Detecting Depression using Decision Trees

[[1]Manik Jain, [2]Dr Stuti Saxena, [3]Dr Sudarshan Goswami]

[1,2,3, Department of Computer Science & Engineering
Echelon Institute of Technology, Faridabad, Haryana, India]

[jainmanik103@gmail.com
stuti.saxena26@gmail.com
sudarshangoswami@eitfaridabad.co.in]

Abstract

Depression has been one of the leading factors of death in recent years, especially in youngsters' lives. According to WHO, Approximately 280 million people have depression and 50%it is more commonly detected in women as compared with men.. This research paper aims to use different Supervised machine-learning techniques namely KNN, SVM(Support Vector Machine), Logistic Regression,Decision Trees and Random Forest for detecting the depression rate and the suicide factor related to it. Our main objective is to determine whether the patient is diagnosed with depression. The Machine Learning algorithms were applied to the obtained dataset and the algorithms were evaluated using Accuracy as the performance measure. The obtained accuracy scores after applying KNN, SVM, LR, DT, RF are 90.7%, 87.2%, 86.2%, 99.89%, 98.8% respectively.Early detection of this upcoming fatal diesease is very important as on the later stage it can be one of the reasons of performing life-threatening tasks like suicide attempts, etc. . With the help of various ML techniques, we can detect Depression more efficiently and effectively.

1. Introduction

This paper deals with the analysis of various machine-learning techniques on the Depression. Depression is one of the life-threatening diseases. If we are able to detect the Depression at an early stage we can save many lives and improve the survival life expectancy ratio. It's estimated that 1 in 3 women and 1 in 5 men will experience major depression in their lives. Other conditions, such as schizophrenia and bipolar disorder, are less common but still have a large impact on people's lives**[11]**. Much research is going on in this field. The scope of this paper is to analyze different machine learning techniques namely Support Vector Machines, Logistic Regression, Linear Regression etc. regarding this problem. This paper deals with detecting whether the person had depression is detected with Depression or not. It will help the doctor to control the mood/Behaviour in this part if they know it in the earlier stage. This

can be cured if it will be diagnosed in an early stage with proper medication and treatment. By developing a computer-based algorithm we can help the doctor to detect the patient's condition earlier. There are 5 stages of Depression, higher the stage the rate of survival will be less. Table 1 represents the various stages of Depression.

Table 1. Stages of Depression

Stage	Description
1. Normal Mood	At this stage, a person experiences normal mood and functioning without any significant depressive symptoms.
2. Sadness	Mild feelings of sadness, low energy, and general unhappiness may start to emerge. This stage can be a normal reaction to life stressors or challenges.
3. Mild Depression	Depressive symptoms intensify, leading to persistent sadness, changes in sleep patterns, appetite disturbances, and decreased motivation. Social and occupational functioning may be affected.
4. Moderate Depression	Symptoms become more severe and persistent. The individual may experience increased isolation, difficulty concentrating, and a marked decrease in daily functioning. Suicidal thoughts may emerge.
5. Severe Depression	This stage involves a profound and debilitating depression with a high impact on all aspects of life. It can lead to severe physical and emotional symptoms, such as a complete loss of interest in life, self-neglect, and a high risk of self-harm or suicide. Professional help is critical at this stage.

1. Literature Survey

The research work done by other authors in the field of Depression detection is summarized in Table 2.

