Mental Health Sickness Prediction using Machine Learning Algorithm

Sk.Heena Kousar[1], M.Maneesha[2], M.Prasanna[3] , U.Sai Anusha[4]

1, 2, 3Student, Department of CSE, Narasaraopeta Engineering College, Narasaraopet, Guntur (D.T) A.P, India

4Asst.Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopet, Guntur (D.T) A.P, India

shaikheenakousar4@gmail.com[1] maneeshamodugula@gmail.com[2] prasannamallabattini@gmail.com[3] uppalasaianusha@gmail.com[4]

1. **ABSTRACT :** Mental Illnes is becoming a serious health problem now-a-days. There is a growing number of patients suffering from different mental health disorders. So we need to find solution to tackle this issue. The main goal of this research project is to develop the machine learning models to predict the mental illness of the users(people with poor mental health). Here we used some of the machine learning models to solve this issue. For to observe clearly, we took datasets of the mental health records of the people to train our models. And also we need to do preprocessing on that dataset because the dataset may have noisy data or maybe invalid data. The data is being subject to various machine learning techniques to obtain labels.

KEYWORDS : Mental Health, Machine Learning , Self-Analysis, Mental Health Counselling

2. INTRODUCTION

Machine Learning is a subfield of Artificial Intelligence that primarily concentrates on using data and the algorithms to mimic how human learns and eventually improves its accuracy. So ml is useful to build the model using different algorithms and then testing and training performed on this huge data. Advanced probabilistic and statistical techniques are used in ml to build computers that can learn from data. It is divided into 3 types:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

This study has focussed on the advantages of machine learning for better mental health identification. Here we used some of the best ML algorithms like Random Forest, KNN, Decision Tree, Logistic Regression and Stacking. All these algorithms are best fit to build a model. Overall, machine learning has the potential to improve the clinical and research efficiency and also providing fresh insight into mental health.

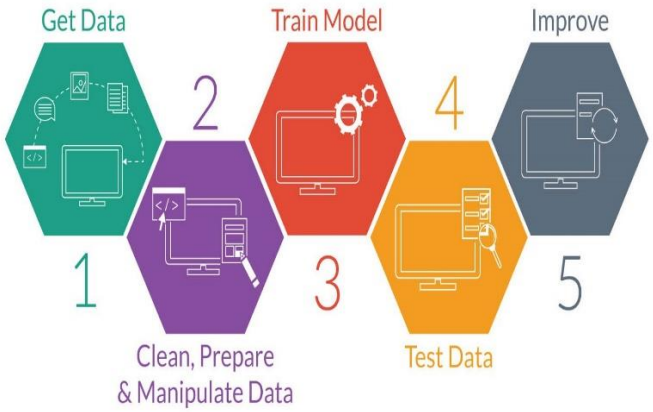


figure : process of building a model

3. LITERATURE REVIEW

This study evaluated the effectiveness of various ml algorithms that categorise the dataset into distinct mental health disorders. This system was developed to assess a person's mental health status and prediction models were built using this framework.

Now-a-days many people are afflicted with mental health conditions like depression, stress, and anxiety as well. Mainly the people who are under the employment sector is mainly facing this issue. For this we need to implement something so that it will be so easy for them to get proper knowledge about their mental health issues. The main aim of this project is to check whether a person needs a treatment or not so that with this result they can proceed to next level treatments. Here we used Random Forest, KNN, Logistic Regression, Decision Tree, AdaBoost, Gradient Boosting, and XGB Classifier. Among these models, XGB performed well.

Hence, in order to obtain correct results for this project, we can utilize this classifier. The XGB has 86% accuracy rate. 29 columns and 1259 entries make up the dataset. The tests revealed that the performance of GD Boost, Decision Tree, and Random Forest was substantially identical.

