Polarity Detection & Analysis Of Suicidal Notes On Sentiment Rich Data

[1]FARZIZ AKTAR AHMED
Faculty of Engineering and Technology
Assam Down Town University
Guwahati, India
farzizkhan786@gmail.com

[2]NITUMANI SARMAH
Faculty of Engineering and Technology
Assam Down Town University
Guwahati, India
nitumani.s@inurture.co.in

[3]JUNUMONI KHAKHLARI
Faculty of Engineering and Technology
Assam Down Town University
Guwahati, India
junumonikhakhlari608@gmail.com

ABSTRACT

This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. No numbering required for Abstract. Suicide becomes an unavoidable issue for the modern society. Every year 800000 peoples lose their lives worldwide. Uses of Smartphone have rapidly increased, and people are interacting with others using social media and other platforms have increased as a result. These social media data can be use as raw data to perform sentient analysis to understand the sentiment. In recent few years sentiment analysis has gain so much of popularity. Many studies have been done already to analyze and understand sentiment behind a statement. And there are still many research in progress to analyze the sentiment. Detecting sentiments of these peoples is a new challenge for many researchers. It can be achieved by analyzing their statements made by the victims before they commit any suicidal actions, with the help of machine learning, natural language processing etc. In this work, we have used a dataset which contains 232074 unique values collected posts from “Suicide Watch” and “depression” subreddits of the Reddit platform, to develop different machine learning model to analyze the sentiments of these data. We developed several types of machine learning model to compare the accuracy and find out the best and suitable algorithm for the project of detecting people's sentiment. The accuracy we able to be achieved, SVM 57.24%, Naive Bayes (Gaussian) 54.69%, Random Forest67.67%, Decision tree 70.95. Along with these algorithms we have also developed different versions of Naïve Bayes model algorithm where Naïve Bayes (Bernoulli), Naïve Bayes (Multinomial) and Naïve Bayes (Gaussian) able to achieve an accuracy of 49.92%, 51.65%, 54.69% accordingly. Here we have found that Decision Tree is providing best accuracy compared to another model algorithm. In addition, among all the versions of Naive Bayes model algorithms Bayes (Gaussian) is providing the best accuracy.

Keywords—Sentiment Analysis, Suicide attempts, Decision Tree algorithm, Machine Learning, Support Vector Machine, Random Forest, Naive Bayes, Deep Learning.

I. INTRODUCTION

Suicide becoming a leading issue in healthcare sector around the world. Due to many reasons every year 703000 [1] of peoples themselves end their life globally and most of them are youth, ages in between 19-25 [1] as per 2019 (World health organization, Updated as on 17 June 2021) data. Among them 77% [1] of the people belong to the countries where people’s incomes are in between middle-lower level. Out of 100 more than one (1.3%) people commit suicide 2019 [2]. A study published recently on depression and anxiety, 67000 college student so more 100 institutions, where one out of five have had thought of suicide at least for once [3]. Among them 9% students making an attempt and 20% of them are reported a self-injury and one out four reported that they are diagnosed with a mental illness. In another recent survey from US “Youth Risk Behaviors Survey 2019” showing that 8.9% of youths with grades 9-12 made a suicide attempt in last 12 months [4]. Among them American Indian or Alaska native student sreported25%andwhite students reported 7.9% of suicide attempt. Risk factor for these suicide victims vary individual to individual.

Here are some risk factors that have been identified are A prior suicide attempt, A sense of isolation and lack of support, Access to a suicide method, Impulsivity issues, Major depression, Physical illness, Poor coping skills, Serve Personality disorder, Substance use issues, Traumatic or stressful life events[3].

Suicidal behaviors are a complex process that can vary from suicidal ideation (communicated verbally or non-verbally) through planning, attempting, and, in the worst-case scenario, committing suicide. Interacting biological, genetic, psychological, social, environmental, and situational factors influence these behaviors [31]. Inequity, social marginalization, and socioeconomic disadvantage have all been associated to suicide [32]. It is a massive problem that is generating unneeded human pain and tremendous societal costs. These are some of the major reasons, which lead people to suicidal ideation, attempting phase and at the end it leads people to committed suicide.

These suicide attempts can be prevented if it could able to know the mental health and conditions of those suicide victims. In most of the cases suicide victim's mental health reflects through their actions. Nowadays almost every person have a access to internet and social media, and most the people's posts, message and comments are simply a reflection of their emotions and Mental health. If a person is felling happy and another person is feeling sad and depressed then their posts, message and comments are respectively going to be positive and negative. In 10% to 43% of the times suicide victims leave a suicide note behind [5]. There are some clinical notes also written by psychiatrists at the time of admission and discharge that includes patients' background, mental health, family history, current circumstances and many more to know the patients mental conditions, and this is very helpful to know the mental health of a patients. As mentioned above these social media's message, comments, posts, tweets, suicidal notes, clinical notes can be use to know the sentiments of a patients using Sentiment Analysis method.

The other name of Sentiment Analysis is called as opinion mining or Emotion AI. Sentiment Analysis is used to extract the emotions of a text or a block of texts. Sentiment analysis is useful in decision making. It uses Natural Language Processing (NLP) & Machine learning to analyze or a statement sentence whether the statement is Positive, Negative or Neutral. In addition Sentiment analysis also analyse the emotions of a sentence whether the statement is happy, sad, angry, etc. Its used in different domains like psychologically, socially, machine learning, etc. The source for Sentiment Analysis is mostly social communications, mental people's communication.

II. Introduction to Sentiment Analysis.

Sentiment analysis is a machine learning tool which is used to extract the sentiments/emotions from a specific block of texts. It identifies the emotions behind the texts. Sentiment analysis explains the filtering the people's emotions like depressed, happy, and sad, along with the positive, negative or neutral state. Also it can recognize the mental state of the depressed people. It is also known as "opinion mining" which is used in different domains such as machine learning, psychologically, socially, etc. From the data of social platforms we can analyze the mental state of that person whether he/she is happy sad or angry. It has different field like Machine learning (ML), Natural language processing (NLP), Computational Linguistics, Bags of words.

This strategy is another technique which presents a Redesign of Bags-of-words model to address major Inadequacies of the Bag – of –words model in assessment Appraisal. It depends upon the word level of feeling examination in one branch of knowledge. In addition, we can isolate components and expressions of the space to arranges sentiments reviews and show up at the exactly meaning of each review. The propose strategy in like manner can introduce deals with any consequences regarding feeling assessment to additionally foster position.

