Forecasting Depression in Bipolar Disorder Using Machine Learning Algorithms

By

Abdul Baqi Malik
(1886-2015)

Muhammad Anas Bin Munzir

(1614-2016)

A project presented to the

Faculty of Engineering Sciences and Technology

Hamdard Institute of Engineering and Technology

In partial fulfillment of the
requirements for the degree of

Bachelors of Science
In
Computer Science

Faculty of Engineering Sciences and Technology
Hamdard Institute of Engineering and Technology
Hamdard University Islamabad Campus, Pakistan

PROJECT IN BRIEF

Project Title:	Forecasting Depression in Bipolar Disorder Using Machine Learning Algorithms
Objective:	To find the best accuracy rate which will help diagnose bipolar disorder and be helpful in their treatment
Undertaken By:	Abdul Baqi Malik, Muhammad Anas Bin Munzir
Supervised By:	Mr. Muhammad Shafiq (Supervisor)
Date Started:	February 2021
Date Completed:	December 2021
Tools Used:	Jupyter Notebook (Anaconda 2.0), Python 2.7, MS Word 2013, Diagram Tool, MS Excel 2013
System Used:	*Processor Core i3, 8GB RAM, 500GB HDD, 4th Gen*

ABSTRACT

The human mind is the most complicated and irreplaceable thing in the world. Mental illness is a severe issue that is needed to be predicted and resolved beforehand. BpD is a complicated disorder that affects people all around the world. With the practice of ML and DL, can help both doctors and patients for better identification of the illness. Mental illness and mood swings have always been the major problem for humans.

The goal of our project to create a prediction model by applying different ML and DL algorithms on symptom-based patient data. This model would help psychiatrists to decide whether their patients might be towards a depression episode or mania episode, or staying in a euthymic state.

The major purpose of this research is to develop a model that will analyze and use Machine and Deep Learning Algorithms on the patients given data to predict their mental illness/mood swings. The main purpose is by applying different algorithms of Machine and Deep Learning on different data sets and finding more prediction accuracies than previous works which will be a help to the medical field in diagnosing and treating such conditions before they turn into a serious issue.

This research will be done in two parts. The first part will be the collection of data about the concerned individual from various reliable resources and knowledge and understanding of it. This project includes many processes like the process of gathering the data, cleaning the data, and visualizing the data from patients with BpD. It includes a comprehensive analysis of the preparation the data which contains snippets of code and plots that help know the data better and detect the possible relationships and dependences among them.

The other part contains the predictive examination of the data. In this part, different algorithms of ML and DL are analyzed and functional to the patient data to equate prediction accuracies and select the algorithms that suit the problem.

1. Introduction

Within the medical division of psychology, within the region of brain disarranges, one of the current disarranges is BpD. This particular disorder is categorized by the faltering of the patient's temperament among two states, mania and depression [1], which ordinarily come gone to by distinctive appearances (features), both physical and psychological.

The background of the categorical method to psychiatric illness lies in the classic work of the originators (founders) of modern psychiatry, such as Emil Kraepelin. Kraepelin proposed a contrast between psychiatric illnesses considered by regularly periodic episodes of prominent changes in effect; and illnesses categorized by abnormal cognitions, beliefs, and experiences (i.e. psychotic indications), which are usually established in early adulthood and continued throughout life [2]. Kraepelin referred to the first category as "manic—depressive psychosis", including illnesses that we now refer to as affecting disorder (e.g. bipolar disorder).

This project was done to get a better understanding of forecasting in data science and its functions. It is designed to help for forecasting bipolar disorder using machine and deep learning. With the help of data science tools, predict accuracies on different data sets and visualize through different mapping schemes like heat maps, histograms … etc.

Machine Learning:

ML is getting to be slowly more existing in all frameworks that collect and prepare endless sums of information, being nearly a imperative necessity within the improvement of modern software applications. One of the areas of work that might advantage from ML is, past trouble, the field of medicine. The utilize of ML calculations permits the methodology of both classification and regression models that offer assistance with the determination of diverse maladies, proposal of drugs, programmed supervision of drugs, etc.

ML is the method by which confident models are made, with the assistance of distinctive calculations that anticipate values based on different highlights and ended up ever way better in making these expectations the more information they train on [3]. For developing a machine learning model, a prescient (predictive) model is built utilizing ML algorithms such as linear regression and decision trees … etc.

We divide data sets into two parts. The judgment of the model is based on the ML algorithm. ML model is based on training data and testing information. The capability of the model is assessed through demonstrate evaluation and optimization. At that point, an result is anticipated after performing parameters tuning and progressing the accuracy of a model through forecast.

1.1. Research Context:

This project determines from the bipolar forecasting project [5], which thinks about the presence of calamity in patients with BpD to predict them. The most point of the Bip4cast extend is to be able to respond in time and maintain a strategic distance from the symptoms before the patients begin to suffer from them.

Numerous other research projects have been exceptionally supportive for this extend, such as the ‘Machine Learning and Big Data analytics in bipolar disorder’ [6]: A location paper from the ‘International Society for Bipolar Disorders Big Data Force’. A task constrain was organized to examine and integrate discoveries from the scientific works related to ML and big data-based considers to clarify terminology and to characterize challenges and possible applications within the field of BpD.

The project has been likely much obliged to projects just like the 'Design of BD Architecture for Prediction of Crises in Bipolar Disorder' [8] and 'Classifying Bipolar Personality Disorder utilizing artificial ML neural networks', which have made a difference out facilitate the gathering of all the patient's information. The objective of the particular project described in this has been to undertake to predict the state a patient is in with the help of ML and DL algorithms like Decision Tree Algorithms, Random Forest, Logistic Regression, and K-Nearest Neighbor.

The trouble of this task is that not all patients behave similarly when in a state or another, so in case the prediction isn't totally precise, it at slightest gives the psychiatrist a moment conclusion almost the state a patient may be in or tend towards. In this project, the information picked up during the Computer Science and Engineering degree has been exceptionally present, especially the one accomplished with subjects like "Data Science", "Probability and Statistics", as well as "Data Structures and Algorithms", since of the assistance it has passed on within the understanding of algorithm structures and ML strategies.

