

Depression Analysis Using Supervised Models

¹M.Shanmugaraj, ²Dr.V.S.Akshaya*, ³Prabakaran N, ⁴Dr.P.Rajasekaran

^{1,3}Assistant Professor, ²Professor, ⁴Associate Professor

1St. Martin's Engineering College, Secuderabad

²Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore

³School of Computer Science and Engineering, Vellore Institute of Technology, Vellore

⁴Department of Computing Technologies, SRM Institute of Science and Technology, Chennai

*Corresponding Author E-mail: Vsakshaya@gmail.com

Abstract— Depression is one of the severe and grave health disorders that affects the steadiness of mind. It has become a serious issue in the present generation. The total number of cases has been increasing day-by-day due to a number of reasons like stress at school, college, work, personal life, other diseases, etc. Although it has become one of the most common disease, people are still reluctant to talk about it openly due to the fear that others might consider them lunatic. The introduction of Machine Learning into the field of Medicine and Health industry has provided diagnostic tools that are able to enhance the precision and accuracy while reducing the difficult tasks which require the intervention of humans. There is promising evidence that Machine Learning has the capability not only to detect but also significantly enhance the treatment of compound mental conditions such as depression by developing a framework. In the past, Machine Learning Algorithms have been proved to be fairly supportive where researchers worked on the data from social media to foresee the number of persons suffering from this ailment on the basis of their initial symptoms. The main aim is to help those patients who suffer from depression in the early recognition of symptoms of this disease which can prove to be valuable not only to them but also to their families.

Keywords: Depression, Machine Learning, Naïve Bayes Algorithm, Support Vector Machine, Decision Tree, Random Forest, K Nearest Neighbour Classifier.

I. INTRODUCTION

Depression is considered to be a very lethal and distinctive medical sickness which is related to the psychological health of an individual. It is also known by another name, Major Depressive Disorder (MDD). It is a disorder that affects a large number of people across the globe, around 264 million people are affected by depression worldwide. The characteristics of this disorder are: losing concentration in several activities, extreme mood swings, sleeping illnesses, demolishing sadness which is then followed by reluctance in getting up in the morning, exhaustion, disinterested in the daily routine, abundant crying, etc. Depression can seriously affect the day-to-day life of a person from the least possible work to any major task. These days people even consider depression as a taboo and choose not to discuss about it with anyone or in

fact, they even hold back from considering this a disease or a grave disorder that is required to be cured. As a consequence of this, people who are going through this condition from are fearful to even talk about it. This not only makes them suffer quietly, without getting any sort of medical help but also worsens their condition further [1].

Depression has become one of the grave conditions that has triggered the suffering among numerous people worldwide. This has become one of the leading causes of worsening in their day-to-day life. Early discovery of this illness is very important as it can help an individual to get the desired medical aid in time, which can help him/her lead a normal healthy life and make his/her condition better instead of making it worse. The objective is given as:

1. Perceiving in advance the chances of a person to suffer from depression by asking him/her certain standard questions.
2. Finding the most accurate machine learning algorithm by comparing several algorithms and checking their accuracy.
3. Knowing the depressions' severity level for a particular person.

II. LITERATURE SURVEY

To appreciate the role of machine learning in mental illness diagnosis, a number of papers and journals have been presented. The study on this matter started in 1980's. Roland H.C. Yap, David M. Clarke portrayed in their paper a skilled system, MILP (Monash Interview for Liaison Psychiatry) that can identify mental condition based on DSM-III-R, DSM-IV and ICD-10 using constraint-based reasoning. Constraint Logic Programming (CLP) language was used to develop this system [2].

There is a dataset from Reddit that was published in 2017 which is publicly accessible. It was taken from Reddit Inc. API, that encompassed features such as: id, title, writing, date. Typically, it comprised of reports

that were provided by persons diagnosed with depression. It was considered to be a characterized data of depression. A control group also shared their data that was considered to be non-depressed characterized data. A total of 135 subjects were found to be depressed out of 887 total subjects [3].

There is another dataset where the data from a depression survey was joined with the already present public social media sources. The sources included self-declared depression cases, different forums, self-reported surveys and post level annotation [4].

There is another publicly available dataset that is used. It is from Open Sourcing Mental Illness (OSMI) survey that was conducted in 2017. The survey was mental health tech survey which consisted of a total of 750 responses with 68 attributes [5].

