

Prediction of Bipolar Disorder from Mental Episodes Using Machine Learning Approach

by

Sumaiya Tasmeeem

18101397

Motiur Rahaman

17301210

Karishma Meherin Khan Piasha

17301101

Samin Yasar

17241004

Murshida Akter Dina

18101233

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
May 2022

© 2022. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Sumaiya Tasmeem
18101397

Motiur Rahaman
17301210

Samin Yasar
17241004

Karishma Meherin Khan Piasha
17301101

Murshida Akter Dina
18101233

Approval

The thesis/project titled “Prediction of High-risk Bipolar Disorder and its variant of people and Primary Treatment using Machine Learning.” submitted by

1. Sumaiya Tasmeeem (18101397)
2. Motiur Rahaman (17301210)
3. Samin Yasar (17241004)
4. Karishma Meherin Khan Piasha (17301101)
5. Murshida Akter Dina (18101233)

Of Spring, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 23, 2015.

Examining Committee:

Supervisor:
(Member)

Tanvir Rahman
Lecturer
Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi
Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

A precious gift to mankind is the ability to express their emotions or feelings and also to realize. Sometimes, an omnipresent sustained feeling or emotion can dominate a person's behavior and affect his perception which can also be defined as mood. There can be illnesses of mental health like any other diseases. A bipolar disorder is one of them which is also known as manic-depressive disorder where people feel overly happy and energized sometimes and feel very sad, hopeless and unmotivated other times. It can be thought of the highs and lows as two poles of mood and this is why it is named as bipolar disorder. There are many factors which work as the main reason for this disorder such as chemical imbalance in the brain, genetic issues, periods of high stress, over uses of drugs or alcohol and many others. Now-a-days cases of bipolar disorder are increasing at an alarming rate. If it can be predicted at the primary stage, the number of cases can be reduced. Technology plays a vital role in the health sector as it is used to lessen the complication and fasten the treatment. The aim of this research is to apply different Machine Learning algorithms to symptoms-based data of patients in order to help to build a model for prediction. This model will not only focus on detecting the disease but also will provide the primary treatment to the patient. We will develop a diagnostic algorithm based on an online questionnaire. Then, a trained dataset and machine learning algorithms will be used to recognize individual bipolar disorder patients. After that, to train and validate our diagnostic model we will use an extreme gradient boosting and cross validation. Another algorithm will be used which is called Light Gradient Boosting Machine Algorithm for ensuring the best result to fulfil our main goal. Last but not the least, some random forest algorithms will be used for detecting and differentiating between the types of BD accurately so that the cases of mistreatment can be brought down.

Keywords: Bipolar Detection; Random Forest; machine learning; Extreme Gradient Boosting; Decision Tree; SVM; MDQ; Logistic Regression; CatBoost; Light-GBM; XGBoost.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our supervisor Tanvir Rahman sir for his kind support and advice in our work. He helped us whenever we needed help.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Acknowledgment	iv
Table of Contents	v
List of Figures	vii
List of Tables	1
1 Introduction	2
1.1 Thoughts behind the Prediction Model	2
1.2 Bipolar Disorder	2
1.2.1 Definition	3
1.2.2 Causes of Bipolar Disorder	3
1.2.3 Variants	3
1.2.4 Symptoms	4
1.2.5 How BPD Affects Our Life	5
1.2.6 Statistical Analysis	5
1.3 Machine Learning	5
1.3.1 Definition	6
1.3.2 Supervised Learning	6
1.3.3 Unsupervised Learning	6
1.4 Detection Of BPD Through Machine Learning	7
2 Related Work	9
3 Data Collection And Processing	14
3.1 Data Input	14
3.2 Data pre-processing	15
3.3 Combination of Dataset	20
4 Research Methodology	23
4.1 Algorithms	23
4.1.1 XGBoost	24
4.1.2 LGBM	29

4.1.3	CATBoost	35
4.2	Solution and Deployment	40
5	Comparison	44
6	Future Direction and Conclusion	47
	Bibliography	49
	Appendix B Overleaf: GitHub for L^AT_EX projects	50

List of Figures

1.1	Supervised and Unsupervised Learning	7
3.1	Interview dataset	14
3.2	Episodes dataset	15
3.3	Episodes dataset before Pre-processing	16
3.4	Episodes dataset after Pre-processing	16
3.5	Interview dataset before Pre-processing	16
3.6	Interview dataset after Pre-processing	17
3.7	Method for Best Function	17
3.8	HeatMap of Interview Dataset	18
3.9	Scatterplot of Interview Dataset	18
3.10	Marginal Plot of Interview Dataset	19
3.11	Density Plot of Interview Dataset	19
3.12	Method For Combine Datasets	20
3.13	Scatter Plot of Combination Datasets	20
3.14	BarPlot of Combined Data	21
3.15	BarPlot of Sleep and Anxiety	22
3.16	Point Plot of Combine Data	22
4.1	Confusion Matrix Before Tuning XGboost	28
4.2	Confusion Matrix After Tuning XGBoost	28
4.3	Accuracy Compare of XGboost	29
4.4	Light Gradient Boosting	30
4.5	Confusion Matrix Before Tuning LGBM	33
4.6	Confusion Matrix After Tuning LGBM	33
4.7	LGBM Accuracy	34
4.8	LGBM Run Time Comparison	35
4.9	Confusion Matrix Before Tuning Catoost	38
4.10	Confusion Matrix After Tuning Catboost	38
4.11	CatBoost Accuracy	39
4.12	CatBoost Run Time Comparison	40
4.13	Normal state of BPD	41
4.14	Manic State	42
4.15	Depression State	43
4.16	Primary Solution Features	43
5.1	Algorithms' Results Compare	44
5.2	Execution Time Compare	45

List of Tables

4.1	XGBoost hyperparameters and their value range	26
4.2	Optimal hyperparameter values after tuning by Random Search	26
4.3	Measure for each class before tuning XGboost	28
4.4	Measure for each class after tuning XGboost	29
4.5	Parameter for LGBM	32
4.6	measure for each class after tuning LGBM	34
4.7	Parameters for catboost	37
4.8	Measure For Each Class Before Tuning Catboost	38
4.9	Measure For Each Class After Tuning Catboost	39

Chapter 1

Introduction

1.1 Thoughts behind the Prediction Model

Improved computer and internet technology has increased the amount and quality of data in a variety of industries, including the healthcare industry. Despite advances in technology, the field of mental illness remains underserved. The nature of the mental illness has long been a source of contention. Plato was the first to use the term "mental health" in ancient Greece, promoting a mentalist definition of mental illness as reason aided by temper and ruling over passion [7]. Similarly, Hippocrates, who took a more physicalist approach, defined various mental conditions as imbalances between different types of "humors" around the same time [18]. Griesinger was the first to state nearly two centuries ago that "mental illness is a brain illness," an expression that has provided a powerful impetus to the more recent medical conception of mental illness[2]. Studying mental disease led us to bipolar disorder, a psychiatric ailment that creates outlandish emotional outbursts with powerful highs and lows. It can be thought of the highs and lows as two poles of mood and this is why it is named bipolar disorder. When we studied its background we come to know that, Major Depressive Disorder (MDD) and bipolar disorder (BD) are two of the most well-known mindset issues and influence 16.6% and 3.9% of the worldwide populace. These numbers have been consistently expanding since the 1990s and the two conditions are as of now among the 20 driving reasons for handicap around the world, with MDD positioned second and BD 17th [12]. The number of instances can be minimized if it can be predicted at the first stage. Many individuals can profit from this if it can be recognized using technology since machine learning model predictions allow the medical business to produce extremely precise projections. Anxiety disorder, mood disorder, eating disorder, personality disorder, post-traumatic stress disorder, psychotic disorder, and many more types of mental diseases exist. In this study, we will emphasize the common mood condition "Bipolar Disorder". Our major objective is to create a prediction model to determine people's mental states in order to advise them with appropriate counsel. So that people are aware of their mental status and can receive appropriate treatment from the start.

1.2 Bipolar Disorder

Mood disorders are a type of mental illness that mostly affect how a person feels. It is a mental condition in which a person experiences mixed feelings which are at

the extreme level when they are either very happy or very sad, or both. It is a normal phenomenon for any person whose mood can change based on what is going on around them. However, for a mood disorder to be diagnosed, the symptoms must last for at least a few weeks. Mood disorders can change how people act and make it hard for them to do normal things like go to work or school. There are two most common mood disorders such as depression and bipolar disorder and we will discuss Bipolar disorder and its impact on our life.

1.2.1 Definition

Mental sickness is a mental disorder that affects our thought, mood, and behavior. There are several types of mental disorders or ailments in the globe. One of them is Bipolar Disorder. Bipolar Disorder is a mental health disease that generates erratic emotional episodes that include intense highs and lows. It was formerly known as manic depression. This mental disorder is also distinguished by recurring depression and mania/hypomania. [8].

When a person gets discouraged due to this mental condition, he or she may feel miserable or unhappy and lose interest or delight in various exercises. He or she may feel exhilarated, eager to go, or weirdly cantankerous when his or her temperament shifts to intensity or hypomania. These kind of mood swings or mental illnesses might produce a variety of issues for him or her, including a loss of sleep. It might also make him or her exhausted all of the time. Not only that, but it can have a significant impact on his or her judgment, behavior, and ability to think. These types of emotional crises might occur infrequently or frequently throughout the year. While the great majority will have some intense expressions between scenes, others will not.

1.2.2 Causes of Bipolar Disorder

There are several numbers of factors which are believed to cause Bipolar Disorder. They include factors like genetics, environment, physical illness and substances. Bipolar Disorder can be frequently inherited with the genetic factors which accounted for approximately 80% of the condition. If one of the parents has the disease then the chances of developing the illness in their child is near about 10%. There is a 40% chance the child will develop the illness if both of the parents have bipolar disorder [23]. Another cause that is said to be accountable for having BPD is the environment surrounding the person. If the person always faces difficulties and stressful situations then it can cause him to develop this illness. Though physical illness is not a direct cause of having BPD but indirectly it can trigger a person who is suffering from it as the person is facing difficulties to handle his or her emotion. Substances can be another factor for developing BPD as certain substances which include drugs, overuses of antidepressants, hormonal medicine and a large amount of caffeine works like a catalyst for developing the illness.[23]

1.2.3 Variants

The symptoms of bipolar disorder include alternating periods of extreme depression and mania. It is not possible to fully recover from having bipolar disorder in any

circumstance. If you don't get treatment, you'll probably experience more episodes of manic or depressive behavior. People who have bipolar disorder may continue to experience symptoms even after they have received treatment. There are several subtypes of bipolar disorder, including: The mood can swing from extreme highs to extreme lows, which is a defining characteristic of bipolar I disorder. The elevated mood associated with bipolar II disorder is not as severe. In this disorder, episodes of hypomania come before episodes of severe depression.

There are phases of hypomanic and depressive symptoms that occur in a person who has a condition known as cyclothymia. These symptoms are relatively mild when contrasted with the more extreme manifestations of a full-blown hypomanic or depressive episode.

It is possible for a person to exhibit symptoms of two different moods at the same time while experiencing mania, hypomania, or depression. This condition is characterized by a lot of energy as well as by insomnia and racing thoughts. In the meantime, they might be experiencing feelings of desolation, hopelessness, and rage, in addition to the desire to hurt themselves. Rapid cycling is referred to as having four or more mood shifts within a given year. In order for two episodes to be considered separate from one another, there has to be a certain amount of time that passes between them. It is not unheard of for a person's mood to go through extreme highs and lows within the span of a single week, or even within the span of a single day. As a direct consequence of this, there may be an absence of the full spectrum of symptoms that define individual episodes. For example, the person may not have a decreased need for sleep. The behavior known as "ultra-rapid cycle," which is frequently referred to as bipolar disorder, is something that psychiatrists do not fully understand. It is possible for a rapid cycling pattern to take place at any time during an illness; however, there is some speculation that it may take place more frequently as the illness continues to progress. In most cases, women are more likely than men to engage in activities that involve rapid cycling. Regular cyclists have a greater risk of experiencing severe depression and making suicidal attempts. Antidepressants have the potential to either cause rapid cycling or to prolong existing cycles, depending on the patient.

1.2.4 Symptoms

Bipolar I disorder is defined by a manic episode lasting at least one week, whereas bipolar II disorder or cyclothymia is defined by hypomanic episodes. However, many persons with bipolar disorder have both hypomanic and manic episodes. Changes in mood do not necessarily follow a predictable pattern, and depression does not usually follow manic times. Before experiencing the opposite mood, a person may experience the same mood state numerous times, with periods of euthymia in between. Mood swings in bipolar disorder can last for weeks, months, or even years. The fact that the mood changes are a departure from your usual self and that the mood change lasts a long period is crucial. Mania might last for several days or weeks, whereas despair can last for several weeks or months. The severity of depressive and manic episodes varies from person to person and from time to time within the same person.

1.2.5 How BPD Affects Our Life

Changes in energy, thought, behavior, and sleep habits are all symptoms of bipolar illness. When you suffer bipolar mood swings, it's difficult to carry out everyday responsibilities, work, go to school, and maintain relationships. A manic episode makes you feel tremendously energetic, productive, and even invincible. Friends and relatives are typically concerned about these sudden behavioral changes. A depressed episode, on the other hand, causes a person to feel overly miserable, hopeless, and weary. They may avoid meeting friends and relatives, as well as participating in their usual activities. Bipolar disorder affects millions of adults. Bipolar disease is most commonly diagnosed in people in their teens or twenties. It can, however, happen at any age. People who have a family history of bipolar disease, have had a stressful incident, or have used drugs or alcohol excessively are more likely to develop the condition.

1.2.6 Statistical Analysis

As mentioned earlier, Major Depressive Disorder (MDD) and Bipolar Disorder (BD) are two of the most well-known mindset issues and influence 16.6% and 3.9% of the worldwide populace, separately, all through their lifetime [5]. In 2017 alone, around 163 million individuals (2.1% of the worldwide populace) experienced MDD and 46 million (0.6%) were influenced by BD, representing 32.8 million years lived with a handicap (YLDs) on account of MDD and 9.3 million YLDs for BD [20]. These numbers have been consistently expanding since the 1990s and the two conditions are as of now among the 20 driving reasons for handicap around the world, with MDD positioned second and BD 17th [10]. In England, the direct financial weight of overseeing the state of mind issues, enveloping medical care, casual consideration, and equity framework administrations, is assessed at £1.68 billion yearly for dependency and £1.64 billion for bipolar range problems, while backhanded expenses related to lost work usefulness add up to £5.82 billion and £3.57 billion, separately, and are required to grow [6].

Even though bipolar disorder is a chronic illness, following a treatment plan may help you regulate your mood swings and other symptoms. Medication and psychological counseling are utilized to treat bipolar illness (psychotherapy) in most cases.