The field of medicine is beginning to focus more on artificial intelligence, which will aid in the study and application of mental health. To fully enact the pledge Scientists, physicians, patients, and regulators are just a few of the diverse professions involved in mental health research and care who must connect and communicate with AI. In this study, discourse analysis was used to analyse the data in order to better understand the representational practises used in machine learning (HCML). Using this , they have discovered a collection of 55 interdisciplinary research that predict the mental health state based on social media data. The methods employed on our dataset are listed below along with their accuracy levels..

So these are the algorithms which is helpful to predict the mental health of a person.

Methods	Accuracy (%)
Logistic Regression	82
Random Forest	85
Decision Tree	85
Gradient Boosting	85
XG Boost	86
SVM	84

4.PROPOSED SYSTEM

4.1 DATASET DESCRIPTION :

In this dataset we have 27 columns and 1259 entries which are Timestamp, Age, Gender, Country, State, ……

1.Timestamp – It describes about time of an event that a computer records.

2. Age – It is about the age of a person

3. Self_employed – whether a person is employed or not

4. work_interfere – whether the work is interferes or not

5. family_history – if the person is having any family background regarding the mental health issue

6. remote_work – does remote work is effects the person or not

7. benefits – does the company provides the benefits or not

8. mental_health_consequence – does the person will think that he will have negative consequence if he/she shares his condition

9. phys_health_consequence – does the person will think that he will have negative consequence if he/she shares his condition

10. coworkers – does the person is willing to share his condition with their coworkers

ge	Gender	Country	state	self_empl	family_his	treatment	work_inte	no_empic	remote_
37	Female	United Sta	IL	NA	No	Yes	Often	Jun-25	No
44	M	United Sta	IN	NA	No	No	Rarely	More thar	No
32	Male	Canada	NA	NA	No	No	Rarely	Jun-25	No
31	Male	United Kir	NA	NA	Yes	Yes	Often	26-100	No
31	Male	United Sta	TX	NA	No	No	Never	100-500	Yes
33	Male	United Sta	TN	NA	Yes	No	Sometime	Jun-25	No
35	Female	United Sta	MI	NA	Yes	Yes	Sometime	01-May	Yes
39	M	Canada	NA	NA	No	No	Never	01-May	Yes
42	Female	United Sta	IL	NA	Yes	Yes	Sometime	100-500	No
23	Male	Canada	NA	NA	No	No	Never	26-100	No
31	Male	United Sta	OH	NA	No	Yes	Sometime	Jun-25	Yes
29	male	Bulgaria	NA	NA	No	No	Never	100-500	Yes
42	female	United Sta	CA	NA	Yes	Yes	Sometime	26-100	No
36	Male	United Sta	CT	NA	Yes	No	Never	500-1000	No
27	Male	Canada	NA	NA	No	No	Never	Jun-25	No
29	female	United Sta	IL	NA	Yes	Yes	Rarely	26-100	No
23	Male	United Kir	NA	NA	No	Yes	Sometime	26-100	Yes
32	Male	United Sta	TN	NA	No	Yes	Sometime	Jun-25	No

