A Machine Learning Approach for Identifying the Determinants of Bipolar Disorder among University Students in Bangladesh

https://github.com/ashib11/Bipolar-Disorder/

S M Ashibur Rahman
221-15-5137
CSE, Dept. of CSE
Daffodil International University
Birulia, Savar, Dhaka-1216

***Abstract*—Mental health disorders, particularly bipolar disorder, are a growing concern among university students in Bangladesh, significantly impacting their academic performance and daily lives. This study employs machine learning techniques to identify the key determinants of bipolar disorder within this demographic. A dataset was collected through survey responses, covering factors such as social life, academic stress, depressive symptoms, and lifestyle habits. This research provides critical insights into the underlying contributors to bipolar disorder, aiding early diagnosis and intervention efforts among university students in Bangladesh.**

***Index Terms*—Determinants of Bipolar Disorder Synthetic Minority Oversampling Technique (SMOTE) Random Forest Academic Pressure Mood swings**

I. Introduction

Mental health is an important issue that affects many individuals, especially university students. One of the most common and challenging mental health conditions is bipolar disorder. This disorder is characterized by significant mood swings, ranging from manic highs to deep depressive lows, which can profoundly impact a person's daily life. University students often face unique pressures from academic demands, social situations, and the transition into adulthood. These pressures can exacerbate mental health challenges, making students particularly vulnerable during this critical period.

Research has shown that untreated mental health issues can lead to academic difficulties, strained social relationships, and long-term health concerns. In Bangladesh, as in many developing countries, addressing mental health issues presents substantial challenges due to limited awareness, persistent stigma, and insufficient mental health care resources. University students often experience heightened stress from academic responsibilities, financial concerns, and family expectations. However, there is a lack of comprehensive research focused on the specific factors contributing to bipolar disorder among these young adults. Understanding these factors is crucial for implementing early interventions and preventive strategies to impact their long-term well-being positively. Thus, more research on the determinants of bipolar disorder in this population is essential to mitigate the long-term effects of this condition.

Machine learning has the potential to analyze large datasets and identify complex patterns, making it a valuable tool for understanding the factors that contribute to mental health disorders. By employing advanced algorithms, researchers can pinpoint key elements associated with bipolar disorder, leading to valuable, data-driven insights. Additionally, using interpretable models enables translating intricate machine learning results into practical, actionable guidance for mental health practitioners. This approach enhances our understanding of mental health and assists professionals in applying these insights effectively in their practice.

This study aims to explore the determinants of bipolar disorder among university students in Bangladesh using a machine-learning approach. We collected a survey-based dataset that captures factors such as academic stress, social interactions, depressive symptoms, and lifestyle habits. To ensure a balanced analysis, we applied the Synthetic Minority Oversampling Technique (SMOTE) and utilized classification algorithms, including Random Forest, to analyze the data.
The findings of this study highlight critical factors associated with bipolar disorder and provide actionable insights for mental health interventions. This research seeks to uncover the determinants of bipolar disorder among university students in Bangladesh, contributing to improved mental health outcomes for this population.

II. Research Gap

- **Knowledge Gap:**
  Despite extensive research on bipolar disorder, there's a significant gap in understanding its specific determinants among university students in Bangladesh. Most studies focus on general populations or regions with strong mental health support, overlooking the unique challenges these students face. Utilizing advanced machine learning techniques to explore these factors presents an oppor-

tunity to gain valuable insights, ultimately enhancing mental health awareness and support in this community. This focus could lead to meaningful improvements for both students and researchers.

- **Population Gap:**
  In addition to the knowledge gap, there is a notable lack of research specifically focusing on university students in Bangladesh regarding the factors that contribute to bipolar disorder within this demographic. University students in Bangladesh experience unique stressors, such as academic pressure, financial burdens, and familial expectations, all of which are compounded by limited mental health awareness and stigma surrounding mental illness. These challenges have yet to be addressed in the context of bipolar disorder, resulting in a significant gap in understanding the mental health needs of this group.

This study aims to fill both the knowledge and population gaps by focusing on the Bangladeshi university student population and using advanced machine learning techniques to identify the critical determinants of bipolar disorder. By doing so, this research will contribute novel insights that are crucial for developing targeted interventions and improving mental health support for students in Bangladesh.

III. Literature Review

The mental health of college students is an important concern, with studies highlighting the connection between stress, academic performance, and well-being. Research focuses on identifying factors affecting mental health and using predictive models to address challenges like stress, anxiety, and depression. This literature review summarizes findings from ten papers, exploring different methodologies and outcomes.

Stress from finances, relationships, and academic pressures significantly impacts mental health among college students. The World Health Organization's International College Student Initiative found that 93.7% of first-year students reported stress in at least one area. Stress prevention interventions could reduce mental disorder prevalence by up to 80% over a year [1]. However, the cross-sectional design of these studies limits causal interpretations, underscoring the need for longitudinal research and targeted interventions to tackle chronic stressors.