1.2. Problem Statement:

BpD, a difficult disorder in the brain has affected many millions of people around the world. This brain disorder is identified by the rate of the fluctuations of the patient's changing mood. The mood swing between two states i.e. depression and mania. This is a result of different psychological and physical features. A set of psycholinguistic features like behavioral changes, mood swings, and mental illness are observed to provide feedback on health and wellness. The study is a target measure of identifying the stress level of the human brain that could improve the harmful effects associated with it considerably.

In the project, we present the study prediction of indications and behavior of a commonly known mental health illness, bipolar disorder using ML and DL Techniques.
Therefore, we extracted data from online resources that were studied and analyzed by using ML and DL techniques. In this project execution of different algorithms like Logistic Regression, Decision Trees, Random Forest, SVM, and K-Nearest Neighbor are studied and analyzed for identifying the state of the patients.

1.3. Work Planning:

The strategy that has been taken after in this project is: data gathering, cleaning, exploratory data investigation, algorithm comparison, determination, and predictions with randomized feature values.

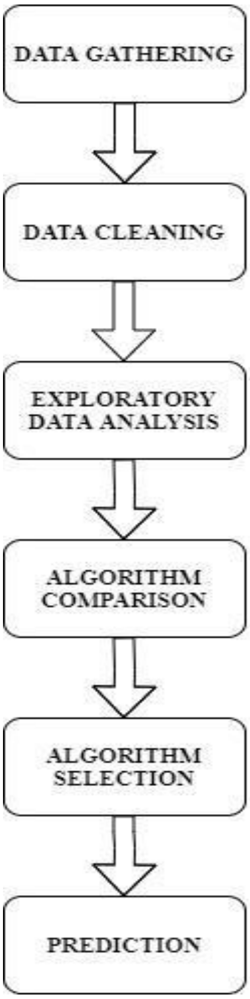


Figure 1.1 Project Process Diagram

The primary portion of the project was to gather all the data required for the method of Machine and Deep Learning, to afterward be able to clean and visualize it. After having assembled all the information required for the project, another step was to clean it. This portion is considered fundamental, as the data in this kind of project nearly continuously lacks entries or has empty values.

The another portion of the project was the Exploratory Data Analysis (EDA), which made a difference us find the variables that were connected and see the distribution of the data within the distinctive data sets. This portion moreover includes the identification of outliers, which are data points that are exceptionally distant from other points that can negatively affect the prediction.

Having a common understanding of how the information were represented within the data sets, another portion was to include different combinations of the data sets to find the features that may suit the problem the most. The portion where the algorithms were applied included the method of testing different standard Machine and Deep Learning algorithms on the different data set combinations. This portion was crucial to choose the algorithms that were getting to be utilized for the forecasts or predictions. The selection was based on their accuracy, which may be a great way to measure their performance.

The final portion of the project comprised of the model forecast. For this, we chosen the algorithms that had the finest testing performance and applied them to randomized patient data to observe how they worked. This section moreover included an application to test data comparing to different states in which the patient may be. The amount of information given for the project was not sufficient to create models with a high level of confidence, but the implementation has been planned as a blueprint for future projects that include bigger amounts of data.

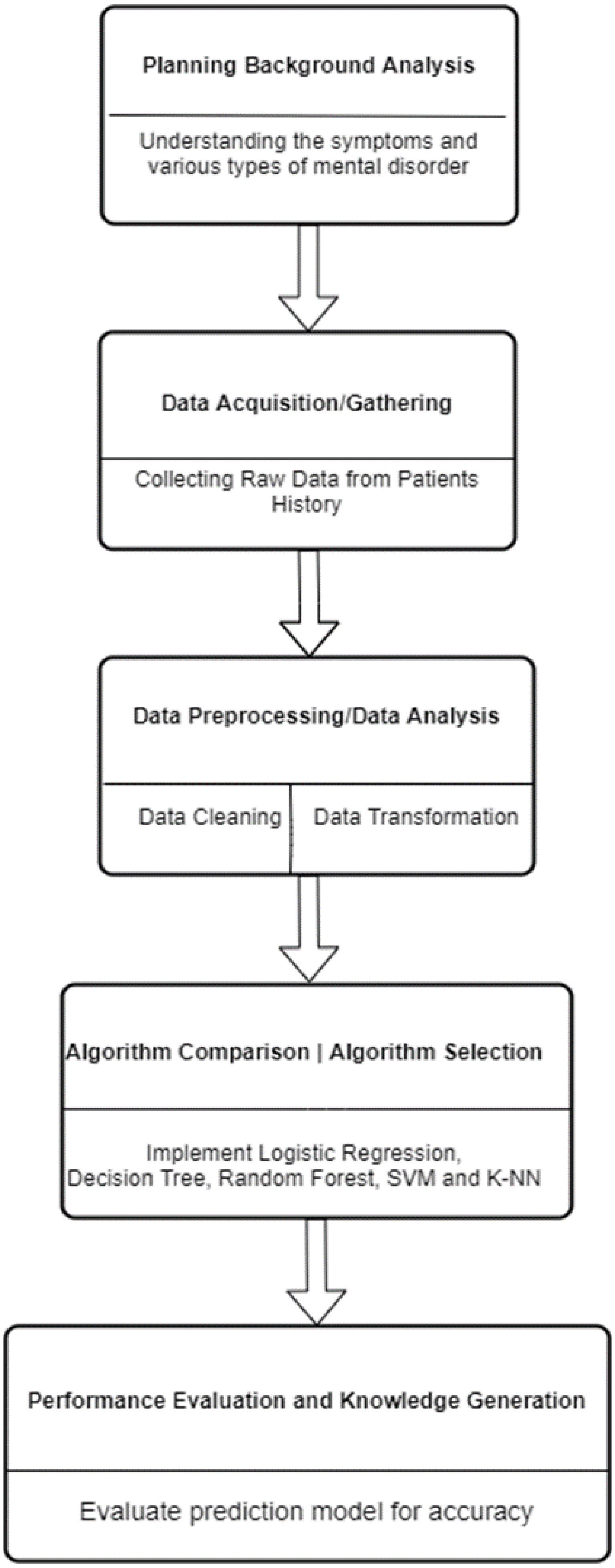


Figure 1.2. Work flow diagram

1.4. Usage of Data:

The data utilized for this project is anonymized patient data gathered by psychiatrists collected through online sources. All the data was accessible in an Excel records with different sheets. The data has been assembled during medical appointments with four different patients that have Bipolar Disorder, but the goal for the future is that it be both recorded by the psychiatrists in appointments and with the assistance of mobile applications [9]. This way, the patients can effectively take an interest in their diagnosis.