It also has three major levels-

i. Word level.
ii. Sentence level.
iii. Document level.

The three level of the sentiment analysis determines the task required for the process. The first level which is word level it is the most difficulty level in carrying out the analysis, whereas the second level and the third level of sentiment analysis which is sentence level and documents level is very similar respectively. Two major techniques which are used for the review of sentimental analysis are Machine learning and Semantic-based Analysis.

And our work is related to analyze the sentiment of all the individuals, those who are posting, sharing, commenting negative statements or it is reflecting that their mental health is not good and they've thinking for suicide at least for once. In this case sentiment analysis a major tool to identify those people, who are posting such type of negative post, comments, messages, tweets or anything, related suicidal indication.

III. Sentiment Analysis on Suicidal notes.

Suicide is one of the most common death reasons in youth around the globe. All the people who have already committed suicide or all those who're going to commit suicide almost all of them are try to express their feeling through various methods. Some of them use to leave suicide notes, and some people try to express their feeling by posting sad or depressed posts on social platforms, also some of them are try to communicate with their closest one through messaging. In all these methods, suicide victims are expressing their state of mind to others via communications. Twitter is one of the most popular social media platform where billions of people share their feelings, opinions, thoughts, incidents happening their life on daily basis. If there's any such statement which is a suicidal indication it can be detect with the help of sentiment analysis. Also, Sentiment Analysis can detect or classify, whether the sentiment of that statement of the victim is negative, positive or neutral.

Sentiment analysis can be done with methods like Machine learning (ML), Natural Language processing (NLP) in order to get the sentiment of any statement. The key purpose of using use NLP techniques, especially semantics and word sense disambiguation, to extract the opinions more accurately. Determining the meaning of a word in NLP is the ability to determine what word meaning is activated by using the word in a given context.

Two major approaches to extract the sentiment automatically of a particular context are Lexicon-based approach and machine-learning approach [7][8].

III.A. Lexicon-based approach

Lexicon-based approach uses predefined collections of word where each and every word associated with a specific meaning. This technique calculates the sentiment orientations of the entire document or set of sentence(s) from semantic orientation of lexicons. The dictionary of the lexicon can be prepared in both ways manually as well as automatically. The wordnet dictionary is the most popular and mostly used dictionary among the researchers.

III.B. Machine-learning approach

Machine learning mostly rely on supervised classification approach, where detection of sentiment is represented as binary which are positive or negative [9]. To train the classifiers machine-learning approaches needs labeled data. There are various types of techniques and complex algorithms to extract the emotions from a particular statement of suicide victims. Some of the algorithms proved to give best performance with maximum accuracy such as Naïve Bayes (NB), Max Entropy (MaxEnt) and Support Vector Machines (SVM) [6]

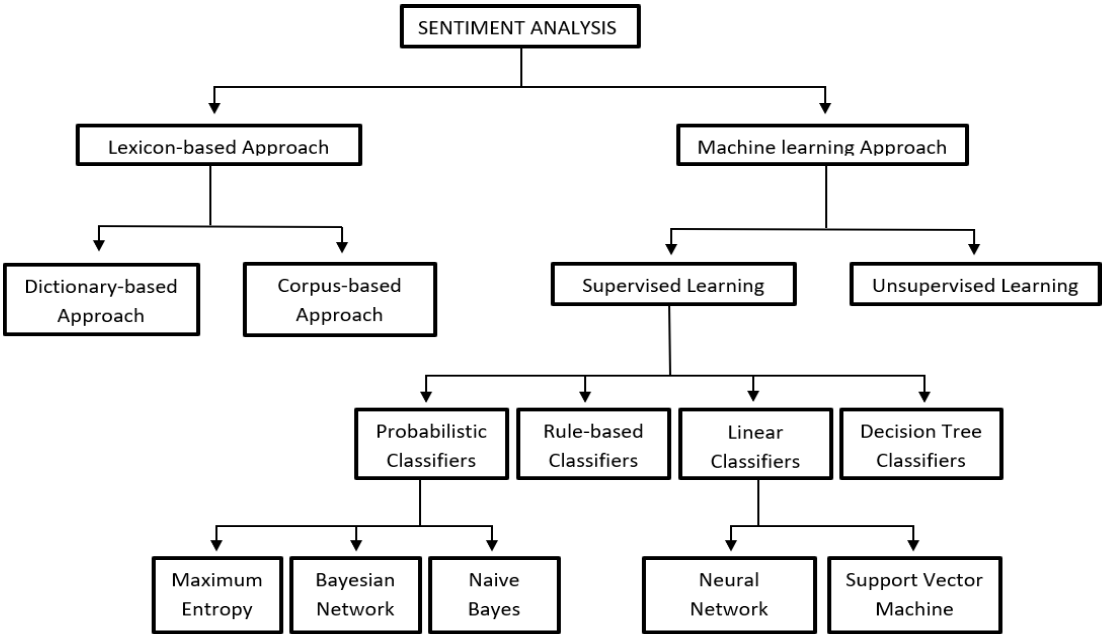


Figure 1: - Classifications of Sentiment Analysis [30]

IV. LITERATURE REVIEW

The use of social media has grown unexpectedly in these last few years. Almost all the people are connected through social media those who are using smart phone. In various social media platform people are communicating through messages, expressing their opinion through comments and also they sharing their day-to-day life incidents. In case of one, two or for few statements are possible to analyse the sentiment manually but when the opinion, comments, texts messages are in a huge numbers like in billions it becomes impossible to analyse the sentiments of those data manually and here the concepts arise to analyse the sentiments of people that what actually they want to express. Are their statement is reflecting positive sentiment, negative sentiment or neutral sentiment. It is also possible to recognize their emotions whether their emotions is happy, sad or angry.

There have been many studies and researches done on this that to analyse the sentiment of an individual's statement or a huge number of opinions that people are generating through social media. There are many different ways to analyses the sentiment of peoples and they are achieved a great accuracy.

Different researchers use different methods to analyze the sentiment more and more accurately. Such as Naïve Bayes, Support Vector machine (SVM), Decision Tree, Know your neighbor (KNN), Random Forest, Logistic Regression, etc. These are some most popular algorithm for sentiment analysis among the Researchers, and using these algorithms they are able to achieve a quite good accuracy in their researches.