One of the common practices in today's world is the use emojis and slangs on social media sites which makes it difficult to interpret. To resolve this issue, both slangs and emojis were replaced with their descriptive text by making the use of an SMS dictionary and emojiopedia. To reduce the words to their respective root words, stemming was done and for the purpose of filtering out various tokens, tokenization was done [6]. Some researchers did the analysis of depression on the Facebook data which is available publicly. They carried out their research on the basis of linguistic and emotional style of word usage [7]. The classification was done by making the use of SVM algorithm with different kernels and the researchers showed that the SVM algorithm outperformed with better accuracy [8]. One of the researchers named Nadeem conducted experiment in 2016 on Major Depressive Disorder (MDD) via twitter data. He made the use of Naïve Bayes and SVM algorithm. His research concluded with Naïve Bayes algorithm outperforming SVM algorithm [9].

In one of the studies, a hybrid model of machine learning was implemented on the twitter data for the detection of depression. The SVM-Naïve Bayes hybrid model showed a great accuracy for the task of sentiment analysis [10].

III. PROPOSED SYSTEM

The existing models that have been created have usually worked on the datasets that were based on the social media like twitter and Facebook. These datasets are not usually reliable as people tend to be different on social media then what they actually are in real life. The posts on the social media sites are usually copied from another person's account or from some other source. So, a proper diagnosis cannot be made on such a situation. Therefore, a different approach has been

taken in this work where a person is asked a series of standard questions and based on his responses to these questions, a prediction is made by the model which has been trained using a dataset which contains about 16000 entries. Five Machine Learning Algorithms were used to train the model:

- SVM Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Naïve Bayes Classifier
- KNN Classifier

SVM Classifier achieved the best accuracy and thus was implemented in the application.

1. Support Vector Machine (SVM) Classifier

SVM is the most prevalent Supervised Learning algorithms. Owing to its simplicity, it surpasses all the other algorithms when it comes to its usage [1]. It finds its use in both Classification and Regression purposes. However, in Machine Learning, it is mainly used for Classification problems. The objective of SVM is to create a decision boundary that can separate n-dimensional space into classes so that a new data point can be inserted in the correct category in the future. The best decision boundary created by SVM is called the hyperplane. The algorithm is called Support Vector Machine as it chooses two extreme points/vectors called support vectors which help in the creation of the hyperplane [11].

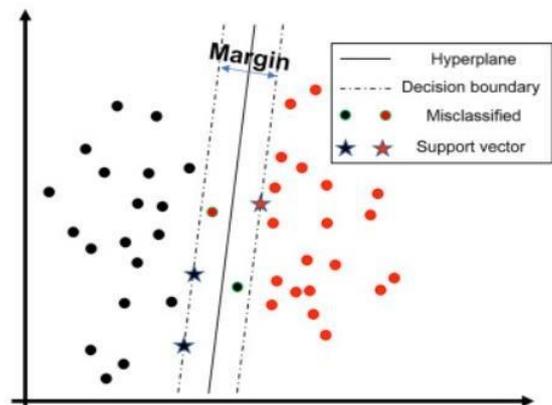


Fig. 1. SVM Classifier [12].

SVM can be of two types:

1. Linear SVM: It is used for Linearly Separable data, which means if a dataset can be classified into two classes by using a single straight line
2. Non-linear SVM: It is used for non-linearly separable data, which means a dataset that cannot be separated by using a single straight line.

2. Decision Tree Classifier

It is a Supervised Learning classifier which is preferably used to solve classification problems. As its

name suggests, Decision Tree has the structure of a tree in which the branches signify the rules of decision, internal nodes signify the features of the dataset and each leaf node characterizes the outcome of a decision. There are two types of nodes in the decision tree, Decision node and leaf node. Decision nodes have multiple branches and are used to make any decision. Leaf nodes do not have any branches as they are the outcome of the decision of decision nodes.

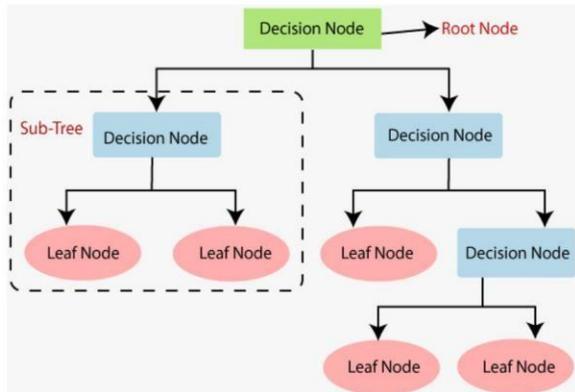


Fig. 2. Decision Tree Classifier [13].