1.5. Technologies Used in this Project:

The programming language utilized in this project is Python [10], which includes a lot of libraries that make information cleaning and visualization less complicated, as well as applying ML and DL algorithms. The environment utilized is Jupyter Note pad [11].

➢ **Python:**

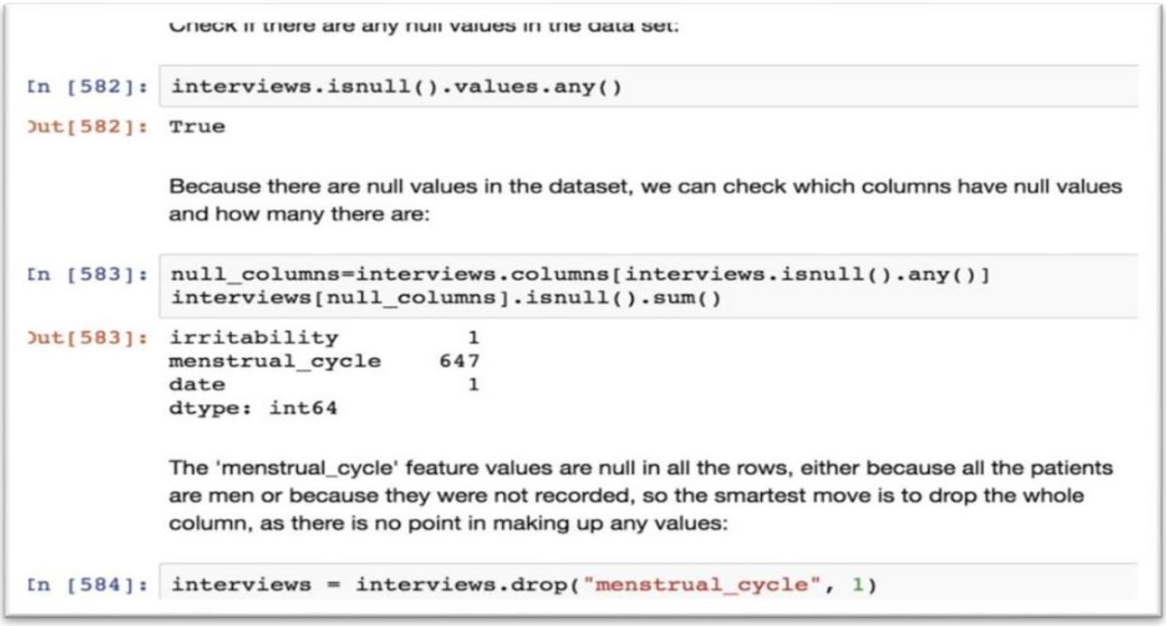
Python is a programming language that's widely utilized in Machine and Deep Learning. It is particularly useful for data cleaning and plotting as the language is designed for clarity and ease of use. There are very some libraries planned for this task, like Pandas [12], which incorporates a lot of strategies for data frame handling, or Seaborn and Matplotlib [13], which are data plotting libraries. Scikit-learn [14] is the Machine Learning library of choice for this project since it includes preprocessing and cross-validation tools as well as all the known standard Machine Learning algorithms.

➢ **Jupyter Notebook:**

Jupyter Notebook, the logo of which can be seen in Figure 1.2, may be a web application consisting of “notebooks” that include live code. It is an intelligent and interactive environment that contains both code and Markdown language to depict each code snippet.

Figure 1.3. logo of Jupyter Notebook

The most reason this environment was chosen is the possibility it has for fast prototyping since you'll be able visualize the comes about very quick and in a cleaner way than with a terminal. The use of Markdown language is very valuable as well since it gives the choice to clarify each code snippet independently, as seen in Figure 1.3.



Check if there are any null values in the data set:

```
In [582]: interviews.isnull().values.any()
Out[582]: True
```

Because there are null values in the dataset, we can check which columns have null values and how many there are:

```
In [583]: null_columns=interviews.columns[interviews.isnull().any()]
          interviews[null_columns].isnull().sum()
Out[583]: irritability          1
          menstrual_cycle     647
          date                  1
          dtype: int64
```

The 'menstrual_cycle' feature values are null in all the rows, either because all the patients are men or because they were not recorded, so the smartest move is to drop the whole column, as there is no point in making up any values:

```
In [584]: interviews = interviews.drop("menstrual_cycle", 1)
```

Figure 1.4. Markdown and Python code Sample

3. Problem Statement and System Analysis:

3.1. Problem Statement:

In this report, we recommended Bipolar disorder is a continuing mental illness that occurs due to episodes of Mania and depression. Even after taking treatment for the illness, people continue to have symptoms of it. Each type of disorder is recognized and treated differently depending upon the type of disorder. BpD disorder has indications of mood episodes from Mania to Depression periodically and can be diagnosed if manic episodes are seen for at least seven days or occur alternatively with severe depression.

ML is the process of generating certain models and algorithms to predict values based on different features. This paper analyzes ML and DL techniques for bipolar disorder. The issue to be addressed is to improve the percentage of correctly predicted depressive disorder. This work is supposed to utilize the recent advancement in machine learning for aiding the correct forecasting of a patient's condition.

In this project, we present the study prediction of symptoms and behavior of a commonly known mental health disorder, BpD using different ML techniques. Therefore, we extracted data from online resources that were studied and analyzed by using different techniques. In this project execution of different algorithms like Logistic Regression, Decision Tree Algorithm, Random Forest, SVM, and K-Nearest Neighbor are studied and evaluated for detecting the state of the patients.

