Article

Data Mining to Detect Depression

Aafia Iqbal [1], Mardiyah Khadijah [2] and Safura Ateeq [3]

[1] f20180061@dubai.bits-pilani.ac.in

[2] f20180257@dubai.bits-pilani.ac.in

[3] f20180080@dubai.bits-pilani.ac.in

Abstract: Issues like mental health have become a common topic in recent times. Nearly one in five people in the US have been detected with some form of mental illness. This project aims at detecting Mental illness using NLP and Sentiment Analysis using feature analysis. In these tough times of COVID-19, people experienced different emotions staying away from their loved ones and staying still in one place, thus taking a toll on their mental health. Our project is aimed to help people by testing algorithms to detect depression.

Keywords: K-Means; Naive Bayes Classifier; Logic Regression; Random Forest Classifier; Precision; Recall; F1-score; Confusion Matrix.

---

1. Introduction

Data mining is the practice of working with large datasets with potentially valuable insight. You get a large amount of data & the data mining algorithms to start looking for patterns. It becomes an extremely strong tool that can be applied in healthcare specifically in psychology in this paper to detect the increase in depression in users due to various factors.

Mental Health, when the term mental health comes up all we think of it as serious mental/psychological disorders. But rather mental health includes our emotional, psychological, and social well-being [1]. Every person's mental health is equally as important as physical health. It can be visualized on a continuum ranking, from none to severe.

One of the major consequences of neglecting mental health is Depression. Depression affects more than 264 million people worldwide which is equal to 1 in 5 people. It is an extremely

complex disease. It can happen to anyone; may it be a person going through a serious illness/daily anxiety.
It takes nearly 2 weeks to notice early symptoms of Depression (NIMH, 2018). Sentiments like hopelessness, sadness, emptiness are usually associated with Depression.

It has been predicted by WHO, that depression will become one of the major medical burdens by the year 2020. The analysis of depression is done by a questionnaire & interviews which are not efficient. Although, many people don't get depression diagnosed & treated which can then affect their physical health too. According to recent statistics, nearly less than 25% of people receive and follow treatment for depression (NAMI, 2018). Diagnosing depression through a medical test is difficult and not 100% accurate. Relating the two worlds of medicine (psychology) & technology can solve this issue to a great extent.

Artificial Intelligence is transforming the world in every walk of life. It is an extremely strong tool that can be used in medicine; mental health to detect depression easily & faster. The main objective is to apply data mining techniques to psychology, specifically in the field of depression. Collect different attributes and train & test a model on the best algorithm and classify users on the basis of attributes if they are tested positive or negative for depression.

Literature surveys led to observations that text and audio attributes have sufficient details to indicate emotions and psychological states of people due to their strong relations with statistical Correlations with Depression. Thus, the focus of this paper is to analyse user text attributes only and create a machine learning model that determines their depression inclination. The ultimate goal is to build a machine learning model that receives a user-generated content as input and should output a prediction regarding the user's susceptibility to depression.

In the rest of this paper, we perform analysis of the existing discoveries from the literature Survey conducted to analyse and understand the research field and find if any gaps or additional discoveries can be done. Then, contains the description of the selected dataset. In the next section the background on all the algorithmic approaches is presented in the proposed methodology section,which covers description of the algorithms compared and tested. The next section shows the working of the algorithms and lastly, sums up the discussions and shows the results.

2. Related Works

In this section we review papers related to the topic. It is divided into two subsections, In section 2.1, we present works on Sentiment Analysis and in section 2.2 we present research on Machine Learning.

2.1. Sentiment Analysis

Lilliam Lee and Bo Pang [2] presented an informative review on sentiment analysis. In order to predict the interpretation, they examined the percentage of positive terms to overall words.

The task is quite stimulating due to the intrinsic ambiguity of the natural language structures– various modes of hate, different types of objectives and different ways of representing the same word. The majority of the previous work focuses mostly around manual extraction of features [3] or using methods of representation learning accompanied by a linear classifier [4]. Recent deep learning approaches, however, have demonstrated advances in precision across a wide range of diverse problems in voice, vision and text applications.

Wiebe et al. [5] researched on the manual annotation of private states including emotions, views, and sentiment in a 10,000-sentence corpus (the MPQA corpus) of news articles. Expressions of feelings in messages have also been explored within the Appraisal Framework [6], a functional theory of the language philosophy that is used to express attitudes, opinions and feelings [7].

In other related work, Liu et al. [8] have used real-world understanding about affect drawn from a common-sense knowledge base. They seek to grasp text grammar in order to recognize feelings at the level of the sentence. They begin from extracting the basis of information from those sentences that provide any valuable information. This knowledge is used in constructing effective text templates that are used to mark each sentence with a six-tuple corresponding to the six fundamental emotions of Ekman. Neviarouskaya et al. [9] have also used a rule-based methodology to evaluate Ekman's fundamental emotions in the phrases in blog posts.

Mihalcea and Liu [10] in their work, they have concentrated on two specific feelings, joy and sorrow. They focused on blog posts that were self-interpreted with positive and negative mood marks by the blog authors.

A recent study [11] assessed the use of social meaning in sentiment analysis literature, which found that context-based methods worked better than conventional analysis without social context (i.e. contextless methods). It also presented a taxonomy of approaches focused on the types of features used in the context: contextless approaches do not use social context; micro approaches only include user and content features; meso approaches provide user and content features, as well as connections between various users and content; and macro approaches also take advantage of other outlets, such as knowledge graph. Sentiment analysis is not a new field, or the method of examining attitudes conveyed in text, but its importance has increased as opinion-rich tools such as online review pages and personal blogs are increasingly accessible and popular [12].

While substantial research work on the identification of multimodal emotions using audio, visual and text modalities has been carried out [13], significantly less work has been committed to conversational emotion recognition (ERC). The lack of a broad multimodal conversational dataset is one key explanation for this. According to [14], the ERC poses many obstacles that make the process more challenging to handle, such as conversational contextual modeling, emotional transition of the interlocutors, and others. Latest study suggests multimodal memory network-based solutions [15].