The working of a decision tree is similar to a normal tree which starts with the root node and then expands further by branching itself and forming a tree-like structure. In order to build the tree, the algorithm which we used is called Classification and Regression Tree (CART) algorithm. In simple terms, a decision tree asks a question and based on the answer which is Yes or No, it splits further into the subtrees.

3. Random Forest Classifier

Random Forest Classifier is a prevalent machine learning algorithm which comes under the supervised learning technique. In ML, it is used for both Classification and Regression problems. The basis for this algorithm is the notion of ensemble learning, which can be defined as a process of combining numerous classifiers to solve a composite problem and also to improve the performance of the model.

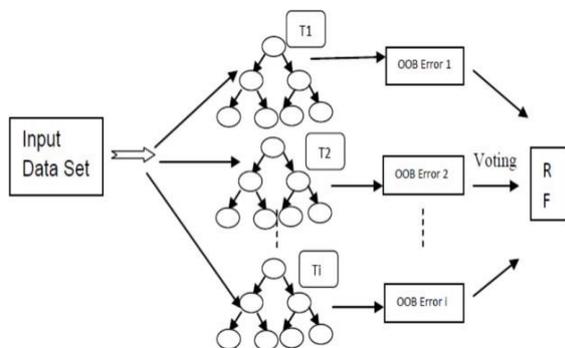


Fig. 3. Random Forest Classifier [14].

A Random Forest algorithm produces several decision trees in place of a single tree. When a new input from a given sample is fed to random forest for classification

purpose, each of its trees is given that same input to perform the classification. The classification from each tree is called “votes” for classified class. The classification which receives maximum number of votes is selected [14]. The more the number of decision trees in the random forest, the higher is the accuracy of the algorithm and also the problem of overfitting is reduced. There is a possibility that some of the decision trees in the forest may predict the incorrect output. But since there are a number of trees in the forest, other trees predict the correct output and hence the final output predicted is correct.

4. Naïve Bayes Classifier

It is a supervised algorithm which is based on Bayes theorem. It is used for classification problems mainly in text classification which includes a high-dimensional training dataset. It is one of the most simple and effective classification algorithms that helps us to build fast Machine Learning models that can make quick predictions. Since this classifier predicts on the basis of the probability, it is called probabilistic classifier. It is composed of two words: Naïve and Bayes which are described as [15]:

Naïve: as it assumes that a certain feature’s occurrence is independent of the occurrence of other features.

Bayes: as it depends on the principle of Bayes’ Theorem.

The Naïve Bayes theorem is shown in fig. 4.

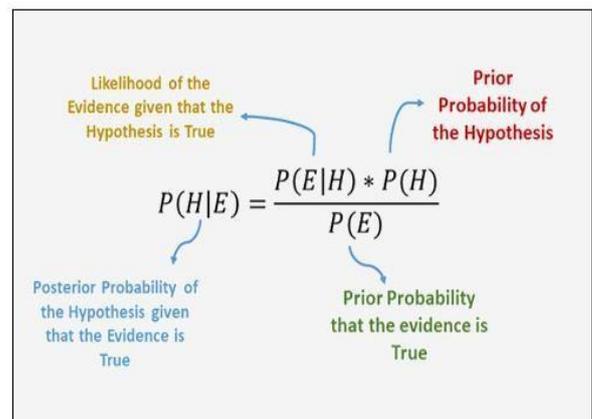


Fig. 4. Naïve Bayes Classifier [16]

In the field of Probability and statistics Bayes theorem is considered to be vital. It defines the probability of the occurrence of an event based on some conditions that are known prior to the user.

5. KNN Classifier

K-Nearest Neighbour is a supervised learning technique-based algorithm. It is an algorithm that works on the assumption that there is a similarity between the new data and the available data. A new case/data is put into the category which is most similar

to the available categories. k-NN stores all the existing data and a new data point is categorized based on the similarity which means that when a new data appears, it can be classified into a well suite category by using k-NN algorithm. This algorithm is mainly used for classification but it can also be used for regression purposes. It is a non-parametric algorithm that does not make any assumption on the original data. Since this algorithm does not immediately learn from the training data set, it is sometimes referred to as lazy learner algorithm. In the training phase it just only stores the dataset and when it gets the new data then it performs the classification on the data and classifies it into a category that is much alike to the new data.