Goet.al.[10]inthejournalpaper"TwitterSentimentAnalysis"statedthattheyhaveperformedasentimentan alysis on a set of twitter data to analyze the sentiments of user. They collected their own data using twitter API; all those twitter messages that have emotions. They have used Naïve Bayes classifiers which they have built from scratch and Third-party library were used for Maximum Entropy and Support Vector Machine (SVM). Using these methods, they are able get a accuracy of 73.913% with Support Vector Machine and with Naïve Bayes they get a accuracy of 44.9% and for Maximum Entropy its didn't contribute much to get a higher accuracy.

Qaiseret.al.[11] in their research paper "Sentiment Analysis of Impact of Technology on Employment from Text on Twitter" they tried to analyse the sentiment of people about the impact of technology on unemployment and technical advancements. In this study they have found that 65% of the people have a negative sentiment about the impact of technology on unemployment and technical advancement. They used twitter data related to their topic keywords. In this paper they have they have trained Naïve Bayes machine learning classifier to classify the data according to the user's sentiments. Using Naïve Bayes and Support Vector Machine (SVM), they able to be achieved an accuracy of 87.18% and 82.05% respectively.

Pak et. al. [12] in their studies "Twitter as a Corpus for Sentiment Analysis and Opinion Mining" they have showed an automatic collection of corpus that they can use to train sentiment classifier. For this method they have used twitter data and through the use of Twitter API they collected data for their sentiment analysis. In this paper they have used multinomial Naive Bayes classifier that uses N-gram and POS-tags as features to classify the content. Also they have used TreeTagger for POS-tagging to observe the difference in distributions among positive, negative and neutral sets.

Altrabsheh et. al. [13] in their paper "SA-E: Sentiment Analysis for Education" they have done a research on a new topic sentiment analysis on education. For this they have collected students feedbacks from social media like twitter, to analyse the sentiments of students to understand whether the students are positive, negative or having any other emotions. In this study this paper they tried to present that Naïve Bayes and Support Vector Machine (SVM) can be combined to analyze the students feedback in Real-time, and it holds a great potential. Along with this they have introduced a new system architecture termed as System Analysis for Education (SA-E).

RamyaSri et. al. [14] in their work "Sentiment Analysis of Patients' Opinions in Health care using Lexicon-based Method" they have done research on sentiment analysis.Which includes analysing the sentiment of patients based on their opinion to improve the healthcare services. To perform this, they have used patients' opinion data from 92 web pages which contains patients' opinions on Southern California Orthopaedic Institute which is located in California, USA. To perform sentiment analysis on this data they have used Lexicon Based method. Sentiment analysis tools like "VADER" and "TextBlob" are used for classifications.Using these methods, they are able to achieve an accuracy of71.9% in VADER lexicon-based approach and 73.0% in the TextBlob lexicon-based approach. But on performing comparative analysis considering precision, recall, and F1-score they have found that VADER lexicon-based approach performs better than the TextBlob lexicon-based approach.

Munezero et.al.[15] in their paper "Exploiting Sentiment Analysis to Track Emotions in Students' Learning Diaries" mentioned that they present a functional system for analyzing and visualizing students emotions expressed in their diaries. This system allows the instructors to extract the emotions expressed in students' diaries. They used a dataset of students Diaries from the Newman. Also, they have used "Stop word removal procedure", "Poter's stemming algorithm".

Graveset.al.[16] in their paper "Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online" says that they used NHS Choices Datasets to analyse the sentiment of people. They have applied machine learning techniques to all 6412 online comments about hospitals on the

English National Health Service website in 2010 using Weka data-mining software. They have compared the results which were obtained from the sentiment analysis with their paper-based national in patient survey results at the hospital level using Spearman rank correlation for all 161 acute adult hospital trusts in England. In this paper researcher have used Naïve Bayes, multinomials, Decision Tree, Bagging, Support Vector Machine (SVM) algorithms. By studying this paper [16] we came to know that using Naïve Bayes Multinomial's algorithm they got the accuracy of 88.6% and using Decision trees the accuracy they got is 80.8%. Using Bagging they got the accuracy 82.5% and lastly by using AVM they got 84.6%.

Saifet.al.[17] in the paper "Semantic Sentiment Analyse of Twitter" the researcher has used Stanford Twitter Sentiment Corpus (STS), Health Care Reform (HCR), Obama-McCain Debate(OMD) Datasets to classify the sentiment of people. In their paper, they have introduced an approach for adding semantics as an additional feature into their training set for analyzing the sentiment of the people. For each extracted entity (e.g., iPhone) from the tweets, which were they added its as semantic concept (e.g., "Apple product") which as an additional feature, and for measuring the correlation of the representative concept with positive, negative or neutral sentiment of the people. They have applied this approach for the prediction of sentiment for different Twitter datasets. Also, they have approach for Naïve Bayes algorithm to perform their analysis on sentiment. In this research paper by performing Stanford Twitter Sentiment Corpus, they got accuracy of 80.7%, with Health Care reform they are able to get the accuracy 71.1% and also by approaching to Obama-McCa in method the researcher are able to get the accuracy of 75.4%.

Neriet.al.[18] in this journal paper "Sentiment Analysis on Social Media" the researcher has used the dataset as 1000 posts-by focus crawling of Facebook. In this journal paper the researcher have used Recall and Precision algorithm to filter the sentiment of the comments which are posted in social media by the people to recognize which are positive, negative or neutral. They have performed their Sentiment Analysis over various Facebook posts about newscasts, comparing the sentiment for Rai-the Italian public broadcasting service-towards the emerging and more dynamic private company La7. Their study maps study the results with observations made by the Osservatoriodi Pavia, which is an Italian institute of research specialized in the media analysis at the or etical and empirical level, engaged in their analysis in the mass media of political communication. Their study also takes in account the data provided by Auditel regarding news cast audience, correlating the analysis of Social Media, of Facebook in particular, with measurable data, available to public domain. The researcher got 87% accuracy by using recall algorithm and they got accuracy of 93% by using Precision algorithm.

Sarlan et.al. [19] the paper named "Twitter Sentiment Analysis" stated that they have used twitter data as dataset, also they used Natural Language Processing (NLP), Case-Based Reasoning (CBR), Artificial Neural Network (ANN), Support Vector Machine (SVM) algorithm to extract the data of sentiments whether its positive, negative or neutral. By using Support Vector Machine (SVM) they got the accuracy as81.3%.