In another study, we found a strong correlation between academic stress and mental health [2]. Research conducted during the COVID-19 pandemic revealed significant differences in psychological well-being among groups. Non-binary students and second-year students reported the highest levels of stress. Although the use of standardized scales like PAS and SWEMWBS provided reliable measurements, the study's reliance on self-reported data and the limited diversity among participants underscored the need for broader, more inclusive sampling and longitudinal research designs.

Emotional intelligence, social skills, and learning styles are crucial for students' psychological well-being, [3] and it's exciting to see that Spanish university students with high emotional intelligence or those who participated in cooperative learning reported greater well-being! Although anxiety is still a challenge during this vibrant phase of life, this study offers valuable insights. To build on this promising work, future research can broaden its reach by gathering multi-site data, helping to amplify the impact of these findings. Together, we can discover even more ways to support students' growth and happiness!

Physical inactivity and mental health issues are critical public health challenges globally. While regular physical exercise is well-known for its benefits, limited epidemiological research exists on its relationship with mental health problems, particularly regarding suicidality. A large-scale study using data from the SHoT2018 survey in Norway (N = 50,054, ages 18–35) provides insights into this connection. [4]

- **Frequency of Exercise:** This was the most influential factor. Women with low levels of physical activity had nearly three times the risk of experiencing psychological distress and depression compared to those exercising almost daily. Men exhibited even stronger effects.
- **Duration and Intensity:** These factors also correlated with better mental health outcomes but had smaller effect sizes than frequency.

The study reveals a negative association between physical exercise and mental health problems, establishing a dose-response relationship.

Academic performance is an important metric in educational institutions, especially for students with bipolar disorder who encounter distinct challenges due to mood swings, concentration difficulties, and fluctuations in motivation. [5] This study assessed various algorithms, including logistic regression, random forest classifier, decision tree classifier, K-Nearest Neighbors (KNN), AdaBoost, Extra Tree classifier, GaussianNB, and BernoulliNB. The evaluation utilized 13 performance metrics, such as accuracy, precision, recall, F1 score, sensitivity, specificity, and several error rates. The decision tree classifier proved to be the most effective model, surpassing the others in predicting academic performance.
The study identified several attributes influencing academic outcomes: [5]

- **Study Hours and Extracurricular Activities**
- **Internet Access and Family Interaction**
- **Drug Use and Psychological Motivation**

The research highlights machine learning's role in fostering inclusive and supportive educational environments, leading to proactive student support systems.

A study investigated the use of machine learning (ML) algorithms to predict depression and anxiety among university students in Bangladesh, with a focus on identifying the most

effective models. The research found that severe depression and anxiety were more prevalent among female students compared to their male counterparts. Additionally, students aged 21 to 25 years and those living with their families were particularly vulnerable to severe mental health issues. [6]

Another study [7] underscores the significance of mental health and the effects of disorders like depression and anxiety. It highlights how big data analytics and machine learning can be used to predict mental health issues and enhance risk classification. This paper will explore the challenges, effectiveness, and limitations of employing machine learning techniques to identify mental health conditions.

- **Bipolar I Disorder:** This type is characterzied by mood episodes that range from mania to deprssion, with manic episodes lasting at least seven days.
- **Bipolar II Disorder:** This involves milder hypomanic episodes that alternate with server depressive episodes.
- **Cyclothymic Disorder:** This type features breif period of hypomanic episodes accomapnied by depressive symptoms
- **Mixed Features:** This category combines symptoms of hypomania, depression and mania.
- **Rapid Cycling:**

IV. Proposed Methodology

The dataset for this study comprises responses from university students in Bangladesh. All participants completed the questionnaires online through a digital and self-administered format. The questionnaire includes various questions related to psychological and demographic factors. Collected variables include demographic information (e.g., age, gender, academic year), academic stress, social interactions, lifestyle habits, and mental health status.
The target variable is: ”14. Have you been diagnosed with depression?” This variable directly reflects the mental health status of the participants and was used to classify students into two groups: those diagnosed with depression and those not diagnosed. The independent variables consist of stress levels,

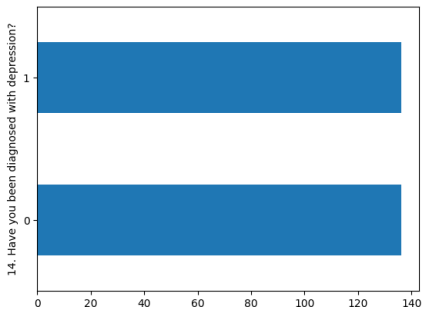

Fig. 1. Target Variable

sleep quality, academic performance, social support, and lifestyle habits. These variables were analyzed to determine their association with the target variable. The dataset was preprocessed to address missing values, encode categorical variables, and normalize numerical data, ensuring it was suitable for statistical and machine learning analyses.

