* 2022 *J. Phys.: Conf. Ser.* **2161** 012021

Predicting Mental Health Illness using Machine Learning Algorithms

Konda Vaishnavi, U Nikhitha Kamath, B Ashwath Rao and N V Subba Reddy

Department of Computer Science & Engg., Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka State, India, PIN-576104

E-mail: konda.vaishnavi@learner.manipal.edu, nikhitha.kamath@learner.manipal.edu, ashwath.rao.b@gmail.com, nvs.reddy@manipal.edu

Abstract. Early detection of mental health issues allows specialists to treat them more effectively and it improves patient's quality of life. Mental health is about one's psychological, emotional, and social well-being. It affects the way how one thinks, feels, and acts. Mental health is very important at every stage of life, from childhood and adolescence through adulthood. This study identified five machine learning techniques and assessed their accuracy in identifying mental health issues using several accuracy criteria. The five machine learning techniques are Logistic Regression, K-NN Classifier, Decision Tree Classifier, Random Forest, and Stacking. We have compared these techniques and implemented them and also obtained the most accurate one in Stacking technique based with an accuracy of prediction 81.75%.

1. Introduction

A person's mental well-being is his or her mental condition, as well as an overview of his or her general environment. Brain chemistry abnormalities are the cause of mental illness. An individual's mental health serves as a barometer for properly addressing his or her diseases. To predict any health-related irregularities, it is critical to keep track of diverse groups' mental health profiles. The community is made up of working professionals, college students, and high school students. There is a widespread belief that stress and sadness affect people of all ages and backgrounds. To avoid serious illness, it is necessary to identify the mental health of different categories at different times. In the next years, healthcare providers will be required to consider a patient's mental health profile to deliver better medication and aid in a speedier recovery.

Some of the major mental health disorders, such as chronic diseases, bipolar disorder, and schizophrenia they don't suddenly arise out of nowhere; they often develop over time and produce symptoms that can be recognized in the early stages. Such disorders could be avoided or controlled more successfully. If abnormal mental states are discovered early in the disease's course when extra treatment and care can be provided. So judging people's mental states based on their looks or conduct is a complex psychological science that has yet to be mechanized. Although screening test solutions exist, due to time and financial constraints, this solution is not feasible for large populations. Furthermore, diagnosis-based procedures have the unintended consequence of discouraging unwell people from taking part. As a result, psychological problems frequently go unnoticed or untreated.

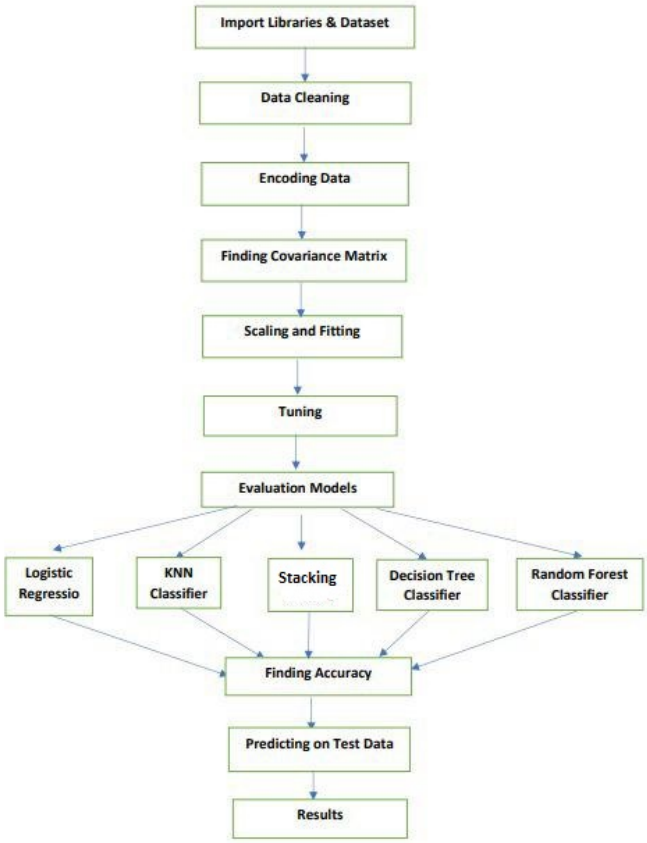

Figure 1. Flow Chart

Anxiety and depression are serious public health issues that affect people all over the world. They affect people of all ages, from children to the elderly, including both men and women. Anxiety and depression disorders have a wide range of effects on health and well-being. They are responsible for a variety of somatic symptoms such as gastritis, acid reflux, palpitation, insomnia or hypersomnia, tremor, significant weight loss or gain, and various psychosocial manifestations such as depressed mood, social withdrawal, decreased workplace productivity, suicidal ideation, or attempt, and lack of concentration. Depression and anxiety are significant risk factors for a variety of other lifestyle disorders, such as ischemic heart disease, hypertension, diabetes, unintentional accidents, and deliberate. Suicidal ideation and depression are intimately linked, and depression can lead to suicide. There are different communicable diseases, such as tuberculosis and HIV, that harm them. Depression and anxiety sufferers are frequently stigmatized by society and socially isolated by their families. In educational institutions and workplaces, they may underperform. As a result, people are losing access to economic and social possibilities, with results in a low quality of life. Economic strain is a massive and often unquantifiable manifestation that contributes to a vicious cycle of poverty and bad health. The majority of those affected are low- and middle-income families

