**Classification and Prediction of Bipolar Disorder using Machine learning Algorithms**

## A PROJECT REPORT

***Submitted by***

**Ms.Jyothi.S - 2016CSE220**

**Mr.Priyanka.B - 2016CSE147**

### *Under the guidance of*

**Mr. Jobin Thomas**

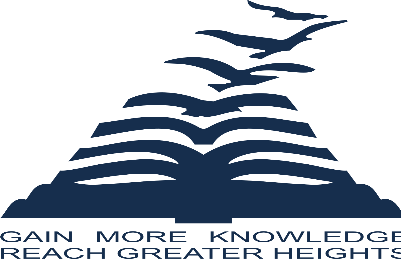
***in partial fulfillment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

**At**



**Department of Computer Science and Engineering**

**School of Engineering**

**PRESIDENCY UNIVERSITY**

**BANGALORE**

**MAY 2020**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SCHOOL OF ENGINEERING**

**PRESIDENCY UNIVERSITY**

**CERTIFICATE**

This is to certified that the Project report **“Classification and Prediction of bipolar disorder using machine learning algorithms”** being submitted by “**Jyothi.S and Priyanka.B”** bearing roll number(s): 2016CSE220 and 2016CSE147, in partial fulfillment of requirement for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** is a bonafide work carried out under my supervision.

|  |  |
| --- | --- |
| **Dr. Mohan K G**  HOD & Professor  Department of CSE  Presidency University | **Mr. Jobin Thomas**  Guide  Assistant Professor  Department of CSE |

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SCHOOL OF ENGINEERING**

**PRESIDENCY UNIVERSITY**

**DECLARATION**

I hereby declare that the work, which is being presented in the project report entitled **Classification and Prediction of bipolar disorder using machine learning algorithms** in partial fulfilment for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering**, is a record of our own investigations carried under the guidance of **Mr. Jobin Thomas ,Asst. Professor, Department of Computer Science and Engineering, School of Engineering, Presidency University, Bangalore.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

|  |  |
| --- | --- |
|  | **Jyothi.S, 2016CSE220**  **Priyanka.B,2016CSE147** |

**Index**

**Index............................................................................................................................ 3**

**Index of figures........................................................................................................... 6**

**Abstract........................................................................................................................1**

**Keywords .....................................................................................................................1**

**1. Introduction .............................................................................................................2**

**1.1. Motivation..............................................................................................................2**

**1.2. Work flow ............................................................................................................. 2**

**1.3. Data used for the project ..................................................................................... 3**

**1.4. Used technologies ................................................................................................. 3**

1.4.1. Python ...........................................................................................................3

1.4.2. Jupyter Notebook...........................................................................................4

**2.Literature Survey ......................................................................................................5**

**2.1. Similar projects and studies ................................................................................ 5**

**2.2. Proposed System.....................................................................................................6**

**3. Data Gathering ..........................................................................................................7**

**3.1. Gathering the data .................................................................................................7**

3.1.1. Episode data set .............................................................................................8

3.1.2. YMRS data set ...............................................................................................8

3.1.3. HDRS data set ................................................................................................8

3.1.4. Interview data set ........................................................................................... 15

3.1.5. Intervention data set ...................................................................................... 17

**4.Data Cleaning..............................................................................................................32**

**4.1.Cleaning the data......................................................................................................18**

4.1.1. Episode data set ............................................................................................. 18

4.1.2. YMRS data set ............................................................................................... 18

4.1.3. HDRS data set ............................................................................................... 19

4.1.4. Interview data set ........................................................................................... 20

4.1.5. Intervention data set ...................................................................................... 22

**5. Exploratory Data Analysis .......................................................................................23**

**5.1. YMRS data set ...................................................................................................... 23**

5.1.1. Summary Statistics.......................................................................................... 23

5.1.2. Histograms ..................................................................................................... 24

5.1.3. Heatmap ......................................................................................................... 25

5.1.4. Scatterplots .................................................................................................... 26

5.1.5. Marginal plot and 2D density plot........................................................ 26

**5.2. HDRS data set ...................................................................................................... 29**

5.2.1. Summary Statistics............................................................................... 29

5.2.2. Histograms ..................................................................................................... 29

5.2.3. Heatmap ......................................................................................................... 30

5.2.4. Scatterplots .................................................................................................... 31

5.2.5. Marginal plot and 2D density plot...................................................... 32

**5.3. Interview data set ................................................................................................. 36**

5.3.1. Summary Statistics.............................................................................. 36

5.3.2. Histograms ..................................................................................................... 36

5.3.3. BoxPlot........................................................................................................... 37

5.3.4. Heatmap ......................................................................................................... 38

5.3.5. Scatterplots .................................................................................................... 39

