Analysing the causes of Mood Disorders : A Comprehensive Study

[[1]Vidushi Gandhi, [2]Stuti Saxena, [3]Dr Vikesh Kumar]
[1,2,3,Department of Computer Science and Engineering
Echelon Institute of Technology, Faridabad, Haryana,
India]

[20cse058vidushi@eitfaridabad.co.in,stutisaxena@eitfaridabad.co.in, vikeshkumar@eitfaridabad.co.in]

Abstract

Depression or Mood Disorder has been one of the most prevalent mental disorders. According to WHO in 2020, more than 380 million people were diagnosed with Depression and more than 1050 thousand deaths have been reported globally. that can cause a loss of interest in general action that can lead to suicidal thoughts. This research paper aims to use different machine-learning techniques namely KNN, SVM(Support Vector Machine), Decision Tree, Naive Bayes and Logistic Regression for depression detection. Early detection and accurate prediction are fundamental to identifying patients who could benefit from the treatment and reduce the mortality rate due to depression. With the help of various ML techniques, we can detect depression more efficiently and effectively. This research's primary contribution is exploring the features and their impact on detecting depression levels.

1. Introduction

This analysis deals with various machine-learning algorithms on the Twitter sentiment analysis dataset.

Depression or Mood Disorder is one of the most common mental disorders and one of the most prevalent mental disorders. According to WHO around 8% of adults are suffering from this disease. Globally more than 350 million patients were diagnosed with this disease. But if we detect this disease, we can save many lives and improve the health condition. Much research is going on in this field. This research paper aims to use different machine-learning techniques namely KNN, SVM(Support Vector Machine), Decision Tree, Naive Bayes and Logistic Regression for depression detection. This paper deals with detecting whether a patient is suffering from depression or not. This will help the doctor to cure the patient of this disease if they know it in the earlier stage. This can be cured if it is diagnosed early with proper medication and treatment. Using the Naive Bayes technique, the accuracy level is greater than 96% among different machine learning techniques. Developing a computer-based algorithm can help the doctor detect the patient's condition earlier.

2. Technology Used

This division deals with the various ML Algorithms we used in the analysis. Five popular ML algorithms namely, K Nearest Neighbours, Support Vector Machines, Naive Bayes, Decision Trees and Logistic Regression have been used to perform experiments. These five algorithms are Supervised Learning algorithms, used for classification.

2.1 KNN

The k-nearest neighbours, popularly called as KNN, is a supervised ML algorithm, that uses the closeness concept to make forcasting for the grouping of a single point. It is mainly used for real-life problems related to classification or regression issues, it is mostly used for categorization problems, based on the belief that similar data could be found near to each other. It identifies data that are split into several classes to forecast the classification of a new data point. It is also known as the 'LAZY' algorithm as it only memorises the process and it does not learn itself. It lags the property of self-learning, and can not take any decision by itself. That's why it is called a lazy algorithm.

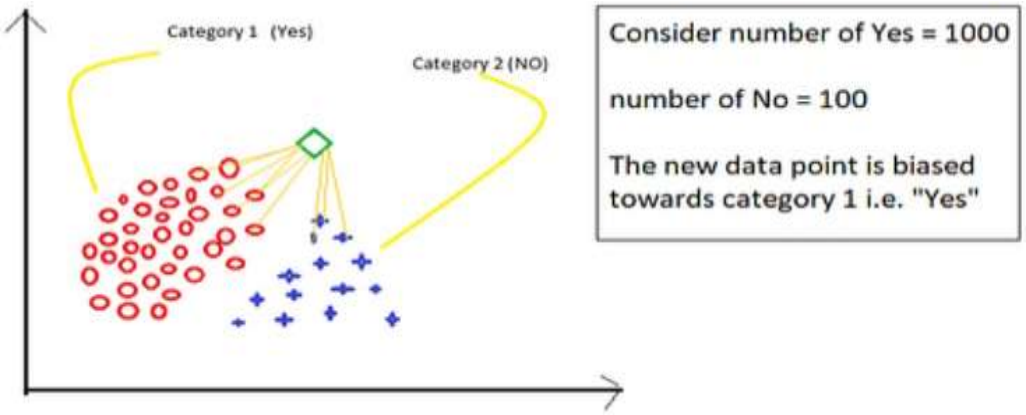

Figure 1. K-nearest neighbours algorithm on a dataset [12]

2.2 SVM

The main aim of this model is to construct the most suitable line which easily sorts out complex space into various categories. This algorithm selects the extreme cases which are called support vectors that create the hyperplane. Hence, the algorithm is termed a Support Vector Machine.