Table 2. Related Work

S.No.	Year Of Publishing	Name of Author	Techniques Used	Accuracy	Dataset Used
1.	2015	Meenal J. Patel , Alexander Khalaf , Howard J. Aizenstein [1]	Machine Learning	88.25%	Depression Dataset
2.	2017	Le Yang, Dongmei Jiang,Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, Hichem Sahli [2]	Convolutionary Neural Network algorithm(CNN)	93.54 %	Depression Dataset

3.	2017	Arkaprabha Sau, Ishita Bhakta [3]	Naïve Bayes, Support vector machines, Decision trees, simple CART etc.	93.07%	Depression Dataset
4.	2018	Md. Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed,Abu Raihan M. Kamal, Hua Wang & Anwaar Ulhaq [4]	Naive Bayes,SVM,Decision Trees, and Logistic Regression	95.04%	Depression Dataset
5.	2019	Ahnaf Atef Choudhury, Md. Rezwan Hassan Khan , Nabuat Zaman Nahim, Sadid Rafsun Tulon, Samiul Islam, Amitabha Chakrabarty[5]	Random Forest,K-NN, and Naive Bayes	92.36%	Depression Dataset.

6.	2019	Xiaowei Li , Xin Zhang , Jing Zhu , Wandeng Mao , Shuting Sun , Zihan Wang , Chen Xia , Bin Hu [6]	Logistic Regression	95.14%	Depression Dataset
7.	2020	Anu Priya ,Shruti Garg, Neha Prerna Tigga [7]	Machine Learning	94.68%	Depression Dataset
8.	2021	Umme Marzia Haque ,Enamul Kabir,Rasheda Khanam [8]	Machine Learning	86%	Depression Dataset
9.	2022	Sri Sweta Bhadra &Chandan Jyoti Kumar [9]	Machine Learning	78%	Depression Dataset
10	2020	Prince Kumar , Shruti Garg, Ashwani Garg [10]	Machine Learning	89.68%	Depression Dataset

3. Techniques Used

In this section , we will understand each and every algorithm we have used in our model which helps in appropriate prediction of Depression. We have used K-NN, SVM(Support Vector Machine), Logistic Regression, Decision Trees, Random Forest.

3.1 K-NN

The KNN, is a supervised Machine Learning algorithm, that uses the closeness concept to make

categories for the grouping of single points. It is typically used for problems related to classification/ regression issues, it is based on the belief that similar data could be found near to each other.

Data Points are splitted into several classes to forecast the classification of a new data point. Its other name is 'LAZY' algorithm as it does not learn anything either on its own nor by the user , it only tries to memorize the process.

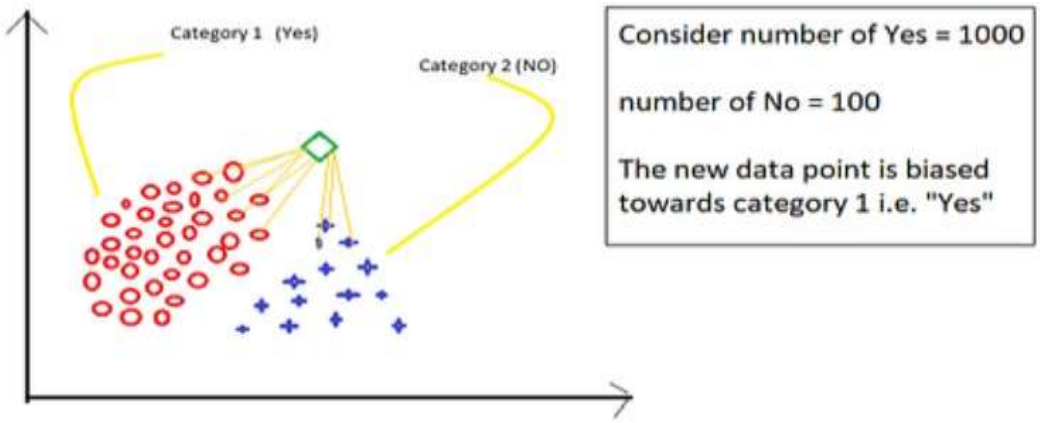


Figure 1. k-nearest neighbor's algorithm on dataset **[12]**

3.2. Logistic Regression(LR)

It is a supervised ML technique where there is a relationship between an Independent and a Dependent Variable. This is majorly used in the fields of Commerce , Share market etc. This is very capable of predicting the upcoming trends that would occur in a share market. This algorithm mainly works on the probability.