Sequence-to-sequence generation models [16] has been successfully extended to large-scale dialogue generation including neural responding machines, hierarchical recurrent models, and many others. These models concentrate on improving the quality of the content of the responses produced, including the promotion of diversity, considering additional information [17], and handling unknown words. However, the emotion aspect in large-scale dialogue generation has not been discussed by any job. There are several studies which produce text that can be controlled from variables [18] suggested a generative model that can generate sentences that depend on certain language attributes, such as feelings and tenses.
Sentiment shifters play a central role in opinion mining, as a collection of terms and phrases that can influence text polarity. Sentiment shifters can reverse the polarity of the text given and are therefore necessary for the precise tasks of classification of polarity. In fact, most medical words such as "pain" and "depression" are pessimistic in the drug domain, but they also appear in positive sentences [19].

Grammatical principles involve very commonly used negations in the text that totally modify the meanings of words. In other words, the recognition of negation and the detection of its reach within a sentence (letter) are important to figure out the feelings of a piece of text. Different techniques can be used such as: Bag of Words, Semantic Relations, Contextual Valence Shifter, Relations and dependency-based, Analysis of Negation [20].

The aim of [21] was to assist patients with depression to track the development of their condition by using random speech to identify the general symptoms of clinical depression to give greater outcomes than using read speech. The findings showed that the use of random speech yielded a stronger outcome and that classes of jitter, shimmer, energy, and loudness features were robust in obtaining general depressive speech characteristics.
Compared to state-of-the-art approaches, the Na¨ive Bayes technique achieves greater precision. The role of deciding in a text the polarity of the present sentiment, assuming it has any. There is also a concern with binary sorting, but the alternative categories are ‘positive' and ‘negative' here. This concept is Emotion Polarity [22].

By examining facial gestures, eye movements, head movements, body movements, heart rate, and body temperature as well as their variation in the multimodal system, the suggested paradigm in [23] employs pattern detection to diagnose depression. Centered on an algebraic representation using multi-dimensional vectors, the model was developed. The premise of this model is that each vector reflects emotions that are equal to the representation of RGB colors.

Computational linguistic (CL) fields [24] are associated with Sentiment Analysis (SA) 's commitment to feeling. Naturally, SA examines the sentiment expressed by a letter, thus separating positive and negative valence at the same time, text-based emotion recognition methods like 1) Keyword spotting techniques, this approach has many common keyword labeling processes from the original text to get emotion detection categories;
2) Lexical affinity methodology, this way explores emotion detection based on relevant keywords; 3) Learning-based methodology and other techniques such as Keyboard Spotting Technique, Lexical Affinity Approach, Learning-Based Method and Hybrid Method.

The authors in [25] asked the candidates to fill out a questionnaire survey form, called PHQ-9, which could be a multipurpose instrument utilized by therapeutic experts to screen, screen, and degree the seriousness of depression. The candidate had to answer every question of the PHQ-9 form and combine the scores of all the answers indicated. The final summed up score is the PHQ-9 score of the candidate. A Score of 5 means mild depression, a Score of 10 means moderate depression, a Score of 15 means moderately severe depression, a Score of 20 means severe depression. The text content of messages of the candidates played an important indicator of depression. The sentence structure and content are the two primary things that were analyzed.

In [26], subjective sentences from online discussions about hearing aids are classified into positive, neutral, and negative, using standard sentiment analysis features, such as the number of subjective terms in the content, the number of adjectives, pronouns and adverbs, and the number of positive, neutral and negative terms, as found in the Subjectivity Lexicon. These features are used to train and test different methods like Naive Bayes, SVM, and logistic regression model. The best result was observed using logistic regression (0.68 F-1).

In recent times, indirect relationships between users are being applied into recommendation systems [27]. The main idea behind working on these projects is that similar users have the similar preferences or behavior habits. However, there is little literature that studies the use of indirect relationships in sentiment analysis.

Borth et al. [28] developed sentiment analysis on visual content with SentiBank, a system taking out mid-level semantic attributes from images. These semantic highlights are yields of classifiers that can foresee the significance of an picture concerning one of the feelings within the Plutchik's wheel of emotions [29]. Encouraged by the growth of deep learning techniques, You et al. [30] utilized convolutional neural networks on Flickr with domain transfer from Twitter for binary sentiment classification. However, image annotation research has exhibited that incorporating text characteristics with images can dramatically boost efficiency, shown by Guillaumin et al. [31] and Gong et al. [32].

2.2. Machine Learning

ML strategies regard sentiment analysis as a text classification challenge [33]. In these methods, features such as unigrams,word embeddings, bigrams are extracted from the text content first, and then features are fed to classification models such as NB, SVM, and deep neural networks (CNNs, RNNs) and so on. Machine learning methods are usually supervised and usually need lots of training data with polarity labels. The accuracy and precision of sentiment categorization is related to the size of the training data.

Traditional ML faces common training issues, like as generalization, model interpretation, and overfitting. So, analysts moved to machine learning strategies that have risen in later years as an effective instrument. This change was mainly because machine learning techniques can solve more complex problems, especially in health data [34].

3. Data

In this section, we provide a brief description of the dataset used for analysis of depression.

3.1. Review

The dataset used comprises about 1500 entries. This dataset includes information like gender, marital status, number of offsprings, education level, their income, expenditure, investments shown in. All these factors are used in determining whether a person can suffer from depression or not [35].

Some of these factors are in the form of ordinal data, some are in the form of nominal data.

	Survey_id	Sex	Age	Married	Number_children	Education_leve	Living_expenses	Other_expenses	Incoming_salary	Labor_primary	Lasting_investmen	No_lasting_investment	Depressed
2	926	1	28	1	4	10	26692283	28203066	0	0	28411718	28292707	0
3	747	1	23	1	3	8	26692283	28203066	0	0	28411718	28292707	1
4	1190	1	22	1	3	9	26692283	28203066	0	0	28411718	28292707	0
5	1065	1	27	1	2	10	397715	44042267	0	0	7781123	69219765	0
6	806	0	59	0	4	10	80877619	74503502	1	1	20100562	43419447	0
7	483	1	35	1	6	10	30696127	11531066	0	0	4442561	76629095	0
8	849	0	34	0	1	9	66730708	10890451	0	0	22562288	55608922	1
9	1386	1	21	1	2	10	80076849	58456101	0	0	33922659	54600174	0
10	930	1	32	1	7	9	30162281	67184479	1	1	14018381	15117619	0
11	390	1	29	1	4	10	26692283	28203066	0	0	28411718	28292707	0
12	540	1	84	0	0	1	26692283	28203066	0	0	28411718	28292707	1
13	557	1	59	0	2	9	262919	30108896	0	1	15714453	20214956	0
14	1280	1	38	1	4	10	10810375	498078	0	0	20745816	15708408	0
15	1195	1	27	1	4	10	21220366	10506083	0	0	62405292	12144989	0
16	603	1	56	1	0	12	34566505	72469551	0	0	12402556	71201668	1
17	729	1	24	1	2	10	38169963	37860336	0	0	23991919	48624439	0
18	770	1	25	1	3	10	40705733	40278656	0	0	11596846	12491988	0
19	76	1	44	1	5	12	26692283	28203066	0	0	28411718	28292707	0
20	1374	1	32	1	4	9	26692283	28203066	0	0	28411718	28292707	0