It is fundamentally based on a statistical approach. There can be several hyperplanes passing through the same point. SVM tries to find the best hyperplane and classifies the two classes perfectly. SVM does this by finding the maximum margin between the hyperplanes which means two classes.

Figure 2 shows the classification done by the SVM algorithm.

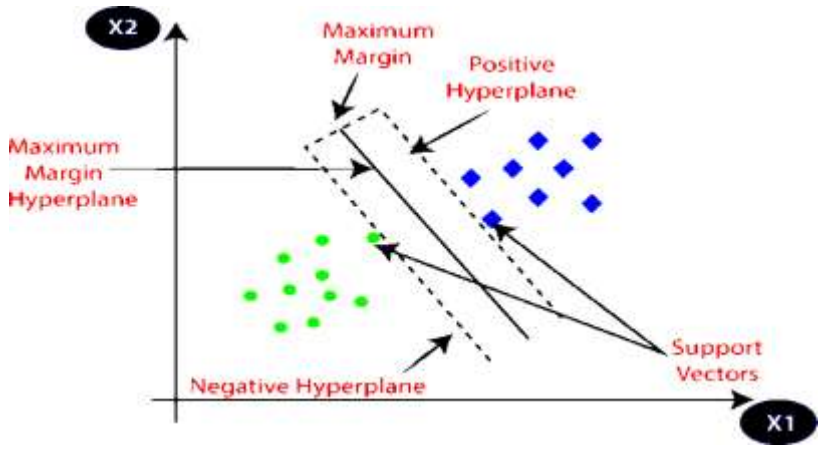

Figure 2. SVM algorithm [14]

2.3 Naive Bayes

The Naïve Bayes algorithm is a supervised learning algorithm, based on the Bayes theorem and used for solving classification problems. It is used in *text classification* that includes a high-dimensional training dataset. Naive Bayes Classifier is one of the simplest and most effective Classification algorithms that help in building fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

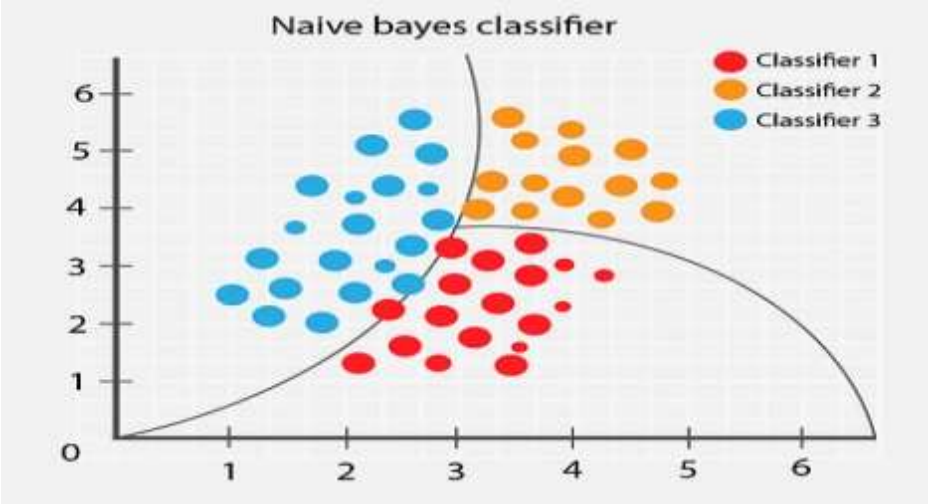

Figure 3. Naive Bayes Algorithm [16]

2.4 Decision Trees

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mainly it is preferred for solving Classification problems. It is a

tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

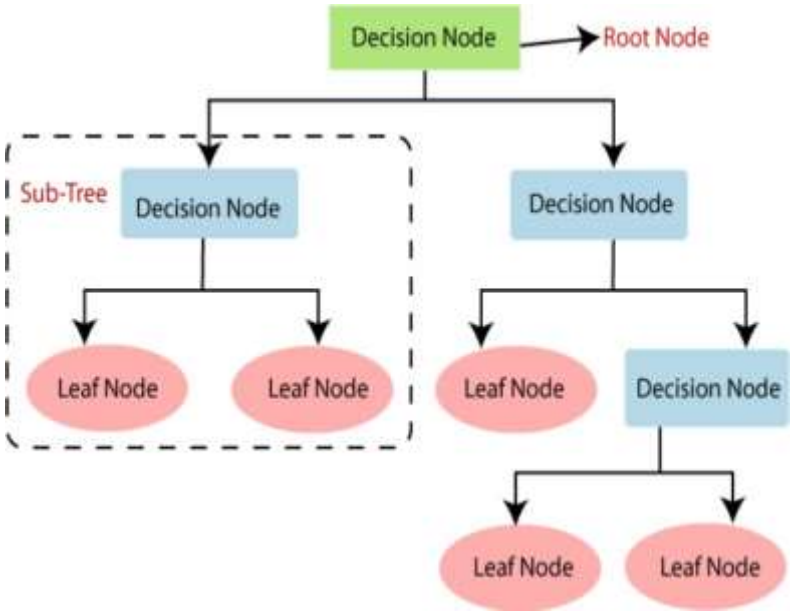

Figure 4. Decision Tree Algorithm [15]

2.5 Logistic Regression

It is a supervised ML technique where a relationship between an Independent and a Dependent Variable is established. It is how one can easily predict how a variable would behave if any of them is changed or removed or something else. This is majorly used in the fields of Financing and Investing, Share market etc. This is very much capable of identifying the trends that would occur in a share market. This algorithm mainly works on probability.

It is used for calculating or predicting the possibility of the occurrence of an event. One example can be to decide if a person's sentiment is positive or not with this dataset. There could be 2 outcomes to this - occurrence or not occurrence - this is known as binary classification.

Figure 5. LR algorithm on a dataset [13]

3. Literature Survey

The research work done by various researchers in the area of Mood Disorder Analysis is summarised in Table 1.

Table 1. Related Work

S.No.	Year Of Publishing	Name of Author	Techniques Used	Accuracy	Dataset Used
1.	2018	Ahmed Husseini Orabi [1]	Deep Learning	87.37	Depression Detection
2.	2018	Md. Rafiqul Islam [2]	Machine Learning	72	Facebook Dataset
3.	2021	Raymond Chiong [3]	LSTM	91	Facebook, Twitter Dataset
4.	2020	Hatoon S. ALSAGRI [4]	Random Forest, SVM etc	86.6	Depression Detection
5.	2018	JT Wolohan [6]	NLP	78	Depression Detection
6.	2021	Rajawat [7]	Neural Network	91.7	Depression Detection

4. Methodology

In this division, we will discuss the progress followed in this research. These figures deal with the process involved in this methodology . In these steps is the collection of the ideal data set pursued with data preprocessing and data cleaning, then data partitioning and implementation of the model by applying various machine learning techniques namely Logistic Regression, KNN, SVM, Naive Bayes and Decision Trees then by result analysis and predictions . Eventually, the model has been evaluated using various parameters. Each step in the process has been described below.

Figure 4 Methodology

4.1 DataSet Collection

For Depression Detection, data selected for this analysis has been taken through the Twitter API. This is the sentiment140 dataset. [17]
It contains 1,600,000 tweets extracted using the Twitter API. The tweets have been annotated (0 = negative, 2 = neutral, 4 = positive) and they can be used to detect sentiment.
It contains the following 6 fields:

1. Target:the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
2. Ids: The id of the tweet (2087)
3. Date: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
4. Flag: The query (lyx). If there is no query, then this value is NO_QUERY.
5. User: the user that tweeted (robotickilldozr)
6. Text: the text of the tweet (Lyx is cool)