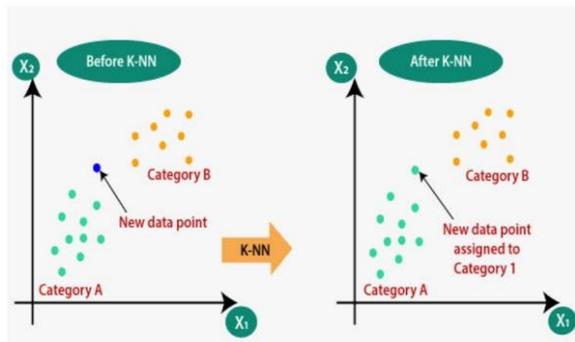


Fig. 5. KNN Classifier [17]

A number of data pre-processing steps are performed on the csv file that is read. Natural language processing has been used for pre-processing methods applied on the extracted data:

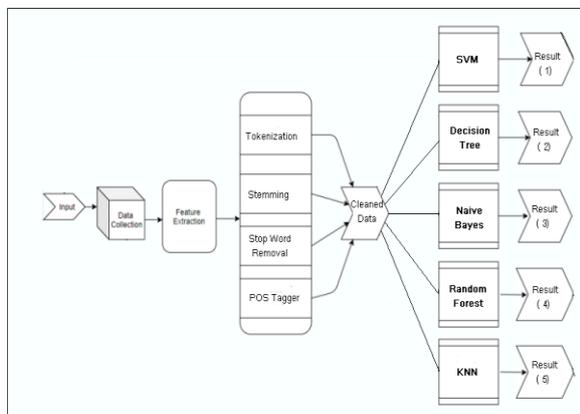


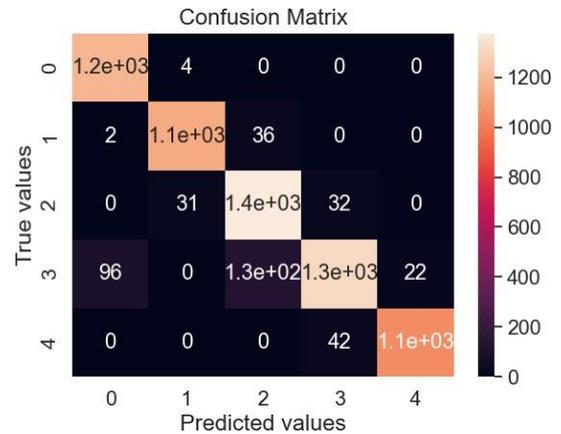
Fig. 6. Workflow Model Architecture.

- Tokenization: Dividing a string into several meaningful substrings like words and sentences.
- Stemming: It involves reducing the words to their root form so as to group similar words together.
- Stop Word Removal: Stop words like a, an, the etc need to be removed since they are of no use.
- POS Tagger: Tokenized words are assigned tags such as nouns, adjectives etc to improve the quality of the trained data.

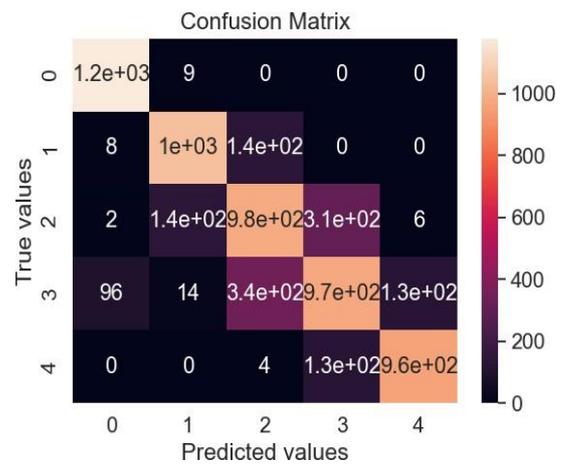
IV. EXPERIMENTAL RESULTS

We trained all the five models and upon testing each produced a different result. The interpretation of the results was done by making the use of confusion matrix (figure 7-11). The labels can be interpreted as:

- 0 - No Depression
- 1- Mild Depression
- 2 – Moderate Depression
- 3 – Moderately Severe Depression
- 4 – Severe Depression