Figure 1. Dataset overview

4. Proposed Methodology

This section presents the proposed methodology for Detection of Mental health using Data Mining.

We collect various information from our audience like gender, marital status, number of offsprings, education level, their income, expenditure, investments and detect if they suffer from depression or not. We have chosen 4 algorithms i.e. K-Means, Naive Bayes Classifier, Logic Regression and Random Forest Classifier, to test and compare the results to find which is the best algorithm to find if any of our audience suffers from depression using factors like the level of precision, recall and F1-score of our predicted result and the result obtained by the algorithms.

This section is further composed of the applied algorithms, with explanations in Section 4.1, and the approach towards the factors used for determining the results in Section 4.2.

4.1. Algorithms Applied

This section is further divided into 4 sections, covering and explaining the algorithms used.

In Section 4.1.1 we discuss K-means Clustering Algorithm, in Section 4.1.2 we discuss Naive Bayes Classifier, in Section 4.1.3 we discuss Logistic Regression and finally in Section 4.1.4 we discuss Random Forest Classifier.

4.1.1. K-means Clustering Algorithm

Clustering is a process of dividing the dataset into various groups, which have similar data points. Clustering is categorized as Unsupervised Machine Learning technique. Clustering has many real life applications like segregating the aisles in a supermarket, organising a class of students according to their ethnicity etc. Clustering is used in recommendation systems. For example, Netflix uses clustering for recommending movies and Noon uses clustering for recommending products according to your past experiences.

K-means is a type for a subcategory of Clustering algorithm namely Exclusive Clustering. K-means is a clustering algorithm which aims to group similar data. The “k” denotes the number of clusters. K-means clustering initially chooses k number of random distinct points and assigns them as the starting point. The starting point is known as Centroid. Then further more elements are chosen and categorized in the Centroid according to their level of similarity.

The K-means algorithm can be applied to numeric or continuous data. Once all the data points are allotted to a cluster, the mean of the clusters is calculated. This process keeps on iterating till the mean value of all the clusters is the same (uniform).

4.1.2. Naive Bayes Classifier

Naive Bayes Classification works on the principle of conditional probability according to the Bayes Theorem. Bayes Theorem gives conditional probability of an event A given another event B has already occurred. Equation (1) below is used for determination of Naive Bayes Classification

(1)

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

Naive Bayes Classification has various applications Face Recognition, Weather Prediction, Diagnosis in Medical field and Classifying the News. Naive Bayes Classification is categorized as Supervised Machine Learning technique.

4.1.3. Logistic Regression

Logistic Regression is categorized as Supervised Machine Learning technique. It uses continuous data. But predicts if something is true or false instead of predicting a continuous value.

Logistic Regression is a method used for predicting a dependent variable, given a set of independent variables, such that the dependent variable is categorical. Equation (2), is used for determination of Logistic Regression

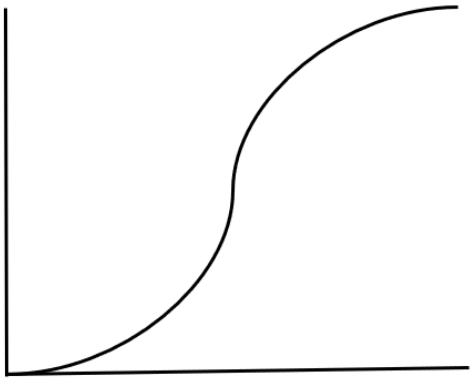

Figure 2 : S-shaped curve for Logistic Regression

(2)

$$\log\left(\frac{Y}{1-Y}\right) = C+B1X1+B2X2\ldots$$

4.1.4. Random Forest Classifier

Random Forest is a method that operates by constructing multiple Decision Trees during the training phase. Decision Tree is a tree shaped diagram used to determine course of action. Each branch of a tree of the tree represents a possible decision, occurrence or reaction.

Random Forest is a supervised machine learning technique. Random Forest Classification uses Ensemble learning which uses results based on combined results of various independent models.

4.2. Factors used for determining results

This section is divided into 4 subsections which discuss the factors used in determination of the best algorithm to determine depression .

In Section 4.2.1 we will discuss Confusion Matrix , In Section 4.2.2 we discuss Precision, in Section 4.2.3 we discuss Recall and in Section 4.2.4 we discuss F1-score.

4.2.1. Confusion Matrix

A confusion matrix is a table containing positive and negative values which have been predicted right or wrong shown in Figure 2. It is used for describing performance of the type of classification algorithm being used.

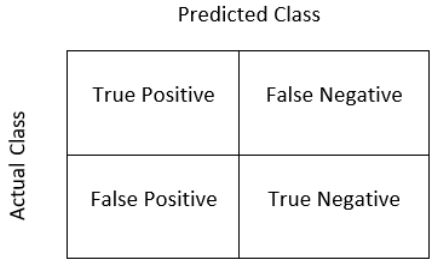

Figure 3 . Confusion Matrix

True Positive - Positive values which were correctly predicted

True Negative - Negative values which were correctly predicted

False Positive - Positive values which were not predicted correctly

False Negative - Negative values which were not predicted correctly

4.2.2. Precision

Precision is a ratio of assumed results that correctly predicted positive observations to the system's total predicted positive observations, both True and False .

It is a measure of exactness. The equation for Precision is given below in Equation(3).