In the Data Preprocessing phase, the dataset was loaded using the pandas library. Any leading or trailing spaces in the columns were removed. The dataset was then inspected for any null or missing values. Since all the questionnaires required responses, there were no null or missing values in the dataset.
As the dataset contained both numerical and categorical values, I utilized a ”Label Encoder” to convert the categorical data into numerical format. Upon testing the dataset, I discovered an imbalance in the target variable. To address this issue, I applied the Synthetic Minority Over-Sampling Technique (SMOTE).

```
Class distribution after SMOTE:
14. Have you been diagnosed with depression?
0    136
1    136
Name: count, dtype: int64
```

Fig. 2. Class distribution after SMOTE

Feature selection was conducted through a combination of domain knowledge and statistical techniques. First, the importance of each feature was assessed based on the feature importance scores from the trained model. The features were then ranked in descending order of their significance for the prediction task. Additionally, univariate feature selection and correlation analysis were employed to identify and retain the most relevant features. This approach ensured that only the most impactful features were included, which improved model performance and reduced the risk of over-fitting.

	Feature	Importance
4	5. How much stress do you experience in the fo...	0.076477
7	18. How many hours do you sleep on average per...	0.075146
2	10. Please indicate how often you experience t...	0.073866
1	21. How would you describe your living environ...	0.071958
3	10. Please indicate how often you experience t...	0.071312
10	20. Over the past two weeks, how often have yo...	0.068067
13	8. Over the last two weeks, how often have you...	0.067443
9	12. How often do you pay attention to your fee...	0.067381
12	8. Over the last two weeks, how often have you...	0.066914
14	22. How often do you go out with friends?	0.065826
6	7. How often do you have a drink containing al...	0.063878
0	5. How much stress do you experience in the fo...	0.063018
8	4. What year of study are you in?	0.061811
5	20. Over the past two weeks, how often have yo...	0.057642
11	15. Have you ever made an attempt to take your...	0.049261

Fig. 3. Features selection

The dataset was divided into training and testing sets to evaluate model performance. Various models, including Logistic Regression, Decision Trees, Linear Regression, Random Forest, K-Nearest Neighbors (KNN), and

Support Vector Machines (SVM), were used to predict outcomes. Scaling techniques, such as Min-Max scaling or StandardScaler, were applied to normalize the features when necessary. The models were then trained on the training set using the selected features.

During the evaluation phase, model performance was assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Among the applied models, the Random Forest (RF) algorithm achieved the highest accuracy of 83.6%. Based on this result, the features were further

```
[30] accuracy_score(ytest, rc.predict(xtest))

0.8363636363636363

[31] print(classification_report(ytest, rc.predict(xtest)))

              precision    recall  f1-score   support

           0       0.79      0.88      0.84        26
           1       0.88      0.79      0.84        29

    accuracy                           0.84        55
   macro avg       0.84      0.84      0.84        55
weighted avg       0.84      0.84      0.84        55
```

Fig. 4. Classification Report for Random Forest

refined, prioritizing those with higher importance as identified by the Random Forest model. To gain deeper insights into the factors influencing the target variable, the LIME (Local Interpretable Model-Agnostic Explanations) technique was applied, highlighting the most impactful features affecting predictions.

V. Result

The results from the analysis indicate that the Random Forest (RF) algorithm achieved the highest accuracy among the applied machine learning models, with a score of 83.6%, as shown in Table 1. Other algorithms, such as Decision Tree (76.3%), KNN (72%), and SVC and Logistic Regression (69%), demonstrated varying levels of accuracy, while Naïve Bayes performed the lowest at 63%.

Algorithm	Accuracy (%)
Decision Tree	76.3
Random Forest	83.6
SVC	69.0
Logistic Regression	69.0
KNN	72.0
Naïve Bayes	63.0

TABLE I
ACCURACY OF DIFFERENT MACHINE LEARNING ALGORITHMS

The classification report for the Random Forest model, presented in Figure 2, provides additional insights into its performance. The model achieved a precision, recall, and F1-score of 0.84 for both classes (0 and 1). This consistency indicates the model's ability to balance false positives and false negatives effectively, with a macro average and weighted average of 0.84 across all metrics.

Additionally, the feature importance analysis highlighted the significant predictors influencing the target variable. LIME visualizations, as shown in Figure 1, provided interpretability by illustrating how individual features impact the model's decisions.

VI. Discussion

The study successfully applied machine learning techniques to predict mental health conditions, achieving optimal results with the Random Forest algorithm. The model's high accuracy and balanced performance metrics suggest its suitability for analyzing the dataset and addressing the research question.