Sometimes late and incorrect treatment of the MDD and BD, creates a great trouble for the patient. Although the occurrence of manic (BD I) or hypomanic (BD II) episodes distinguishes BD from MDD, these periods are frequently overlooked because individuals are more likely to seek medical care during a depressed episode [4]. As a result, because the depressed phases in bipolar disorder are essentially comparable to those in MDD, bipolar disorder is commonly misdiagnosed as MDD, even when the depressive symptoms are preceded by a manic/hypomanic episode. Indeed, 37% of people with bipolar disorder (BD) who come in after their first manic/hypomanic episode are misdiagnosed as having MDD. [3].

1.3 Machine Learning

Machine learning has several applications nowadays. One of the most well-known instances of machine learning in operation is the recommendation engine in the

Facebook news feed.

Facebook employs machine learning to personalize each user's feed based on their interests and preferences. If a member often pauses to read articles from a certain group, the recommendation engine may begin to prioritize them. Behind the scenes, the engine is attempting to reinforce patterns of online behavior identified in the user. If the member's habits change and he or she stops reading messages from that specific group in the future weeks, the news feed will be changed.

1.3.1 Definition

Machine learning (ML), an artificial intelligence (AI) technology, enables software programs to grow increasingly accurate at predicting events without needing to be programmed to do so. Machine learning algorithms utilize previous data as input to predict future values. Machine learning (ML), an artificial intelligence (AI) technology, allows software programs to increase their capacity to anticipate events without needing to be programmed to do so. Machine learning algorithms use previous data as input to anticipate new output values.

Machine learning is commonly used in the development of recommendation engines. Other common applications of machine learning include fraud detection, spam filtering, malware threat identification, business process automation (BPA), and predictive maintenance.

1.3.2 Supervised Learning

The data scientist must train the algorithm with both labeled inputs and intended outputs in supervised machine learning. The following tasks benefit from supervised learning algorithms:

- **Binary classification:** It categorizes the data into two groups.
- **Multi-class classification:** It selects answers from more than two categories.
- **Regression modeling:** It is used to obtain continuous values.
- **Ensembling:** It is used to make an accurate prediction by merging the predictions of many ML models.

1.3.3 Unsupervised Learning

Unsupervised machine learning methods do not require data labels. Their task is to hunt for patterns in unlabeled data so that it may be divided into digestible bits for subsequent analysis. Unsupervised algorithms, such as neural networks, are examples. Unsupervised learning techniques are preferable for the following tasks:

- **Clustering:** Using similarity to divide the dataset into categories. It is used to divide a dataset into groups based on similarities.
- **Anomaly detection:** It is used to identify any out-of-the-ordinary data points in a data set.

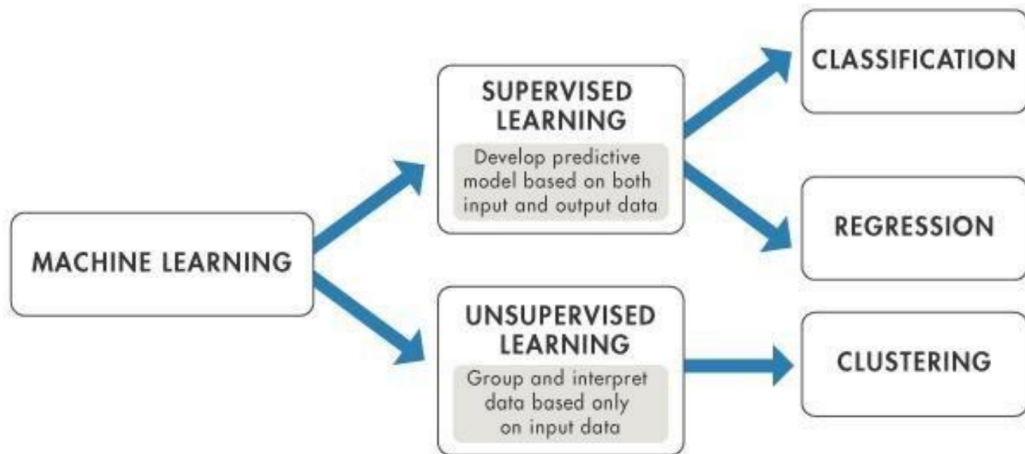


Figure 1.1: Supervised and Unsupervised Learning

- **Association mining:** It is used to detect groups of elements in a data collection that appear often together.
- **Dimensionality reduction:** It is employed in order to reduce the number of variables in a data set.

1.4 Detection Of BPD Through Machine Learning

When bipolar disorder was first discovered in the 1800s, it was extremely difficult to diagnose in patients. There are four main types of these disorders. Cyclothymic, as well as the other. In clinical settings, the most difficult to detect is BP 2. It takes 5-10 years for most people experiencing a depressive episode to be diagnosed with bipolar disorder, and only 20% are diagnosed within the first year of seeking therapy for the disease [8]. Distinguishing unipolar depression from bipolar disorder I or II can be challenging, especially in individuals who come during a depressive episode and do not have a clear history of either mania or hypomania. Patients with bipolar disorder type II, in particular, are more usually than not misdiagnosed with unipolar depression, a disease in which a manic episode does not exist. For example, giving the inappropriate drugs and therapies to treat bipolar illness type I or type II, such as antidepressants instead of mood stabilizers, might result in mania and expensive treatment expenditures for the patient [1]. New self-administered and clinician-administered rating scales have made distinguishing between unipolar and bipolar illness easier. It is, however, a time-consuming and error-prone operation. Our ML system aims to properly detect bipolar disease in its early stages in huge populations, avoiding the long-term negative repercussions of misdiagnosis or overlooking.

Patients are asked questions to gather information about their mental health issues. Fortunately, digital platforms can now play an important and crucial role in gathering this type of data, which is both readily available as well as inexpensive. Using machine learning algorithms on these data, we can also make predictions about the

patient's future health and start treatment right away. Algorithms based on Machine Learning can now be used to diagnose a variety of mental health issues such as autism. So, we'll propose a model that can predict a person's mental state and provide a primary treatment for those who are predicted to have bipolar disorder.

Chapter 2

Related Work

If left untreated, bipolar illness can be lethal. It can worsen the condition's severity and may lead to suicide, since persons with the disorder have a high suicide rate. When addressed, the symptoms can be managed and patients can live a more stable existence. Keeping this in mind, this research will emphasize the fact that can identify the levels of BD since if the diagnosis is incorrect, the therapy will also be incorrect, which is highly destructive to the patient's health. For a long time, the healthcare sector has adapted to and profited immensely from technology innovations. Machine learning, a subset of artificial intelligence, now plays an important role in many health-related fields, including the refinement of existing therapeutic techniques, dealing with patient information and records, and treating chronic illnesses. "Just as machines made human muscles a thousand times stronger, machines will make the human brain a thousand times more powerful," computer scientist Sebastian Thrun told the New Yorker in a recent story headlined "A.I. Versus M.D." [11] Different machine learning approaches, such as disease forecasting, infection categorization, and therapeutic image recognition processes, have been widely used in pharmaceuticals.

Using technology such as sensors incorporated into cellphones, several studies are attempting to identify a patient's mood and detect changes in people with Bipolar Disorder. [9] Grunerbl, A., Muaremi, A. and colleagues proposed a smartphone-based system capable of detecting state changes in bipolar illness patients and distinguishing depressive and manic states. The success of bipolar disorder medication is important to its administration at the start of a patient's shift into a new state. This suggested technology delivers therapeutic effects by automatically identifying state transitions.

[22] Other study by Li,H., Cui,L., and others has employed a support vector machine (SVM) model using a mix of structural and functional MRI to accurately identify individuals with BPD. In this study, an SVM model containing VBM and ReHo measurements in gray matter volumes was built to distinguish patients with BPD from healthy controls (HC). They assessed the model's classification skills and discovered brain regions important for distinguishing between BPD and HCs. This is the first research to use an SVM classifier based on ReHo and VBM studies to discriminate between patients with BPD and HCs. The combination of structural and functional MRI data may be useful in the development of SVM classifiers for reliable identification of BPD. [15] Belizario, Salvini, and colleagues employed machine learning (ML) algorithms to properly assess a patient's Predominant polarity (PP) without tak-

ing into account the number or polarity of previous episodes, as well as investigate correlations between PP and demographic and clinical factors . Demographic and clinical characteristics from 148 BD patients were obtained in this study utilizing a tailored questionnaire and the SCID-CV. The Random-Forest method was used to categorize patients into depressed or manic predominate polarity and to determine which characteristics were connected with the specifier. According to the findings, the ML method may be useful in establishing a patient's predominant polarity (PP), which is important in predicting bipolar illness.

[14]Another study used signature-based machine learning models to identify bipolar illness by Perez Arribas, I., Goodwin, G. M., and others. This model is used to reanalyse data from a clinical research. The model is specifically intended to categorize participants' diagnoses based on their changing emotions. This algorithm may also forecast their mood the next day. In this study, 130 samples with bipolar disorder (48 samples) or borderline personality disorder (31 samples) and 51 healthy people used a special smartphone app to track their moods on a daily basis for up to a year . Participants used a 7-point Likert scale to score their mood on a daily basis across six distinct areas (anxiety, exhilaration, sorrow, anger, impatience, and energy). They created input-output pairs with the shortened signature of order n to predict a participant's mood. They obtain their model by applying random forest regression to these pairings of inputs and outputs. Healthy volunteers had the best mood prediction accuracy of about 89–98% , followed by bipolar illness at approximately 82–90% and borderline personality disorders around 70–78%. In terms of diagnostic categorization and future mood prediction, the signature technique is a successful approach to mood data inspection.

The aim of this reasearch was to see how different data pre-processing procedures affected the performance of ML models used to differentiate neurological patients who had fallen from those who had not for future fall risk assessment. A random forest (RF) classifier trained on path signature-prepossessed data achieved optimum classification accuracy of 98 percent, with 99 percent sensitivity and 98 percent specificity. Still, It's difficult to extract therapeutically valuable information from the complicated time series data offered by these technologies, and the current implications for patient care are unknown . [17]The goal of this work was to create a new Bipolar Diagnosis Checklist in Chinese (BDCC) by shortening the Affective Disorder Evaluation scale (ADE) utilizing machine learning and an examination of registered Chinese multisite cohort data. In this study, five standard machine learning methods were used: a random forest approach, support vector regression (SVR), the least absolute shrinkage and selection operator (LASSO), linear discriminant analysis (LDA), and logistic regression. A case control study involving 360 bipolar disorder (BPD) patients, 255 major depressive disorder (MDD) patients, and 228 healthy (no psychiatric diagnosis) controls (HCs) was carried out across nine Chinese health facilities participating in the Comprehensive Assessment and Follow-up Descriptive Study on Bipolar Disorder (CAFÉ-BD). Current clinical state (11 questions), lifetime clinical trials (5 questions covering abrupt onset and rage episodes of prior depression, dysthymia, age at first use of antidepressant medicine, and lifetime euphoria), and past psychiatric history were all questioned (1 questionnaire on suicide attempts). The data was then analyzed using five machine learning algorithms, and the best performance of each machine learning technique was compared. The random forest algorithm outperformed all others.

We discovered some further studies in which several machine learning methods were used to improve accuracy. Mar'a Victorias' research [13] examined the accuracy of Decision Tree, Random Forest, SVM, and Logistic Regression on various data sets. Logistic Regression outperformed the others in terms of average accuracy.

[25] This research, published by Tomasik, J., Han, S. Y. S., Barton-Owen, G., and others, primarily focused on persons who had been diagnosed with MDD during the previous 5 years. The participants ranged in age from 18 to 45 years old and were from the United Kingdom. Participants had to be at least somewhat depressed, not pregnant or nursing, and not suicidal. Approximately 635 questions were asked of participants via social media or website, divided into six modules: (1) demographic information, (2) manic and hypomanic symptoms, (3) depressive symptoms, (4) personality characteristics, (5) psychiatric history, and (6) other psychiatric diseases. The greatest number of questions that could be asked of a person was 382, with an average of 284. These questions were made with the help of experienced psychiatrists and also follow the existing diagnostic interviews as well.

Participants were requested to provide a blood sample following the interview. employing a evidenced targeted proteomic technique, these blood samples were examined for medical specialty biomark levels. Participants who answered the queries and provided blood samples were invited to require participate within the World Health Organization's interview. They were requested to participate over the phone. This analysis enclosed three hundred patients who had no previous identification of a mental condition. In addition, forty persons with emotional disturbance were included during this study to validate the methodology. standard question information were remodeled to rankings, while categorical information were born-again to dummy variables. Duplicate, bijective, or constant feature values were eliminated. Peptides that weren't discovered within the blood were additionally deleted from the analysis (N=9). Before analysis, biomark level measurements were log2-transformed. the foremost recent range of highlights analyzed was 1151, including 957 things from an internet emotional well-being survey and 194 super-molecule amide estimates. The variable quantity was the CIDI diagnosis. The diagnostic algorithmic program was designed mistreatment EGB, a choice tree-based machine learning approach. They also used the Nested cross-validation (CV) technique. the suitable classification cut-off with balanced sensitivity and specificity resolve mistreatment Youden' J statistics. These models achieved a sixty three percent improvement over the baseline accuracy of 0.60. whereas we tend to foreseen the AUROC edge for clinical significance at 0.8019, a score of quite 0.90 is regarded 'excellent' or 'practically optimal' for deciding psychological well-being issues. The residual error between the computation and therefore the CIDI results is expected given the general symptomatic vulnerability as well as mental illnesses, during which even the 'gold standard' assessments deviate in a very few situations.

This study includes a range of benefits compared to previous studies reaching to distinguish Bipolar Disorder however there are some disadvantages as well. The researchers needed to make a diagnostic rule supported a web psychological state form and blood bio-marker knowledge to find Bipolar Disorder patients among freshly diagnosed MDD patients during this study. This step made us interested to create a dataset by an online questionnaire and predict BPD. But initially, a blood biomarker can lessen the interest of people to approach. On the other hand, for the research purpose average of 284 questions were asked to an individual, which can be demoti-

vating too. Besides, The examined population may be unbalanced and not realistic due to the use of the internet to recruit participants and to achieve specified research recruitment targets. Furthermore, a person who does not experience manic or hypomanic episodes may be missed by this approach. So we thought to improve these parts in our study.

[19] In this paper written by Passos, I. C., Ballester, P. L., Barros, R. C., and others, the role of the ISBD (Big Data Task Force of the International Society for Bipolar Disorder) is to uncover the role played by ML techniques in projecting beneficial outcomes in the diagnosis and prevention of BD as well as providing solutions for the treatment of individuals. Here they have 1) outlined the challenges that needed to be considered in machine learning-based investigations and described big data and machine learning approaches; 2) Conducted and overhauled a thorough review of published works in BD using machine learning and big data to show where the field is now ; and 3) recognized the barriers to adopting these approaches in BD, and strategies to overcome them have been proposed. .

The review criteria were research involving patients under the age of 18 and those that used a machine learning approach. Two of the researchers gathered data on the year the study was published, the data used in the machine learning model (– for example, characteristics in population demography, biomarkers, neuro-imagery ,clinical features etc etc), ML algorithm as well as type of diagnoses assessed, size of sample and the usual statistical measures of model performance such as accuracy, precision, auc (area under curve), true and false positive/negatives.