(3)

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

4.2.3. Recall

Recall is the ratio of assumed results that correctly predicted positive observations to all the observations that are positive.

Recall is also known as Sensitivity or True Positive Rate. It is a measure of completeness. Equation(4) is used for calculating Recall.

(4)

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

4.2.4. F1-score

F1-score is the harmonic mean precision and recall. It takes True Positives, False Positives and False Negatives into account. Equation(5) is used for determining the F1-score.

(5)

$$\text{F1 Score} = \frac{2 \text{ x (Precision x Recall)}}{\text{Precision + Recall}}$$

5. Implementation

In this section, we discuss our implementation environment in detail. This section includes 2 sub-sections. In 5.1 the experimental setup is explained, 5.2 explanation of data collection process and data description.

5.1 Experimental Setup

The setup can be summarised in Table 1. The Operating System used is Windows 10, with 32 GB memory coded in Jupyter Notebook using Anaconda's open source network in the language Python 3.6.

Table 1. System components and specifications.

System components	Specifications
Operating System	Windows 10
Memory	32 GB
Platform	Anaconda 1.10
Application	Jupyter Notebook 6.0.3

5.2 Experimental Data

The data considered for experiment, contains 12 attributes as follows:

1. Sex
2. Age
3. Married
4. Number_children
5. Education_level
6. Living_expenses
7. Other_expenses
8. Incoming_salary
9. Labor_primary
10. Lasting_investment
11. No_lasting_investment
12. Depressed

Most dominant attributes for evaluation of depression are: Age, Sex and Married.

The observations are displayed in Figure 4, 5, 6 respectively.

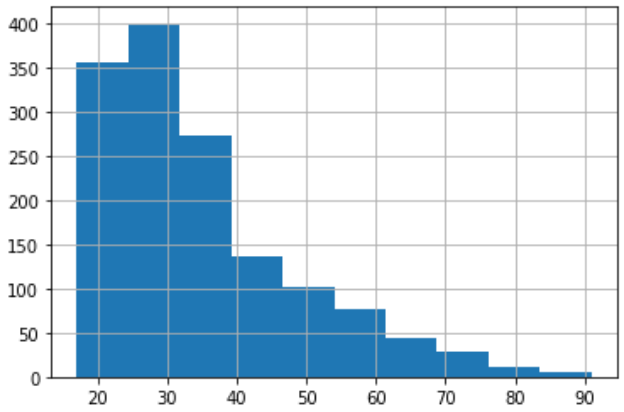

Figure 4. Histogram depicting age distribution with positive depression.

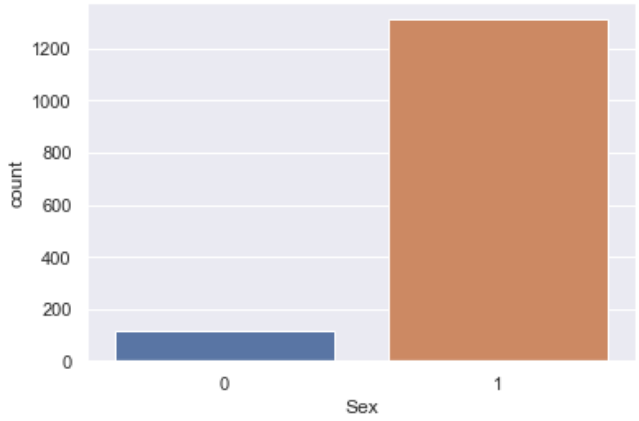

Figure 5. Bar graph of sex(Female=1 and Male=0) distribution with positive depression

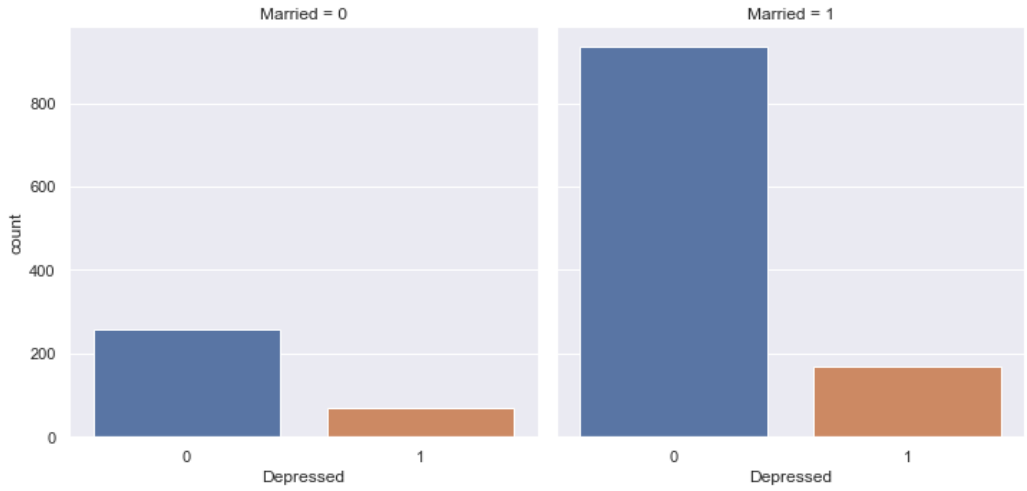

Figure 6: Bar graph of marital status with positive depression

From the above plots it is observed that, people between the ages of 20-35, female and married are more prone to be diagnosed with depression.

6. Results

In this section the data is tested using 4 different algorithms: K-Means, Naive Bayes Classifier, Logic Regression and Random Forest Classifier. In section 6.1, results on K-Means clustering are presented. In section 6.2 Naive Bayes Classifier results are presented, 6.3 Logic Regression results and in 6.4 Random Forest classifier results followed by precision, recall, and F1 score of confusion matrix calculated for each. In section 6.5, the accuracy of each model is discussed.

6.1 K-Means Clustering

For detecting depression we first used the KMeans clustering model to train and test the data importing from sklearn package `sklearn.cluster`. KMeans helps reduce the size of the dataset. It uses the partitional approach. First, the data is partitioned into 1 which results in (1429, 12) results into the cluster size to be equal to 1429. Columns: 'Living_expenses', 'Other_expenses', 'Lasting_investment', 'No_lasting_investmen' are separated to create better clusters. Furthermore, the data is partitioned into 3 which results in (1429,3). The mean or centroids of the dataset are displayed in Figure 7.