3.2. System Analysis:

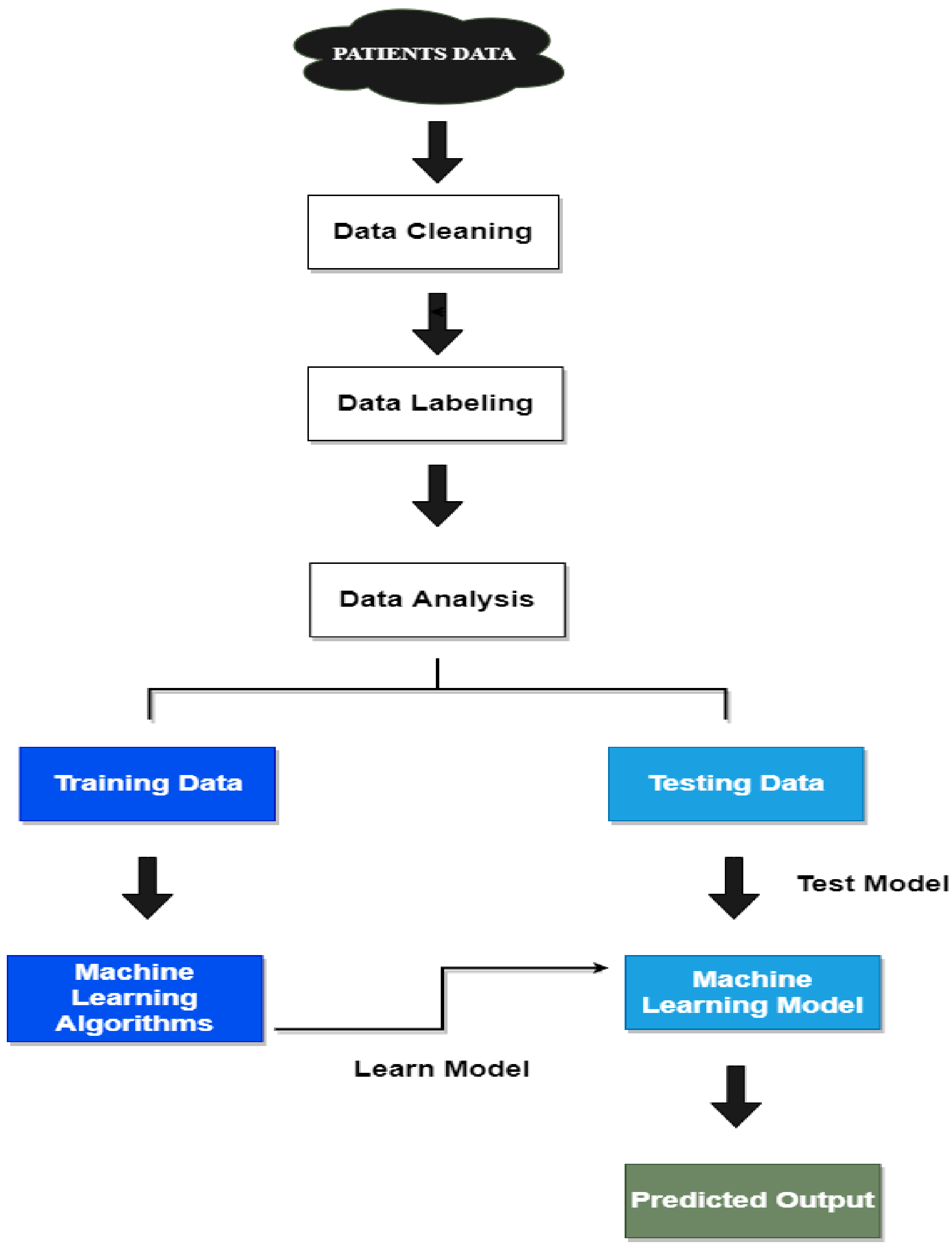


Figure 3.1. System Analysis Diagram

- **Patients Data:**

Patient data consists of different datasets like episodes, YMRS (young mania rating scale), HDRS (Hamilton data rating scale), interviews, and interventions.

- **Data Cleaning:**

The raw data needed to be preprocessed into a comprehensible format. Because of recording unrelated data in data sets, data cleaning is required. Collected data was transformed into proper forms so that the machine can recognize the format in the implementation process. In our data sets, we cleaned up all null and missing data and checked. We handle missing values in ML datasets by deleting rows, replacing them with mean/median/mode, assigning an individual value, and predicting a missing value.

- **Data Labeling:**

In ML, data labeling is the process of identifying raw data and adding one or more significant and informative labels to provide relation (context) so that an ML model can learn from it.

- **Data Analysis:**

ML is a method of data analysis that mechanizes analytical model building. It is a branch of AI-based on the knowledge that systems can learn from data, identify patterns and make decisions with minimal human intervention.

- **Training Data:**

Training data is the data you use to train an algorithm or ML model to predict the result to you, design a model to forecast. In this study, we projected Logistic Regression, Decision Tree, Random Forest, SVM, and K-Nearest Neighbor algorithms to generate the prediction model. Descriptions are in the related algorithm section.

- **Testing Data:**

Test data is used to measure the performance, such as accuracy or efficiency, of the algorithm you are using to train the machine. In the evaluation of performance and knowledge generation, the results/outcomes are calculated. For better performance, the constraints (parameters) of different algorithms are taken into consideration.

7. Application of the Algorithms:

The two most common types of ML that are utilized these days are supervised learning and unsupervised learning. The supervised learning method comprises in creating a model, with the help of an algorithm, which can predict output for a certain input. It frequently counts with a set of data for which the input and output are known. This set of information or data is called the training set.

Supervised learning problems can be of two types, regression, and classification. Regression could be a problem in which the output is quantitative, like predicting stock prices based on already gotten results. Classification may be a problem in which the output is qualitative. The issue that's displayed in this project could be a classification issue, because it aims to predict whether the patient is having a mania or depression episode, or is remaining in a euthymic state.

Unsupervised learning is used when the output isn't known for the input values. The most sort of problem in unsupervised learning is clustering, which comprises of grouping the input data in clusters. Clusters are gigantic collections of data that have similar features [27]. In this portion, we tested different classification algorithms on the data sets that we had already obtained. The target or value that we attempted to predict was the state a patient is in (Depression, Mania, or Euthymic), so all the data sets had to have a column with the episode related with the rest of the features, which we ensured within the data set combination section. The data sets on which we tested the algorithms were:

- **YMRS:** Young Mania Rating Scale data (young_episodes).
- **HDRS:** Hamilton Depression Rating Scale data (hamilton_episodes).
- **Interviews:** interview data (interviews_episodes).
- **Interventions:** intervention data (intervention_episodes).
- **YMRS-HDRS:** The combination of the YMRS and HDRS data (ymrs_hdrs).
- **Interviews-Interventions:** The combination of the interview and intervention data (interviews_interventions).

The algorithms that we utilized for this part are Logistic Regression [31], Decision Tree [28], Random Forest [29], Support Vector Machines [30] and K- Nearest Neighbor. The reason why these algorithms have been chosen for this project is explained down below within the section that corresponds to each algorithm.