This study found that, because to the high morbidity and death rates among BD patients, it is advised that techniques for calculating is prediction, tailored therapy, and prognosis be improved at a faster rate. It’s also worth mentioning that under universal health-care systems, each doctors/researchers has access to a plethora of untapped and still available patient-specific data that may be used to develop diagnostic tools. Individual information is presently underutilized to its full potential, and the information value of episode sequence and length is undeveloped. So, while machine learning and Big Data analytics are still relatively new, they have removed the limits that were behind the analysis of big and complicated datasets that would have been incredibly time consuming to deal with using traditional approaches. Machine learning and big data analytics technology enable us to study the ”real patient” with all of its inherent complexities . There’s an overview of how ML and big data analytics can help the field by providing individual specific prediction models. The research also discovered that the most intriguing examples were those carried out by smaller studies.

[26] Nisha Agnihotri wrote this review study to examine the ML approach for BD and the clinical process related to it, data from healthy and unwell persons are examined from a specific survey. The purpose of this study is to examine and summarize the research on ML techniques for predicting bipolar disease . The purpose of this literary review is to find ways to lessen the occurrence and incidence of anxiety disorders by using effective early prediction. . This significantly minimizes hospitalization, improves their quality of life, and significantly lowers their healthcare costs.

Another study [16] we found was done by Ranjana Jadhav and his team where the goal was helping the basic investigation to detect whether a person is suffering from bipolar disorder by using a mood disorder questionnaire (MDQ). They considered

MDQ based on symptoms of mania and Based on their answers the team analyzed and using a certain methodology predicted whether the individual is suffering from bipolar disorder and whether he should go for necessary treatment. The experiment interface differed from the questionnaire in the last question where the original questionnaire provided 4 fields namely Low, Medium, High, and Severe. The team clubbed Low + Medium as “Moderate” and High + Severe as “Severe”. They said that this did not affect the credibility of the data as the original questionnaire produced the same effect for the clubbed types. They also added that this modification also helped in collecting clean data as it is internally stored as 0 or 1 for “No” or “Yes” in the first fourteen questions and 0 or 1 for “Moderate” and “Severe” in the fifteenth question respectively. The 983-dataset was partitioned 60:40, with 60% of the data being used to train the model and 40% being used to test the model. The model was created in Python using the Sklearn Package’s Decision Tree Classifier. The CART algorithm is used to implement the Decision Tree in the Decision Tree Classifier class

Chapter 3

Data Collection And Processing

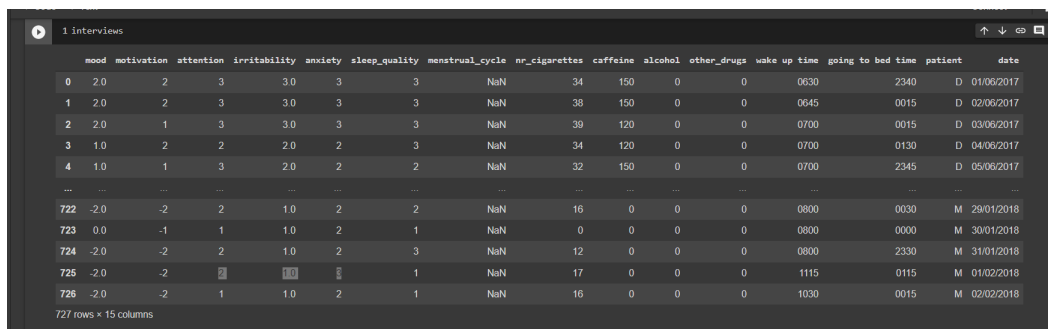
In this paper, we have worked with some datasets which deal with bipolar disorder.

3.1 Data Input

As it is a psychiatrist problem, the records for bipolar ailment isn't a lot to be had. The records used for this task is anonymized affected person records amassed via way of means of psychiatrists at Clínica Nuestra Señora de la. Paz in Madrid. All the records become to be had in an Excel record with special sheets. The records has been amassed throughout scientific appointments with 4 special sufferers which have bipolar ailment, however the intention for the destiny is that it's miles each recorded via way of means of the psychiatrists in appointments and with the assist of cell applications [13]. This way, the patients can actively participate in their own diagnosis. The raw dataset can be downloaded from:

<https://github.com/AxelJunes/BDCP/tree/master/data>

For this project the informationset used is principally an interview dataset that has been gathered by the medical specialist in their session with the patient. It contained each physical and psychological parts. Psychological elements represent a lot of subjective data, comparable to anxiety, irritability, or issues concentrating, whereas physical elements embrace more objective data, such as the amount of cigarettes the patient smoke-cured during a day. additionally include the data of once the person wakes up from the bed and goes to sleep. From this we will calculate the time when he's active in his or her work. Figure 3.1 shows the interview dataset which we used for this project.



	mood	motivation	attention	irritability	anxiety	sleep_quality	menstrual_cycle	nr_cigarettes	caffeine	alcohol	other_drugs	wake up time	going to bed time	patient	date
0	2.0	2	3	3.0	3	3	NaN	34	150	0	0	0630	2340	D	01/09/2017
1	2.0	2	3	3.0	3	3	NaN	38	150	0	0	0645	0015	D	02/06/2017
2	2.0	1	3	3.0	3	3	NaN	39	120	0	0	0700	0015	D	03/06/2017
3	1.0	2	2	2.0	2	3	NaN	34	120	0	0	0700	0130	D	04/06/2017
4	1.0	1	3	2.0	2	2	NaN	32	150	0	0	0700	2345	D	05/06/2017
...
722	-2.0	-2	2	1.0	2	2	NaN	16	0	0	0	0800	0030	M	29/01/2018
723	0.0	-1	1	1.0	2	1	NaN	0	0	0	0	0800	0000	M	30/01/2018
724	-2.0	-2	2	1.0	2	3	NaN	12	0	0	0	0800	2330	M	31/01/2018
725	-2.0	-2	2	1.0	2	1	NaN	17	0	0	0	1115	0115	M	01/02/2018
726	-2.0	-2	1	1.0	2	1	NaN	16	0	0	0	1030	0015	M	02/02/2018

Figure 3.1: Interview dataset

We also used episode data which includes different episode periods of the patient such as whether he or she is in a depressed mood or in the maniac episode. Figure 3.2 shows the episodes dataset.

	patient	start	end	episode
0	D	01/07/2017	24/07/2017	D
1	D	15/08/2017	11/09/2917	D
2	G	24/07/2017	07/08/2017	D
3	G	04/09/2017	01/11/2017	M
5	M	07/06/2017	01/07/2017	M
6	M	14/07/2017	30/07/2017	D
7	M	25/09/2017	10/10/2017	D

Figure 3.2: Episodes dataset

For this project, we used these two datasets. We combined these two datasets and applied the algorithm in it.

3.2 Data pre-processing

Pre-processing of the data is major part of data-classification. It is an important process of transforming raw data into an understandable format to an algorithm. Pre-processing helps to clean, format and organize the raw data which helps to avoid misleading predictions by the model.

Data is often taken from multiple sources which are not too reliable and often has many errors due to the flaw in the data collection process. First step of data pre-processing is to clean the data. Data cleansing is that the method to get rid of incorrect knowledge, incomplete data and inaccurate data from the datasets. This step also handles missing values of the dataset. Missing values can be handled by deleting the rows or columns having null values.

Encoding categorical features is an important step in data pre-processing, Machine learning algorithm only works with numerical data. It is necessary to convert categorical value to numerical value in order to fit and evaluate model.

As we are using Spanish dataset here, we have translated dataset to English in order to work efficiently with the dataset. After doing necessary processing with episodes and interview dataset, our final dataset will look like this:

Pre-processing of Episodes Dataset:

Here we can drop patient with code 'O' from the data set as there are no clear episodes of Mania or Depression. We renamed the columns in order to work with the dataset in a better way. The updated dataset is been shown in figure 3.4.

	patient	start	end	episode
0	D	01/07/2017	24/07/2017	DEPRESIÓN
1	D	15/08/2017	11/09/2917	DEPRESIÓN
2	G	24/07/2017	07/08/2017	DEPRESIÓN
3	G	04/09/2017	01/11/2017	MANIA
5	M	07/06/2017	01/07/2017	MANIA
6	M	14/07/2017	30/07/2017	DEPRESIÓN
7	M	25/09/2017	10/10/2017	DEPRESIÓN

Figure 3.3: Episodes dataset before Pre-processing

	patient	start	end	episode
0	D	01/07/2017	24/07/2017	D
1	D	15/08/2017	11/09/2917	D
2	G	24/07/2017	07/08/2017	D
3	G	04/09/2017	01/11/2017	M
5	M	07/06/2017	01/07/2017	M
6	M	14/07/2017	30/07/2017	D
7	M	25/09/2017	10/10/2017	D

Figure 3.4: Episodes dataset after Pre-processing

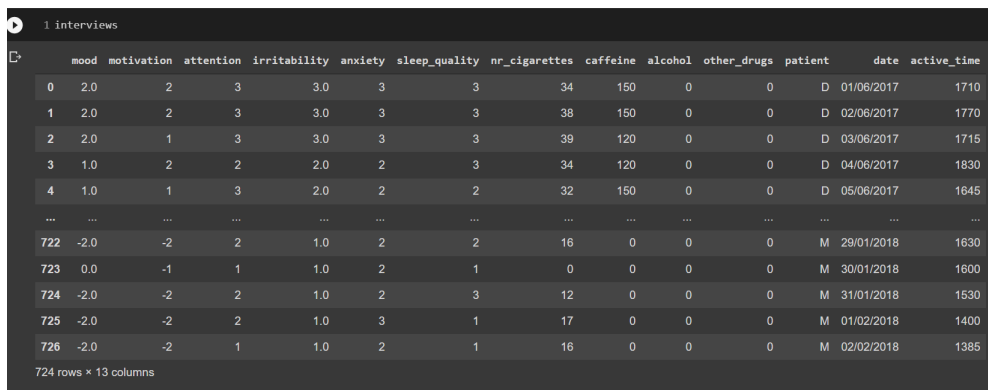
Pre-processing of Interview Dataset:

	Estado de ánimo	Motivación	Problemas de concentración y atención	Irritabilidad	Ansiedad	Calidad del sueño	Ciclo menstrual	Número de cigarrillos	Cafeína	Alcohol	Otras drogas	Hora de despertar	Hora a la que te dormiste	Código	Fecha
0	2.0	2	3	3.0	3	3	NaN	34	150	No	No	06:30	23:40	D	01/06/2017
1	2.0	2	3	3.0	3	3	NaN	38	150	NO	No	06:45	00:15	D	02/06/2017
2	2.0	1	3	3.0	3	3	NaN	39	120	NO	No	07:00	00:15	D	03/06/2017
3	1.0	2	2	2.0	2	3	NaN	34	120	No	No	07:00	01:30	D	04/06/2017
4	1.0	1	3	2.0	2	2	NaN	32	150	No	No	07:00	23:45	D	05/06/2017
722	-2.0	-2	2	1.0	2	2	NaN	16	0	No	No	08:00	00:30	M	29/01/2018
723	0.0	-1	1	1.0	2	1	NaN	0	0	No	No	08:00	00:00	M	30/01/2018
724	-2.0	-2	2	1.0	2	3	NaN	12	0	No	No	08:00	23:30	M	31/01/2018
725	-2.0	-2	2	1.0	3	1	NaN	17	0	No	No	11:15	01:15	M	01/02/2018
726	-2.0	-2	1	1.0	2	1	NaN	16	0	No	No	10:30	00:15	M	02/02/2018

Figure 3.5: Interview dataset before Pre-processing

We can now map the Yes/No features, giving them a numerical value (No=0, Yes=1). As the napping time overlaps exceptional days, we will calculate what number of hours the affected person has been energetic in place of the hours of sleep. If the affected person has long gone to mattress after midnight, we ought to upload 24 hours in order to get the appropriate range of energetic hours. After calculating the quantity of time, the affected person has been energetic, we will drop the 'awaken time' and 'going to mattress time' columns. Later we can take a look at if there are any null values with inside the information set. As there are null values with inside the dataset, we will take a look at which columns have null values and delete the ones columns to address lacking values.

Finally, the clean interview dataset will look like this in Figure 3.6



	mood	motivation	attention	irritability	anxiety	sleep_quality	nr_cigarettes	caffeine	alcohol	other_drugs	patient	date	active_time
0	2.0	2	3	3.0	3	3	34	150	0	0	D	01/06/2017	1710
1	2.0	2	3	3.0	3	3	38	150	0	0	D	02/06/2017	1770
2	2.0	1	3	3.0	3	3	39	120	0	0	D	03/06/2017	1715
3	1.0	2	2	2.0	2	3	34	120	0	0	D	04/06/2017	1830
4	1.0	1	3	2.0	2	2	32	150	0	0	D	05/06/2017	1645
...
722	-2.0	-2	2	1.0	2	2	16	0	0	0	M	29/01/2018	1630
723	0.0	-1	1	1.0	2	1	0	0	0	0	M	30/01/2018	1600
724	-2.0	-2	2	1.0	2	3	12	0	0	0	M	31/01/2018	1530
725	-2.0	-2	2	1.0	3	1	17	0	0	0	M	01/02/2018	1400
726	-2.0	-2	1	1.0	2	1	16	0	0	0	M	02/02/2018	1385

Figure 3.6: Interview dataset after Pre-processing

To implement the algorithm, at first we need to determine the characteristics of the data set which we used for our study. Because of that we plot so many graphs to the ‘Interview’ dataset which will help us to understand the characteristics of the data.

Before plotting the graphs of the interview dataset we removed the ‘patient’ column which contains the mood or stage of the patient and we also removed the column ‘date’ which contains the time period of that stage. We removed those columns because we can not plot them in our graphs. As we all know that we cannot plot singular matrices, we need to know which variable has multiple values. Because of that we create a function which will tell us which variable is best for plot.

```
def get_plottable_columns(df):
    for column in df:
        print (column, ": ")
        n_values = len(df[column].unique())
        if n_values > 1:
            print ("Yes, ", n_values, " values" )
        else:
            print ("No, 1 value")
```

Figure 3.7: Method for Best Function

Now we plot different types of graphs. We plot:

HeatMap: HeatMap is a way to represent various data. It uses a color coding system which is a representation of the different types of value. This kind of representation is used for different purposes but mainly used to demonstrate the user behavior of any website or webpage. This kind of plot is mainly used to determine the relationship between two variables which is plotted on both axes. From this heatmap we can determine whether any pattern exists between the variable.

We see from figure 3.8 the relationship between mood and motivation is high. This is so much understandable because unmotivation is a sign of depression.