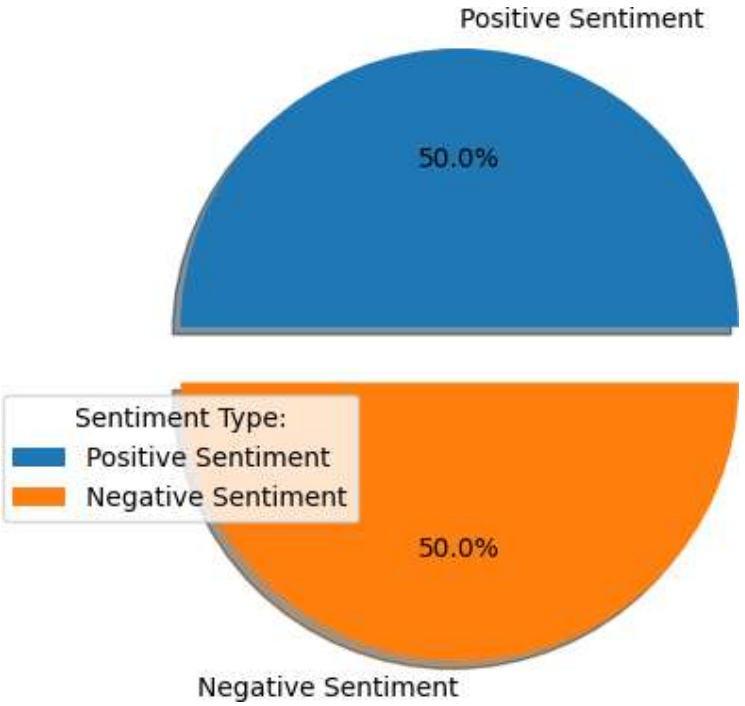

Figure 5. Percentage of Positive and Negative sentiment identified out of 1600000 tweets

4.2 PreProcessing And Data Cleaning

This process involves checking the datasets, removing absent values , finding out outliers present in this twitter datasets . This process also involve removing of punctuations, emoji, hyperlinks from the textual data and converting this into a suitable form for applying the algorithm and selecting attribute.

Feature Selection : Out of 6 features we have selected 2 features for the analysis

4.3 Data Partitioning

In this research, the dataset is partitioned by a 95% -5 % rule into two different division train datasets and test datasets. On this training dataset which is 95% machine learning algorithm is applied and the test data is used for testing the algorithm. The first image is of the testing dataset { X= x_test & Y=y_test}The second image is of the training dataset{ X= x_train & Y=y_train}.

4.4 Machine Learning Technique

In this model was made using different Machine Learning techniques — KNN, Naive Bayes, LR, Decision Trees and SVM. These models were applied to the training and testing dataset.

The analysis of the results was obtained after using algorithms, and it was concluded by calculating accuracy in both training and testing datasets. Initially, the total no. of records was 1600000 tweets. The research dataset was split according to 95-5% in training data the records were 1520000 and for testing data 80000 tweets.

4.5 Result Analysis

These sections show us the results and analysis of different machine-learning techniques used in this research study

Table 2 shows us the results obtained in this research work

Table 2. Results obtained

Model Used	Training Accuracy	Testing Accuracy
KNN	95.78	93.68
SVM	95.51	-
Logistic Regression	81.3	79.6
Naive Bayes	84.3	77.6
Decision Tree	87.6	85.56

The results show the accuracy achieved by different machine learning algorithms applied both on training and testing datasets. The highest accuracy received in the training dataset is 95.78% and the highest accuracy received in the testing dataset was 93.68%

4.6 Prediction

This research work helps in detecting depression at the early stage which can help doctors improve the patient's health and save the life of the people [8]. By using this model it can help them to detect the mood disorder and analysis their sentiments

5. Conclusion and Future Scope

The ML approaches can be used to assist in diagnosing mental health conditions. PTSD, schizophrenia, depression, ASD, and bipolar diseases lie in the domains of mental disorders. Social media data, clinical health records, and mobile device sensor data can be analyzed to identify mood disorders. This research paper aims to provide information about basic concepts of ML algorithms frequently used in the mental health domain, specifically for depression and their practical application. According to the research work, applications based on machine learning provide a significant potential for progress in mental healthcare, including the prediction of outcomes and therapies for mental illnesses and depression.

In future work in this research, other ML techniques and Deep Learning techniques can be applied for better prediction accuracy. Moreover, a dataset consisting of many instances can be collected and used for performing experiments.