Figure – dataset image

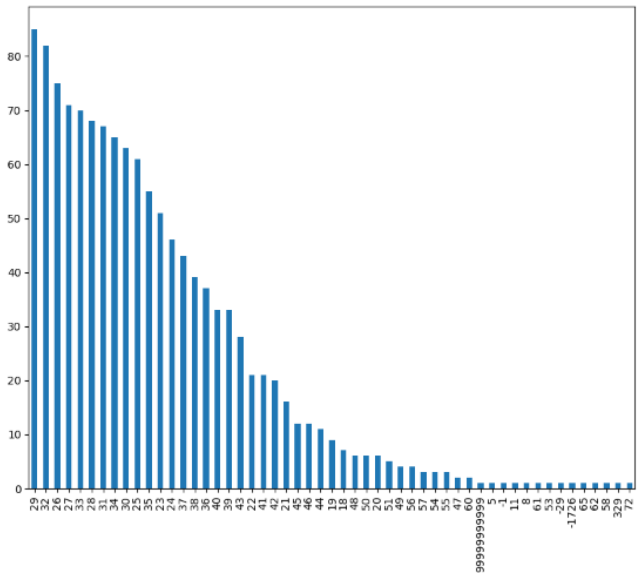
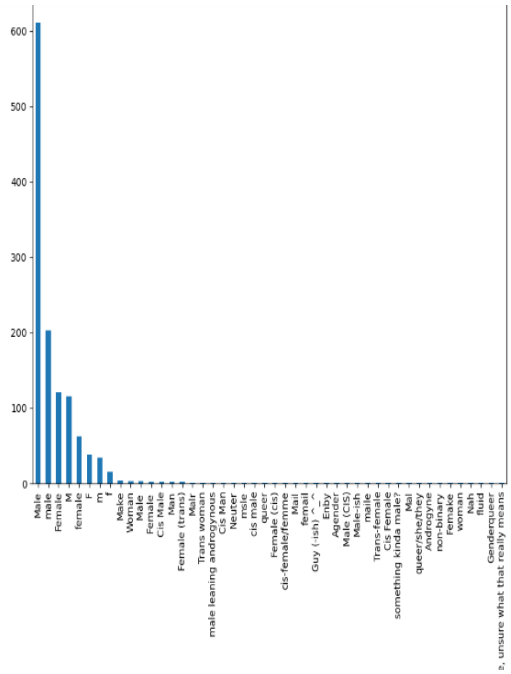


Figure – age distribution



The figure is about the dataset and we have some invalid data. So we need to do pre-processing on that dataset. Here we will remove the invalid data from the age column because the age of a person can not be a negative value and '9999999' cannot be the age of ay person. The columns which are removed from dataset is 'comments' because that column has majority null values so we need to delete the whole column in order to get the accurate results.

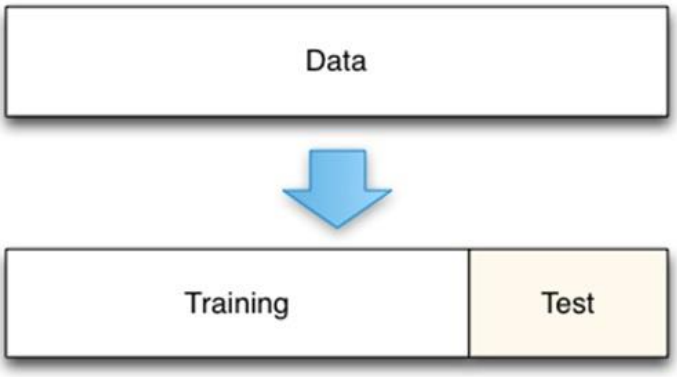
4.2 DATA PREPROCESSING :

Preparing the data is crucial stage in machine learning. It is crucial that we pre-process our data before fitting it into our model since the valuable information that can be deduced from it and the quality of the data directly influence how well our model learns. Data engineering and feature engineering are both involved in the pre-processing of the data for ml. The process of transforming unprepared data into prepared data is known as data engineering. The data that has been prepared is then tuned as a result of feature engineering to provide the features that our ML model anticipates.We must first determine whether our dataset contains any null values as the first stage in the pre-processing procedure. If we find more number of null values in any column then simply we can drop that column. If we check our age and gender column, we have invalid data.

33	M
37	male
23	Male
43	M
32	Male
26	Male
32	Male
37	F
29	Male
34	Male
27	male
30	male
29	Male
32	Male
-1726	male
25	Male
37	male
29	Male
27	Man

Here in both the age and gender column we have invalid data. For to remove those values from our dataset we need to perform pre-processing. After filtering our data we need to do Ordinal Encoding on our attributes because majority of the attributes are categorical values. By using ordinal encoding, it will convert the categorical values to numerical values so that it can be easy to train our model. And Label Encoding is performed only on the target value i.e. treatment.

4.2.2 Splitting the data :



Splitting the data makes sense since we are looking for a pattern that will accurately reflect the majority of the data points while minimising error when we train a machine learning model. Two common mistakes occur throughout this process: overfitting and underfitting.

4.3 MODEL SELECTION

Model selection can help in choosing better hyperparameters. Model selection is straightforward.