Before applying each algorithm, as is essential for each classification problem, the initial data required to be split into a training and testing set (see Figure 7.1). Afterward, the training data would be utilized to train the prediction models and the testing data would be used to compare the output of the model with the real targets by cross-validation, a strategy displayed to begin with by M. Stone in 1974 [34] that's used widely in Machine Learning for algorithm performance comparison [35].

Machine Learning Algorithm Application Process Diagram:

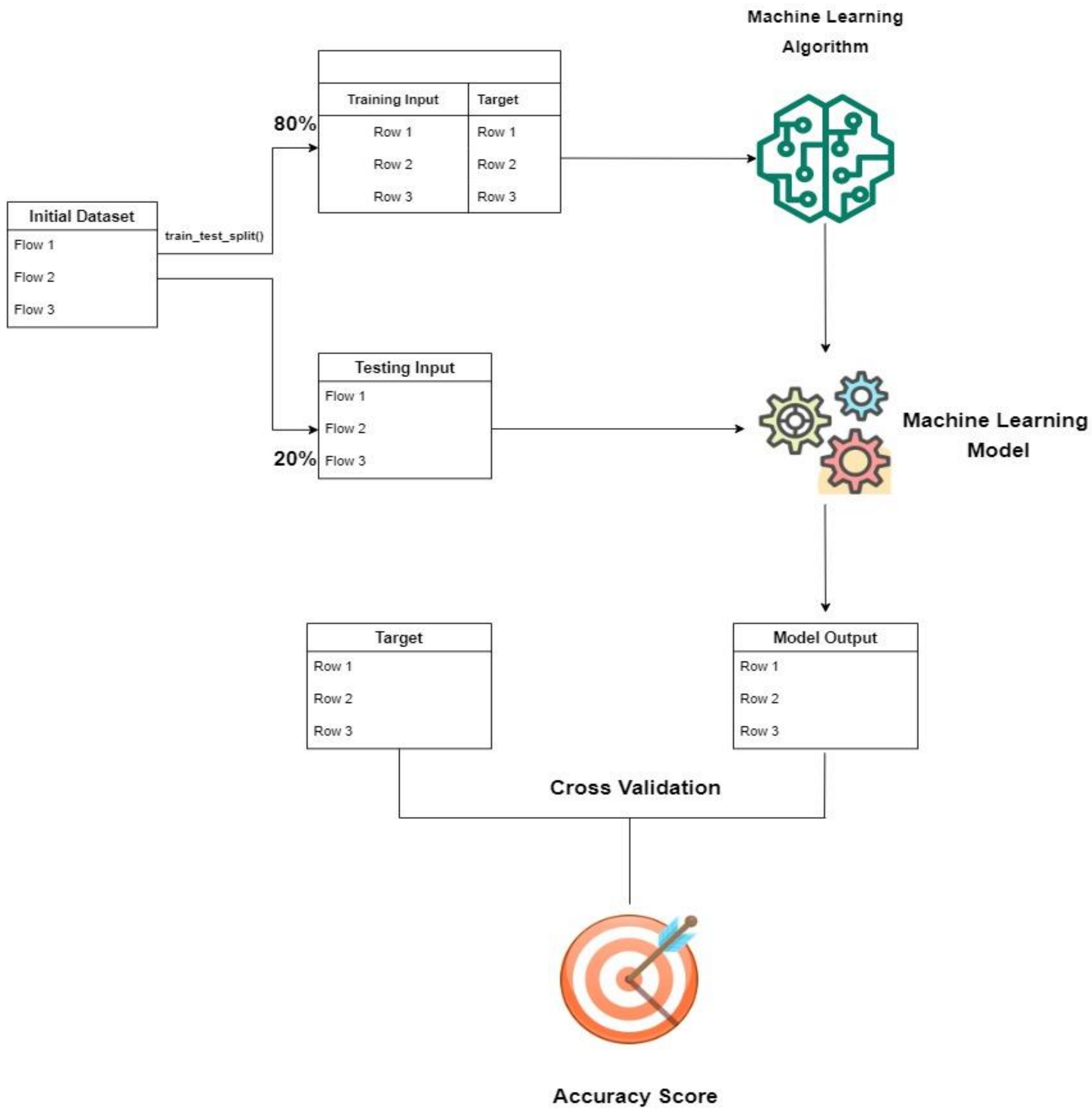


Figure 7.1. ML algorithm application process diagram

7.6. Tests on Randomized Data

To check if the predictions made by the distinctive models obtained with the algorithms made any sense, we continued to randomize a few patient data to test the models. This was done by randomizing the values of the different things in the Interview data set. Every item has a different range so, for case, the mood values that we produced had to be between -3 and 3, the range in which those values were in the interviews performed by the psychiatrist. We did this with a circle that generated meet data for fifty random patients to be able to see which algorithms made the most reasonable forecast.

```
variables = {}
variables['mood'] = randint(-3, 3)
variables['motivation'] = randint(-3, 3)
variables['attention'] = randint(0, 4)
variables['irritability'] = randint(0, 4)
variables['anxiety'] = randint(0, 4)
variables['sleep_quality'] = randint(0, 4)
variables['nr_cigarettes'] = randint(0, 40)
variables['caffeine'] = randint(0, 250)
variables['active_time'] = randint(200, 1800)

test = np.array([variables['mood'], variables['motivation'], variables['attention'],
            variables['irritability'], variables['anxiety'], variables['sleep_quality'],
            variables['nr_cigarettes'], variables['caffeine'], variables['active_time']])
```

Figure 7.6. Random patient interview data generation

The Random Forest classifier appeared to make very sensible predictions. Within the case of a patient that feels down and sleeps a lot, this algorithm predicted that he or she might be tending towards a Depression episode. Test of randomized data are:

```
--------------------------------------------
PATIENT  2
--------------------------------------------
Mood:  -3
Motivation:  0
Attention:  2
Irritability:  1
Anxiety:  2
Sleep quality:  1
Number of cigarettes:  10
Caffeine:  221
Active time:  702
--------------------------------------------
PREDICTIONS
--------------------------------------------
- Logistic Regression prediction:  The patient is in a Euthymic state
- Decision Tree prediction:  The patient is in a Euthymic state
- Random Forest prediction:  The patient is in a Euthymic state
- K-Nearest Neighbour prediction:  The patient could be tending towards a Depression episode
- SVM prediction:  The patient is in a Euthymic state
```