Aladağ et.al.[20] in their journal paper "Detecting Suicidal Ideation on Forums: Proof-of-Concept Study" they said that they have used the dataset as Reddit dataset (2008 -2016). In their paper they have said that they have used Logistic regression, Random Forest, SVM, Baseline ZeroR algorithms to perform the classification of sentiment of the people whether the sentiment of the comment is happy, sadorangry.They have used method as a total of 508,398 Reddit posts were posted between 2008 and 2016 on SuicideWatch, it has longer than 100 characters in their posts. In their paper Depression, Anxiety, and Shower Thoughts subreddits were downloaded from the publicly available in Reddit dataset. 10,785 posts were randomly selected and 785were manually annotated as suicidal or non-suicidal in their paper. Some features were extracted using term frequency-inverse document frequency, linguistic inquiry and word count, and sentiment analysis on post titles and bodies. Logistic regression, Random Forest, and support vector machine (SVM) classification algorithms were applied on resulting for performance evaluation of corpus and prediction. The researcher has got the accuracy of 80% by using Logistic regression, using Random Forest they got the accuracy as 92%, they got the accuracy of 50%by using Support Vector Machine (SVM), lastly by using baseline ZerorR algorithm they got the accuracy of 66%.

McCart et. al. [21] in their studies "Using Ensemble Models to Classify the Sentiment Expressed in Suicide Notes" in their journal they have used the dataset consisted of 900 suicide notes collected over a 70-year period (1940–2010). Their team has explored multiple approaches combining regular expression-based rules, statistical text mining (STM), and an approach that applies weights to text while accounting for multiple labels. Their best submission used an ensemble of both rules andSTMmodelstoachieveamicro-averagedF1score of0.5023, slightly above the mean from the 26 teams that competed (0.4875). Also, they have used algorithm as Decision trees, KNN, SVM to filter out the sentiment of the people from various statement which are posted in different social media platforms.

Pestian et. al. [22] in their paper "sentiment Analysis of suicide notes: A shared Task" stated that they have used 1319 people suicide notes (1950-2011, CHRISTINE) as a dataset in their paper. They have done their research using a shared task in biomedical domain which includes two features one is the Anonymized clinical texts and annotated suicide notes and the other one is it requires categorization large set of labels. In this paper they have describe about the challenges to classify the emotions found in notes left behind by those who have died by suicide in 2011. In total 106 scientists who have comprised 24 teams responded to the call for the participation. This paper's results were presented at the Fifth i2b2/VA/Cincinnati Shared-Task and Workshop: Challenges in Natural Language Processing for Clinical Data in Washington, DC, on October 21–22, 2011, as an American Medical Informatics Association Workshop.

George et. al. [23] in their research paper "Application of Aspect-based Sentiment Analysis on Psychiatric Clinical notes to Study Suicide in Youth" they said that they haveused1559suicidalnotes (H18-01402;June2018) as a dataset in their paper. They propose to address the lack of terminological resources related to suicide by a method of constructing a vocabulary associated with suicide. For a better analysis, they used Weka as a tool of data mining that can extract useful information from Twitter data collected by Twitter4J. Also, they said that they have used Logistic regression, Random Forest algorithm to classify the sentiment of the people by the suicidal notes.

Birjaliet.al.[24]in the paper named "Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks" that they 892 TWEETS (Using Twitter4J API)used as a dataset to classify the feelings of people. Also, they have used algorithm IB1, J48, CART, SMO, NAÏVEBAYES.

Mbarek et. al. [25] in the paper "Suicidal Profiles Detection in Twitter" in this journal paper the researcher have stated that they have used 115 suicidal profiles, 172 not suicidal profiles (Using TWITTER HEREAFTER Site) as a dataset. They have used Random Forest, BayesNet, Adaboost, J48, SMO these algorithms to figure out the statement which are positive, negative, or neutral. By using Random Forest algorithm, the researcher has got 77% accuracy; using SMO the researcher has got 74% accuracy.

Sohnet.al.[26] in this journal paper "A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes" the researcher has used datasets as 600 actual suicide notes. The researcher has used two algorithm named as NAÏVE BAYES and RIPPER.

Ji et.al. [27] in this paper "Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications" the researcher has stated that they use datasets as TEXTDATA (Reddit, Twitter, ReachOut), EHR, Mental Disorders(questionnaires III-A suicide notes III-C suicide blogs III-Celectronic health records III-Bonline social textsIII-D). Also, they have used algorithm as Machine Learning and DEEP LEARNING to extract the sentiment of the data which are available in social platform.

Glenn et.al. [28] In the paper named "Can Text Messages Identify Suicide Risk in Real Time? A With in-Subjects Pilot Examination of Temporally Sensitive Markers of Suicide Risk" the researcher has stated that they have used DIGITAL TEXT DATA (Social media) as their datasets. To filter the sentiment of the people from their comments posting on social media platforms they have used Machine Learning algorithm whether the statements are positive, negative or neutral.

Sharma et.al. [29] in their journal paper named "Analyzing the depression and suicidal tendencies of people affected byCOVID-19's lockdown using sentiment analysis on social networking websites" the researcher has used the Twitter data as a datasets. Also, they have used the algorithm as Machine Learning (Unsupervised) to filter the sentiment of the statement of people which are stated in various social media platforms to classify whether it's happy, sad or angry.

V. METHODOLOGY

Attempting suicide now a day's becomes a global problem. And detection of suicidal sentiment using social media message, comments, posts, suicidal notes has drawn attention of many researchers. Many researchers all over the world trying to build sentiment analysis model to detect the sentiment behind some particular texts, message, and social media posts. Different techniques are used by the researchers to detect the suicidal sentiment of people. For example, clinical methods with patient-clinic interaction [33] and automatic detection from user-generated content (mainly text) [34], [35], [36]. There is many research has been done in this particular field using different models to get more accurate results. Most of the researchers used Machine learning algorithms like Naïve Bayes, Support Vector Machine, Logistic Regression, Natural Language Processing (NLP) etc. But there are very few works has been done using methods like Ensemble model. This is the field that still needs to explore more by the researchers to reach, a step ahead to get more accuracy.

An Ensemble model is a machine learning model which combines two or more than two different models or a model that trained on different datasets to get better accuracy. In this model different algorithms contribute to the ensemble to predict an outcome more accurately. Basically, Ensemble model have two techniques Bagging and Boosting. Bagging also called as Bootstrap Aggregation. Bagging includes Random Forest and in Boosting it has three techniques Adaboost, Gradient Boosting and XGBoost.

