

# Student Mental Health Risk Classification Using Random Forest and BERTopic: A Tabular and Text Analysis Approach

Fairuz Arya Bagus Pramana<sup>1</sup>, Rumadi<sup>2</sup>, Dian Anggraini<sup>3</sup>

<sup>1</sup> Military Mathematic Student, FMIPA, Indonesian Defense University, Bogor

<sup>2</sup> Research Center For Data And Information Sains, Badan Riset Dan Informasi Nasional

<sup>3</sup> Military Mathematic Lecturer, FMIPA, Indonesian Defense University, Bogor

fairuz.pramana@mipa.idu.ac.id

Diterima: 31-08-2026 ; Direvisi: 15-09-2026; Dipublikasi: 22-09-2026

## Abstract

Student mental health has become a critical issue affecting academic performance and quality of life. Many previous studies have only used numerical data and have not explored textual data that reflects students' subjective conditions. This study develops a classification model using Random Forest on tabular data to predict the risk of student mental health issues. The dataset consists of 500 students with a training-to-testing data split of 80:20. The model uses features such as academic performance, sleep patterns, and psychological scores. The results show that the model achieves a macro F1 Score of 0.90 despite class imbalance. The most influential variables are depression scores, stress levels, and anxiety. In addition, the BERTopic method was used separately to analyze text data and identify thematic patterns from student responses. However, not all topics generated were directly related to mental health. The model also has limitations in detecting minority classes due to data imbalance. This study provides a classification approach supported by text analysis to comprehensively understand students' mental health conditions.

**Keywords:** Random Forest; BERTopic; Mental Health

## Abstrak

Kesehatan mental mahasiswa telah menjadi isu kritis yang memengaruhi prestasi akademik dan kualitas hidup. Banyak penelitian sebelumnya hanya menggunakan data numerik dan belum mengeksplorasi data teks yang mencerminkan kondisi subjektif mahasiswa. Penelitian ini mengembangkan model klasifikasi menggunakan Random Forest data tabular untuk prediksi risiko kesehatan mental mahasiswa. Dataset yang digunakan terdiri dari 500 mahasiswa dengan pembagian data latih dan uji sebesar 80:20. Model menggunakan fitur performa akademik, pola tidur, dan skor psikologis. Hasil penelitian ini menunjukkan bahwa model mencapai macro F1 Score 0.90 dengan kelas yang tidak seimbang. Variabel paling berpengaruh adalah skor depresi tingkat stres, dan kecemasan. Selain itu, metode BERTopic digunakan secara terpisah untuk menganalisis data teks dan mengidentifikasi pola tematik dari jawaban mahasiswa. Namun, tidak semua topik yang dihasilkan berkaitan langsung dengan kesehatan mental. Model juga memiliki keterbatasan dalam mendeteksi kelas minoritas akibat ketidakseimbangan data. Penelitian ini memberikan pendekatan klasifikasi yang didukung analisis teks untuk memahami kondisi kesehatan mental mahasiswa secara komprehensif.

**Kata Kunci:** Random Forest; BERTopic; Kesehatan Mental.

## 1. INTRODUCTION

The mental health of students is an important issue that directly impacts their quality of life and academic performance. Academic pressure, disciplinary demands, and high social burdens can trigger an increased risk of mental disorders among students (Figueira et al., 2020). This condition, if not detected early, can lead to a decrease in learning motivation and an increase in dropout rates. Conventionally, detecting the risk of mental health disorders often relies on research methods that are limited to a single type of data (Russ et al., 2019). Therefore, a modern system capable of detecting the risk of student mental health disorders with more data and accuracy is needed.

Regarding this matter, the development of machine learning technology opens new opportunities in the development of predictive models to analyze mental health conditions. One of the algorithms widely used for modeling and prediction is Random Forest (Mohana et al., 2021). This algorithm has a high capability in handling large and complex data and producing good classification accuracy (Tiwari et al., 2020). Most studies using numerical data with the Random Forest algorithm have not provided an understanding of the factors influencing mental health. For example, the research conducted by Tadesse et al. (2019) used Random Forest to detect depression based on social media data, specifically the Twitter platform. The results showed that Random Forest was able to classify users who were potentially experiencing depression with a high level of accuracy. However, this study focuses on classification without further analyzing the topics that emerge in the text data.