Figure 7.7. ***Prediction of a Depression episode on a patient with the lowest mood value (Sample)***

7.7. Tests on Real Data

How to utilize these algorithms on real patients, we executed a small Python program that imports the Random Forest classifier and makes predictions based on specific input from a terminal. This input contained all the factors that can be show in a meet made by the specialist. The tests weren’t run on genuine patients, but the input data used in them corresponded to different states that the patient may be in. The program returns the prediction in a human-readable format and colors the output messages depending on the seriousness of the diagnosis. In case the model predicts that the patient might be tending towards a Mania or Depression episodes or Euthymic state, the messages are appeared in yellow.

```
(DS27) C:\Users\abdul\FBPD - master>python state_prediction.py


----------------------------------------
|  Forecasting BiPolar Disorder        |
----------------------------------------
|        Final Year Project             |
|          BS - Computer Sciences       |
----------------------------------------
| Author : Abdul Baqi, Anas Munzir      |
----------------------------------------


Please indicate each of the values gathered during the interview:
* Mood (-3 to 3)? 1
* Motivation (-3 to 3)? 3
* Attention (0 to 4)? 2
* Irritability (0 to 4)? 1
* Anxiety (0 to 4)? 4
* Sleep quality (0 to 4)? 2
* Number of cigarettes smoked? 5
* Amount of caffeine ingested? 100
* When did the patient wake up (hh:mm)? 09:30
* When did the patient go to bed (hh:mm)? 23:40


▯[33mThe patient could be tending towards a DEPRESSION episode▯[0m
```

Figure 7.08. Depression episode prediction

```
(DS27) C:\Users\abdul\FBPD - master>python state_prediction.py


----------------------------------------
|  Forecasting BiPolar Disorder        |
----------------------------------------
|        Final Year Project             |
|          BS - Computer Sciences       |
----------------------------------------
| Author : Abdul Baqi, Anas Munzir      |
----------------------------------------


Please indicate each of the values gathered during the interview:
* Mood (-3 to 3)? 2
* Motivation (-3 to 3)? -2
* Attention (0 to 4)? 1
* Irritability (0 to 4)? 2
* Anxiety (0 to 4)? 1
* Sleep quality (0 to 4)? 3
* Number of cigarettes smoked? 20
* Amount of caffeine ingested? 0
* When did the patient wake up (hh:mm)? 10:00
* When did the patient go to bed (hh:mm)? 01:00


□[33mThe patient could be tending towards a MANIA episode□[0m
```

Figure 7.09. Mania episode prediction

```
(DS27) C:\Users\abdul\FBPD - master>python state_prediction.py


----------------------------------------
|  Forecasting BiPolar Disorder        |
----------------------------------------
|        Final Year Project             |
|          BS - Computer Sciences       |
----------------------------------------
| Author : Abdul Baqi, Anas Munzir      |
----------------------------------------


Please indicate each of the values gathered during the interview:
* Mood (-3 to 3)? -2
* Motivation (-3 to 3)? -2
* Attention (0 to 4)? 4
* Irritability (0 to 4)? 3
* Anxiety (0 to 4)? 3
* Sleep quality (0 to 4)? 0
* Number of cigarettes smoked? 30
* Amount of caffeine ingested? 200
* When did the patient wake up (hh:mm)? 05:00
* When did the patient go to bed (hh:mm)? 20:30


□[36mThe patient is staying in a EUTHYMIC state□[0m
```

Figure 7.10. Euthymic state prediction

8. RESULTS

The results gotten amid this project are separated into different sections. Each section contains a summary of the most important results that were obtained within the corresponding part of the project.

8.1. Data Gathering Results

The data gathering may be a small part of the project but one of the most important, because it lays the ground for the rest of the project. With the assistance of the results gotten, we could observe that the data sets were generally really small, but for the Interview data set which had an satisfactory size in comparison with the others. The fact that the data sets had such a small amount of data has obviously been the most challenging portion of the project since the rest of the sections depended on the amount of data that was available.

```
print "Number of rows in HDRS dataframe:", len(hamilton.index)
print "Number of rows in YDRS dataframe:", len(young.index)
print "Number of rows in interview dataframe:", len(interviews.index)
print "Number of rows in intervention dataframe:", len(interventions.index)
```

```
Number of rows in HDRS dataframe: 36
Number of rows in YDRS dataframe: 35
Number of rows in interview dataframe: 647
Number of rows in intervention dataframe: 59
```

Figure 8.1. Length of the data sets after importing the data from the files

8.2. Data Cleaning Results

In the data cleaning portion, there were very few uncompleted values within the data sets, and the empty ones did not display a very big challenge as the distribution was very similar within the data set and they may easily be filled. The entirety process of data cleaning returned a clean data set for each data set that was gathered: "Episode dataset, Young dataset, Hamilton dataset, Interviews dataset and Interventions dataset".

Clean data set:

```
episodes                    #print table of episodes
```

	patient	start	end	episode
0	D	2017-07-01 00:00:00	2017-07-24 00:00:00	D
1	D	2017-08-15 00:00:00	2917-09-11 00:00:00	D
2	G	2017-07-24 00:00:00	2017-08-07 00:00:00	D
3	G	2017-09-04 00:00:00	2017-11-01 00:00:00	M
5	M	2017-06-07 00:00:00	2017-07-01 00:00:00	M
6	M	2017-07-14 00:00:00	2017-07-30 00:00:00	D
7	M	2017-09-25 00:00:00	2017-10-10 00:00:00	D

Figure 8.2. Sample of clean EDS

code	date	euphoria	hyperactivity	sexual_impulse	sleep	irritability	verbal_expression	language	thought	aggressiveness	appearance	illness_awareness
G	2017-06-06 00:00:00	0	1	0	0	2	2	0	0	0	0	0
G	2017-06-16 00:00:00	0	0	0	0	0	2	0	0	0	0	0
M	2017-06-21 00:00:00	0	1	0	0	0	2	0	0	2	0	0
M	2017-06-21 00:00:00	0	1	0	0	0	2	0	0	2	0	0
D	2017-06-26 00:00:00	1	1	0	0	0	2	0	0	0	0	0