7. SVM Classifier.



8. DT Classifier.

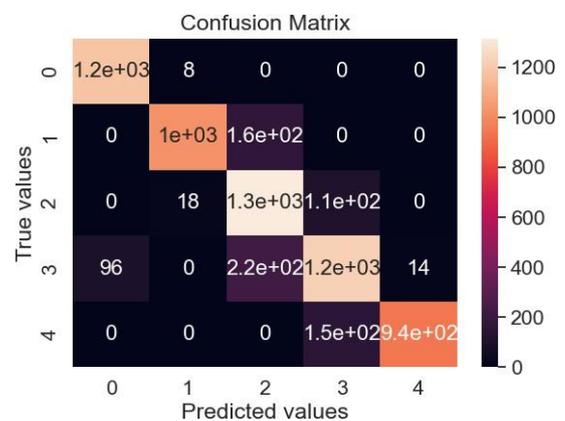


Fig. 9. Random Forest Classifier.

IV. CONCLUSION

Since depression is a mental health ailment which is spread across the world, immediate dealing with it has become necessary. Some of the most critical steps to deal with this disorder are an early detection of this disease through common symptoms and more widespread knowledge about it. These two steps can aid the people in getting better treatment and can also save many lives. This work was initiated with the goal of helping the people in need by predicting the depression in its early stages so that they can get cure in time. In this work, different machine learning algorithms were used, and also various feature datasets to train the model. Preparation of data and its alignment, labelling of data and feature extraction and selection are a few of the pre-processing procedures. There are some of the other existing systems that are based on numerous other techniques like - Support Vector Machines with an accuracy of 85.71%, Boosting which gives an accuracy of 75%, Convolution Neural Network-based prototype with diverse features fed into the model with an accuracy of 95%, and another model which uses random forest, having an accuracy of 81.04%. Data analysis was done thoroughly to define the participants' behaviour based on many features of their PHQ9 question responses. On the basis of the conclusions drawn from the trained model, the result of the research done was divided into five labels according to the severity of the depression. Different machine learning algorithms that have been used in this paper are as follows: Naïve-Bayes theorem, Support Vector, k-Nearest Neighbour, Decision Tree and Random Forest. The accuracies achieved in ascending order are Naive Bayes classifier with 96.6%, Decision Tree classifier with 99.2%, Random Forest with 99.5%, KNN classifier with 99.6% and SVM with 99.7%.

V. FUTURE ENHANCEMENTS

Depression is a perplexing mental health disorder i.e.; it is very hard to understand everything about it. Sometimes the symptoms are very common and confined only to some health issues that are very basic but other times they may be very obvious to the person. This limit sometimes makes it really hard for the person to diagnose and get appropriate treatment. There is a compromise in both the excellence as well as the size of the dataset. As a consequence of this, usually the work has to be done on a small dataset. In the future any work done on this subject should be done on a dataset that is both large with a greater number of attributes and also the quality should be better so that it

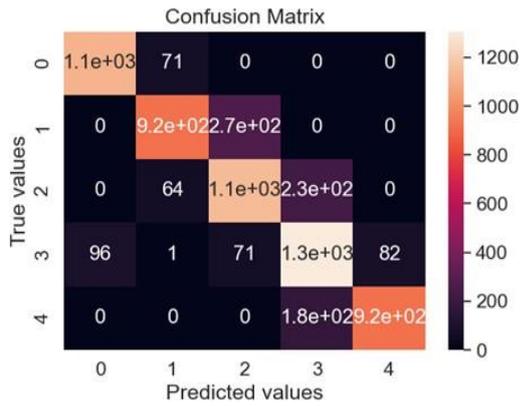


Fig. 10. Naïve Bayes Classifier.

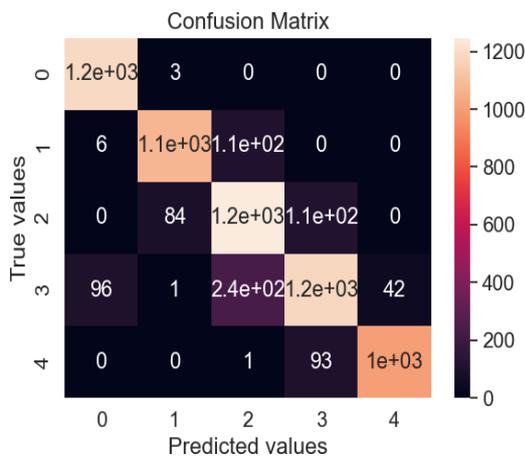


Fig. 11. KNN Classifier.

The results obtained by different algorithms are given in table 1 and its graphical representation is given in Fig. 12.