It is basically used for calculating the possibility of the occurrence of an event. One example can be to decide if a person is Suffering From Depression or not. There could be 2 outcomes to this - occurrence or not occurrence - this is known as binary classification [10].

Figure 2. LR algorithm on dataset [13]

3.3. Support Vector Machines

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 help in creating the hyperplane. Hence, the algorithm is termed as Support Vector Machine {S.V.M.}.

SVM is based on statistical approaches. There can be n-number of 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 two classes. Figure 3 shows the classification done by the SVM algorithm.

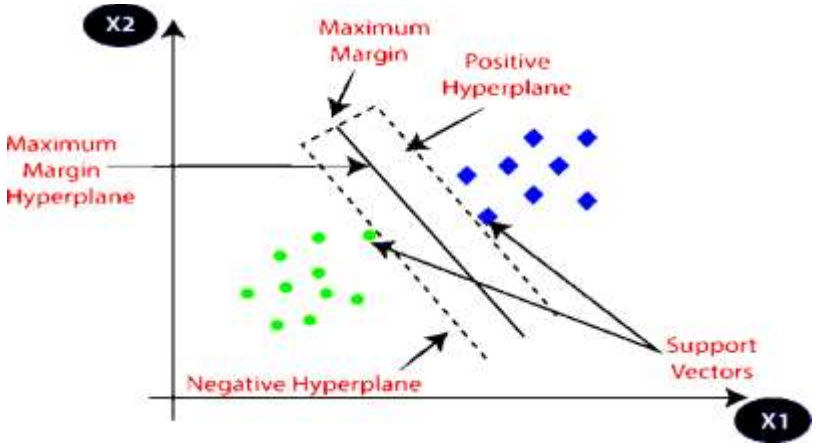


Figure 3. SVM algorithm [14]

3.4. K - Nearest Neighbors

This model is basically used for the purpose of Classification as well as Regression problems.It relies on the idea that similar data points tend to have similar labels or values.During the training phase, the KNN algorithm stores the entire training dataset as a reference. When making predictions, it calculates the distance between the input data point and all the training examples, using a chosen distance metric such as Euclidean distance.**[15]**

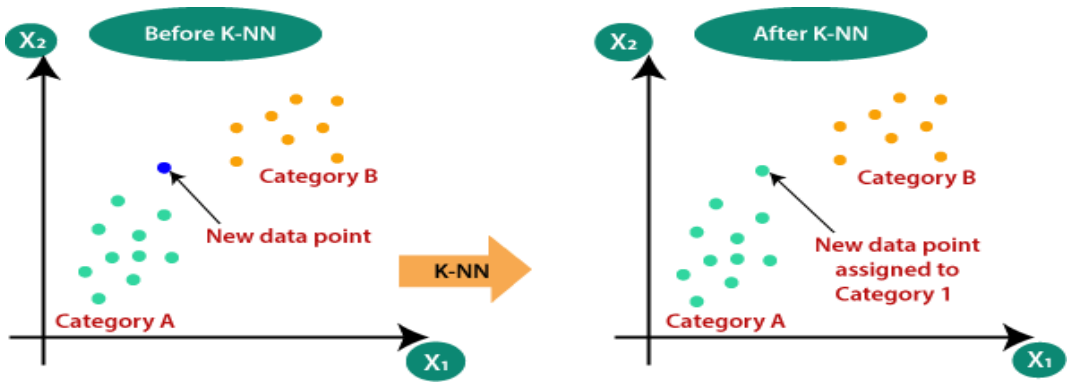


Figure 3. K-NN algorithm [16]

3.5. Random Forest

Random Forest is a machine learning ensemble technique used for both classification and regression tasks. It's a popular and powerful algorithm that combines multiple decision trees to make more accurate predictions. It is very easy to use and very much flexible .

Random Forest produces the best output of all the time as compared to other algorithms even if less cleaning had been provided to it.