VI.A. MACHINE LEARNING

Machine Learning is the domain of studies of different computer algorithms that can improve through experience. Simon has defined the Machine Learning as “the process of a change and enhancement in the behaviors through exploring new information in time”. It mainly used for to solve the complex problems by using the previous data.[37] Machine Learning can be examined in four parts as follows:

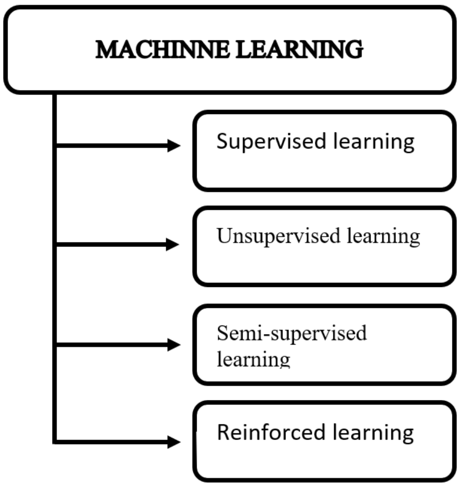


Figure 2:-Classification of Machine Learning [38]

VI.B. SUPERVISEDLEARNING

Supervised Learning defines the process of providing input data and giving the correct output to the machine learning model. It also helps us to solve various real-world problems. It includes the following classifications which are given below:

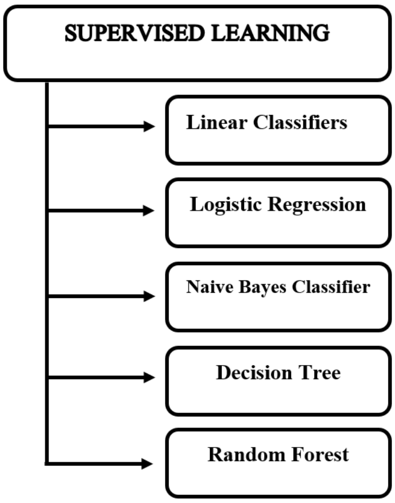


Figure 3:-Classification of Supervised Learning [30]

VI.C. DECISION TREE CLASSIFIERS

It is the most popular method to classify the techniques in data mining. [40] It has the capability to handle the large amount of information. It has several types of Decision Tree algorithms such as:

a) IterativeDichotomies3(ID3),

b) SuccessorofID3(C4.5),

c) Classification And Regression Tree (CART)[42],

d) CHi-squared Automatic Interaction Detector (CHAID)[43],

e) Multivariate Adaptive Regression Splines(MARS)[44],

f) Generalized, Unbiased, Interaction Detection and Estimation (GUIDE), Conditional Inference Trees (CTREE) [45] [46]

g) Classification Rule with Unbiased Interaction Selection and

h) Estimation(CRUISE), Quick, Unbiased and Efficient

i) Statistical Tree(QUEST).[41]

VI.D. SUPPORT VECTOR MACHINES

Support Vector Machines is related to classical multi-layer perception neural networks. Support vector machines (SVMs), which were introduced by Vapnik and his coworkers in the early 1990's (Cortes, Vapnik 1995; Vapnik 1996, 1998), these are proved to be effective techniques for data mining (Peng et al. 2008; Yang, Wu2006).[47][48]. The main motive of SVM is to divide the datasets in to classes to find a maximum marginal hyperplane (MMH).

Important concept in SVM is as following: -

i. Support Vectors: - Support vectors are all the data points which are closest to the Hyper line.

ii. Hyper line: - Hyper line is the decision plane or space which is divided between a set of objects having different classes.

iii. Margin: - Margin is the gap between two lines on the closet data points of different classes.

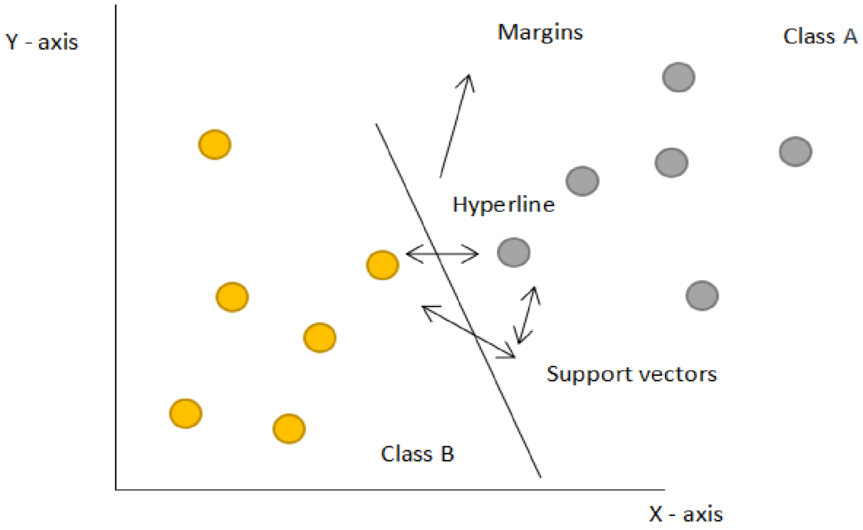


Figure4:-Working graph of SVM [39]

VI.E. NAIVE BAYES

Naïve Bayes is the simplest and most powerful machine learning algorithm which is used to solve classification problem. Basically, it's used for text classification. Naïve Bayes methods are a set of supervised learning algorithms which is based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. This dataset is divided into two features -

i. Feature Matrix

ii. Response Vector

Naive Bayes theorem explains the probability of an event occurring given the probability of another event that has already occurred. The equation of Naive Bayes theorem is given below-

$$\boldsymbol{P(A|B|)} = \frac{\boldsymbol{P(B|A|)P(A)}}{\boldsymbol{P(B)}}$$

Where it defines A and B are the events and P(B)≠0.

VI.E.I. Gaussian Naive Bayes

Gaussian Naive Bayes is a different version of Naive Bayes which follows the Gaussian normal distribution and supports continuous data containing many non-prior knowledge. A way to deal with make a straightforward model is to expect that the information is described by a Gaussian method with no independent dimensions between dimensions.

VI.E.II. Bernoulli Naive Bayes

Bernoulli Naive Bayes is a different version of Naive Bayes. It is used for discrete data which works on Bernoulli distribution. It depends on the Bernoulli distribution and acknowledges binary values, which is 0 or 1.

VI.E.III. Multinomial Naive Bayes

The Naive Bayes method is a strong tool for analyzing text input and solving problems with numerous classes. Multinomial Naive Bayes algorithm is a probabilistic learning strategy that is utilized in Natural Language Processing (NLP).