ALGORITHMS USED	ACCURACY
SUPPORT VECTOR MACHINE (SVM)	99.74
DECISION TREE CLASSIFIER	99.29
RANDOM FOREST CLASSIFIER	99.59
NAIVE BAYES CLASSIFIER	96.63
KNN CLASSIFIER CLASSIFIER	99.60

Table 1. Accuracies obtained from various Machine Learning algorithms.

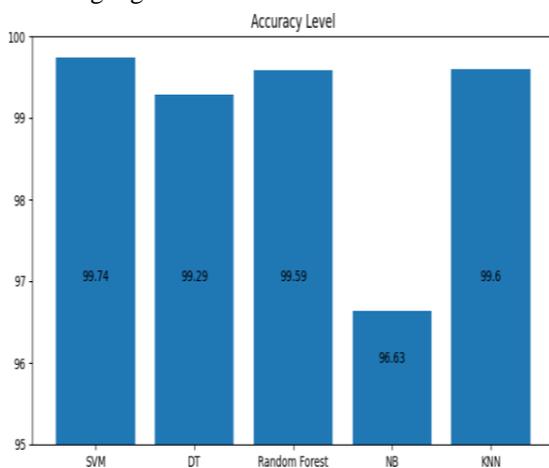


Fig. 12. Graphical Representation of results.

may be trustworthy and can achieve a more promising result. Neural network-based models can also be built as an improvement to the present work to check their performance and precision.

REFERENCE

- [1] G. Geetha, G. Saranya, Dr.K.Chakrapani, Dr.J.Godwin Ponsam, M.Safa, Dr.S.Karpagaselvi. "EARLY DETECTION OF DEPRESSION FROM SOCIAL MEDIA DATA USING MACHINE LEARNING ALGORITHMS"2020 2nd International Conference on Power, Energy, Control and Transmission Systems
- [2] RH Yap and David M Clarke. "An expert system for psychiatric diagnosis using the dsm-iii-r, dsm-iv and icd-10 classifications." In Proceedings of the AMIA Annual Fall Symposium, page 229. American Medical Informatics Association, 1996.
- [3] Fidel Cacheda, PhD; Diego Fernandez, PhD; Francisco J Novoa1, PhD; Victor Carneiro, PhD,"Early Detection of Depression: Social Network Analysis and Random Forest Techniques",2019.
- [4] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar and Johannes C Eichstaedt, "Detecting depression and mental illness on social media: an integrative review", 2017.
- [5] Kali Cornn, Department of Statistics, Stanford University, "Identifying Depression on Social Media",2018.
- [6] Akshi Kumar, Aditi Sharma, Anshika Arora, "Anxious Depression Prediction in Real-time Social Data".2019.
- [7] M. R. Islam, A. R. M. Kamal, N. Sultana, R. Islam, M. A. Moni, and A. Ulhaq, "Detecting Depression Using K-Nearest Neighbors (KNN) Classification Technique," Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. IC4ME2 2018, pp. 4–7, 2018, doi: 10.1109/IC4ME2.2018.8465641.
- [8] H. S. ALSAGRI and M. YKHLEF,"Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features," IEICE Trans. Inf. Syst., vol. E103.D, no. 8, pp. 1825–1832, 2020, doi: 10.1587/transinf.2020edp7023.
- [9] M. Nadeem, "Identifying Depression on Twitter," pp. 1–9, 2016.
- [10] M. Gaikar, J. Chavan, K. Indore, and R. Shedje, "Depression Detection and Prevention System by Analysing Tweets," SSRN Electron. J., pp. 1–6, 2019, doi: 10.2139/ssrn.3358809.
- [11] support-vector-machine-algorithm
- [12] Siddharth Misra, Hao Li, "Non-invasive fracture characterization based on the classification of sonic wave travel times" Machine Learning for Subsurface Characterization, 2020.
- [13] Muhammad Asfand Hafeez, Muhammad Rashid, Hassan Tariq, Zain Ul Abideen, Saud S. Alotaibi and Mohammed H. Sinky "Performance Improvement of Decision Tree: A Robust Classifier Using Tabu Search Algorithm"
- [14] Veena N. Jokhakar and S.V. Patel "A Random Forest Based Machine Learning Approach for Mild Steel Defect Diagnosis" 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICIC)
- [15] <https://www.javatpoint.com/machine-learning-naive-bayes-classifier>
- [16] <https://medium.com/analytics-vidhya/na%C3%AFve-bayes-algorithm-5bf31e9032a2>
- [17] <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>