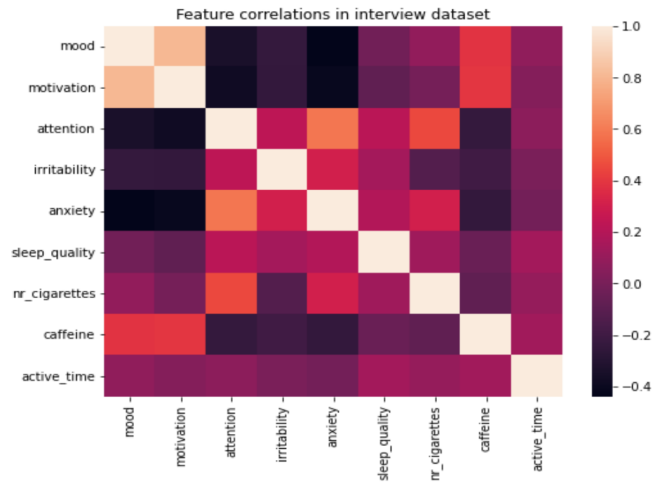


Figure 3.8: HeatMap of Interview Dataset

Scatterplot: It is a data display method which helps us to understand the relation between two numerical values. To realize if two variables mean something else when they are together, this plot helps us a lot. Not only this, it also helps us to determine if there is any kind of potential relationship between them. It uses dots to represent values.

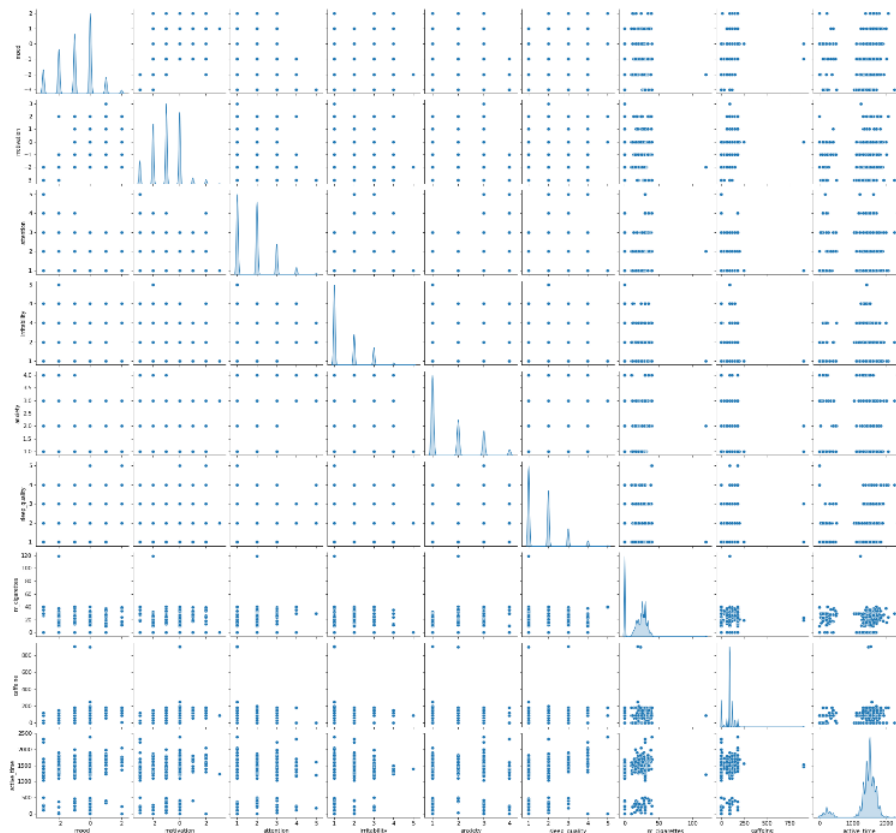


Figure 3.9: Scatterplot of Interview Dataset

From figure 3.9 we can see the relationship between so many variables. We can see the relationship between caffeine and motivation; anxiety and sleep quality and so on.

MarginalPlot: This is a plot which could be a scatterplot that has histograms, box plots, or dot plots within the margins of the x- and y-axes. It permits finding out the connection between two numeric variables. the bottom plot visualizes the correlation between the x and y axes variables. it's typically a scatterplot or a density plot.[5].

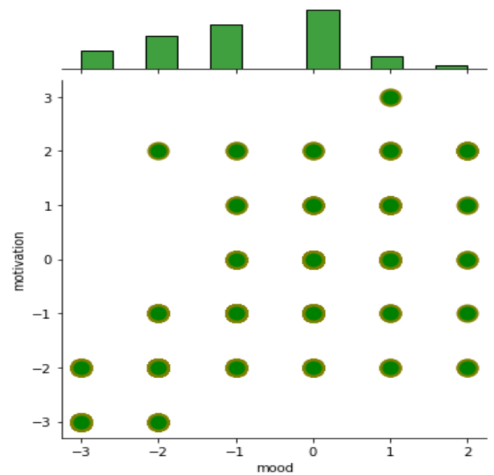


Figure 3.10: Marginal Plot of Interview Dataset

We also plot the marginal plot of motivation and mood in figure 3.10, which will help us to understand the relationship between mood and motivation better. For further understanding we draw the 2-d kernel density plot of the two variables of mood and motivation.

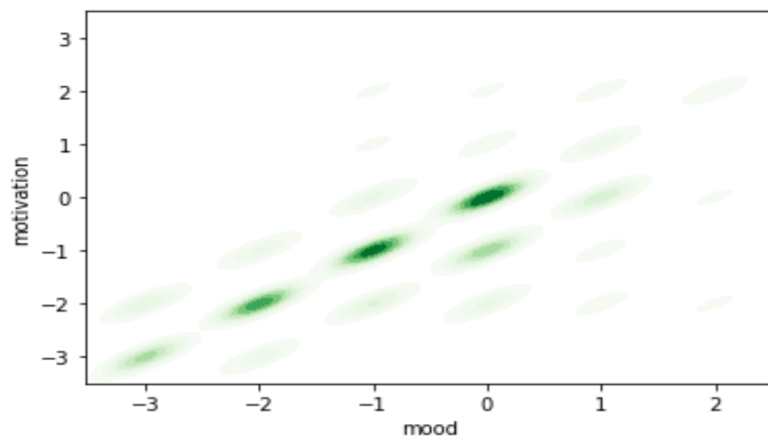


Figure 3.11: Density Plot of Interview Dataset

It helps us to understand the way of the distribution of the data and where the data is mostly concentrated. The data is concentrated in the point where the both

variables have the same value. In the plot where we see darker, it means that both the variables are 0 in that place.

3.3 Combination of Dataset

Now we will combine the dataset 'Interview' with the dataset 'Episode' so that the algorithms which we used for this study can get enough data to process. Our main motive is to predict the state of the patient. We replace the entries or rows which do not contain any record with the value of neutral. Because of that, we get three states of the patient. They are Depression (D), Mania (M), Neutral (N). to combine the two datasets we also create a method which helps us to combine the two datasets (figure 3.12).

```
[ ] 1 def checkEpisode(date, patient):
2     episode = 'N'
3     ep = episodes.loc[episodes['patient'] == patient]
4     for index, row in ep.iterrows():
5         if date >= row.start and date < row.end:
6             episode = row.episode
7     return episode
```

Figure 3.12: Method For Combine Datasets

We also draw some plots of this combination data to understand the data, to visualize how the values of the various variables are distributed. For plotting we did the same thing here meaning we drop the 'Patient' and 'Date' column for the same reason.

We plot a scatter plot to see the distribution of the data in different episodes (figure 3.13).

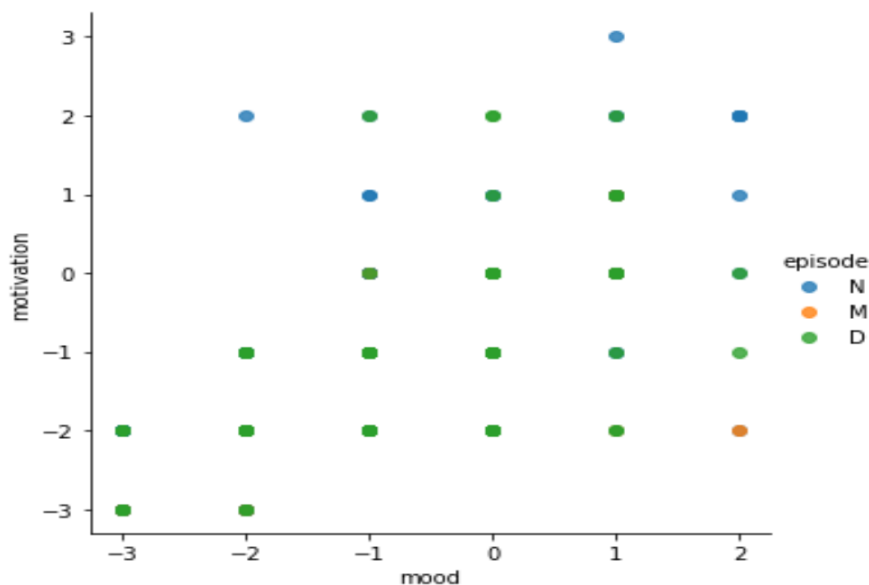


Figure 3.13: Scatter Plot of Combination Datasets

But from this plot it is difficult to understand the patterns of the different episodes. For further understanding we plot Barplot (figure 3.14).

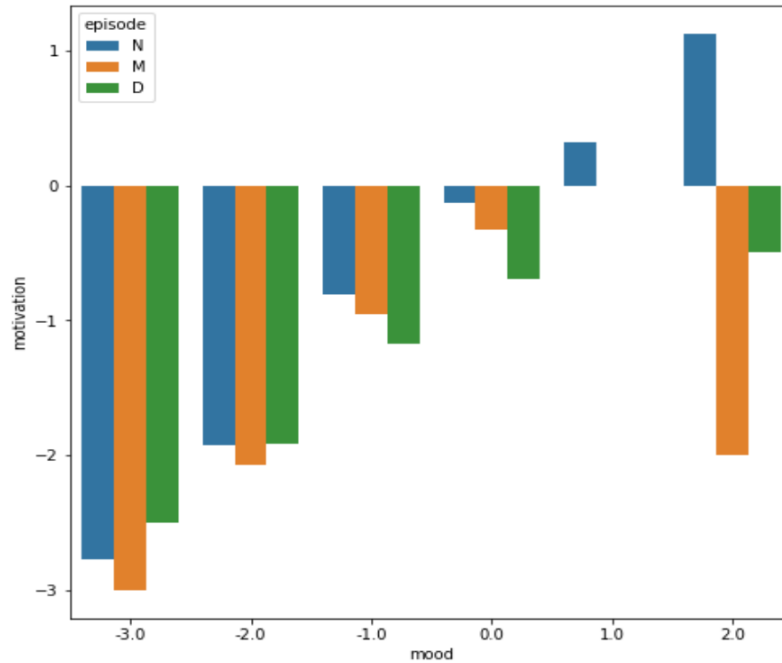


Figure 3.14: BarPlot of Combined Data

From this bar plot we can observe so many different things. From this plot we can see that, all the combination of the two variables which is mood and motivation exists in the negative side. Only a person who is motivated in a neutral mood is positive.

We also make a barplot for sleep quality and anxiety to see if these two features can define the state of the patient (figure 3.15).

From the graph see that the data representation is similar. Because of that, we cannot use it to be the representative feature. The reason behind is that the data is not enough.

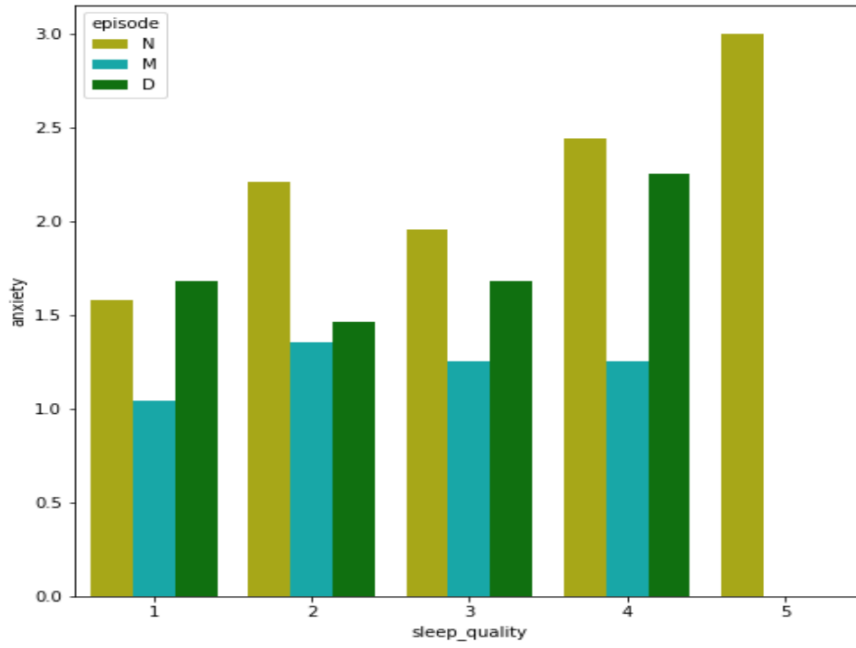


Figure 3.15: BarPlot of Sleep and Anxiety

We also plot a point plot of the 'sleep quality' , 'anxiety' and 'episode' variable (figure 3.16).

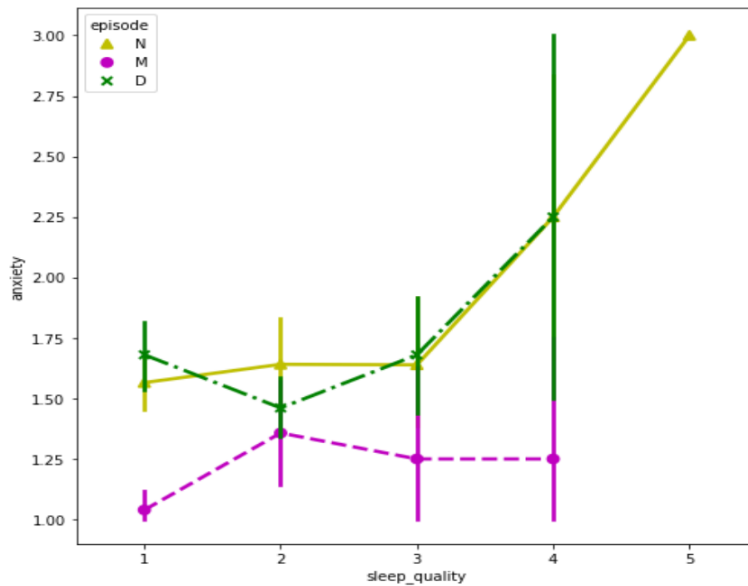


Figure 3.16: Point Plot of Combine Data

Chapter 4

Research Methodology

In the past, we had to do everything physically especially when it comes to the matter of treatment. The introduction of advanced technology in this area was a relief as a lot of time can be saved and the physicians and related workers can be focused on solutions rather than finding problems. Among many technologies, Machine Learning is one of the best methods to use for predicting and we are using this in our research.