The algorithm depends on the Bayes theorem and predicts the tag of a text. It computes the probability of each tag for a given example and afterward gives the tag with the highest probability as result.

Table no 3: Differences of various Naive Bayes algorithms.

Gaussian Naïve Bayes	**Multinomial Naïve bayes**	**Bernoulli Naïve Bayes**
It follows normal distribution.	It follows Bayesian Learning approach in Natural language Processing (NLP)	It follows the Bernoulli distribution.
It is good at handling continuous values	It is good at handling discrete values.	It is good at handling Boolean/binary attributes.
It tells the assumption of each class whether it is normally distributed or not.	This model tells two facts that whether the word occur in a document or not as well as its frequency in that document	The term occurs in a document.

VI.F. Random Forest

Random Forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. Random Forest, like its name proposes, contains several numbers of individual decision trees that work in coordinates.[49] Each-and-every individual tree in the irregular forest lets out a class forecast and the class with the most votes, which we select as our model for prediction.

VI. IMPLEMENTATION

This project includes dataset where all the data is taken from the posts of “Suicide Watch” and “depression” subreddits of the Reddit platform etc., and the dataset contain the comment of the post, their own post. After importing the dataset first work to get the size of the dataset, for that we use “. shape” function and we get shape as (232074, 3). Since the dataset contain various emoji’s and stop-words and individual character hence it is particularly important to clean the data in the first place. For that we use “neattext.functions” which is use to clean the text in column wise. Now we get the clean data and now it was ready for splitting. But as in this paper we are implementing different model algorithm so in the first place we must convert string dataset to an integer dataset. For that we must perform level encoding to convert the dataset. Now the dataset is ready for splitting. We split the data according to the calculation so that we get maximum accuracy. Now, the main data set is split into 7:3 ratio, where 70% is training dataset and where 30% testing dataset. After that we remove the target column from both the part and store differently in the different dataframe and the remaining data on a separate dataframe. As we are working on a label dataset, so we implemented supervised learning model. In this dataset the target column having a binary value, for that we mistrusted classification model only. The first model we use is Decision Tree Algorithm where we use decision tree classifier as the data set was binary target column. Decision tree is structured tree classifier in which internal node represents the feature in the dataset, the branches represent the decision rule and leaf node represents the decision. In simple words root node represent the entire dataset, and the leaf node represents the output. As decision tree simple in decision making, so it is one of the fastest algorithms to work on. Using this algorithm, we got high accuracy in the model.

a. Second model we use Support Vector Classifier which import from sklearn library. This model works on a linear kernel function to classification. It works exceptionally well with large data size. After implementing this model, we got high accuracy.

b. Third algorithm we use in Naïve Bayes. In this algorithm we use Gaussian NB, as data is categorical, and we must perform classification model. Naïve Bayes works on Probability distribution. Gaussian Distribution is also known as normal distribution. In this model we got near about 50% accuracy.

c. Fourth algorithm we use is Random Forest. It consists of Several Decision tree algorithm. The algorithm operates by constructing a multiple of decision trees at training time and outputting the mean or mode of prediction of the individual trees. Hence it has the highest accuracy.

That is how we implement our coding part. Where, we use multiple models to get the highest accuracy. As suicide is an important topic to work on so we want to get the highest accuracy.

VIII. RESULT ANALYSIS

Suicide is one the burning topic in this fast-moving world, as society does not have time to discuss about it. So, we want to track suicidal thought on the various social media platform. Because of that we want to develop a model which can detect suicidal thought by looking at their comment, post, and text etc., and we develop the model. As it is very sensible topic to work on so accuracy is most important, that is why implement various model until we get satisfactory accuracy. The highest accuracy found in Decision Tree model. The highest accuracy was achieved was 70.95%. It means if we implement this model in real life then we will be able save more than 70.95%human resource from suicide. Along with this some other algorithm also been tasted in this research work, those as Support Vector Machine (SVM) 57.24%, Naïve Bayes (Gaussian)54.69%, Random Forest 67.67% etc. An accuracy comparison graph has been included [Figure 5].

Table no 2: - Accuracy Representation.

Model	Percentage (%)
Support Vector Machine	57.24%
Gaussian Naïve Bayes	54.69%
Multinomial Naïve Bayes	51.65%
Bernoulli naïve Bayes	49.92%
Random Forest	67.67%
Decision Tree	70.95%

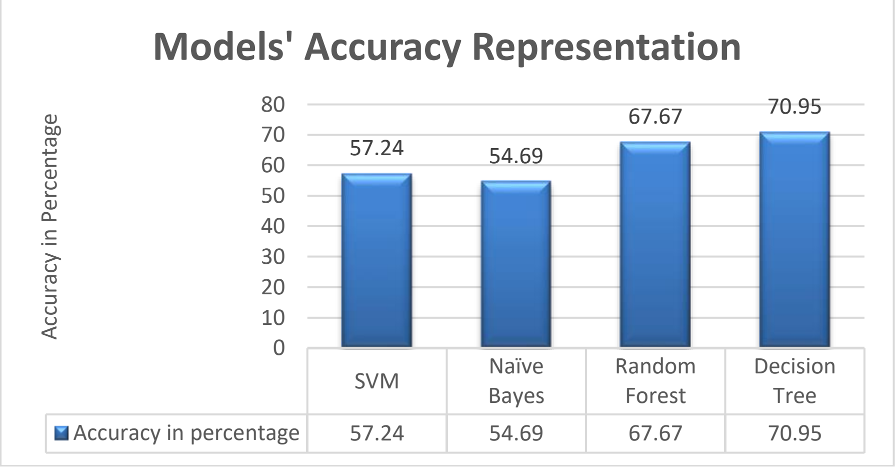
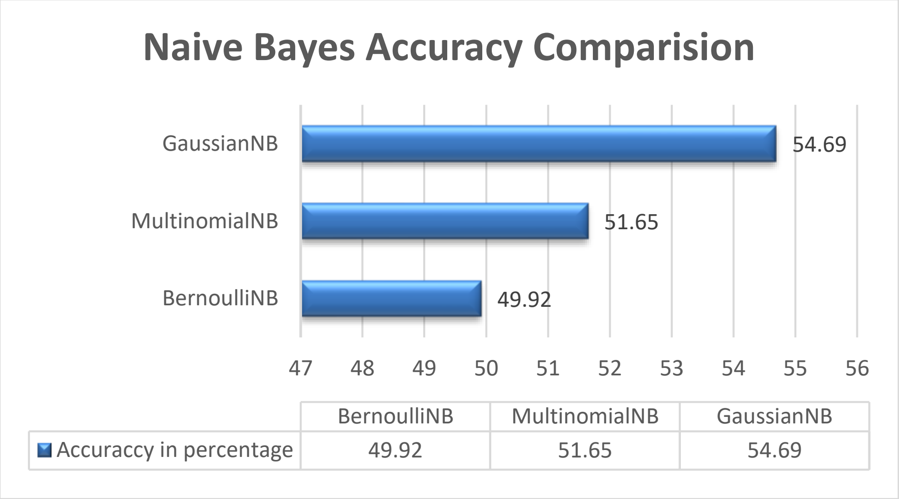