The evaluation metrics, particularly the F1-score and AUC-ROC, indicate the robustness of the Random Forest model in distinguishing between individuals diagnosed with depression and those who are not. Moreover, the integration of LIME provided actionable insights, making it possible to identify key factors influencing the diagnosis. For instance, features related to stress levels, social interactions, and lifestyle habits emerged as significant predictors.

Figure 5 illustrates the Local Interpretable Model-Agnostic

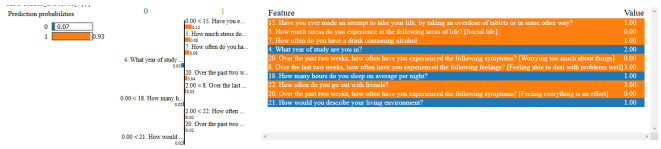

Fig. 5. Classification Report for Random Forest

Explanations (LIME) visualization for a prediction made by the Random Forest model. The left side shows the prediction probabilities for the target variable, where the model assigns a 93% probability for the positive class and 7% for the negative class.

On the right side, the contributing features are listed with their respective values and importance. Features like "Have you ever made an attempt to take your life?" and "How much stress do you experience in your social life?" are identified as highly impactful, with their contributions shown in orange for the positive class and blue for the negative class. The numeric values associated with each feature reflect their influence in determining the prediction.

This analysis highlights the key psychological and demographic factors contributing to the model's decision-making, providing insights into the primary drivers behind the classification of the target variable.

However, the study is not without limitations. While Random Forest delivered promising results, further validation on external datasets is essential to generalize the findings. Additionally, while LIME enhances model interpretability, the use of larger datasets or more advanced explanation techniques could provide deeper insights into causal relationships.

REFERENCES

[1] K. Chanasit, E. Chuangsuwanich, A. Suchato, and P. Punyabukkana, "A real estate valuation model using boosted feature selection," *IEEE Access*, vol. 9, pp. 86938–86953, Jan. 2021. Available: doi: 10.1109/access.2021.3089198.

[2] P.-F. Pai and W.-C. Wang, "Using machine learning models and actual transaction data for predicting real estate prices," *Applied Sciences*, vol. 10, no. 17, p. 5832, Aug. 2020. Available: doi: 10.3390/app10175832.

[3] P.-Y. Wang, C.-T. Chen, J.-W. Su, T.-Y. Wang, and S.-H. Huang, "Deep learning model for house price prediction using heterogeneous data analysis along with Joint Self-Attention Mechanism," *IEEE Access*, vol. 9, pp. 55244–55259, Jan. 2021. Available: doi: 10.1109/access.2021.3071306.

[4] N. H. Zulkifley, S. A. Rahman, N. H. Ubaidullah, and I. Ibrahim, "House Price Prediction using a Machine Learning Model: A Survey of Literature," *International Journal of Modern Education and Computer Science*, vol. 12, no. 6, pp. 46–54, Dec. 2020. Available: doi: 10.5815/ijmecs.2020.06.04.

[5] M. Thamarai and S. P. Malarvizhi, "House price prediction modeling using machine learning," *International Journal of Information Engineering and Electronic Business*, vol. 12, no. 2, pp. 15–20, Apr. 2020. Available: doi: 10.5815/ijieeb.2020.02.03.

[6] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique," *Procedia Computer Science*, vol. 199, pp. 806–813, Jan. 2022. Available: doi: 10.1016/j.procs.2022.01.100.

[7] F. Lorenz, J. Willwersch, M. Cajias, and F. Fuerst, "Interpretable machine learning for real estate market analysis," *Real Estate Economics*, vol. 51, no. 5, pp. 1178–1208, May 2022. Available: doi: 10.1111/1540-6229.12397.

[8] J. M.-T. Wu, Z. Li, N. Herencsar, B. Vo, and J. C.-W. Lin, "A graph-based CNN-LSTM stock price prediction algorithm with leading indicators," *Multimedia Systems*, vol. 29, no. 3, pp. 1751–1770, Feb. 2021. Available: doi: 10.1007/s00530-021-00758-w.

[9] Q. Zhang, "Housing price prediction based on multiple linear regression," *Scientific Programming*, vol. 2021, pp. 1–9, Oct. 2021. Available: doi: 10.1155/2021/7678931.

[10] P.-Y. Wang, C.-T. Chen, J.-W. Su, T.-Y. Wang, and S.-H. Huang, "Deep learning model for house price prediction using heterogeneous data analysis along with Joint Self-Attention Mechanism," *IEEE Access*, vol. 9, pp. 55244–55259, Jan. 2021. Available: doi: 10.1109/access.2021.3071306.