Technological advancements such as smartphones, social media, neuroimaging, and wearables have enabled researchers of mental health and doctors to gather a tremendous amount of information at a rapid rate. Machine learning has developed as a reliable tool for analyzing these data. Machine Learning is the application of advanced probabilistic and statistical techniques to create computers that can learn from data on their own. This allows data patterns to be

Methods	Accuracy (%)
Logistic Regression	79.63
KNeighbors Classifier	80.42
Decision Tree classifier	80.69
Random Forests	81.22
Stacking	81.75

Figure 2. Accuracy of all Classifiers

more easily and correctly discovered, as well as more accurate predictions from data sources. Natural language processing, speech recognition, computer vision, and Artificial intelligence have all benefited from machine learning, which allows developers and researchers to extract crucial data from datasets, deliver personalized experiences, and develop intelligent systems. In domains like bioinformatics, ML has aided substantial progress by allowing for quick and scalable analysis of complicated data. Mental health data is also being investigated using similar analytic tools, with the potential to improve patient outcomes as well as improve understanding of psychological diseases and their management.

2. Literature Review

The study assessed the performance of eight different machine learning algorithms which classify the dataset into various issues of mental health. Their results show that the three classifiers tested, namely the Multiclass Classifier, Multilayer Perceptron, and the LAD Tree, generate results that are more accurate than the others [4]. This paper explains that mental health analysis in terms that are intuitive to different target groups. They have created a system for determining an individual's mental health status and prediction models were built using this framework. Clustering methods were also been used to determine the number of clusters before developing models. MOS was used to validate the class labels produced, which were then used to train the classifier. The trials showed that KNN, SVM, and Random Forest performed nearly equally well. The usage of ensemble classifiers was also discovered to considerably increase the performance of mental health prediction with a 90% accuracy rate [6]. In this, the research has concentrated on the benefits of machine learning in improving mental health identification and diagnosis of Alzheimer's disease, depression, and other mental illnesses schizophrenia. Overall, machine learning has the potential to increase clinical and research efficiency while also providing fresh insight into mental health and wellbeing [5]. The key contribution of this paper is that the ILIOU preprocessing method can be utilized to significantly improve the performance of classification algorithms in similar datasets and it can also be used to forecast different types of depression. Depression prediction is critical for patients to receive the most appropriate treatment as soon as feasible [3]. Artificial intelligence is becoming a bigger aspect of medicine, and it will help with mental health research and practice. To realize the full promise of AI, a varied community of specialists involved in mental health research and care, including scientists, clinicians, patients, and regulators, must communicate and interact [2]. They have analyzed by using discourse analysis in this study to better understand the practices of representation in human-centered machine learning (HCML). From this, case prediction of mental health status on social media data, they have found a dataset of 55 interdisciplinary studies. Their findings show that opposing the discourses of interaction throughout the dataset to construct and grant agency for the humans. Their findings demonstrate how the five discourses produce a paradoxical object and subject views of the human, potentially dehumanizing it accidentally [1].