Figure 5: - Accuracy Representation of various models

VIII.A. NAIVE BAYES

Along with these above-mentioned algorithms, this research works explores some other versions of Naïve Bayes algorithm also. There are three distinct types of Naïve Bayes, Gaussian Naive Bayes, Bernoulli Naive Bayes, and Multinomial Naive Bayes. This research work able to identify the better version of Naïve Bayes is Gaussian Naïve Bayes which records highest accuracy with 54.69% among all the versions of Naïve Bayes [Figure 6].



	BernoulliNB	MultinomialNB	GaussianNB
▪ Accuraccy in percentage	49.92	51.65	54.69

Figure 6: - Accuracy Representation of Naïve Bayes.

IX. CONCLUSION AND FUTURE WORK

Prevention of suicide is becoming most important task. By detecting the sentiment of people in early time we can prevent suicide. To detect the suicidal sentiment there are several algorithms Naive Bayes, Super Vector Machines, Random Forest, Deep Learning, etc. We have seen that to detect the sentiment of the people we can approach through various datasets such as machine learning models and ensemble models.

We implemented Decision Tree, Naïve Bayes (Multinomial) Naïve Bayes (BernouliNB), Naïve Bayes (GaussianNB), Support Vector Machine, Random Forest, a Machine learning model, which are applied to detect suicidal tendency from different social media. Our implementation exhibited that our model performing Machine Learning models and Supervised learning approaches with a dataset containing 232074 unique values. After implementing these algorithms this research works has able to achieve a proficient level of accuracy.

There is a higher possibility to improve this accuracy score using some other machine learning model. The method proposed in this work to improve the accuracy is Ensemble Model. Where multiple models combined to build the classifier which can lead this research works to a new achievement with a higher accuracy compare to the traditional models being used before. This research works prefers Ensemble Model as a future work to improve the accuracy.

REFERENCES

[1] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press. 1. Worldhealthorganisation.Data17 J une2021-https://www.who.int/news-room/fact-sheets/detail/suicide

[2] Internationassociationforsuicideprevention.WSPDFacts&FiguresInfographic(iasp.info)

[3] AmericanCollegeHealthAssociation(ACHA).Verywellmind.StatisticsonCollegeandTeen Suicides(verywellmind.com)

[4] AmericanFoundationforSuicidePrevention.AFSP.Suicidestatistics|AFSP

[5] Pestian, John & Matykiewicz, Pawel & Linn-Gust, Michelle & South, Brett & Uzuner, Ozlem & Wiebe, Jan & Cohen, Kevin & Hurdle, John & Brew, Chris. (2012). Sentiment Analysis of Suicide Notes: A Shared Task. Biomedical informatics insights. 5. 3-16. 10.4137/BII.S9042.

[6] Berardinelli, Nabeela & Gaber, Mohamed & Haig, Ella. (2013). SA-E: Sentiment Analysis for Education. Frontiers in Artificial Intelligence and Applications. 255. 10.3233/978-1-61499-264-6-353.

[7] M.Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “LexiconBased Methods for Sentiment Analysis,” Association forComputational Linguistics,2011.

[8] S. Sharma, “Application of Support Vector Machines for Damagedetection in Structure,” Journal of Machine Learning Research,2008.

[9] A.Sharma,andS.Dey,“PerformanceInvestigationofFeatureSelectionMethodsandSentimentLexiconsforSentimentAnalysis,”AssociationfortheadvancementofArtificialIntelligence,2012

[10] Twitter Sentiment Analysis. Alec Go (alecmgo@stanford.edu) LeiHuang (leirocky@stanford.edu) RichaBhayani(richab86@stanford.edu) CS224N - Final Project Report June 6,2009,5:00PM(3LateDays)

[11] Sentiment Analysis of Impact of Technology on Employment fromTextonTwitterhttps://doi.org/10.3991/ijim.v14i07.10600ShahzadQaiserCapitalUniversityofScienceandTechnology(CUST),Islamabad,PakistanNooraini Yusoff UniversitiMalaysiaKelantan,Kelantan,Malaysianooraini.y@umk.edu.myFarzanaKabirAhmad,RamshaAli UniversitiUtaraMalaysia, Kedah,Malaysia

[12] Pak, Alexander & Paroubek, Patrick. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of LREC.10.

[13] Berardinelli, Nabeela & Gaber, Mohamed & Haig, Ella. (2013). SA-E: Sentiment Analysis for Education. Frontiers in Artificial Intelligence and Applications. 255. 10.3233/978-1-61499-264-6-353.

[14] Ramyasri, V & Niharika, Ch & Maneesh, K & Ismail, Mohammed. (2019). Sentiment Analysis of Patients' Opinions in Healthcare using Lexicon-based Method. 2249-8958. 10.35940/ijeat.A2141.109119.

[15] Munezero, M., Montero, C.S., Mozgovoy, M., & Sutinen, E. (2013). Exploiting sentiment analysis to track emotions in students' learning diaries. Koli Calling '13.

[16] Greaves, Felix & Ramirez-Cano, Daniel & Millett, Christopher & Darzi, Ara & Donaldson, Liam. (2013). Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. Journal of medical Internet research. 15. e239. 10.2196/jmir.2721.

[17] Saif, Hassan & Alani, Harith. (2012). Semantic Sentiment Analysis of Twitter. The Semantic Web--ISWC 2012. 7649. 508-524. 10.1007/978-3-642-35176-1_32.

[18] Neri, Federico & Aliprandi, Carlo & Capeci, Federico & Cuadros, Montse & By, Tomas. (2012). Sentiment Analysis on Social Media. 10.1109/ASONAM.2012.164.

[19] Sarlan, Aliza & Nadam, Chayanit & Basri, Shuib. (2014). Twitter sentiment analysis. 212-216. 10.1109/ICIMU.2014.7066632.