4.5.1 RANDOM FOREST

The primary idea behind the random forest is that it blends many trees that are utilised to determine the dataset's class. Also, this technique employees three concepts: randomly selecting training data when creating trees; selecting particular subsets of variables when splitting each node.

4.5.2 LOGISTIC REGRESSION

The likelihood of a dependent or a target variable is predicted using supervised machine learning. We can establish and predict the associations between the dependent variable and one or more independent variables using logistic regression.

4.5.3 DECISION TREE

A supervised learning approach called a decision tree is utilised for both classification and regression tasks. A root node serves as the first node in a decision tree and has no outgoing branches. It uses a greedy search to implement a divide and conquer approach.

4.5.4 SUPPORT VECTOR MACHINE

Regression and classification issues are both addressed by it. The SVM's objective is to construct the optimal line that can divide n-dimensional space into classes.

4.5.5 ADABOOST

AdaBoost is a particular kind of boosting that functions as an ensemble model to advance any machine learning technique's prediction.

4.5.6 GRADIENT BOOSTING

It is a method of machine learning that is applied to classification and regression issues. It provides a very ineffective ensemble of models for prediction in the form of a prediction model.

4.5.7 XGBOOST

It is an enhanced distributed gradient boosting library created for effective and scalable machine learning model training. In an ensemble model, weak predictions from various weak models are combined to yield a stronger forecast.

5. RESULTS

Thus, we applied a variety of machine learning models in this case, including Random Forest, Logistic Regression, Decision Tree, SVM, Adaboost, Gradient Boosting, and XGboost. Among these models, XGboost has the highest accuracy and best results thanks to its 86% accuracy.

Methods	Accuracy (%)
Logistic Regression	82
Random Forest	85
Decision Tree	85
Gradient Boosting	85
XG Boost	86
SVM	84

In the above table we can see that how much accuracy that each models has and also random forest, decision tree and gradient boosting is having same accuracy.

6. OUTPUT SCREENS

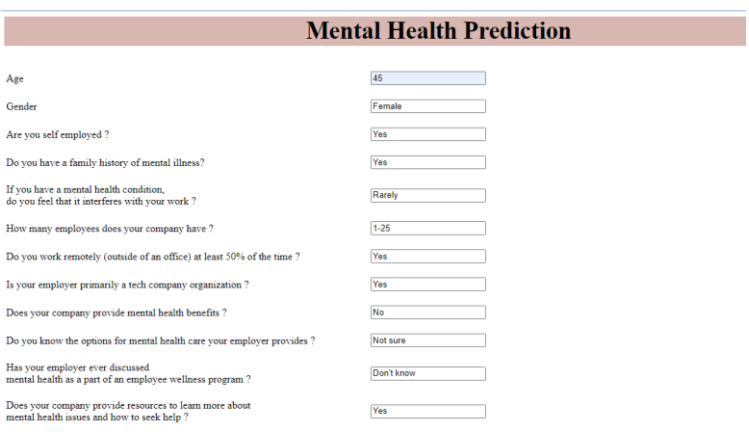


Figure – filling form

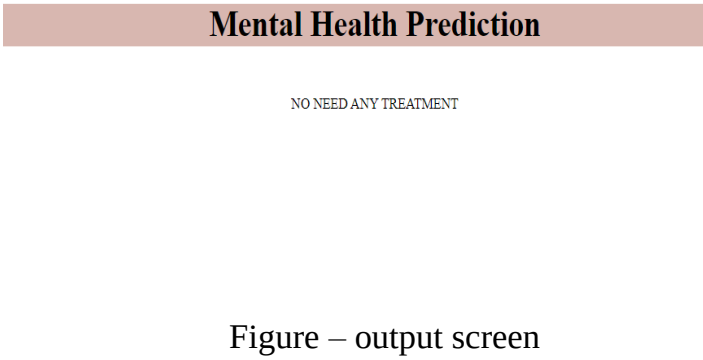


Figure – output screen

7. 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.