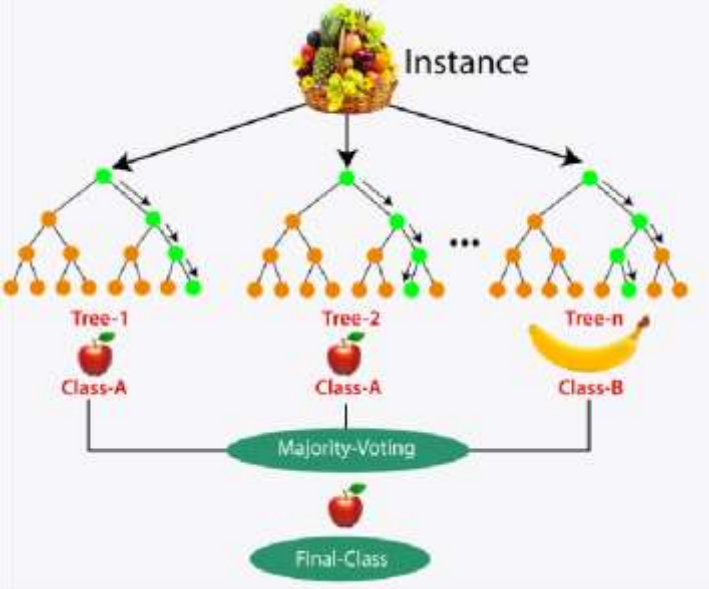
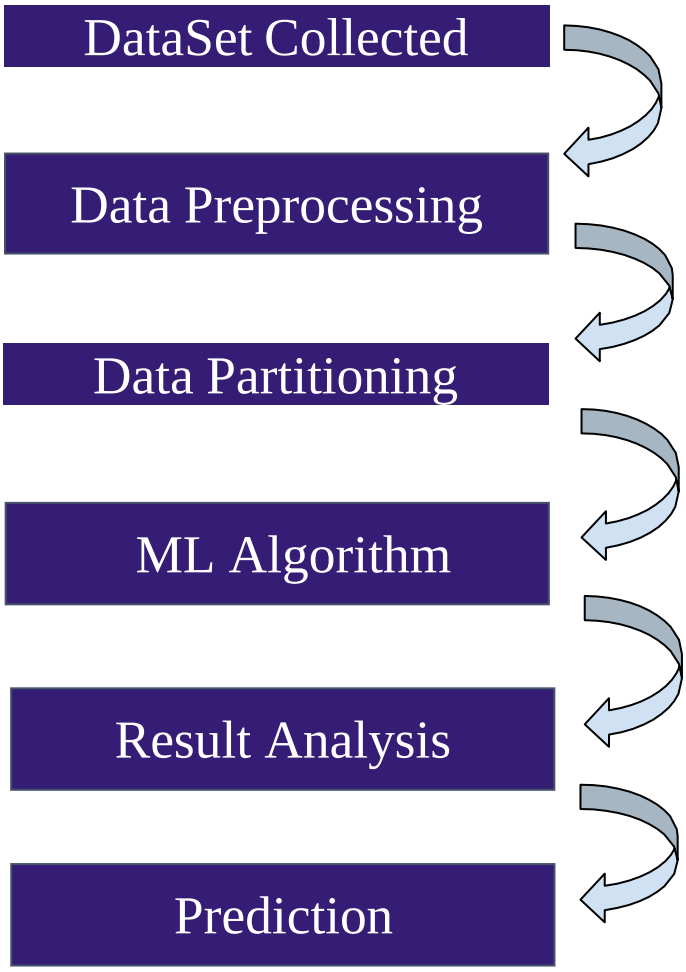


Fig. 5 Random Forest[17]

4. Methodology

This part deals with the variety of steps involved in our study, Figure displays different stages of methods and procedures used. The methodology proceeds with the dataset collection,data pre-processing, then data partitioning and implementation using different ML techniques used namely KNN, SVM, Logistic Regression, Decision Trees, and Random Forest followed by result analysis. Eventually, the model has been tested and trained in diverse conditions. Each step in this technique has been described below.