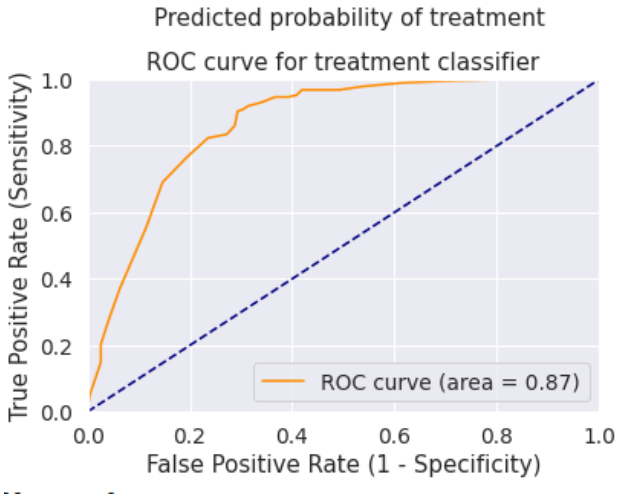

Figure 3. ROC of KNN

3. Methodology

Knowledge Discovery from Data as shown in Figure.1. includes Data Collection, Data Cleaning, encoding data, Finding Co-variance matrix, Scaling and Fitting, Tuning, Evaluation models, Finding Accuracy, Predicting Data and Results. First, we consider the dataset with 27 columns and 1259 entries. The following phase is data cleaning, which is the process of finding incomplete, erroneous, unneeded, or missing data and then modifying, replacing, or eliminating it based on the specific necessity. We found that three columns have the missing data. Not a Number, or NaN, is a special value in Data Frames and Numpy arrays that represents a cell with no value. The next step is data encoding. When the categorical feature is identified ordinal, we apply this categorical data encoding strategy. In this case, it is important to retain the order. Hence the sequence should be reflected in encoding. Each label will be turned into a value of an integer value during label encoding. After that, we'll look for the covariance matrix. In data science and machine learning, it is one of the most significant matrices. It offers information on feature co-movement (correlation). In the matrix of variance-covariance, the variances of the variables will be in the main diagonal and the covariances will be between each pair of variables of the other matrix places, while the mean vector contains the means of each variable.

In Feature scaling, we put the data's independent features into a set range. It handles significantly changing values or units and magnitudes during data pre-processing. Next, we split the dataset into training and testing data set. The next step is feature importance. Feature selection is critical in machine learning since it is a fundamental strategy for directing variable usage to what is most efficient and effective for a certain machine learning system. The next step is tuning. Tuning is the process of enhancing a model's performance while avoiding overfitting or excessive variance. This is performed in machine learning by picking appropriate hyperparameters. Hyperparameters can be thought of as a machine learning model's "dial" or "knobs". The models are then evaluated using a variety of machine learning methods, such as stacking, logistic regression, K-nearest neighbor classifier, decision tree classifier, and random forest classifier.

3.1. Logistic Regression

Logistic regression is a prominent machine learning algorithm that comes under the supervised learning approach. In this method, we predict a specific dependent variable from a set of unbiased variables. Logistic regression is used to expect the output of specific structured variables. So

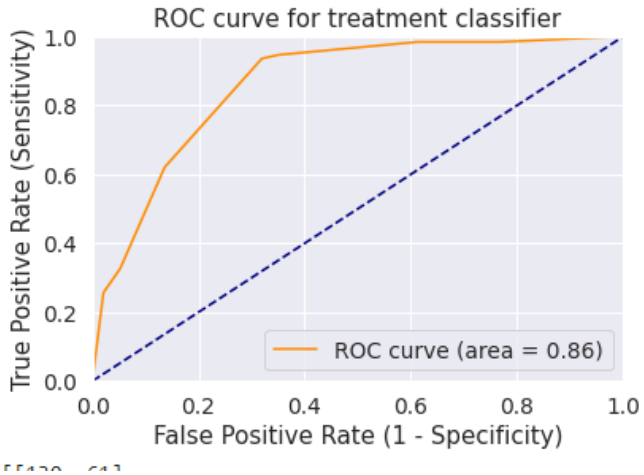

Figure 4. ROC of Decision tree

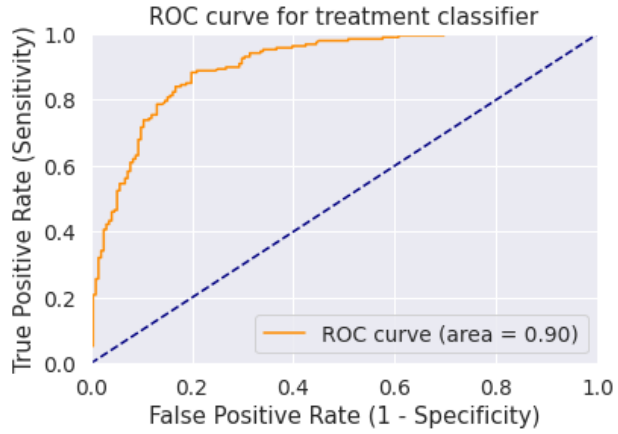

Figure 5. ROC of Random forest

the result should be a categorical or discrete value. It can be 0 or 1, Yes or No, true or false, and so on, but it delivers probabilistic values that are somewhere between 0 and 1 instead of giving exact values like 0 and 1.

3.2. K nearest neighbour classifier

The K-Nearest Neighbour is a basic machine learning algorithm that is based on the Supervised Learning technique. In the K-NN method, the existing cases and new case/data will be similar. KNN is a non-parametric algorithm that doesn't make any assumption of its underlined data or its distribution. And also it works with multiple classes.

3.3. Decision tree classifier

Decision Tree is the widely used supervised machine learning technique that is used in data mining. A decision tree is a diagram that individuals use to illustrate a statistical likelihood or to determine the sequence of events, actions, or outcomes.