Figure 8.3. Sample of clean YDS

code	date	depressed_mood	guilt	suicide	precocious_insomnia	medium_insomnia	verbal_expression	language	thought	...	retardation	agitation
G	2017-06-06 00:00:00	0	0	0	0	0	0	0	0	...	0	0
G	2017-06-16 00:00:00	0	0	0	0	0	0	0	0	...	0	0
M	2017-06-21 00:00:00	0	0	0	0	0	0	0	0	...	0	0
M	2017-06-21 00:00:00	0	0	0	0	0	0	0	0	...	0	0

Figure 8.4. Sample of clean HDS

	mood	motivation	attention	irritability	anxiety	sleep_quality	nr_cigarettes	caffeine	alcohol	other_drugs	patient	date	active_time
274	0	1	1	1	1	2	0	90	0	0	G	2007-07-17 00:00:00	1470
230	-1	-1	2	1	1	3	24	120	0	0	D	2017-01-02 00:00:00	1790
241	0	-1	2	1	1	1	24	90	0	0	D	2017-01-15 00:00:00	1545
0	2	2	3	3	3	3	34	150	0	0	D	2017-06-01 00:00:00	1710

Figure 8.5. Sample of clean IDS

	code	date	gaf	relief
0	D	2017-06-01 00:00:00	80	5
1	D	2017-06-01 00:00:00	80	5
20	G	2017-06-06 00:00:00	80	1
68	M	2017-06-07 00:00:00	60	3

Figure 8.6. Sample of clean IVDS

8.3. Data Analysis Results

In data analysis showed some interesting relationships between certain features in the data sets that we visualized. Especially, the Heatmap of feature correlations within the hamilton data set (see Figure 5.12), appeared a possible relationship between the work/activities performed by the patient and a depressed mood, which we confirmed with the minimal plot (see Figure 5.13). Within the comparing section, usually a relationship that creates a lot of sense because when a patient feels depressed, he or she usually dedicates less time to work and other activities. This investigation helped us understand the data much better and visualize the most important connections between features in all the data sets gathered for the project with the assistance of diverse types of plots and graphs.

8.4. Data Combination Results

The data combination resulted in some exceptionally small data sets, just like the combination of Interviews and Interventions, which only had fourteen rows, as seen in Figure 8.7. The rest of the combinations did not have a lot of entries either, but they were interesting for future implementations, as they contained a lot of features.

The size of the data sets made it conceivable to select the ones that we might use for the algorithm application and comparison, which in the end were the YMRS, HDRS, Interviews and Interventions data sets, as well as the combination of YMRS and HDRS and the combination of Interviews and Interventions.

```
print "Length of the combined data set (interviews and interventions): ",
len(interviews_interventions)
```

```
Length of the combined data set (interviews and interventions):

14
```

Figure 8.7. Length of the combination of Interviews and Interventions

8.5. Implementation of Algorithm Results

With the help of the study of the different algorithms and their executions, we were able to compare the data sets to see which ones performed better and may be further applied for foreseeing randomized test data. The Interview data set performed very well, accompanied by a prediction accuracy over 60% with all the algorithms used. The rest of the data sets did not perform as well as the Interview data set, which we might conclude is because of the lack of data. In future usage of these algorithms, the forecast accuracies will most likely be higher thanks to new data. The best way to compare the accuracies obtained by the different algorithms with all data sets were to make an algorithm performance table, which is appeared within the figure.

Algorithms/Datasets	Logistic Regression	Decision Tree	Random Forest	SVM	KNN	Average
YMRS	83%	53%	52%	60%	64%	**63%**
HDRS	70%	45%	45%	51%	57%	**54%**
Interviews	73%	68%	69%	69%	59%	**68%**
Interventions	67%	33%	50%	45%	40%	**47%**
YMRS-HDRS	61%	70%	63%	63%	38%	**60%**
Interviews-Interventions	100%	58%	41%	42%	66%	**62%**
Average	**76%**	**55%**	**54%**	**55%**	**54%**	

Table 8.1. Performance Algorithm Table

This table showed that, on average, the data set that returned the leading forecast accuracy (68%) was the Interview data set, as seen on the right column. The calculation that performed the best, moreover in average (76%), was the Logistic Regression algorithm, as seen on the row farthest down. Even the algorithm that had the best accuracy in average was the Logistic Regression, we stated that the Random Forest algorithm made the most exact predictions, within the sense that they were very sensible given the behavior of the patients, which we tested with randomized data. These predictions made possible the implementation of a small program with the Random Forest classifier that we obtained, which can be utilized by doctors with real patients.

9. CONCLUSION AND FUTURE WORK

9.1. Conclusions

In conclusion, we can state that having a deep understanding of the data we are working with Machine Learning project focused on a branch of medical science like psychiatry, where knowing behavior is normal and abnormal in the patients can help us create much more precise prediction models.

With different parameters or indicators easily identify actual condition of patient and its helpful for psychiatrists to diagnose actual state of patient either depression or mania or euthymic state.

In this project, proposed techniques or algorithms are helpful to get good accuracy scores from datasets. In the same way, the data is important, having deep perception of the theory behind each algorithm used and their different implementations were crucial in order to get the models to perform in the best way possible.

9.2. Future Work

The most immediate use of the results obtained in this project would be to train the same algorithms used with larger amounts of data, in order to see if they perform in a similar way.

Gathering objective data from devices like mobile phones or wristbands is something that can be accomplished quite easily. The goal of this would be to compare the performance of different algorithms on the objective data gathered from these devices with the performance results obtained on the subjective data used in this project.

As for other more indirect applications of the results gotten during this project, the implementation of a drug recommending system for patients with Bipolar Disorder might be made by predicting the states in which the patients are in during a certain period of time. These predictions may be stored in a database which also contains the medicine that these patients have been prescribed with during the same period of time, hence giving the possibility of checking which drugs are better for the patient.

BIBLIOGRAPHY AND REFERENCES

Bibliography:

1.M. Bonsall et al., Bipolar disorder dynamics: affective instabilities, relaxation oscillations and noise, vol. 12, Journal of The Royal Society Interface, The Royal Society, 2015.

2.Kraepelin E. Manic-depressive insanity and paranoia. New York: E & S Livingstone; 1921

3.https://searchenterpriseai.techtarget.com/definition/machine-learning-ML

4.https://searchenterpriseai.techtarget.com/definition/deep-learning-deep-neural-network

5.www.bip4cast.org (Last accessed: 23/05/2018).

6.J. Anchiraico, Diseño de una Arquitectura Big Data para la Predicción de Crisis en el Trastorno Bipolar, Trabajo de Fin de Master, E-Prints Universidad Complutense de Madrid, 2016.