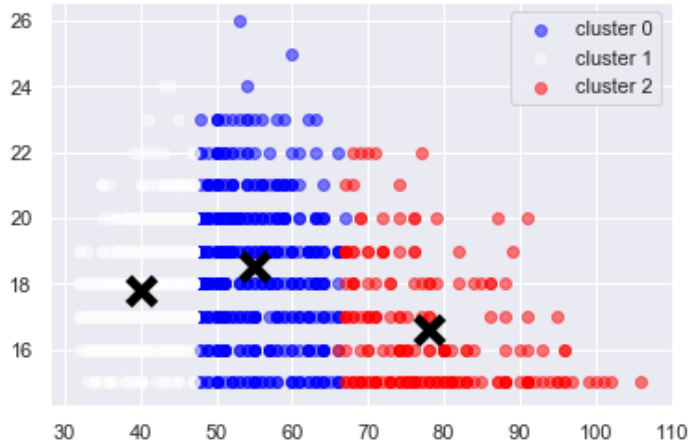

Figure 7: Scattered plot of Centroids

The confusion matrix for K-Means displayed the true and false predictions in Figure 8.

We perceive that 280 observations have been predicted correctly whereas 78 observations have been predicted wrong out of 358.

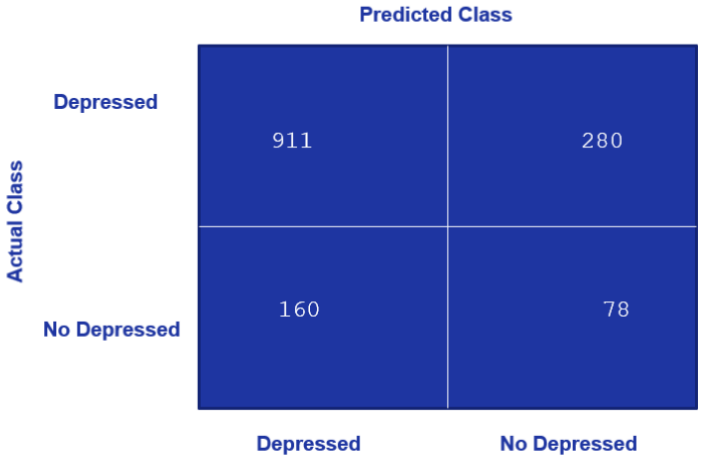

Figure 8. Confusion Matrix for K-Means

Observations from Confusion Matrix in Table 2:

The measure of purity for clusters is 0.7494751574527642.

Precision	Recall	F1- Score
0.53	0.54	0.53

Table 2. KMeans Results

6.2 Naive Bayes Classifier

Another model called GaussianNB is used to test the same data for better results, it is imported from sklearn package `sklearn.naive_bayes`. For a naive bayes classifier, the data is divided into training and testing, 70% for training and 30% for testing. This is fed onto the GaussianNB() model. The confusion matrix for Naive Bayes classifier displayed the true and false predictions in Figure 9.

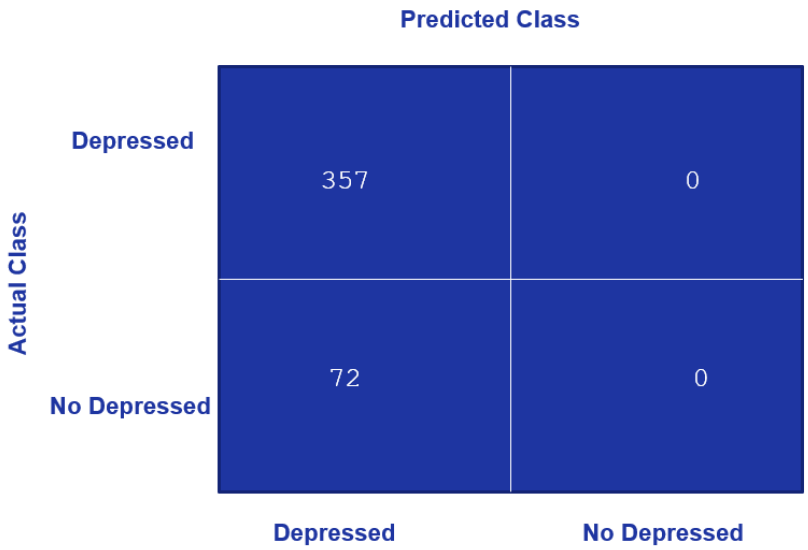

Figure 9. Confusion Matrix for Naive Bayes Classifier

Observations from Confusion Matrix in Table 3:

	Precision	Recall	F1- Score
0	0.83	1.00	0.91
1	0.00	0.00	0.00

Table 3.Naive Bayes Results

6.3 Logistic Regression

Logistic Regression inputs values (x) combines linearly using weights or coefficient values to predict an output value (y).The model is imported from sklearn package `sklearn.linear_model`.The data is fed into the Logistic Regression Classifier again using the 7:3 ratio. The confusion matrix observed for Logistic Regression is shown in Figure 10.

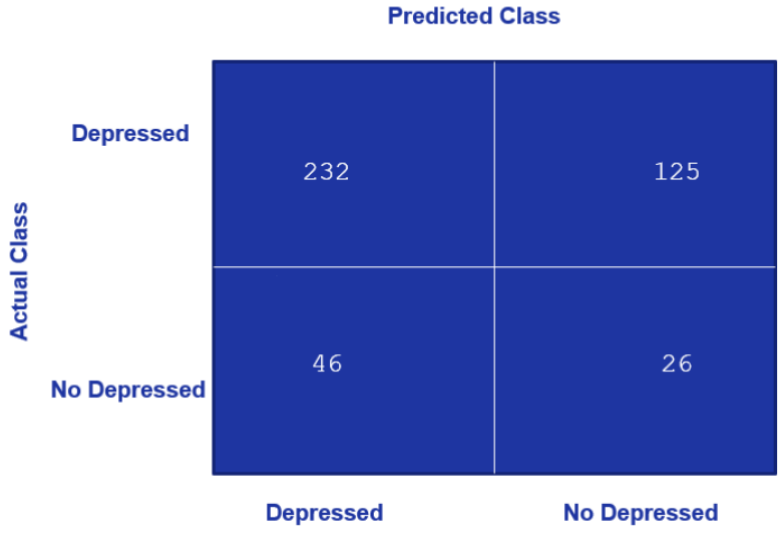

Figure 10: Confusion Matrix for Logistic Regression

Observations from Confusion Matrix in Table 4:

	Precision	Recall	F1- Score
0	0.83	0.65	0.73
1	0.17	0.36	0.23

Table 4. Logistic Regression Results

6.4 Random Forest Classifier

Random Forest Classifier is imported from sklearn package `sklearn.ensemble`. The data is fed into the RandomForestClassifier() using the 7:3 ratio.