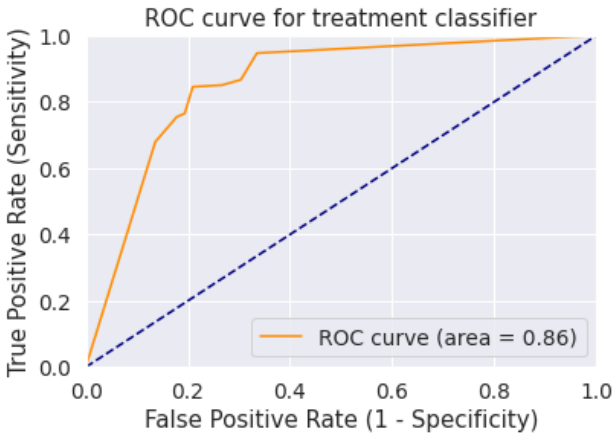

Figure 6. ROC of Stacking

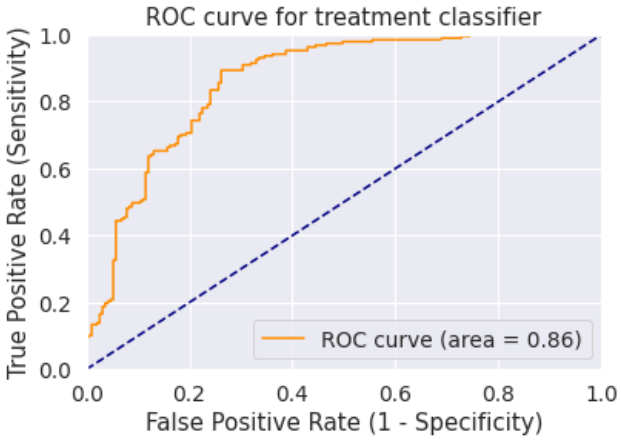

Figure 7. ROC of Logistic Regression

3.4. Random forest classifier

Random forest is a technique that solves classification as well as regression problems and it is based on the supervised machine learning approach. However, it is frequently used for classification. Because it combines many decision trees to create a "forest" and feeds random features from the input dataset to them, it is called a random forest.

3.5. Stacking

Stacked Generalization, or simply "Stacking," is a machine learning ensemble algorithm. Like bagging and boosting, it entails aggregating predictions from many machine learning models on the same dataset.

4. Results

This study identified five machine learning techniques i.e k nearest neighbor classifier, logistic regression, decision tree, and stacking, and random forest. And we assessed their accuracy in identifying mental health issues. First, we executed the classifiers which included all the 27 attributes which were identified from the text documents and then we executed it by including

8 attributes were selected using the algorithm of feature selection. The Accuracy of a given

test set for a classifier is the percentage of test set instances that are classified correctly by using the classifier. Figure 2 shows the accuracy of stacking is more accurate compared to other classifiers. The accuracy of any classifier will depend upon how well the classifier will classify the data set which is being tested. We measured that by using the area under the Receiver Operating Curve. In the ROC area, a perfect test will represent an area of 1 and a worthless test will represent an area of 0.5. Figure 3-7 illustrates the graph of five classifiers on ROC Area values. We observed that the classifiers were more accurate in predicting the condition of mental health than other classifiers because the ROC area of all classifiers used is between 0.8 and 0.9.

5. Conclusion

As there are many available techniques of machine learning, it is very important to compare those techniques and then identify the best among them that will suit the domain of interest. Nowadays, we have many special programs in the medical field that predict disease very accurately in advance so that treatment can be done effectively and efficiently. In this proposed work we have compared five different techniques of machine learning which are used to classify the dataset on various problems of mental health. It is very clear from the results that all the five machine learning techniques give more accurate results. The accuracy of all the classifiers are above 79%. The data set used in the research is very minimal and in the future, a large data set can be used and the research can be applied on the same for more accuracy.

References

[1] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32, 2019.

[2] Sarah Graham, Colin Depp, Ellen E Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V Jeste. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21(11):1–18, 2019.

[3] Theodoros Iliou, Georgia Konstantopoulou, Mandani Ntekouli, Christina Lymperopoulou, Konstantinos Assimakopoulos, Dimitrios Galiatsatos, and George Anastassopoulos. Iliou machine learning preprocessing method for depression type prediction. *Evolving Systems*, 10(1):29–39, 2019.

[4] T Nagar. Prediction of mental health problems among children using machine learning techniques.

[5] Adrian BR Shatte, Delyse M Hutchinson, and Samantha J Teague. Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine*, 49(9):1426–1448, 2019.

[6] M Srividya, S Mohanavalli, and N Bhalaji. Behavioral modeling for mental health using machine learning algorithms. *Journal of medical systems*, 42(5):1–12, 2018.