4.1 Dataset Collection

For Depression Detection, The dataset we have selected for the analysis purpose is from Kaggle. It consists of the records of 919 Depressed patients which are present in the Govt. Repository. The features were detected when the doctor had a 1 to 1 interaction with the patient.It is very essential to have a word because it is the only source of gathering information regarding the same.

The characteristics of the taken dataset are:-

1. Age
2. Sex
3. Chest Pain
4. Blood Pressure { B.P.}
5. Fasting Blood Sugar {FBS}
6. Electrocardiography {ECG}

7. Maximum Heart Rate
8. Exercise Angina
9. Old Peak
10. ST_Slope
11. Heart Disease

We have used the above mentioned factors to determine whether a person had Depression or not.

4.2 Data Pre-Processing

This is a process which is made up for examining the dataset for absent values , finding and getting rid of any outliers, converting this into an appropriate form for applying the algorithm and selecting attribute

4.3 Data Partitioning

It was separated according to the 80%–20% rules in two subdivisions: testing and training dataset. For dividing the dataset, arbitrary sampling was done which resulted in 80% of training data, as well as the remaining 20% being employed for the algorithm testing. The first is of the testing dataset { X= x_test & Y=y_test} and the second is of the training dataset.{ X= x_train & Y=y_train}.

4.4 Machine Learning Algorithms

This model was developed using various ML techniques — KNN,LR, SVM,DT and also RF. These models were applied on the testing and training dataset.

The analysis of the results was done after implementing these algorithms, and it was finalized by calculating the accuracy in both training and testing datasets. Initially, the total number of records was 919 patients. The dataset was divided according to 80-20% in training data the records were 735 and for testing data 184 records.

4.5 Result Analysis

Each experiment in this research paper was done using Python. Five types of machine learning algorithms using KNN, Logistic Regression, Random Forest, Decision Trees and SVM have separately experimented on this particular dataset. The results are presented in the following section.

After applying the Machine Learning techniques, the result of Logistic Regression the accuracy result of testing data is 86.2% and for training data is 83.3%,

For KNN the accuracy of testing and training data is 90.7 %, For SVM the accuracy of testing and training data is 87.2%. For Decision Trees the accuracy of testing and training data are 99.8 %, For Random Forest the accuracy of testing and training data are 98.8%. Decision Tree's results are much Satisfactory than the other algorithms in terms of accuracy.The accuracy we

have achieved in our research is better as compared to the earlier researches done in this sector

4.6 Prediction

This is the final step of the model at which the Doctor provides the output of the finding that one had after their consultation and complete checkup. Using those values, the user can check whether they had Depression or not.

5. Experiments & Results

This section explains the output we got while conducting the research and the results obtained out of those. For performing experiments, the algorithms were applied to the data, using Python. Table 6 represents the results obtained after performing the algorithms.

Model Used	Training Accuracy	Test Accuracy
Logistic Regression	86.2 %	83.3 %
SVM	87.2%	83.3%
Random Forest	98.8%	82.6%
Decision Trees	99.8%	78.0%
K-NN	90.7%	82.6%

The results show the accuracy achieved on Training dataset and Testing dataset, by applying Machine Learning techniques. It can be seen that an accuracy of 86.2% has been achieved when Logistic Regression was applied on the Training dataset, whereas the accuracy achieved was around 83.3% on Testing dataset. An accuracy of 87.2% was achieved by applying the SVM algorithm. The application of the KNN algorithm resulted in 90.7% accuracy. Figure 6 shows the comparison between the accuracies of all algorithms..