The confusion matrix observed for Random Forest is shown in Figure 11.

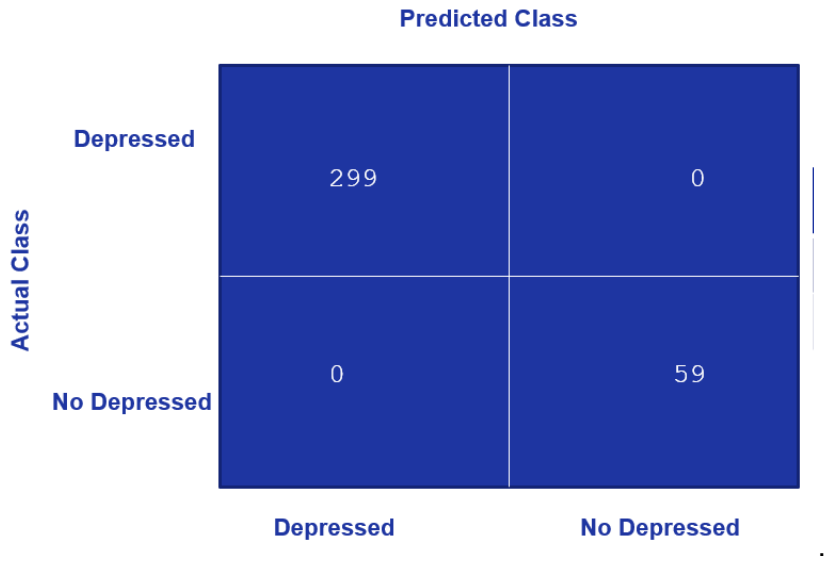

Figure 11. Confusion Matrix for Random Forest

Observations from Confusion Matrix in Table 5:

	Precision	Recall	F1- Score
0	1.00	1.00	1.00
1	1.00	1.00	1.00

Table 5. Random Forest Results

6.5 Accuracy

The 4 models are finally analysed on the basis of Accuracy; the degree to which the result of a measurement, calculation, or specification conforms to the correct value or a standard also known as the testing error.

Accuracy of K-Means = `0.74`

Accuracy of Naive Bayes Classifier = `0.83`

Accuracy of Logistic Regression = `0.60`

Accuracy of Random Forest Classifier = `1.00`

Accuracy is visualised in Figure 12.

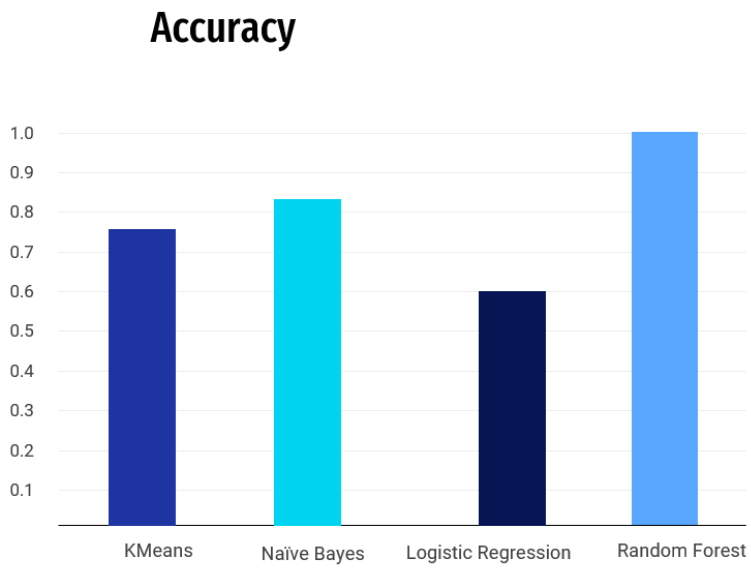

Figure 12. Accuracy of Classifiers

The accuracy of the algorithms in descending order: Random Forest Classifier, Naive Bayes Classifier, K-means and Logistic Regression.

7. Discussion and Challenges

The purpose of this paper was to provide a state of art overview in research about Data Mining techniques and algorithms applied to detecting depression. The Data Mining techniques have recently become a predominant field of research with wide applications in medical healthcare, financial services, telecommunications, natural sciences, etc. It is a process to discover useful models in data, with the aim of interpreting existing behaviors or predicting future results [36]. The Data Mining approach can significantly help the research into mental illness, to find patterns and knowledge embedded into the data. It requires exploration and analysis of large quantities of data for the purpose of better understanding and deriving knowledge regarding the problem at hand [37].

The approach of using Data Mining techniques in psychiatry has the potential to open a completely new area of research in the detection, diagnosis and classification of psychiatric disorders such as schizophrenia, dementia, depression, anxiety and alcohol abuse. Today, depression occurs in adolescents and suicidal depression number increases every time. Therefore, it is known that depression is a contaminant of morbidity, mortality and economic loss. Although effective treatments for Mental Health conditions such as depression and anxiety have been available for some time, less than half of people with a mental disorder search for a primary care medical or psychiatrist. Understanding the factors that predicting mental health care-seeking behaviors is crucial for the formulation of health policies and the design of interventions to address inequities in access to Mental Health services [38].