One of the major subsets of AI is Machine Learning which permits machines to become extremely accurate an identification and detection of object,even better than humans in alot of cases. Some algorithms are applied to machine learning programs because they are versatile, flexible and user friendly. The algorithms can process a large amount of data as well as can detect patterns. Machine Learning is one of top shelf tools for statistical analysis and advanced technology as it computes very quickly so that we do not have to spend a lot of time to solve any problem. Additionally, machine learning is capable of increasing the speed when datasets are being processed and analyzed. Moreover, it can tune the parameters of the models which will fit the data. If we want to do this work physically, it will take a lot of time.

4.1 Algorithms

There are many predictive modeling algorithms such as Linear regression, logistics regression, naive bayes, random forest, extreme gradient boost, LightGBM and many more. We have chosen XGBoost, LightGBM and CATboost to solve our problems related to bipolar disorder. These algorithms are fast and work with big datasets with maximum accuracy than other algorithms for prediction. Mental health is a very sensitive issue and it should be handled cautiously. This is the main reason why we want the result to be accurate so that the people who are suffering from this illness can get the proper medication. There is plenty of research that proves why these algorithms are best for prediction. We want maximum accuracy in a short time and both of them fulfill these conditions. The main difference between these is how they work. XGBoost follows the growth of level-wise tree which grows horizontally whereas LightGBM follows the growth of leaf-wise tree which grows vertically. However, CatBoost employs oblivious decision trees, in which the same splitting criterion is applied to the entirety of a tree level. Such trees are well-balanced, less susceptible to overfitting, and expedite testing time prediction greatly.

The research shows that the rate of true positivity is very high in these cases. [24]

4.1.1 XGBoost

XGBoost is a tree based gradient boosting framework. It follows the supervised learning approach to combine the predictions from multiple weaker models for accurately predicting the objective variable. Xgboost aims to minimize the loss function and create an add-on extension of the object function and find the best split to minimize computational complexity.

Methodology

XGBoost controls over-fitting with a more regularised model formulation. It is based on decision trees which are graph based models that assess input under various if conditions. The following "if" condition and eventual prediction are influenced by whether the "if" condition is satisfied. To develop a stronger model, XGBoost gradually adds more and more "if" conditions to the decision tree.

It's simple to create a model with XGBoost. However, applying XGBoost to improve the model is tough. XGBoost has a wide range of hyperparameters. Hyperparameters are specific variables or weights that determine the learning process of an algorithm. By fine-tuning XGBoost's hyperparameters, we can get the most out of it. In the XGBoost documentation, there are roughly 35 distinct hyperparameters stated. However, they are not all equally significant; some have a greater impact than others. The Major parameters for Xgboost have been classified into 3 distinct groups:

General parameters: They serve as a guide for the overall operation.

- booster [default = gbtrees]: It divides the data into 2 categories.
 - gbtrees stands for "tree-based models."
 - gblinear is a program that creates linear models.
- silent [default = 0]: Silent mode switched to 1, which means no running messages are displayed. Generally advised to leave it at one for beginners or those who want to get a better understanding of the model.
- nthread [default = maximum number of threads available]: recommended to set the value to whatever the maximum number of cores the system has available for fastest computation.

Booster parameters: At each phase, they guide the specific booster (tree/regression).

Tree boosters and linear boosters are the two types of boosters available. Tree boosting is preferred due to it performing better in majority cases.

- colsample_bytree [default=1]: Set value to determine the fraction of columns that should be randomly sampled. Typical ranges are 0.5-1.
- gamma [default=0]: gamma specifies the minimal loss reduction to result in a split usually when the node has a positive value of the loss function . The algorithm becomes more conservative as a result of this. The value should be fine tuned based on the loss function.

- `max_depth` [default=6]: This is the maximum depth of a tree. Higher the depth, the better the model becomes at controlling overfitting by finding more complex relations within a sample. CV is recommended to use for fine tuning this value. Typical ranges are 3-10.
- `min_child_weight` [default=1]: Defines the child's minimum weighted total of all required observations. It's employed to keep over-fitting in check. Higher values make it difficult for a model to understand relations that are highly particular to the sample used to build a tree. It should be fine-tuned using CV because too high values can result in under-fitting.
- `subsample` [default=1]: The same as the GBM `subsample`. This amount specifies the percentage of data that will be randomly sampled for each tree. Lower values are more conservative and help to avoid overfitting, however smaller values are susceptible to underfitting.

Learning Task Parameters: They specify the parameters that will be used to guide the optimization process.

- `objective` [default=reg:linear]: The loss function that needs to be minimized is this one. The most commonly used values are :
 - `binary:logistic` – Gives a predicted probability (not class) for binary classification using logistic regression.
 - `multi:softmax` – Returns the softmax objective's projected class (not probability) for multiclass classification. .
 - `multi:softprob` – similar to softmax, but calculates the probability of each data point falling into each class. .
- `eval_metric` [default based on objective]: This value is used to verify the accuracy of the data. `Rmse` and `error` are the default settings for regression and classification, respectively. :
 - `rmse` – root mean square error
 - `auc`: Area under the curve
 - `mae` – mean absolute error
 - `merror` – Multiclass classification error rate
 - `mlogloss` – Multiclass logloss
- `seed` [default=0]: It can be used to get reproducible outcomes as well as to fine-tune parameters.

Logical Reasoning

If we talk about XGBoost, it is a tool that is high in flexibility and versatility which can work through most diverse situations and is also user friendly. As the software is open-source, it can be accessed easily and it can be used through different interfaces or platforms. Many formulas are used to make it work accurately and without bugs but in our scenario our main goal is to make sure the result is accurate. This step is

ensured by the help of the formula which lets the algorithm highlight on data that are not classified. Below is the XGB Mean Square Error Formula:

$$F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma) \quad (4.1)$$

$$\arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i - \gamma)^2 \quad (4.2)$$

Implementation of XGBoost

XGBoost is developed in our dataset to use decision trees in such a way that the knowledge received from the previous tree is used to create the next tree, implying the trees are sequential and interrelated. We used Random Search optimization to tune the hyperparameters of the XGB classifier and train the model. . Random Search takes a wide range of hyperparameters and iterates over them randomly for a specified number of times.

Here, the XGBoost model was trained and validated on 80% of the dataset and tested on 20% of the dataset. The model is built on a training dataset initially and after training it is used to predict on validation data.

Though XGBoost contains a larger number of hyper parameters, we emphasise on those that have been demonstrated to have significant impact on model performance in prior studies.

Table 4.1 shows some of the hyper-parameters used in this study and their corresponding range values.

Name	Value Ranges for Hyperparameters
max depth	[3, 4, 5, 6, 8, 10, 12]
gamma	[0.0, 0.1, 0.2, 0.3, 0.4]
learning rate	[0.05,0.10,0.15,0.20,0.25]
min child weight	[0.03, 0.04, 0.05, 0.07,0.08]
subsample	[0.3, 0.4, 0.5, 0.7, 0.8, 0.9,1.0]
colsample bytree	[0.3, 0.4, 0.5, 0.7,0.8,0.9,1.0]

Table 4.1: XGBoost hyperparameters and their value range

Name	Optimal values for Hyperparameters
max depth	10
gamma	0.3
learning rate	0.25
min child weight	0.03
subsample	0.8
colsample bytree	0.8

Table 4.2: Optimal hyperparameter values after tuning by Random Search

The result we find after tuning is:

XGBoost Model Accuracy Score: 85.52

Training set accuracy score: 98.27

Random search is the best approach to identify an optimal set of hyperparameters for a machine learning model. The randomized search meta-estimator is an approach for training and evaluating models that uses random attracts from a collection of hyperparameter distributions. The approach finds the foremost eminent version of the model when coaching N completely different versions of the model with different haphazardly elect hyperparameter mixtures .

Initially, we first establish a dictionary containing various parameters to train on. The keys are simply the parameters, and the values are a list of values for the parameters to be trained on . Random search uses this grid of possible hyperparameter values to iterate a different random combination and provide the optimal combination of hyperparameters.

We call `RandomizedSearchCV()` with the following parameters:

- `classifier` : Defines the dictionary of parameters that we need to optimise - `scoring` : scoring attribute.
- `n_iter` : Number of iterations which indicates how many parameters are sampled. It is set to 10 by default.
- `n_jobs` : This defines the number of jobs to be run in parallel; -1 indicates that all processors should be used.
- `cv` : We pass an integer value here, as it represents the number of splits required for cross validation. It is set to five by default.
- `verbose` : To generate messages while the model is being trained.

Result Analysis of XGBoost

These are the confusion matrix of our datasets before tuning and after tuning.

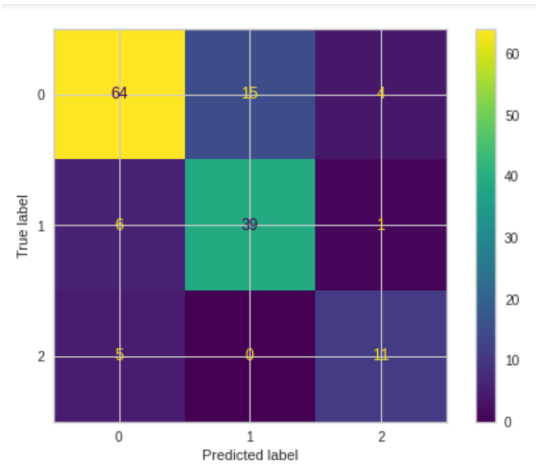


Figure 4.1: Confusion Matrix Before Tuning XGboost

From the confusion matrix (figure 1), we can figure out each measure for each class. Here is a table 4.3 that shows each measure for each class before tuning.

Class	Precision	Recall	F1-score
Normal	0.91	0.91	0.91
Depression	0.88	0.90	0.89
Mania	0.84	0.81	0.82

Table 4.3: Measure for each class before tuning XGboost

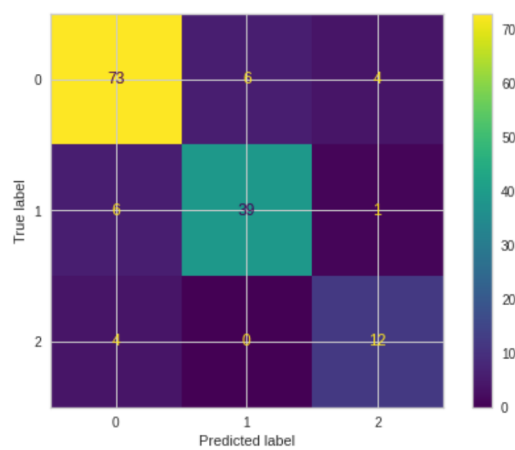


Figure 4.2: Confusion Matrix After Tuning XGBoost

Here is table 4.4 that shows each measure for each class after tuning.

Class	Precision	Recall	F1-score
Normal	0.99	0.98	0.99
Depression	0.98	0.99	0.99
Mania	0.95	0.95	0.95

Table 4.4: Measure for each class after tuning XGboost

Precision is defined as the ratio of successfully expected positive observations to total predicted positive observations. Precision is linked to a low false positive rate. Recall is defined as the proportion of correctly predicted positive observations to all observations in the actual class. The F1 Score is the weighted average of Precision and Recall. As a result, both false positives and false negatives are taken into account in this score. We can see that following parameter adjustment with Random Search, the model has a higher precision score, recall, and f1-score. The accuracy level also increased after tuning the parameter.

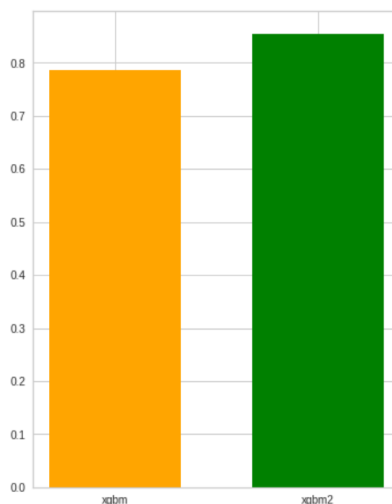


Figure 4.3: Accuracy Compare of XGboost

4.1.2 LGBM

One of the machine learning techniques utilized in our study was LGBM, or light gradient boosting. LGBM is a gradient boosting framework based on a decision tree algorithm that may be used for ranking, classification, and a variety of other machine learning problems. Unlike most other algorithms, lgbm divides the decision tree leaf-by-leaf (best first) rather than level-on-level. This usually results in lower loss over level based trees but can lead to overfitting of the model if the data set is small. This problem is usually overcome by tuning the ‘max depth’ hyper-parameter. Speaking of, LGBM also includes numerous hyperparameters which, although does work fine by default, can be tuned to improve accuracy or process time. The ‘light’ in LGBM is because of the histogram-based techniques that bucket continuous feature data

into discrete bins to speed up the training process and save memory. Histogram models initially have the same time complexity as pre-sort algorithms but once constructed, its complexity decreases to $O(\text{bins})$ instead of $O(\text{data})$ which is normally a smaller set. LGBM has reduced communication cost for distributed learning and reduced memory usage using discrete bins and not having to store additional pre-sort information. LGBM is quite compatible with large datasets.

LGBM has a few optimizations in its distributed learning process that make it compute data faster than traditional algorithms. Since each worker in the LGBM feature parallel holds all of the data, there is no need to communicate with other splits. Workers can find the local best split point and communicate with each other and improve splits until the best split is found. Data Parallel is further optimized by merging histograms of various (non-overlapping) features for separate workers using 'Reduce Scatter'. The workers then find the local best split based on local merged histograms and compare it to the global best split. How DART boosting type to reduce overfitting and find best splits:

In m -th training round, suppose m trees are selected to be dropped.

$$\text{Obj} = \sum_{j=1}^n L(y_j, \hat{y}_j^{m-1} - D_j + \tilde{F}_m) + \Omega(\tilde{F}_m) \quad (4.3)$$

incase of overshoot, below function is used to scale:

$$\hat{y}_j^m = \sum_{i \notin \mathbf{K}} F_i + a \left(\sum_{i \in \mathbf{K}} F_i + b F_m \right) \quad (4.4)$$

From the diagram below (figure 4.4), we can see that decision tree based models are not heavily used in the medical research field and the aim of our research would be to find out why such may be the case and how we can improve upon it.

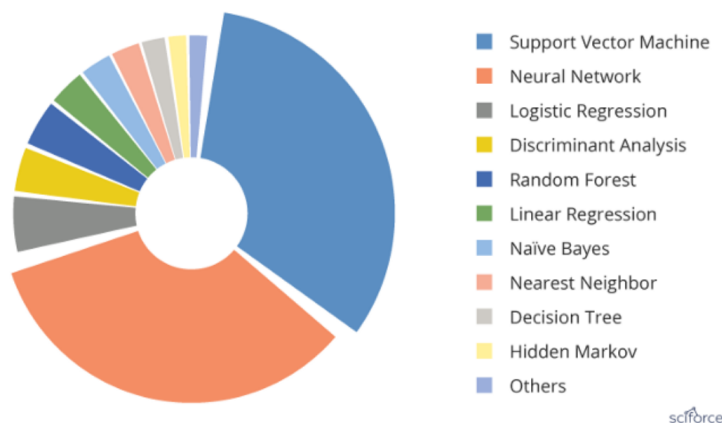


Figure 4.4: Light Gradient Boosting

Methodology

Hyperparameters & Implementation: Despite the fact that the model is called light, because it is prone to overfitting and can readily overfit tiny datasets, this is not a good fit for small datasets. It is usually advisable to use it for datasets with rows in the magnitude of 10,000s. Implementation of LGBM is relatively simple, with only the only difficult bit being the tuning of hyperparameters if necessary.