In text data analysis, the Natural Language Processing (NLP) approach is increasingly evolving with modern topic modeling methods such as BERTopic. The BERTopic modeling approach can be used for the automatic extraction of dominant topics from a collection of unstructured texts (Liu et al., 2016). The application of BERTopic uses transformer-based models, such as BERT, which produce better semantic representations than conventional methods (George & Sumathy, 2023). The advantage of the BERTopic method is its ability to capture contextual meanings and synonyms that are often overlooked by frequency-based approaches (Zhang & Milios, 2023). Additionally, BERTopic is highly flexible and allows for adaptation to various embedding models and clustering techniques (Gao & Sazara, 2023). Therefore, BERTopic is a suitable choice for applying complex text analysis, such as student mental health data. This study uses the BERTopic approach to analyze text data consisting of open-ended responses from students (Tadesse et al., 2019). It is expected that the main topics representing common psychological issues faced by students will be identified. These findings are expected to serve as a basis for decision-making or intervention. By using the data, the model can recognize patterns that indicate the risk of mental disorders, allowing the system to alert the relevant parties about mental health issues (Liao et al., 2021).

Based on these issues, this study has the following research questions:

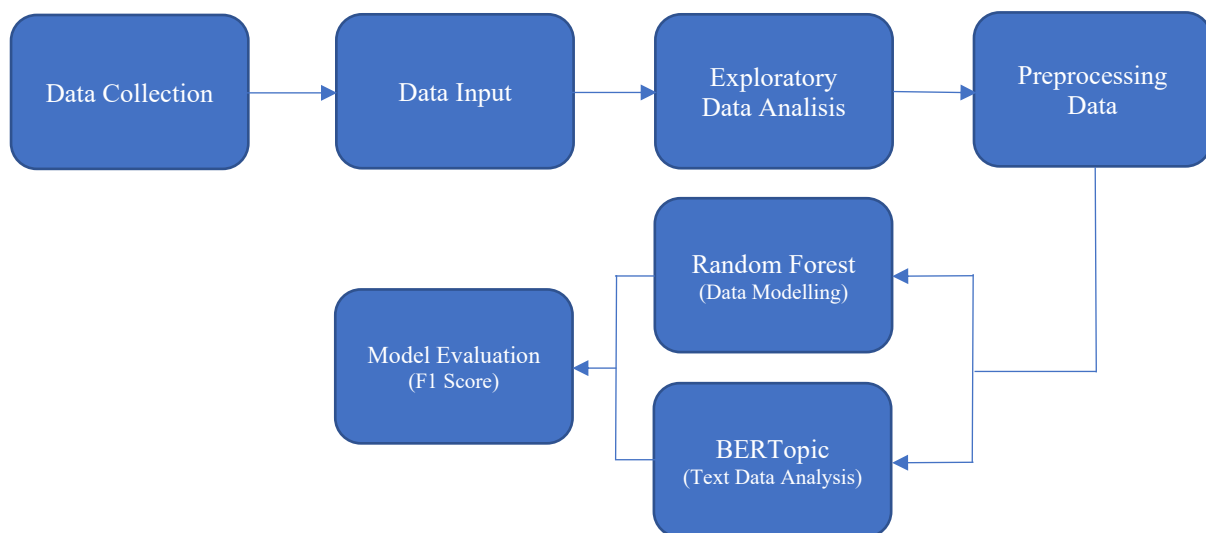
1. What factors or variables are most dominant in influencing the prediction of students' mental health risks?
2. What are the main topics generated from students' open-ended responses (free-text daily reflections) using the BERTopic approach?

This study is expected to classify students mental health conditions using the Random Forest algorithm with high accuracy and identify the main topics that emerge in the text data using the BERTopic algorithm.

## 2. METHOD

This research aims to build a mental health risk classification model for students using tabular data, supported by text analysis to provide additional understanding of students' psychological conditions. The approach used is complementary, where the numerical data-based classification model is combined with text analysis separately.

The stages of the research are illustrated in Figure 1 as follows:



**Figure 1.** Research Stages

### 2.1 Data Collection

The dataset used in this research was obtained from the Kaggle Platform and consists of 500 student respondents. The dataset consists of 11 variables divided into 4 dimensions, namely: Psychological Dimension (Depression Score, Anxiety Score, Sentiment Score, Stress Level, Mood), Physical Dimension (Sleep Hours, Physical Activity in 1 Day), Academic (GPA, Age, Gender), Free-text Daily Reflections (Open statements from students).