5.3.6. Marginal plot and 2D density plot......................................................... 40

**5.4. Intervention data set ............................................................................................ 41**

5.4.1. Summary Statistics ...........................................................................................41

5.4.2. Marginal plot and 2D density plot..........................................................41

**6.Combination of Dataset.............................................................................................. 42**

**6.1. YMRS and Episode data sets .............................................................................. 42**

**6.2. HDRS and Episode data sets ............................................................................... 43**

**6.3. Interview and Episode data sets .......................................................................... 45**

**7.Model training and Building........................................................................................49**

**7.1. Decision Tree ........................................................................................................ 51**

7.1.1. YMRS ............................................................................................................ 51

7.1.2. HDRS ............................................................................................................ 53

7.1.3. Interviews ...................................................................................................... 54

7.1.4. Interventions .................................................................................................. 55

**7.2. Random Forest ..................................................................................................... 56**

7.2.1. Prediction accuracy comparison .................................................................... 56

**7.3. SVM ....................................................................................................................... 56**

7.3.1. Prediction accuracy comparison .................................................................... 57

**7.4. Logistic Regression ............................................................................................... 58**

7.4.1. Prediction accuracy comparison .................................................................... 59

**8. Model Testing on Randomized data........................................................................ 60**

**8.1. Accuracy Table........................................................................................................63**

**9.Model Deployment.................................................................................................... 64**

**9.1. Final test with real data......................................................................................... 65**

**10.Results and Conclusion............................................................................................ 67**

10.1. Results............................................................................................................. 67

10.2. Conlusion.........................................................................................................68.

10.2. Future work..................................................................................................... 68

**11. Bibliography .................................................................................................... 69**

**Index of Figure**

Figure 1.1. Project process .............................................................................................……....2

Figure 1.2. Jupyter Notebook logo .............................................................................................4

Figure 2.1. Proposed System……………………………………………...................................6

Figure 3.1. Sample of the original Excel file with interview data (Part 1) .................................7

Figure 3.2. Sample of the original Excel file with interview data (Part 2) .................................7

Figure 3.3. Mania and depression episodes in patients ...............................................................8

Figure 3.4. Fragment of the YMRS patient data (Part 1) ............................................................11

Figure 3.5. Fragment of the YMRS patient data (Part 2) ............................................................11

Figure 3.6. Fragment of the HDRS patient data (Part 1) .............................................................15

Figure 3.7. Fragment of the HDRS patient data (Part 2) .............................................................15

Figure 3.8. Interview data (Part 1) ...............................................................................................16

Figure 3.9. Interview data (Part 2) ...............................................................................................17

Figure 3.10. Interventions.............................................................................................................17

Figure 4.1. Column name translation of the Episode data set ......................................................18

Figure 4.2. Formatting the dates in the Episode data set ..............................................................18

Figure 4.3. Dropping Observaciones column from the YMRS data set........................................18

Figure 4.4. Column name translation of the YMRS data set ........................................................19

Figure 4.5. Column name translation of the HDRS data set .........................................................19

Figure 4.6. Filling empty values in the HDRS data set .................................................................20

Figure 4.7. Column name translation of the Interview data set .....................................................20

Figure 4.8. Completing missing values in the Interview data set ..................................................21

Figure 4.9. Column name translation of the Intervention data set .................................................22

Figure 4.10. Dropping entries from interventions that patients did not attend to...........................22

Figure 4.11. Taking the numerical value from the relief column ..................................................22

Figure 5.1.Describing the ymrs data frame....................................................................................24

Figure 5.2. Histogram distribution in the YMRS data set ..............................................................24

Figure 5.3. YMRS data set correlation heatmap ............................................................................25

Figure 5.4. YMRS scatterplot matrix .............................................................................................26

Figure 5.5. Marginal plot of hyperactivity and irritability in the YMRS data set .............27

Figure 5.6. 2D kernel density plot of irritability and hyperactivity in YMRS data set ……..........27

Figure 5.7. Marginal plot of verbal expression and euphoria in the YMRS data set .....................28

Figure 5.8. 2D kernel density plot of verbal expression and euphoria in the YMRS data set .......28

Figure 5.9. Describe the HDRS Data frame.....................................................................................29

Figure 5.10. Histogram distribution in the YMRS data set .............................................................29

Figure 5.11. HDRS data set correlation heatmap ............................................................................30

Figure 5.12. Scatter plot for HDRS dataset......................................................................................31

Figure 5.13. Marginal plot of depressed mood and work in the HDRS data set .............................32

Figure 5.14. 2D kernel density plot of depressed mood and work in the HDRS data set……........32

Figure 5.15. Marginal plot of depressed mood and psychic anxiety in the HDRS data set ........... 33

Figure 5.16. 2D density plot of depressed mood and psychic anxiety in the HDRS data set..........33