Nowadays, the main challenge in studying emotional disorders is to master the patient's emotional changes. The research on user emotion detection mainly includes three categories: emotions recognition based on audio-visual signals, physio-logical signals and multimodal data [39]. Continuous monitoring of a person's stress levels is essential to understand and manage personal stress [40]. A series of physiological markers are widely used for the stress assessment, which include: skin galvanic response, various characteristics of heartbeat patterns, blood pressure and respiratory activity [41]. Suicide is another of the most feared consequences of mental illness. Early recognition and accurate diagnosis of depression are essential criteria to optimize treatment selection and improve outcomes, thus reducing the economic and psycho-social burdens that result from hospitalization, lost work productivity and suicide [42].

Table 6. The studies found related to the Data Mining techniques and algorithms in patients with depression

Authors	Year of publication	Study proposal	Techniques and Algorithms	Results
		neurodegeneration. The work is focuses on the identification of depression symptoms that coexist with the cognitive deterioration, the correlation of the examined neurophysiological features with the geriatric depression combined with cognitive impairment.		Random Forest being the most accurate (95.5%).
Kim et al. [59]	2017	They propose a simple and discreet detection system that uses passive infrared sensors to monitor the daily life activities of elderly who live alone.	Neural networks, DT C4.5, Bayesian networks, SVM	- Neural networks surpasses the other algorithms, followed by C4.5 DT and is effective to detect normal conditions and mild depressions with up to 96% accuracy.

In [43], the authors have proposed a new approach using Data Mining techniques to predict the stress level of a patient using a logistic model tree and know different factors that affect the Mental Health of the patient efficiently. Stress prediction and generated rules will act as a support tool to assist medical experts provide treatment and to consult patients to take precautions to prevent future complications. It also will reduce the cost of several medical tests and facilitate patients to take preventive measures well in advance.

Information technologies have the power to positively trans-form the way patients are treated, and help us advance knowl-edge more quickly [44]. Patients can receive highly personalized treatments, therapists will receive help in making evidence-based decisions, and the scientist will be able to search newknowledge that reveals the true causes of Mental Health illnesses while developing more effective treatment approaches.

8. Conclusion

Depression has been a serious mental illness since past decades which negatively affects human's health. It is difficult to confirm human's depression symptoms from their behaviors via restricted clinic records. Our proposed methods and experiments illustrate that user details provide rich information for depression symptoms extraction from a distinctive perspective. We have demonstrated the potential of using data mining as a tool for measuring and predicting depressive disorder in individuals. Our aim was to establish a method by which recognition of depression through analysis of large-scale records of a user's personal details, and we yielded promising results with good accuracy.

The primary contributions of this study are as follows:

1. Collect an initial set of dataset consisting of 1500 entries where factors are in the form of ordinal and nominal data.
2. The data contained 12 attributes among which the dominant attributes for evaluation of depression are Age, Sex and Married.
3. The plots depicted through the histogram and bar graph showed people between the ages of 20-35, female and married are more prone to be diagnosed with depression.
4. The data is tested and the confusion matrix for each of the 4 different algorithms: K-Means, Naive Bayes Classifier, Logic Regression and Random Forest Classifier is displayed. The precision, recall, F1-score and support is observed from their confusion matrix.
5. Accuracy of each of the algorithms is analyzed based on the standard value or testing error.
6. The accuracy of the algorithms in descending order: Random Forest Classifier, Naive Bayes Classifier, K-means and Logistic Regression.

In the future, we will collect other types of data, e.g. image and video from other social networks to detect depression. Additionally, advanced entity selection techniques would be used to select more accurate and meaningful depression symptoms. Among future directions, we hope to have the ability to estimate, extrapolate, and interpret daily variations in depression and may prove itself as a useful tool for identifying depression prior to mild onset, and therefore expand its potential to save lives. Determining techniques that may be used in a medical context to identify clinical depression from the behavior of social media users' is an important task to benefit the populace.

This effort addresses an automated device for detecting depression from acoustic features in speech. The tool is aimed at lowering the barrier of entry in seeking help for potential mental illness and supporting medical professionals' diagnoses.
Early detection and treatment of depression is essential in promoting remission, preventing relapse, and reducing the emotional burden of the disease. Current diagnoses are primarily subjective, inconsistent across professionals, and expensive for the individual who may be in dire need of help. Additionally, early signs of depression are difficult to detect and quantify.

These early signs have a promising potential to be quantified by machine learning algorithms that could be implemented in a wearable artificial intelligence (AI) or home device. The motivation behind is for detecting depression and is aimed at lowering the barrier of entry in seeking help for potential mental illness and supporting medical professionals' diagnoses.

Author Contributions: All the three authors collectively conceived the idea for the paper and wrote a collective Literature Survey and designed the experiments. Aafia worked on Abstract, the Dataset and proposed Methodology. Safura worked on Introduction, Implementation and Results. Mardiyah worked on Discussion, Challenges and Conclusions. All the authors have proof-read the manuscript and agreed to the published version of the manuscript.
Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgments: We express our sincere and heartfelt gratitude to Prof. R. N. Saha, Director, BITS Pilani, Dubai Campus for allowing us to apply and understand our engineering concepts in a practical atmosphere. We would like to thank our project guide Dr. J Angel Arul Jothi for her instinct help and valuable guidance with a lot of encouragement throughout this project work, right from the selection of topic work up to its completion. We are also thankful to all teaching and non-teaching staff members, for their valuable suggestions and cooperation for the completion of this Assignment. Also, sincere gratitude for all the people who directly and indirectly helped in completing this assignment.

References

[1] Koldijk, Saskia, Mark A. Neerincx, and Wessel Kraaij. "Detecting work stress in offices by combining unobtrusive sensors." *IEEE Transactions on Affective Computing* 9.2 (2016): 227-239.

[2] Pang, Bo, and Lillian Lee. "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. CoRR." *arXiv preprint cs.CL/0409058* (2004).

[3] Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." *Proceedings of the NAACL student research workshop*. 2016.

[4] Djuric, Nemanja, et al. "Hate speech detection with comment embeddings." *Proceedings of the 24th international conference on world wide web*. 2015.

[5] Wiebe, Janyce, Theresa Wilson, and Claire Cardie. "Annotating Expressions of Opinions and Emotions in." *To appear in Language Resources and Evaluation* 1 (2004): 2.

[6] Martin, James R., and Peter R. White. *The language of evaluation*. Vol. 2. London: Palgrave Macmillan, 2003.