6. Reference

1. Orabi, Ahmed Husseini, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. "Deep learning for depression detection of twitter users." In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pp. 88-97. 2018.
2. Islam, Md Rafiqul, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M. Kamal, Hua Wang, and Anwaar Ulhaq. "Depression detection from social network data using machine learning techniques." *Health information science and systems* 6 (2018): 1-12.
3. Chiong, Raymond, Gregorius Satia Budhi, Sandeep Dhakal, and Fabian Chiong. "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts." *Computers in Biology and Medicine* 135 (2021): 104499.
4. AlSagri, Hatoon S., and Mourad Ykhlef. "Machine learning-based approach for depression detection in twitter using content and activity features." *IEICE Transactions on Information and Systems* 103, no. 8 (2020): 1825-1832.
5. Deshpande, Mandar, and Vignesh Rao. "Depression detection using emotion artificial intelligence." In *2017 international conference on intelligent sustainable systems (iciss)*, pp. 858-862. IEEE, 2017.
6. Wolohan, J. T., Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. "Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP." In *Proceedings of the first international workshop on language cognition and computational models*, pp. 11-21. 2018.
7. Rajawat, Anand Singh, Romil Rawat, Kanishk Barhanpurkar, Rabindra Nath Shaw, and Ankush Ghosh. "Depression detection for elderly people using AI robotic systems leveraging the Nelder–Mead Method." In *Artificial Intelligence for Future Generation Robotics*, pp. 55-70. Elsevier, 2021.
8. Joshi, Manju Lata, and Nehal Kanoongo. "Depression detection using emotional artificial intelligence and machine learning: A closer review." *Materials Today: Proceedings* 58 (2022): 217-226.
9. AlSagri, Hatoon S., and Mourad Ykhlef. "Machine learning-based approach for depression detection in twitter using content and activity features." *IEICE Transactions on Information and Systems* 103, no. 8 (2020): 1825-1832.
10. Yating GU, Chi ZHANG, Xiaojian JIA, Shiguang NI. "Text-based depression detection dataset (Chinese)." doi: 10.21227/zsfc-av84
11. Giachanou, Anastasia, and Fabio Crestani. "Like it or not: A survey of twitter sentiment analysis methods." *ACM Computing Surveys (CSUR)* 49, no. 2 (2016): 1-41.

12. https://in.images.search.yahoo.com/search/images;_ylt=AwrPrEzP.2Rk19QbYnq9HAx.;_ylu=c2VjA3NlYXJjaARzbGsDYXNzaXN0;_ylc=X1MDMjExNDcyMzAwNQRfcgMyBGZyA21jYWZlZQRmcjIDc2EtZ3Atc2VhcmNoBGdwcmlkA3pJTTZGdGQ1VE1pWmJCN3hPSDFmRUEEbl9yc2x0AzAEbl9zdWdnAzEwBG9yaWdpbgNpbi5pbWFnZXMuc2VhcmNoLnlhaG9vLmNvbQRwb3MDMgRwcXN0cgNrbm4gBHBxc3RybAM0BHFzdHJsAzIzBHF1ZXJ5A2tubiUyMGluJTIwbWFjaGluZSUyMGxlYXJuaW5nBHRfc3RtcAMxNjg0MzM5NzgwBHVzZV9jYXNlAw--?p=knn+in+machine+learning&fr=mcafee&fr2=sa-gp-search&ei=UTF-8&x=wrt&type=E211IN826G0#id=3&iurl=https%3A%2F%2Fstatic.javatpoint.com%2Ftutorial%2Fmachine-learning%2Fimages%2Fk-nearest-neighbor-algorithm-for-machine-learning5.png&action=click

13. https://in.images.search.yahoo.com/search/images;_ylt=AwrPqBw6_WRkkowWxAu7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3BpdnM-?p=logistic+regression&fr2=piv-web&type=E211IN826G0&fr=mcafee#id=5&iurl=https%3A%2F%2Fmiro.medium.com%2Fmax%2F1280%2F1*iKo3KI4kqkZ47W7pmmH4cw.png&action=click

14. https://in.images.search.yahoo.com/search/images;_ylt=Awr1Skrf_WRksZgVe8.7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3BpdnM-?p=svm+in+machine+learning&fr2=piv-web&type=E211IN826G0&fr=mcafee#id=0&iurl=https%3A%2F%2Fstatic.javatpoint.com%2Ftutorial%2Fmachine-learning%2Fimages%2Fsupport-vector-machine-algorithm.png&action=click

15. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.javatpoint.com%2Fmachine-learning-decision-tree-classification-algorithm&psig=AOvVaw2AS04ztWZMIyhlwMNK7AJ4&ust=1717413756196000&source=images&cd=vfe&opi=89978449&ved=0CAQQjB1qFwoTCODt2vjmvIYDFQAAAAAdAAAAABAE

16. https://www.researchgate.net/publication/372837562/figure/fig1/AS:11431281194657281@1696198174384/Naive-Bayes-Classifier-The-Naive-Bayes-algorithm-is-a-statistical-method-in-machine.png

17. https://www.kaggle.com/datasets/kazanova/sentiment140/data