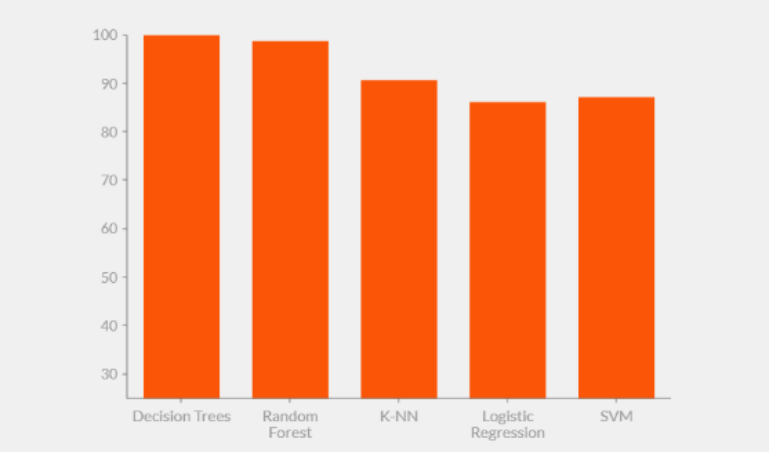


Figure 6. Accuracy comparison graph

The analysis shows that the random Forest algorithm is better and more accurate in detecting Depression.

6. Conclusion & Future Scope

The objective of this research was to detect the beginning and to detect the root cause of depression. For this purpose,the Dataset has been collected . It was pre-processed to prepare it . Then, the dataset was split into 80%-20% partitions for Training and Testing respectively. Further, popular ML algorithms, Logistic Regression, SVM, KNN, Decision Trees and Random Forest were applied. The results obtained after experiments show that the DT technique resulted in a classification accuracy of 99.98%.

In Order to get better accuracy and better Predictive results, We can also implement some other ML algorithms and Deep Learning algorithms in the future. Moreover, a dataset consisting of many instances can be collected and used for performing experiments.

References:

1. https://pubmed.ncbi.nlm.nih.gov/26759786/

2. https://researchportal.vub.be/en/publications/hybrid-depression-classification-and-estimation-from-audio-video-

3. https://pubmed.ncbi.nlm.nih.gov/28261491/

4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6111060/ 5. https://www.researchgate.net/publication/338938497_Predicting_Depression_in_Bangladeshi_Undergraduates_using_Machine_Learning

6. https://pubmed.ncbi.nlm.nih.gov/31606115/

7. https://www.sciencedirect.com/science/article/pii/S1877050920309091 8. https://link.springer.com/chapter/10.1007/978-981-99-7108-4_1

9. https://pubmed.ncbi.nlm.nih.gov/35129401/

10. https://www.researchgate.net/publication/341906771_Assessment_of_Anxiety_Depression_and_Stress_using_Machine_Learning_Models

11. https://ourworldindata.org/mental-health#introduction

12. https://in.images.search.yahoo.com/search/images;_ylt=AwrPrEzP.2Rk19QbYnq9HAx.;_ylu=c2VjA3NlYXJjaARzbGsDYXNzaXN0;_ylc=X1MDMjExND

cyMzAwNQRfcgMyBGZyA21jYWZlZQRmcjIDc2EtZ3Atc2VhcmNoBGdwcmlkA3pJTTZGdGQ1VE1pWmJCN3hPSDFmRUEEbl9yc2x0AzAEbl9zdWdnAzEwBG9yaWdpbgNpbi5pbWFnZXMuc2VhcmNoLnlhaG9vLmNvbQRwb3MDMgRwcXN0cgNrbm4gBHBxc3RybAM0BHFzdHJsAzIzBHF1ZXJ5A2tubiUyMGluJTIwbWFjaGluZSUyMGxlYXJuaW5nBHRfc3RtcAMxNjg0MzM5NzgwBHVzZV9jYXNlAw--?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://image3.slideserve.com/6689896/picture-of-logistic-regression-l.jpg**

14. **https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm**

15. **https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/**

16. **https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning**

17. **https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/**