As mentioned before, LGBM also has various hyperparameters for users to work with. Most of these parameters work fine as default but tuning several, especially for large datasets or for those that understand requirements of each data set, does often yield better results than normal.

Let's start by talking about the parameters used in our research.

- Learning rate: The learning rate's function is to increase the magnitude of changes in the approximation that is updated from each tree's output. The learning rate should be inversely proportional to the model's maximum iterations.
- N_estimators: Also known as number of iterations of the decision tree
- Max_depth: Specifies the depth of the tree to be utilized, as well as data overfitting.
- Num_leaves: Sets the total number of leaves that would appear in the graph
- Max_bin: The maximum number of bins in which the feature values can be bucketed.
- Subsample: Sets the percentage of data that will be used for each iteration, and is commonly used to speed up training and avoid overfitting.
- Boosting_type: select boosting type for LGBM, default is gradient boosting but can also select Random Forest, Gradient-based One-Side Sampling or Dropouts meet Multiple Additive Regression Trees

Other important parameters are 'Objective', which determines the type of classification used, and has numerous options such as binary, regression, multiclass, lambda etc. 'Device' which can be used to select the gpu instead of the cpu to speed up training 'Min_gain_to_split' sets min gain to split leaves from.

Logical Reasoning

LightGBM is a high-performance boosting system based on choice tree calculations that can be used for positioning, classification, and a variety of other machine learning tasks. It splits the tree leaf astute with the best fit because it is based on choice tree calculations, while another boosting calculation splits the tree profundity sharp or level astute instead of leaf wise. When growing on the same leaf in Light GBM, the leaf-wise calculation can reduce more mistakes than the level-wise calculation, resulting in significantly more accuracy than can be achieved by any of the existing boosting calculations of the accurate result on occasion. It can be represented by the following formula:

$$Y = Base_tree(X) - lr * Tree1(X) - lr * Tree2(X) - lr * Tree3(X) \quad (4.5)$$

Implementation of LGB

We will be trying the LGMB algorithm in both default state and with tuned parameters to compare results.. Firstly we did not tune any parameters. We just used the default value of all the parameters. We split the dataset in the test_size =0.2 and got the following results :

LightGBM Model accuracy score: 0.8345

Training-set accuracy score: 0.9637

Training set score: 0.9637

Test set score: 0.8345

exec time: 0.14418400099998507

Now, after tuning parameters via trial and error we have ended up with the following state of parameters.

The below table (table 4.5) shows the value of the parameters which we tuned:

Parameter	Value (Default)
learning_rate	0.5 (0.1)
n_estimators	50 (100)
max_depth	1 (-1)
num_leaves	100 (31)
max_bin	30 (255)
Subsample	0.8 (1.0)
boosting_type	dart (gbdt)

Table 4.5: Parameter for LGBM

After that, the result we find is:

LightGBM Model accuracy score: 0.8414

Training-set accuracy score: 0.9793

Training set score: 0.9793

Test set score: 0.8414

exec time: 0.10369553700002143

Result Analysis of LGB

Figure 4.5 and 4.6 are the confusion matrix of our datasets before and after tuning.

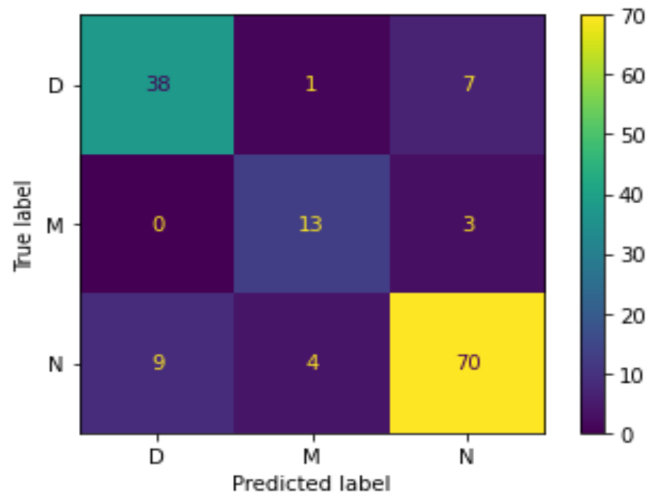


Figure 4.5: Confusion Matrix Before Tuning LGBM

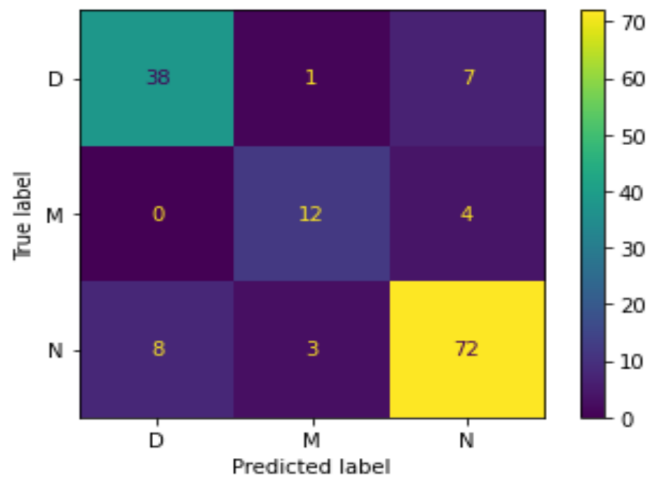


Figure 4.6: Confusion Matrix After Tuning LGBM

Table 4.6 shows each measure for each class after tuning.

Class	Precision	Recall	F1-score
Normal	0.96	0.99	0.97
Depression	0.92	0.89	0.90
Mania	0.97	0.96	0.97

Table 4.6: measure for each class after tuning LGBM

We see that after tuning, the number of correctly identified Mania and Normal has increased.

The accuracy level also increased after tuning the parameters.

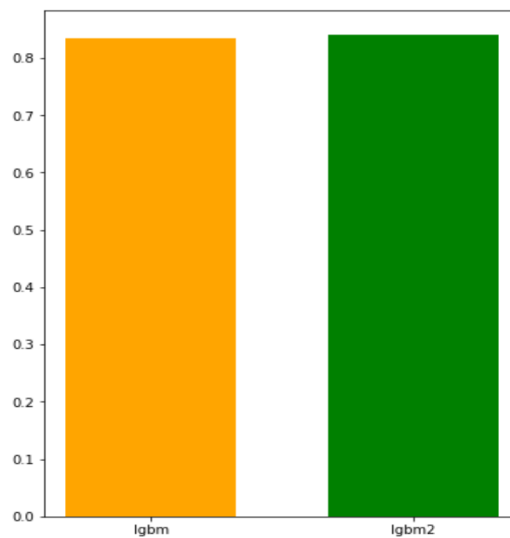


Figure 4.7: LGBM Accuracy

Though the accuracy level increased not much but the execution time of the algorithm decreased rather significantly. We plot that difference in a bar chart as well.

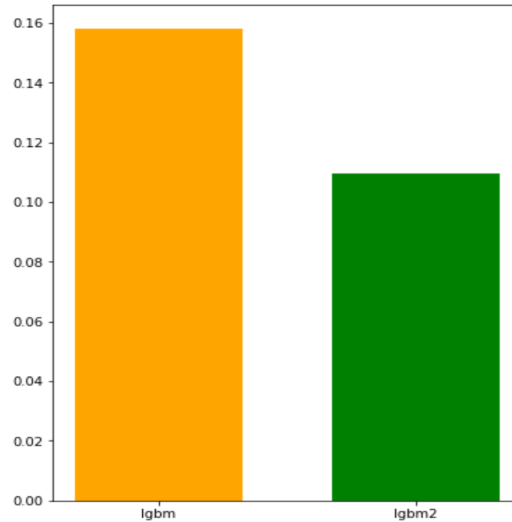


Figure 4.8: LGBM Run Time Comparison

4.1.3 CATBoost

Like the other algorithms which are used in that research, CatBoost is also an algorithm which is used for this research. CatBoost is the open source machine learning algorithm. It has an open source machine learning library with many parameters which helps to implement the gradient boosting algorithm fast and with reliability. This algorithm can be used for so many learning tasks like classification, regression, ranking and others. This algorithm was developed by the developers of Yandex. This algorithm is already in use for the different companies. They use this algorithm for so many purposes like searching for a product, recommending something to the customers, to create a AI based personal assistant, self-driving cars, to predict the weather. This algorithm supports the grid search and the randomized search which gives the best combination of the parameters which gives us the best results. The developers do not need to handle the categorical and the text features separately when they use this algorithm. This algorithm also allows us to run the training dataset in the GPU. Not only that, we can use multiple GPUs with a simple configuration where we can run the training process. Though the CatBoost algorithm is best for handling the categorical features, it is also capable of handling numerical and also text features. It has so many advantages but there are also some problems with this algorithm. When a developer trains a model, he or she wants to control the behavior of the tree which is built by the algorithm. They may want specific features to be treated as categorical. But it is quite difficult, sometimes impossible to do that because plotting the tree and controlling it in the Catboost algorithm is quite difficult for the developer.

The main reason for us to use this algorithm for our research is that this algorithm is fast and easy to implement. Also this algorithm does not apply much in the area of medical or psychology. It has a huge number of training parameters, which provide fine control over the categorical pre-processing [8]. This algorithm has so many parameters that it gives quite a good accuracy with the default parameter settings. We implemented this algorithm and tried to figure out which algorithm gives us the best results. As used to determine the Bipolar Disorder, we need to

take the best algorithm which can provide the best accuracy.

Methodology

Logical Reasoning

It is thought that gradient boosting creates a chain of estimations F^t repeatedly. Here the loss characteristic is $L(y_i, F^t)$ in which it includes enter variables. Here, y_i is the output cost that's anticipated for i^{th} time. It is believed that we're capable of making our approximations enhance y_i through some other characteristic that's

$$F^t = F^{t-1} + \alpha \cdot h^t \quad (4.6)$$

Here,

$\alpha = \text{Stepsize}$

$h^t =$ Base which used to minimize the loss which is expected.

This is chosen from a family of functions. Here this base is

$$h^t = \arg \min_{h \in H} EL(y_i, F^{t-1} + h) \quad (4.7)$$

This component use Taylor Approximation Technique for this, which is:

$$h^t = \arg \min_{h \in H} E(LyF^t - 1 - h)^2 \approx \arg \min_{h \in H} 1n(LyF^t - 1 - h)^2, h \in H \quad (4.8)$$

Cat-boost improves this technique by making some small changes. Suppose there is a Data-set which contains n number of samples. Now, dataset

$$D = \{x_k, y_k \mid |D| = n, x_k \in R^m, y_k \in R\} \quad (4.9)$$

Here,

$x =$ vector where m set of features situated for each sample.

$y =$ real targeted value

$m =$ number of features for each sample.

IMPLEMENTATION

We implement the catboost algorithm in the Google Colab. Google Collab does not have the catboost installed. Because of that we need to install the Catboost first.

We implement the catboost algorithm in two ways. Firstly we did not tune any parameters. We just used the default value of all the parameters. We split the dataset in the `test_size = 0.2` and give the `random state = 42`. By doing that we get the following result:

CatBoost Model accuracy score: 0.8690 Training-set accuracy score: 0.9845 Training set score: 0.9845 Test set score: 0.8690

After that we implement the same algorithm but this time we tune some parameters. We used grid search to determine which combination of the value can give us the best results. Also, we used the CatBoost Classifier method. The parameters which choose to tuning are:

`One_hot_max_size =` This parameter used for one_hot_encoding. `One_hot_encoding` is a technique which helps us to convert the categorical features into the variables which is easy for the algorithm to understand. The value we set for the `one_hot_max_size`

parameter; the algorithm will convert all the categorical features in the randomized different value but not more than the value set by the parameter.

Iterations = We used this parameter to determine the highest number of trees that can be built. This parameter can help us to achieve the best prediction result.

Max_depth = This parameter specifies the maximum number of depth that each tree can traverse.

Learning_rate = this parameter used for reducing the gradient steps. The value we set here will be used for all the loss_function parameter by default if the parameter l2_l1_reg parameter is not set.

Loss_function = As we need to find whether a patient is going through different types of episodic situations or not to determine if he or she has bipolar disorder, that's why it is a multiclass situation. Because of that reason loss_function is MultiClass.

The below table shows the value of the parameters which we tuned:

Parameter	Value (Default)
one_hot_max_size	30
iterations	500
max_depth	7
learning_rate	0.062030
loss_function	Multiclass

Table 4.7: Parameters for catboost

After that, the result we find is:

CatBoost Model accuracy score: 0.8828

Training-set accuracy score: 0.9672

Training set score: 0.9672

Test set score: 0.8828

Result Analysis of CATBoost

These are the confusion matrix of our datasets before tuning and after tuning.

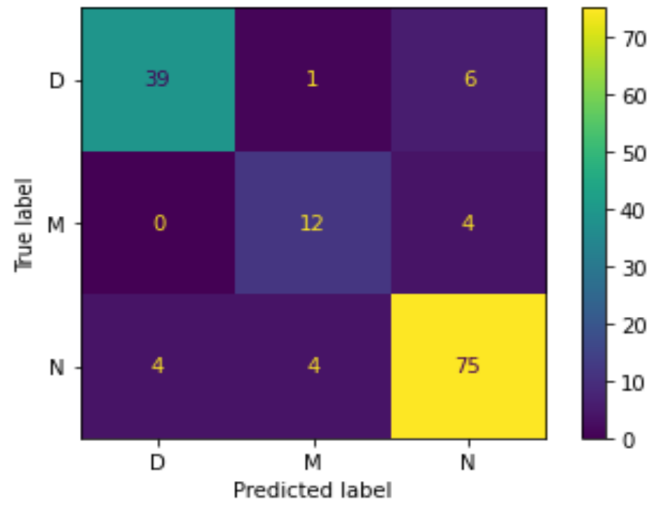


Figure 4.9: Confusion Matrix Before Tuning Catboost

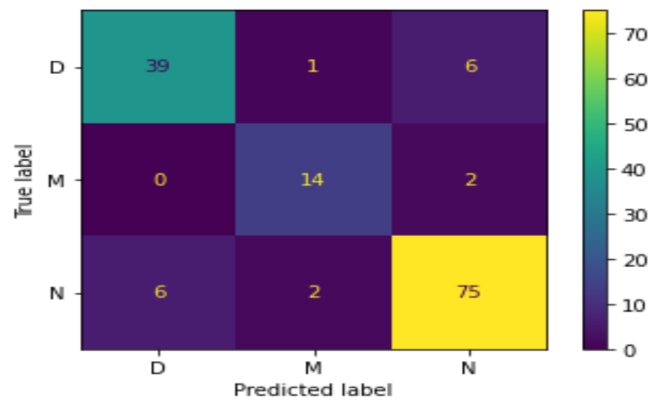


Figure 4.10: Confusion Matrix After Tuning Catboost

From the confusion matrix we can see each measure for each class. Here is a table that shows each measure for each class before tuning.