Figure 5.17. Marginal plot of somatic anxiety and psychic anxiety in the HDRS data set ............34

Figure 5.18. 2D density plot of somatic anxiety and psychic anxiety in the HDRS data set….….34

Figure 5.19. Marginal plot of Somatic General symptoms and genital symptomsin the HDRS data set ........................................................................................................................................................ 35

Figure 5.20. 2D density plot of Somatic General symptoms and genital symptoms in the HDRS data set…………………………………………………………………………………………............35

Figure 5.21. Describing Interview dataset…………………………………………………..........36

Figure 5.22. Histogram for Interview dataset…………………………………....………….........36

Figure 5.23. Finding the outliers using box plot for Interview Dataset..........................................37

Figure 5.24. Checking if there are outliers in the Interview data set .............................................37

Figure 5.25. Changing the value of the outliers in the Interview data set .....................................38

Figure 5.26. Interview data set correlation heatmap ..................................................................... 38

Figure 5.27. Interviews scatterplot matrix .................................................................................... 39

Figure 5.28. Marginal plot of motivation and mood in the Interview data set ..............................40

Figure 5.29. 2D kernel density plot of motivation and mood in the Interview data set ……...….40

Figure 5.31. Describing Intervention Data Frame………………………......................................41

Figure 5.32. Marginal plot of relief and GAF in the Intervention data set ....................................41

Figure 5.33. 2D kernel density plot of relief and GAF in the Intervention data set ..................... 41

Figure 6.1. Merging the YMRS and Episode data sets .....................................................42

Figure 6.2. Distribution of verbal expression and euphoria on different patient states in the YMRS data set .........................................................................................................................................................42

Figure 6.2. Distribution of verbal expression and aggressiveness on different patient states in the YMRS data set ...........................................................................................................................................43

Figure 6.4. Distribution of depressed mood and work on different patient states in the HDRS data set ....................................................................................................................................................... 43

Figure 6.5. Distribution of depressed mood and psychic anxiety patient states in HDRS

data set .......................................................................................................................................... 44

Figure 6.6. Distribution of somatic anxiety and psychic anxiety patient states in HDRS dataset………......................................................................................................................44

Figure 6.7. Distribution of mood and motivation on different patient states in the Interview data set ..........................................................................................................................................................45

Figure 6.8. Mood and motivation bar plot with patient states in the Interview data set …....….....46

Figure 6.9. Anxiety and sleep quality barplot with patient states in the Interview dataset......……46

Figure 6.10. Anxiety and Irritability barplot with patient states in the Interview data set ..........…47

Figure 6.11. GAF and relief bar plot with patient states in the Intervention dataset.….…48

Figure 6.12. GAF and relief bar plot with patient states in Intervention dataset ...............48

Figure 7.1. Diagram of the Machine Learning algorithm application process ...............................50

Figure 7.2. Splitting the YMRS data set into training and testing data sets ...................................50

Figure 7.3. Model training with Decision Tree classifier ...............................................................52

Figure 7.4. Decision Tree graph of YMRS data .............................................................................52

Figure 7.5. Decision Tree graph of HDRS data ..............................................................................53

Figure 7.6. Decision Tree graph of Interview data .........................................................................54

Figure 7.7. Decision Tree graph of Intervention data..................................................................... 55

Figure 7.8. Prediction accuracies of the Decision Tree algorithm ..................................................56

Figure 7.9. Prediction accuracies of the Random Forest algorithm ................................................57

Figure 7.10. Finding the largest possible margin to the hyperplane................................………....57

Figure 7.11. Prediction accuracies of the SVM algorithm ..............................................................57

Figure 7.12. Basic diagram of one vs. all Multi-class Logistic Regression classification ..............59

Figure 7.13. Prediction accuracies of the Logistic Regression algorithm ......................................59

Figure 8.1. Random patient interview data generation ...................................................................60

Figure 8.2. Prediction of a Depression episode on a patient with the lowest mood value .60

Figure 8.3. Prediction of Depression episode on a patient with low mood and motivation values .........................................................................................................................................................61

Figure 8.4. Prediction of a Mania episode on an active patient with high anxiety and irritability levels ...................................................................................................... ………………………..............61

Figure 8.5. Prediction of a Euthymic state on a patient with normal behaviour ...........................62

Figure 8.6. Prediction of a Euthymic state on a patient with clear depression symptoms.............62

Figure 8.7. Accuracy Matrix of overall algorithms ...........................................................63

Figure 9.1 Model deployment………………………………………………………….....64

Figure 9.2. Parameters for Prediction…………………………………………………...65

Figure 9.3. Prediction for a Depression episode….…………………………………......65

Figure 9.4. Prediction of a Mania episode ........................................................................66

Figure 9.5. Prediction of a Euthymic state .....................................................................................66