[7] Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. "Using appraisal groups for sentiment analysis." *Proceedings of the 14th ACM international conference on Information and knowledge management*. 2005.

[8] Liu, Hugo, Henry Lieberman, and Ted Selker. "A model of textual affect sensing using real-world knowledge." *Proceedings of the 8th international conference on Intelligent user interfaces*. 2003.

[9] Neviarouskaya, Alena, Helmut Prendinger, and Mitsuru Ishizuka. "Analysis of affect expressed through the evolving language of online communication." *Proceedings of the 12th international conference on Intelligent user interfaces*. 2007.

[10] Mihalcea, Rada, and Hugo Liu. "A corpus-based approach to finding happiness, in the AAAI Spring Symposium on Computational Approaches to Weblogs." (2006).

[11] Sánchez-Rada, J. Fernando, and Carlos A. Iglesias. "Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison." *Information Fusion* 52 (2019): 344-356.

[12] Pang, B.; Lee, L. Opinion Mining, and Sentiment Analysis. Found Trends R Inf. Retr. 2008, 2, 1–135.

[13] Zadeh, AmirAli Bagher, et al. "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018.

[14] Poria, Soujanya, et al. "Emotion recognition in conversation: Research challenges, datasets, and recent advances." *IEEE Access* 7 (2019): 100943-100953.

[15] Hazarika, Devamanyu, et al. "Conversational memory network for emotion recognition in dyadic dialogue videos." *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. Vol. 2018. NIH Public Access, 2018.

[16] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.

[17] Herzig, Jonathan, et al. "Neural response generation for customer service based on personality traits." *Proceedings of the 10th International Conference on Natural Language Generation*. 2017.

[18] Hu, Zhiting, et al. "Toward controlled generation of text." *arXiv preprint arXiv:1703.00955* (2017).

[19] Rahimi, Zeinab, Samira Noferesti, and Mehrnoush Shamsfard. "Applying data mining and machine learning techniques for sentiment shifter identification." *Language Resources and Evaluation* 53.2 (2019): 279-302.

[20] Asmi, Amna, and Tanko Ishaya. "Negation identification and calculation in sentiment analysis." *The second international conference on advances in information mining and management*. 2012.

[21] Alghowinem, Sharifa, et al. "Detecting depression: a comparison between spontaneous and read speech." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.

[22] Gholipour Shahraki, Ameneh. "Emotion Mining from Text." (2015).

[23] Alabdani, Ohoud Mohammad, Abeer Abdulaziz Aldahash, and Lubna Yousef AlKhalil. "A framework for depression dataset to build automatic diagnoses in clinically depressed Saudi patients." *2016 SAI Computing Conference (SAI)*. IEEE, 2016.

[24] Hulliyah, Khodijah, Normi Sham Awang Abu Bakar, and Amelia Ritahani Ismail. "Emotion recognition and brain mapping for sentiment analysis: A review." *2017 Second International Conference on Informatics and Computing (ICIC)*. IEEE, 2017.

[25]Resom, Adonay Zenebe, et al. "Machine Learning for Mental Health Detection." (2019).

[26] Ali, Tanveer, et al. "Can i hear you? sentiment analysis on medical forums." *Proceedings of the sixth international joint conference on natural language processing*. 2013.

[27] Tang, Jiliang, et al. "Recommendation with social dimensions." *AAAI*. Vol. 2016. 2016.

[28] Borth, Damian, et al. "Large-scale visual sentiment ontology and detectors using adjective noun pairs." *Proceedings of the 21st ACM international conference on Multimedia*. 2013.

[29] Plutchik, Robert. "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice." *American scientist* 89.4 (2001): 344-350.

[30] You, Quanzeng, et al. "Robust image sentiment analysis using progressively trained and domain transferred deep networks." *arXiv preprint arXiv:1509.06041* (2015).

[31] Guillaumin, Matthieu, Jakob Verbeek, and Cordelia Schmid. "Multimodal semi-supervised learning for image classification." *2010 IEEE Computer society conference on computer vision and pattern recognition*. IEEE, 2010.

[32] Gong, Yunchao, et al. "A multi-view embedding space for modeling internet images, tags, and their semantics." *International journal of computer vision* 106.2 (2014): 210-233.

[33] Poria, Soujanya, et al. "Convolutional MKL based multimodal emotion recognition and sentiment analysis." *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016.

[34] Li, Xiang, et al. "Lightweight Attention Convolutional Neural Network for Retinal Vessel Segmentation." *IEEE Transactions on Industrial Informatics* (2020).

[35] Kaggle.com. 2020. *B_Depressed.Csv*. [online] Available at: <https://www.kaggle.com/kimjihyundev/b-depressedcsv> [Accessed 13 December 2020].

[36] Yuan, Cui. "Data mining techniques with its application to the dataset of mental health of college students." *2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA)*. IEEE, 2014.

[37] Hadzic, Maja, Fedja Hadzic, and Tharam Dillon. "Tree mining in mental health domain." *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE, 2008.

[38] Cairney, John, et al. "Exploring the social determinants of mental health service use using intersectionality theory and CART analysis." *J Epidemiol Community Health* 68.2 (2014): 145-150.

[39] Yang, Shiqi, et al. "emHealth: towards emotion health through depression prediction and intelligent health recommender system." *Mobile Networks and Applications* 23.2 (2018): 216-226.

[40] Jena, Lambodar, and Narendra K. Kamila. "A model for prediction of human depression using Apriori algorithm." *2014 International Conference on Information Technology*. IEEE, 2014.

[41] Jung, Yuchae, and Yong Ik Yoon. "Multi-level assessment model for wellness service based on human mental stress level." *Multimedia Tools and Applications* 76.9 (2017): 11305-11317.

[42] Mohammadi, Mahdi, et al. "Data mining EEG signals in depression for their diagnostic value." *BMC medical informatics and decision making* 15.1 (2015): 108.

[43] D'monte, Silviya, and Dakshata Panchal. "Data mining approach for diagnose of anxiety disorder." *International Conference on Computing, Communication & Automation*. IEEE, 2015.

[44] Hadzic, Maja, Fedja Hadzic, and Tharam S. Dillon. "Mining of patient data: towards better treatment strategies for depression." *International Journal of Functional Informatics and Personalised Medicine* 3.2 (2010): 122-143.