[20] Aladağ, Ahmet Emre & Muderrisoglu, Serra & Akbas, Naz & Zahmacioglu, Oguzhan & Bingol, Haluk. (2018). Detecting Suicidal Ideation on Forums and Blogs: Proof-of-Concept Study. Journal of Medical Internet Research. 20. e215. 10.2196/jmir.9840.

[21] McCart, James & Finch, Dezon & Jarman, Jay & Hickling, Edward & Lind, Jason & Richardson, Matthew & Berndt, Donald & Luther, Stephen. (2012). Using Ensemble Models to Classify the Sentiment Expressed in Suicide Notes. Biomedical informatics insights. 5. 77-85. 10.4137/BII.S8931.

[22] Pestian, John & Matykiewicz, Pawel & Linn-Gust, Michelle & South, Brett & Uzuner, Ozlem & Wiebe, Jan & Cohen, Kevin & Hurdle, John & Brew, Chris. (2012). Sentiment Analysis of Suicide Notes: A Shared Task. Biomedical informatics insights. 5. 3-16. 10.4137/BII.S9042.

[23] George, Amy & Johnson, David & Carenini, Giuseppe & Eslami, Ali & Ng, Raymond & Portales-Casamar, Elodie. (2021). Applications of Aspect-based Sentiment Analysis on Psychiatric Clinical Notes to Study Suicide in Youth. AMIA ... Annual Symposium proceedings. AMIA Symposium. 2021. 229-237.

[24] Birjali, Marouane & beni hssane, Abderrahim & Erritali, Mohammed. (2017). Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks. Procedia Computer Science. 113. 65-72. 10.1016/j.procs.2017.08.290.

[25] Mbarek, Atika & Jamoussi, Salma & Charfi, Anis & Ben Hamadou, Abdelmajid. (2019). Suicidal Profiles Detection in Twitter. 289-296. 10.5220/0008167602890296.

[26] Sohn, Sunghwan & Torii, Manabu & Li, Dingcheng & Wagholikar, Kavishwar & Wu, Stephen & Liu, Hongfang. (2012). A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes. Biomedical informatics insights. 5. 43-50. 10.4137/BII.S8961.

[27] Ji, Shaoxiong & Pan, Shirui & Li, Xue & Cambria, Erik & Long, Guodong. (2019). Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications.

[28] Glenn, Jeffrey & Nobles, Alicia & Barnes, Laura & Teachman, Bethany. (2020). Can Text Messages Identify Suicide Risk in Real Time? A Within-Subjects Pilot Examination of Temporally Sensitive Markers of Suicide Risk. Clinical Psychological Science. 8. 216770262090614. 10.1177/2167702620906146.

[29] Sharma, Sparsh & Sharma, Surbhi. (2020). Analyzing the depression and suicidal tendencies of people affected by COVID-19’s lockdown using sentiment analysis on social networking websites. Journal of Statistics and Management Systems. 24. 10.1080/09720510.2020.1833453.

[30] Source:-https://www.researchgate.net/figure/Sentiment-classification-techniques_fig1_261875740

[31] Danuta Wasserman, Ellenor Mittendorfer Rutz, Wolfgang Rutz, and Armin Schmidtke. 2004. Suicide Prevention In Europe. Technical report, National and Stockholm County Council’s Centre for Suicide Research and Prevention of Mental Ill Health.

[32] M. Berk and Henry S. Dodd. 2006. The effect of macroeconomic variables on suicide. Psychol Med, 36(2):181–189.

[33] V. Venek, S. Scherer, L.-P. Morency, J. Pestian et al., “Adolescent suicidal risk assessment in clinician-patient interaction,” IEEE Transactions on Affective Computing, vol. 8, no. 2, pp. 204–215, 2017.

[34] B. O’Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, “Detecting suicidality on twitter,” Internet Interventions, vol. 2, no. 2, pp. 183–188, 2015.

[35] S. Ji, C. P. Yu, S.-f. Fung, S. Pan, and G. Long, “Supervised learning for suicidal ideation detection in online user content,”

Complexity, 2018.

[36] Ji, Shaoxiong & Pan, Shirui & Li, Xue & Cambria, Erik & Long, Guodong. (2020). Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. IEEE Transactions on Computational Social Systems. PP. 1-13. 10.1109/TCSS.2020.3021467.

[37] Çelik, Özer. (2018). A Research on Machine Learning Methods and Its Applications. 10.31681/jetol.457046.

[38] Source:-https://litslink.com/blog/an-introduction-to-machine-learning- algorithms

[39] Source:-https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

[40] "A REVIEW PAPER ON STUDENT PERFORMANCE USING DECISION TREE ALGORITHMS", International Journal of Emerging Technologies and Innovative Research (www.jetir.org | UGC and issn Approved), ISSN:2349-5162, Vol.6, Issue 3, page no. pp336-341, March-2019, Available at:http://www.jetir.org/papers/JETIRAH06058.pdf

[41] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning", JASTT, vol. 2, no. 01, pp. 20 - 28, Mar. 2021.

[42] C. E. Brodley and P. E. Utgoff, "Multivariate decision trees," Machine learning, vol. 19, no. 1, pp. 45–77, 1995.

[43] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," Energy, vol. 32, no. 9, pp. 1761–1768, Sep. 2007, doi: 10.1016/j.energy.2006.11.010

[44] S. Singh and P. Gupta, "Comparative study ID3, cart and C4. 5 decision tree algorithms: a survey," International Journal of Advanced Information Science and Technology (IJAIST), vol. 27, no. 27, pp. 97– 103, 2014.

[45] L. Rokach and O. Maimon, "Top-Down Induction of Decision Trees Classifiers—A Survey," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 35, pp. 476– 487, Dec. 2005, doi: 10.1109/TSMCC.2004.843247.

[46] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms," Machine learning, vol. 40, no. 3, pp. 203– 228, 2000.

[47] Akinsola, J E T. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT). 48. 128 - 138. 10.14445/22312803/IJCTT-V48P126.

[48] Tian, Yingjie & Shi, Yong & Liu, Xiaohui. (2012). Recent advances on support vector machines research. Technological and Economic Development of Economy. 18. 10.3846/20294913.2012.661205.

[49] "A Review On Analysis Of Suicidal Notes Sentiment Rich Data", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.9, Issue 5, page no.f21-f30, May-2022, Available :http://www.jetir.org/papers/JETIR2205604.pdf