7.A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder' published by National Center for Biotechnology Information. link: https://pubmed.ncbi.nlm.nih.gov/30546013/

8.Design of BD Architecture for Prediction of Crises in Bipolar Disorder' link: https://www.mdpi.com/2227-7390/9/11/1174/pdf

9.A. Martínez, Introduction to Big Data and First Steps in a Big Data Project, Trabajo de Fin de Grado, Madrid: E-Prints Universidad Complutense de Madrid, 2016.

10.Python Software Foundation, Python Language Reference, version 2.7 (http://www.python.org).

11.F. Pérez and B.E. Granger, IPython: A System for Interactive Scientific Computing, Computing in Science & Engineering, 2007.

12.W. McKinney, Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 2010.

13. J. D. Hunter, Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 2007.

14. F. Pedregosa et al., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 2011.

15.Research paper by P.J Moore published by National Library of Medicine. Link: https://pubmed.ncbi.nlm.nih.gov/22855220/

16.DL by Depression Detection. Link: https://aclanthology.org/W18-0609.pdf

17.https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0148195&type=printa%5b7%5dble

18.An Item Response Theory Evaluation of the Young Mania Rating Scale in the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD). link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5048523/ . https://www.semanticscholar.org/paper/DeepMood%3A-Forecasting-Depressed-Mood-Based-on-via-Suhara-Xu/a656cab6ec7f825eb54de24f8224246e3acef31b?p2df.

19.https://www.researchgate.net/publication/350639348_Review_on_Machine_Learning_Techniques_to_predict_Bipolar_Disorder

20.https://www.researchgate.net/publication/311336418_Classification_of_Schizophrenia_and_Bipolar_Disorder_by_Using_Machine_Learning_Algorithms?enrichId=rgreq-7be31377809c41f3a35a86e966fbbe1e-XXX&enrichSource=Y292ZXJQYWdlOzMxMTMzNjQxODtBUzo1MDg0MjA2Nzc5MzEwMDhAMTQ5ODIyODM0MzAyMg%3D%3D&el=1_x_2&_esc=publicationCoverPdf

21.Python string to datetime - strptime() – JournalDev

22.https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/

23.Detection of Bipolar Disorder Using Machine Learning with MRI. URL: http://ceur-ws.org/Vol-2786/Paper53.pdf

24.https://www.slideshare.net/MohammedARajab/bipolar-disorder-59525559

25.https://www.slideshare.net/vaishnavisnair/manic-depressive-psychosis-bipolar-disease

26.Implementing machine learning in bipolar diagnosis. URL: https://pubmed.ncbi.nlm.nih.gov/31740657/

27.Garima, H. Gulati and P. K. Singh, Clustering techniques in data mining: A comparison, New Delhi: 2015 2nd International Conference on Computing for Sustainable Global Development, 2015, pp. 410-415.

28.J. R. Quinlan, Induction of Decision Trees, vol. 1, Hingham, Massachusetts: Kluwer Academic Publishers, 1986, pp. 81-106.

29.L. Breiman, Random Forests, vol. 45, Berkeley, California: Machine Learning, 2001, pp. 5-32.

30.C. Hsu, C. Chang and C. Lin, A Practical Guide to Support Vector Classification, Taipei: National Taiwan University, 2003.

31.C. Peng, K. Lee and G. Ingersoll, An Introduction to Logistic Regression Analysis and Reporting, vol. 96, Bloomington, Indiana: The Journal of Educational Research, 2002, pp. 3-14.

32.T. Patil and S. Sherekar, Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, vol. 6, International Journal Of Computer Science And Applications, 2013, pp. 256-261.

33.P. Bühlmann and T. Hothorn, Boosting Algorithms: Regularization, Prediction and Model Fitting, vol. 22, Zürich: Seminar für Statistik, ETH Zürich, 2007, pp. 477-505.

34.M. Stone, Cross-validatory choice and assessment of statistical predictions, vol. 36, London: Journal of the Royal Statistical Society, Royal Statistical Society, 1974, pp. 111-147.

35.P. Refaeilzadeh, L. Tang and H. Lui, Cross-Validation, Encyclopedia of Database Systems, Springer US, 2008, p. 532–538.

36.AT&T Labs Research, GraphViz: Graph Visualization Software

37.B. Li et al., Scalable Random Forests for Massive Data, vol. 7301, Berlin: PAKDD 2012: Advances in Knowledge Discovery and Data Mining, Springer, 2012, pp. 135-146.

38.V. Vapnik, The nature of statistical learning, vol. 6, New York: Springer- Verlag, 1995.

39.L. Auria and R. Moro, Support Vector Machines (SVM) as a Technique for Solvency Analysis, vol. 811, Berlin: Deutsches Institut für Wirtschaftsforschung, 2008.

40.D. Hosmer and S. Lemeshow, Applied logistic regression, vol. 10, New York: Wiley, 1989, pp. 1162-1163.

41.J. W. Tukey, Exploratory Data Analysis, vol. 23, London: Addison-Wesley Publishing Company Reading, 1977, pp. 413-414.

42.M. Hamilton, A rating scale for depression, Journal of Neurology, Neurosurgery, and Psychiatry, 1960, pp. 56-62.

43.R. C. Young et al., A rating scale for mania: reliability, validity and sensitivity, British Journal of Psychiatry, 1978, pp. 429-435.

44.P. Grof et al., Mood Charting and Technology: New Approach to Monitoring Patients with Mood Disorders, vol. 2, Current Psychiatry Reviews, 2006, pp. 423-429.

45.M. Faurholt-Jepsen et al., Smartphone data as objective measures of bipolar disorder symptoms, vol. 217, Psychiatry Research, 2014, pp. 124-127.

46.A. Doryab et al., Impact factor analysis: combining prediction with parameter ranking to reveal the impact of behaviour on health outcome, vol. 19, Personal and Ubiquitous Computing, 2014, pp. 355-365.

47. J. Nicholas et al., Mobile Apps for Bipolar Disorder: A Systematic Review of Features and Content Quality, vol. 17, Journal of medical Internet research, 2015, p. 198.

48.A. Grünerbl et al., Smart-Phone Based Recognition of States and State Changes in Bipolar Disorder Patients, vol. 19, IEEE Journal of Biomedical and Health Informatics, 2014, pp. 140-148.