### 2.2 Exploratory Data Analysis

Exploratory Data Analysis in this research includes the Distribution of Mental Health Status, which encompasses the distribution of students' mental health conditions to provide an overview of the respondents' mental health based on the dataset used. The

purpose of analyzing the distribution is to obtain the percentage of each risk category in mental health. Next, the results of the topic extraction are converted into numerical features and combined with tabular data to form a multimodal dataset used in the modeling process.

### **2.3 Data Preprocessing**

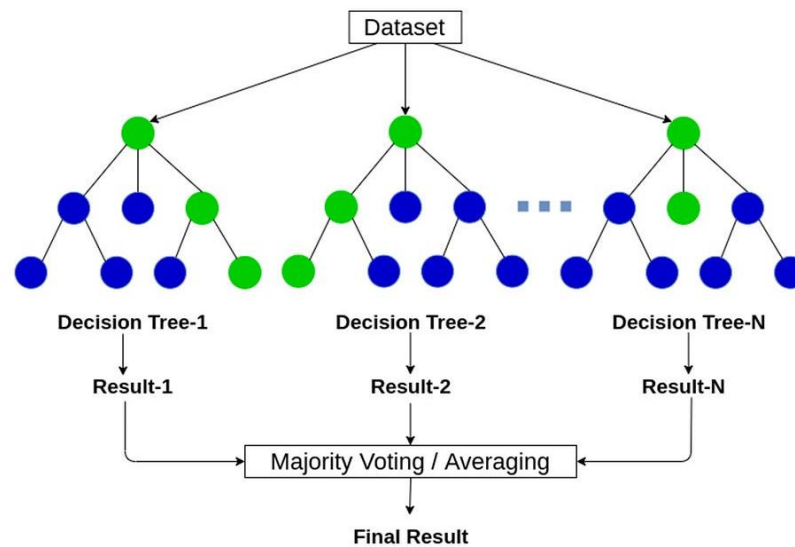
The preprocessing stage in this research is carried out by separating the dataset into two, namely text data and tabular data. The text data derived from Free-text Daily Reflections is used for analysis using the BERTopic method. The tabular data, which includes numerical and categorical variables such as those in the Psychological Dimension, Physical Dimension, and Academic, is used as input for the Random Forest algorithm. The separation of this dataset consists of:

- a. Text Data: Data in the form of open-ended answers or essays regarding the conditions of students, which will later be processed using the BERTopic Method. In this study, the text data consists of open statements from students about their psychological conditions and experiences.
- b. Tabular Data: Data in the form of variables or numerical scores that serve as input and are processed using the Random Forest Method. In the conducted research, the tabular data consists of gender, age, mood, sleep hours, GPA, physical activity in one day, depression score, anxiety score, sentiment score, and stress level.

### **2.4 Classification Modeling With Random Forest**

The classification model in this study is built using the Random Forest algorithm. This model is an ensemble learning method based on decision trees. The first step in the modeling process is to divide the dataset into training data consisting of 400 data points (80%) and test data consisting of 100 data points (20%). After being separated, the model is trained to learn the pattern of relationships between input variables and mental health risk categories. The next stage involves testing the model using the test data to evaluate the model's classification performance.

Random Forest works by building a number of decision trees based on different data samples and combining the prediction results to produce the final decision. This approach uses variables such as gender, age, mood, sleep hours, GPA, physical activity in one day, depression score, anxiety score, sentiment score, and stress level modeled with Random Forest.



**Figure 2.** Random Forest Method

#### 2.4.1 Random Forest

Random Forest is an algorithm that focuses on ensemble learning consisting of a number of decision trees (Mohana et al., 2021). This method involves creating various decision trees using different data samples, followed by synthesizing the prediction results from each decision tree to produce the final decision (Tajima et al., 2024). The benefit of the random forest model is that it helps avoid overfitting due to the combination of multiple decision trees. In addition, this algorithm is capable of handling datasets that contain many variables. This method has the following formula:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B T_b(x)$$

Keterangan:

$\hat{f}$  = Final Prediction results for the Random Forest Model

$B$  = Total number of decision trees

$T_b(x)$  = The prediction results from the decision tree

The Random Forest algorithm was initially configured using hyperparameters to mitigate the risk of overfitting prior to model analysis. Full details regarding the hyperparameter structure used in this modeling are presented in the following table:

**Table 1.** Hyperparameter Random Forest

| <b>HYPERPARAMETER</b> | <b>VALUE</b> |
|-----------------------|--------------|
| <b>n_estimator</b>    | 300          |
| <b>max_depth</b>      | 20           |
| <b>random_state</b>   | 42           |

## 2.5 Text Data Analysis with BERTopic

In the data analysis stage, the BERTopic approach is used to identify and extract main topics based on text data collected from students' open-ended responses. The results of this type of data analysis consist of various topics, marked by the key terms used. Compared to other approaches for analyzing text data, BERTopic can provide more detailed results due to its ability to determine the context of the words used by the students (Figueira et al., 2020).

### 2.5.1 BERTopic

BERTopic is a topic modeling technique based on unsupervised learning that utilizes deep learning algorithms to extract main themes from unstructured text collections (Mersha et al., 2024). BERTopic has been applied to analyze text documents to identify topics in mental disorders. The BERTopic approach uses transformers (such as BERT) to generate document vectors that help in understanding the semantic meaning of the involved words (Xu & Mou, 2020). The BERTopic approach follows the following formula:

$$W_{t,c} = tf_{t,c} \log(1 + \frac{A}{\sum_i tf_{t,i}})$$

Keterangan:

$tf_{t,c}$  = Frequency of word  $t$  in cluster  $c$

$tf_{t,i}$  = Total frequency of word  $t$  across all clusters

$A$  = Average number of words per cluster ( $\frac{\sum_c N_c}{C}$ )

This process begins with the use of a sentence transformer embedding model to convert text into numerical vectors. Next, dimensionality reduction is performed using the UMAP algorithm ( $n\_neighbors = 15$ ,  $n\_components = 5$ ), followed by clustering using HBDSCAN ( $min\_cluster\_size = 10$ ,  $min\_samples = 5$ ). The final results for each cluster are represented using the c-TF-IDF technique to identify the most relevant keywords for each topic that emerges.

## 2.6 Model Evaluation

To determine the performance of the output obtained from the data modeling process, the developed model undergoes an evaluation stage to ensure that there are accurate data predictions. This evaluation includes the use of Accuracy, Precision, Recall, and F1-Score metrics. Additionally, a confusion matrix is used to evaluate the errors present in the predictions generated by the model (Mortaz, 2020). This is essential for the

implementation of an efficient mental health risk prediction system. Measurement can be conducted thru a formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (4)$$

True Positives (TP) are the total correct predictions for a specific category, True Negatives (TN) are the total correct predictions that are not for a specific category (excluding TP from that category), False Positives (FP) are the total incorrect predictions indicating members of a specific category (excluding TP from that category), and False Negatives (FN) are the total incorrect predictions indicating not members of a specific category (excluding rows and columns from that category).

### 3. RESULTS AND DISCUSSION

#### 3.1 Dataset Description

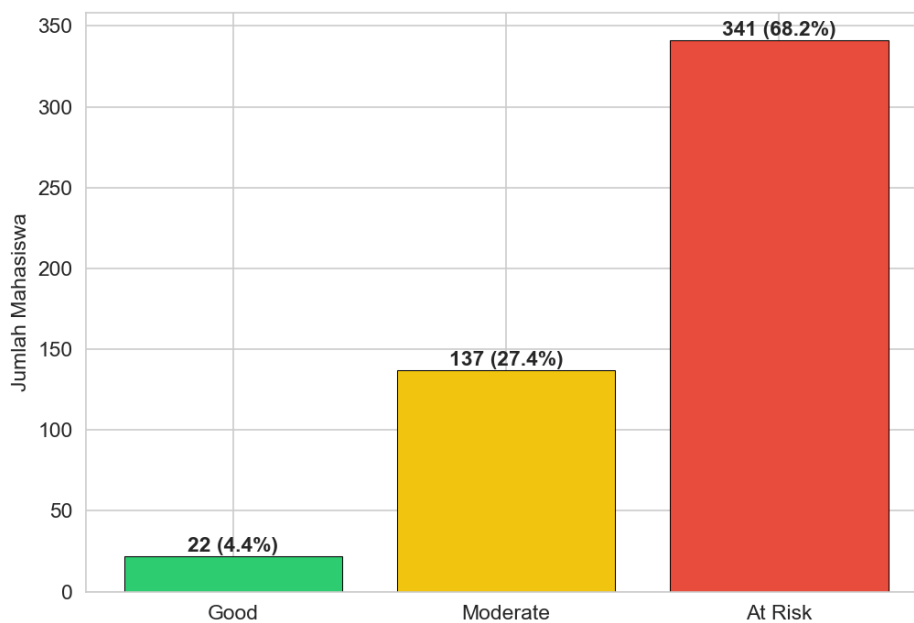
The data source used in this analysis was obtained from the Kaggle website, as it offers a dataset on college students' mental health. The dataset used is synthetic that is, real data that does not contain any personal information and consists of data from 500 respondents, divided into two types of data. The tabular data includes gender, age, mood, hours of sleep, GPA, daily physical activity, depression score, anxiety score, sentiment score, and stress level. On the other hand, the text data consists of open-ended statements from students regarding their psychological condition.