Class	Precision	Recall	F1-score
Normal	0.95	0.95	0.95
Depression	0.98	0.99	0.99
Mania	0.97	0.95	0.96

Table 4.8: Measure For Each Class Before Tuning Catboost

Class	Precision	Recall	F1-score
Normal	0.97	0.97	0.97
Depression	0.96	0.97	0.97
Mania	0.95	0.94	0.94

Table 4.9: Measure For Each Class After Tuning Catboost

We split the dataset 80-20. The meaning of this is 80 percent of the data used for the training of the algorithm and 20 percent of the data we use for the prediction. We see that before tuning the parameter and after tuning the value of prediction of the mania increased. The other values are unchanged.

The accuracy level also increased after tuning the parameter.

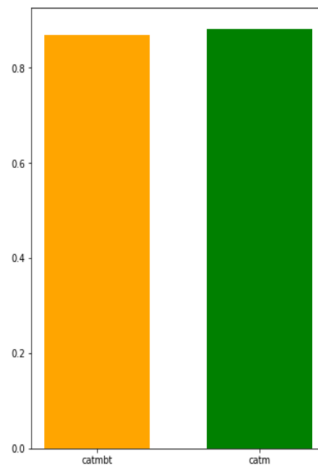


Figure 4.11: CatBoost Accuracy

Though the accuracy level increased not much but the execution time of the algorithm decreased significantly. We plot that difference in a bar chart.

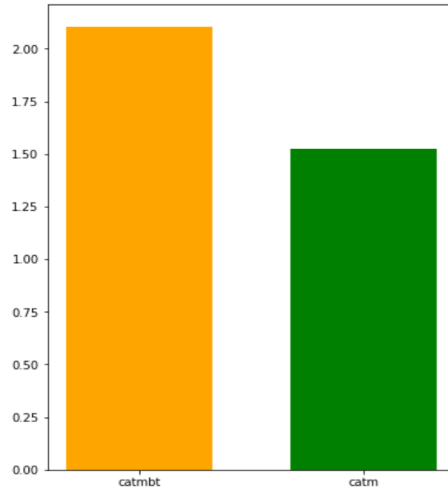


Figure 4.12: CatBoost Run Time Comparison

4.2 Solution and Deployment

We have designed a website as our primary Bipolar Disorder solution. The website will connect the patient with consultants based on the results. People internalizing are commonly considered more introverted, reserved, cold, and stoic than other people living with Bipolar Disorder. They may spend a lot of time attempting to manage or reason their emotions, but they often feel out of control, which worsens their symptoms. For those categorical people, online consultants will be a great help.

The site collects data in digits from users and applies the data from users and shows their mental states (normal/manic/depression). To calculate the state we have used the CATBoost model because this model gave us the highest accuracy among XGB and LGBM. As our two initial datasets were Episodes and Interviews, The two datasets have been merged and loaded in the model as interviews-episodes and applied in the tuned CATBoost classifier to make predictions of the Bipolar Disorder states. For example, after giving inputs of Mood: -2, Motivation:-3, Attention: 3, Irritability: 2, Anxiety: 3, Sleep quality: 1, Number of cigarettes: 30, caffeine: 0, alcohol: 0, Other drugs: 0, Active time: 30, the state shows Normal (Figure 6). Similarly for Mood: 0, Motivation: -1, Attention: 2, Irritability: 1, Anxiety: 1, Sleep quality: 2, Number of cigarettes: 24, caffeine: 120, alcohol: 0, Other drugs: 0, Active time: 1705 the state shows Depression (Figure 7) and for Mood: 1, Motivation: 1, Attention: 1, Irritability: 2, Anxiety: 1, Sleep quality: 2, Number of cigarettes: 20, caffeine: 60, Alcohol: 0, Other Drugs: 0, Active time: 1630 the state shows Manic (Figure 8). The classifier has been stored in a Pickle file and used in the Flask file to calculate the new data. The pickle module keeps track of previously serialized items so that subsequent references to the same item aren't serialized again, allowing for faster execution.

We've observed that the most effective treatment for bipolar illness is a mix of medication and psychotherapy. The majority of people use several drugs, including a mood stabilizer and an antipsychotic or antidepressant. Treatment should, however, be continued even when you feel better in order to keep your mood problems under

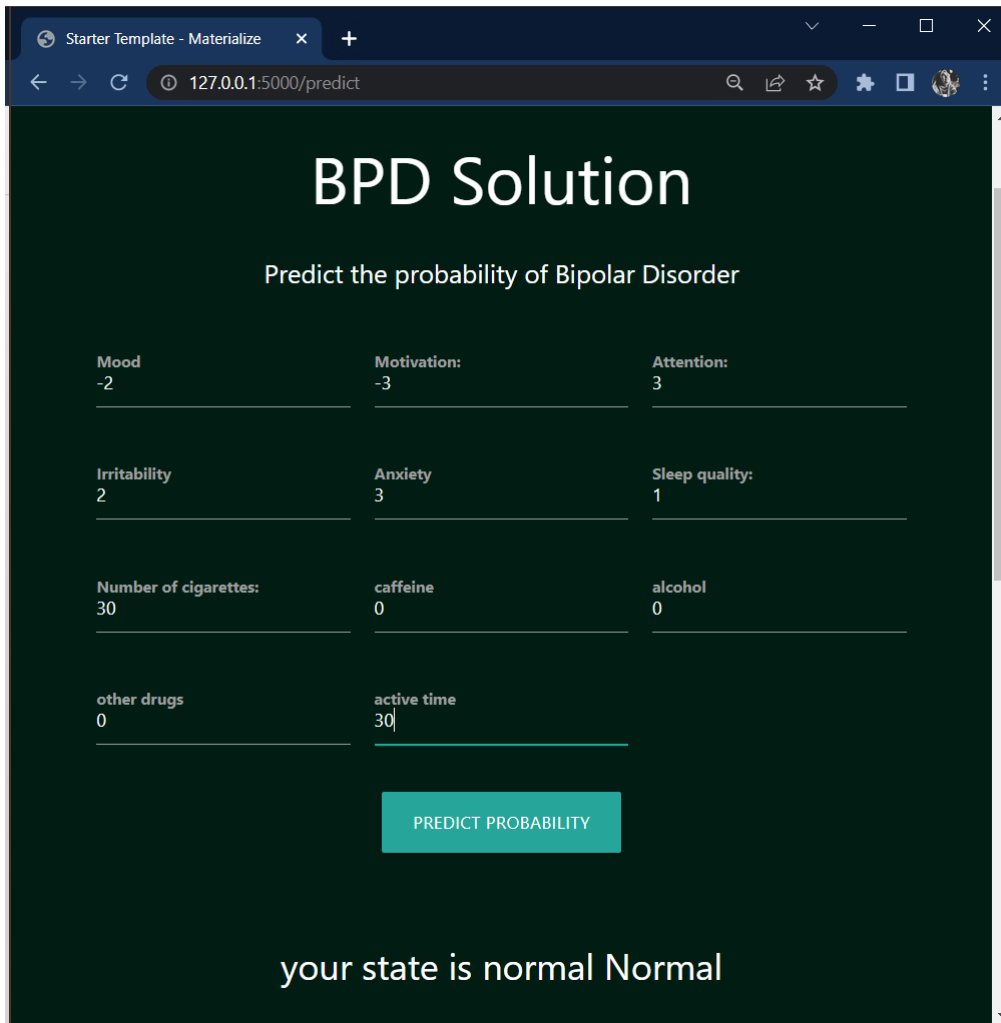


Figure 4.13: Normal state of BPD

control. As Bipolar illness is treated with more than just medication, medication can only be initiated under an expert's supervision, we are focusing on psychosocial treatment. Psychosocial therapy can enhance a patient's medication adherence and knowledge of their disease.

Cognitive Behavioral Therapy is one of the psychosocial therapies that might be effective (CBT). According to the American Psychological Association, attempting to change your mental processes can help with bipolar disorder. In CBT, role-playing is used to prepare for potentially difficult situations, face issues rather than avoiding them, and develop ways to calm and relax the mind and body. Adding cognitive-behavioral therapy to a treatment regimen may also improve the prognosis of bipolar illness, according to studies.[21]

As it is mentioned earlier, trying to modify your thought processes is useful for bipolar illness. To change an individual's thought pattern it is necessary to replace the negative thoughts with positivity. Negative thought habits can cause undue tension and worry, as well as a pessimistic attitude toward life. One of the steps to getting out of negative thoughts is Practicing Mindfulness.

One must first become conscious of their existing thought habits to acquire new positive thinking patterns. He may notice and identify habitual thought patterns, then choose whether or not to engage them, by practicing mindfulness. Mindfulness

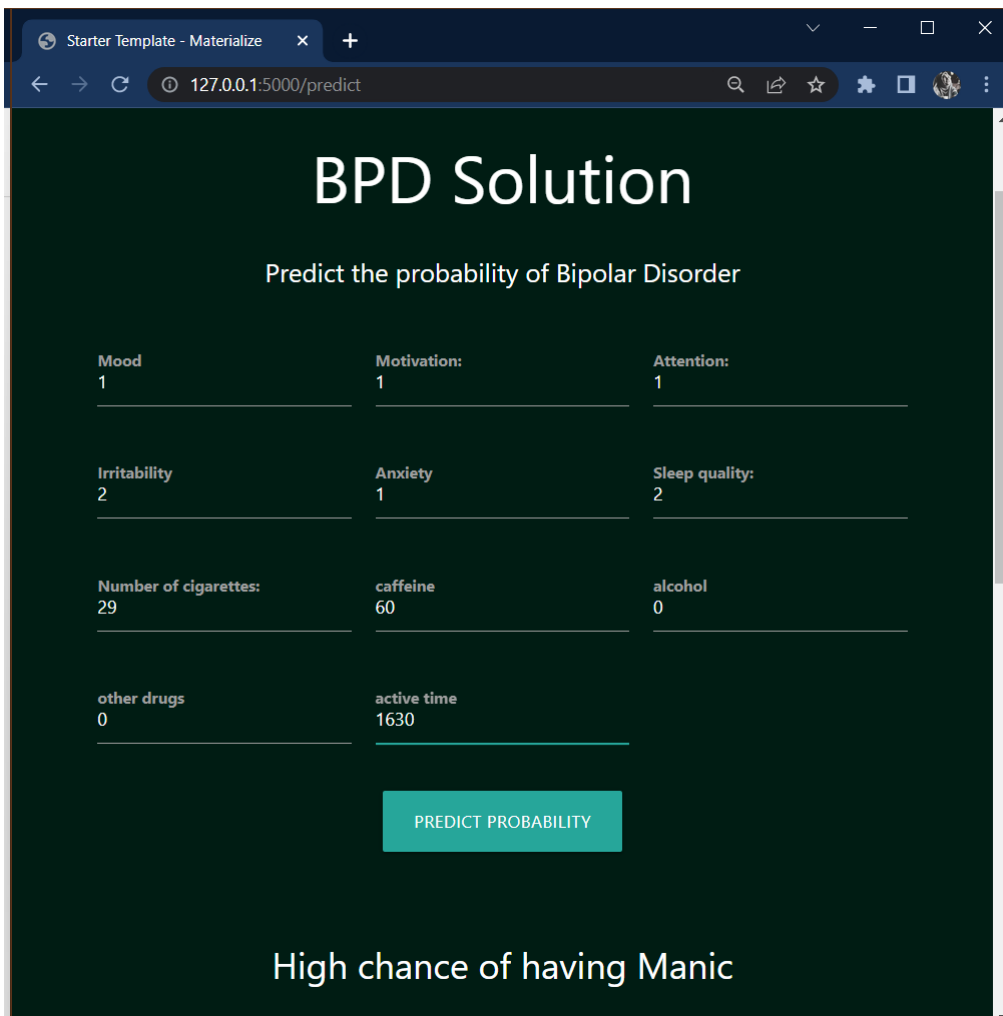


Figure 4.14: Manic State

establishes a separation between an individual and his thoughts, allowing him to see himself as distinct from others, and to do this writing down thoughts is most effective.

Putting emotions through writing is a great approach to offload and learn more about them. People are frequently unaware of how terrible their ideas are. Negative thought habits become automatic over time, usually without their knowledge. A person can more quickly recognize the areas that demand his or her attention by writing them down. One may also question them once they're written down to check if they're correct or relevant. If not, discard them or replace them with more optimistic thoughts.

Keeping all this in mind, we add a thought diary as an immediate solution. They can write their regular thoughts in their thought diary and can keep track of their writings no matter whether they have BPD or not. After that, we are going to add a real-time chat option "Text your Consultant here" so that people who are more comfortable with texting than visiting in person or in audio or video sessions. They can tell their symptoms or difficulties through online conversations. Finally, there is a "Check on Map" option we are going to build, from which they can book a suitable slot and get the location to visit their psychiatrist (Figure 9).

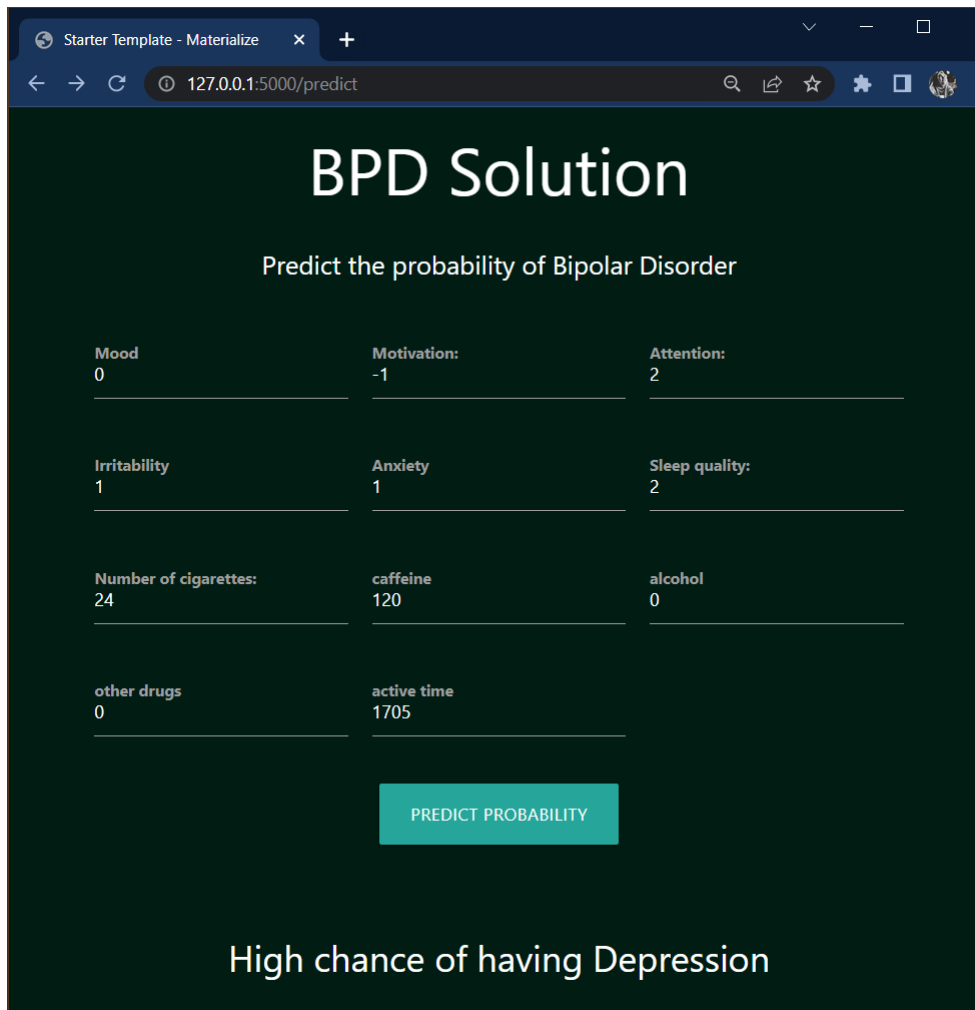


Figure 4.15: Depression State

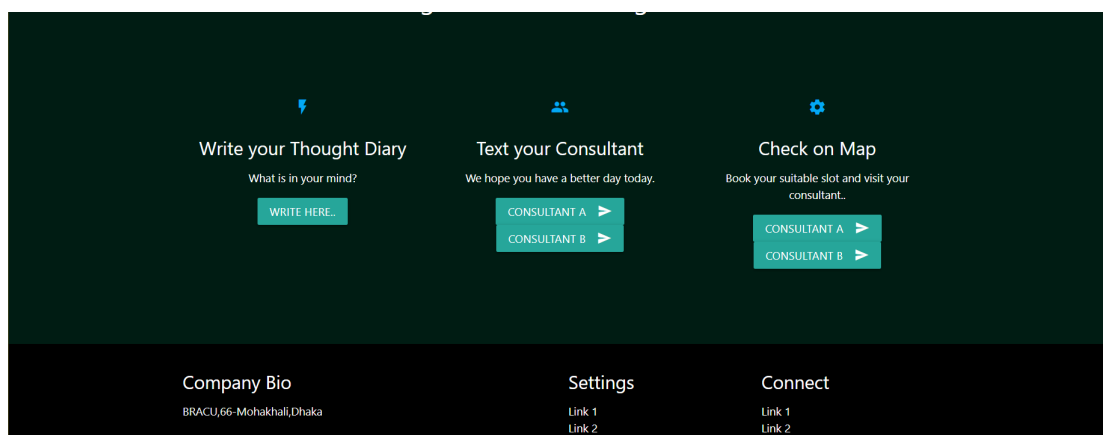


Figure 4.16: Primary Solution Features

Chapter 5

Comparison

Below is the graph which shows the results of our Algorithms which we applied for our study. From this graph we can compare our results.

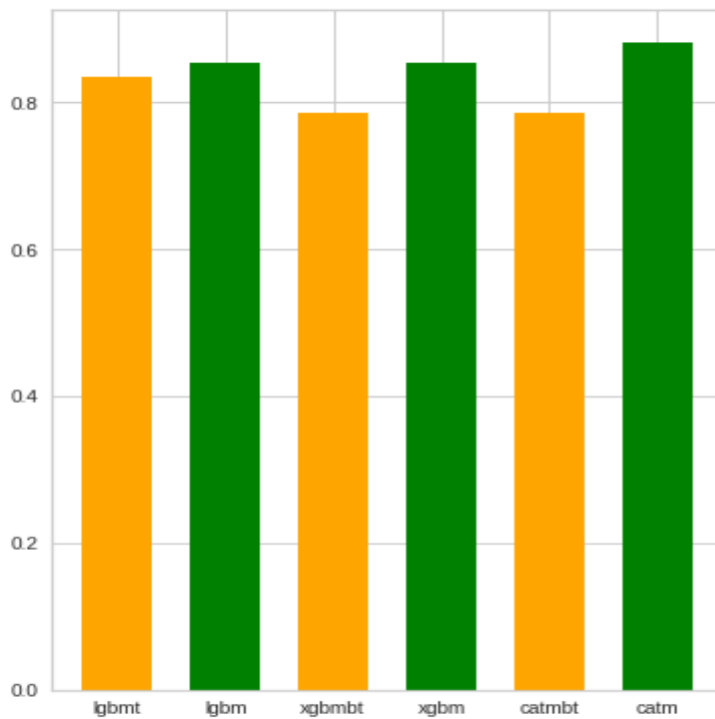


Figure 5.1: Algorithms' Results Compare

Here:

'lgbmt' = Accuracy of Light Gradient Boosting Algorithm before tuning the parameters.

'lgbm' = Accuracy of Light Gradient Boosting Algorithm after tuning the parameters.

'xgbmbt' = Accuracy of XGBoost Algorithm before tuning the parameters.

‘xgbm’ = Accuracy of XGBoost Algorithm after tuning the parameters.
‘catmbt’ = Accuracy of CatBoost Algorithm before tuning the parameters.
‘catm’ = Accuracy of CatBoost Algorithm after tuning the parameters.
From the above bar chart we can see that all the algorithms’ accuracy results increase after tuning the parameters. The accuracy of the Catboost Algorithm is higher than all other algorithms. Because the dataset which we used is small, the difference between the accuracy values among algorithms are not much. Below is the bar chart of the execution time each algorithm took.

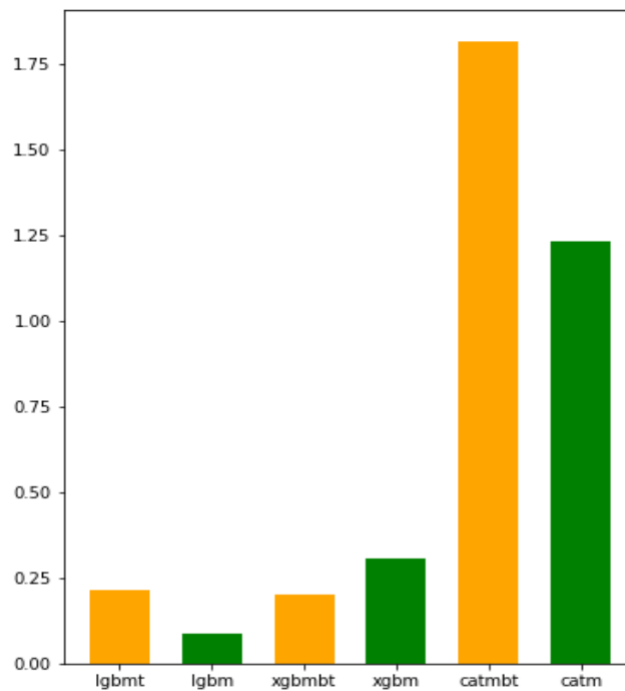


Figure 5.2: Execution Time Compare

Here:

‘lgbmt’ = Execution time of Light Gradient Boosting Algorithm before tuning the parameters.

‘lgbm’ = Execution time of Light Gradient Boosting Algorithm after tuning the parameters.

‘xgbmbt’ = Execution time of XGBoost Algorithm before tuning the parameters.

‘xgbm’ = Execution time of XGBoost Algorithm after tuning the parameters.

‘catmbt’ = Execution time of CatBoost Algorithm before tuning the parameters.

‘catm’ = Execution time of CatBoost Algorithm after tuning the parameters.

From the above graph see that, after tuning the parameter XGBoost algorithm’s execution time increased. Though Ctboost algorithm is faster than two other algo-

gorithms, but we see that Catboost algorithm's execution time is greater than other two algorithms both before and after tuning the parameters.

Chapter 6

Future Direction and Conclusion

In this thesis we have worked with some algorithms which are not only used for maximum accuracy but also for having an easier approach to solve this problem. The datasets of bipolar disorder are not available as other datasets. The dataset which is used for our study is not large. We need to find another dataset and apply this model to see the impact of the results. As previously stated, bipolar disorder is quite variable. There are now five kinds. Bipolar I, bipolar II, cyclothymic disorder, other defined bipolar and associated disorders, and unidentified bipolar and related disorders are the diagnoses. More research is needed to establish which person receives which variant. Though we applied only three algorithms, other available algorithms. We need to study further and apply other available algorithms to compare the data. Moreover, we can add more features for the primary treatment. For example, if the results show that a person has bipolar II disorder then the next possible next step can be an option for consulting a psychiatrist or having an appointment with a doctor for medication.

In this review, we discussed how we might reduce the risk of bipolar illness by applying machine learning prediction. Because Machine Learning model predictions allow the medical industry to generate very accurate projections based on past data, and bipolar illness is one of them. Thus, it is projected in this study that patients at high risk of BPD will be connected with professionals in their area.

To summarize, the symptoms of bipolar illness are severe enough that they can lead to death. Previous studies established how many individuals had BPD or are at high risk, but there was a lack of therapy or competent consultation. Many people are affected by this disease due to a lack of consciousness. We hope that our efforts will help such people.

Bibliography

- [1] A. AC01696967 *et al.*, *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*. World Health Organization, 1993, vol. 2.
- [2] K. Arens, “Wilhelm griesinger: Psychiatry between philosophy and praxis,” *Philosophy, Psychiatry, & Psychology*, vol. 3, no. 3, pp. 147–163, 1996.
- [3] S. N. Ghaemi, E. E. Boiman, F. K. Goodwin, *et al.*, “Diagnosing bipolar disorder and the effect of antidepressants: A naturalistic study,” *J clin Psychiatry*, vol. 61, no. 10, 2000.
- [4] R. M. Hirschfeld, A. R. Cass, D. C. Holt, and C. A. Carlson, “Screening for bipolar disorder in patients treated for depression in a family medicine clinic,” *The Journal of the American board of family practice*, vol. 18, no. 4, pp. 233–239, 2005.
- [5] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters, “Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication,” *Archives of general psychiatry*, vol. 62, no. 6, pp. 593–602, 2005.
- [6] P. McCrone, S. Dhanasiri, A. Patel, M. Knapp, and S. Lawton-Smith, “Paying the price,” *The cost of mental health care in England to*, vol. 2026, pp. 1–165, 2008.
- [7] K. Seeskin, “Plato and the origin of mental health,” *International Journal of Law and Psychiatry*, vol. 31, no. 6, pp. 487–494, 2008.
- [8] M. L. Phillips and D. J. Kupfer, “Bipolar disorder diagnosis: Challenges and future directions,” *The Lancet*, vol. 381, no. 9878, pp. 1663–1671, 2013.
- [9] A. Grünerbl, A. Muaremi, V. Osmani, *et al.*, “Smartphone-based recognition of states and state changes in bipolar disorder patients,” *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 140–148, 2014.
- [10] D. Vigo, G. Thornicroft, and R. Atun, “Estimating the true global burden of mental illness,” *The Lancet Psychiatry*, vol. 3, no. 2, pp. 171–178, 2016.
- [11] S. Mukherjee, *A.I. Versus M.D. What happens when diagnosis is automated?* The Newyorker, 2017, vol. 2.
- [12] W. H. Organization *et al.*, “Depression and other common mental disorders: Global health estimates,” World Health Organization, Tech. Rep., 2017.
- [13] M. V. L. López, *Application of Machine Learning Algorithms for Bipolar Disorder Crisis Prediction*. Universidad Complutense de Madrid, 2018, vol. 0.

- [14] I. Perez Arribas, G. M. Goodwin, J. R. Geddes, T. Lyons, and K. E. Saunders, “A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder,” *Translational psychiatry*, vol. 8, no. 1, pp. 1–7, 2018.
- [15] G. O. Belizario, R. G. B. Junior, R. Salvini, B. Lafer, and R. da Silva Dias, “Predominant polarity classification and associated clinical variables in bipolar disorder: A machine learning approach,” *Journal of affective disorders*, vol. 245, pp. 279–282, 2019.
- [16] R. Jadhav, V. Chellwani, S. Deshmukh, and H. Sachdev, “Mental disorder detection: Bipolar disorder scrutinization using machine learning,” in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, 2019, pp. 304–308.
- [17] Y. Ma, J. Ji, Y. Huang, *et al.*, “Implementing machine learning in bipolar diagnosis in china,” *Translational Psychiatry*, vol. 9, no. 1, pp. 1–7, 2019.
- [18] E. Papyrus, *Ancient Perspectives on Mental Illness: Insight from Hippocrates and Avicenna*. The Dunn Lab., 2019, vol. 3.
- [19] I. C. Passos, P. L. Ballester, R. C. Barros, *et al.*, “Machine learning and big data analytics in bipolar disorder: A position paper from the international society for bipolar disorders big data task force,” *Bipolar Disorders*, vol. 21, no. 7, pp. 582–594, 2019.
- [20] H. Sun, T.-T. Gong, Y.-T. Jiang, S. Zhang, Y.-H. Zhao, and Q.-J. Wu, “Global, regional, and national prevalence and disability-adjusted life-years for infertility in 195 countries and territories, 1990–2017: Results from a global burden of disease study, 2017,” *Aging (Albany NY)*, vol. 11, no. 23, p. 10952, 2019.
- [21] R. Black, “Treatments for bipolar disorder: Cognitive behavioral therapy and more,” *Remedy Health Media*, vol. 8, pp. 2–4, 2020.
- [22] H. Li, L. Cui, L. Cao, *et al.*, “Identification of bipolar disorder using a combination of multimodality magnetic resonance imaging and machine learning techniques,” *BMC psychiatry*, vol. 20, no. 1, pp. 1–12, 2020.
- [23] Anonymus *et al.*, *Bipolar disorder causes*. Black dog institute, 2021, vol. 0.
- [24] J. Jin-sol, “Theragenbio develops ai algorithm predicting immuno-oncology response,” *Korea Biomedical Review(KBR)*, vol. 13, no. 2, pp. 195–206, 2021.
- [25] J. Tomasik, S. Y. S. Han, G. Barton-Owen, *et al.*, “A machine learning algorithm to differentiate bipolar disorder from major depressive disorder using an online mental health questionnaire and blood biomarker data,” *Translational psychiatry*, vol. 11, no. 1, pp. 1–12, 2021.
- [26] N. Agnihotri and S. K. Prasad, “Review on machine learning techniques to predict bipolar disorder,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 13, no. 2, pp. 195–206, 2022.

Overleaf: GitHub for L^AT_EX projects

This Project was developed using Overleaf(<https://www.overleaf.com/>), an online L^AT_EX editor that allows real-time collaboration and online compiling of projects to PDF format. In comparison to other L^AT_EX editors, Overleaf is a server-based application, which is accessed through a web browser.