**Table 2. Mental Health Dataset**

| N<br>o | Ag<br>e | Gend<br>er | GP<br>A  | Stre<br>ss<br>leve<br>l | Anxie<br>ty<br>level | Depress<br>ion<br>Level | Sentim<br>ent<br>Score | Daily<br>Reflectio<br>n                                                          | Slee<br>p<br>Hou<br>rs | Ste<br>ps<br>per<br>Day | Mood<br>Descript<br>ion |
|--------|---------|------------|----------|-------------------------|----------------------|-------------------------|------------------------|----------------------------------------------------------------------------------|------------------------|-------------------------|-------------------------|
| 1      | 23      | Other      | 2.5<br>2 | 5                       | 20                   | 6                       | 0                      | Onto<br>foreign do<br>environme<br>ntal<br>anyone<br>every<br>nearly<br>another. | 6.8                    | 416<br>6                | Happy                   |
| 2      | 19      | Male       | 2.7<br>4 | 5                       | 3                    | 7                       | 0.47                   | Party but<br>others<br>visit admit<br>industry                                   | 5.1                    | 494<br>9                | Tired                   |

|         |      |            |          |     |     |     |       | country<br>very.                                                                                                            |     |          |          |
|---------|------|------------|----------|-----|-----|-----|-------|-----------------------------------------------------------------------------------------------------------------------------|-----|----------|----------|
| ...     | .... | ...        | ...      | ... | ... | ... | ...   | ...                                                                                                                         | ... | ...      | ...      |
| 49<br>9 | 18   | Fema<br>le | 2.4<br>5 | 4   | 14  | 0   | 0..49 | Care can<br>now<br>outside<br>real rest<br>that<br>perform.                                                                 | 6   | 501<br>3 | Stressed |
| 50<br>0 | 18   | Other      | 3.5<br>2 | 1   | 0   | 5   | 0.44  | Get turn<br>Congress<br>list mouth<br>city<br>decision<br>easy grow<br>travel<br>trade start<br>resource<br>home<br>family. | 5.6 | 544<br>0 | Anxious  |

### 3.2 Distribution of Mental Health Status



**Figure 3.** Distribution of Mental Health Status

In Figure 3, there is a graph showing the number of students based on their mental health. The data analysis results indicate that the largest number of students falls into the AtRisk group, with 341 students or 68.2%. Thus, the majority of the analyzed students

are at risk of mental health issues. The second largest group consists of students with Moderate mental health, totaling 137 students or 27.4% of the entire sample. This means there is a chance for these students to experience a decline in their mental condition, but not all of them are aware of it. Lastly, the smallest portion consists of 22 students who have Good mental health. This shows that there is an imbalance in the dataset, where the Risky category is far more numerous than the other categories. This is a significant factor because it can impact the accuracy of the classifier, particularly in predicting the majority category.

### 3.3 Random Forest Classification Results

The main result of the Random Forest algorithm is the Feature Importance Chart, which displays a bar graph with the most influential variables. This visualization provides students with an overview to identify the main factors that trigger mental health disorders. In addition, the Result Visualization also displays the Confusion Matrix, which is a technique to show the results of the prediction model based on the comparison between the predicted values and the actual values. In the Confusion Matrix, there are four main components: TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative). These components provide insights into the accuracy and errors made by the model during predictions. By using the Confusion Matrix, the model's predictive ability in categorizing the risk levels related to the mental well-being of each group of students can be displayed. In addition, the Confusion Matrix allows for the detection of common errors and the improvement of prediction performance.

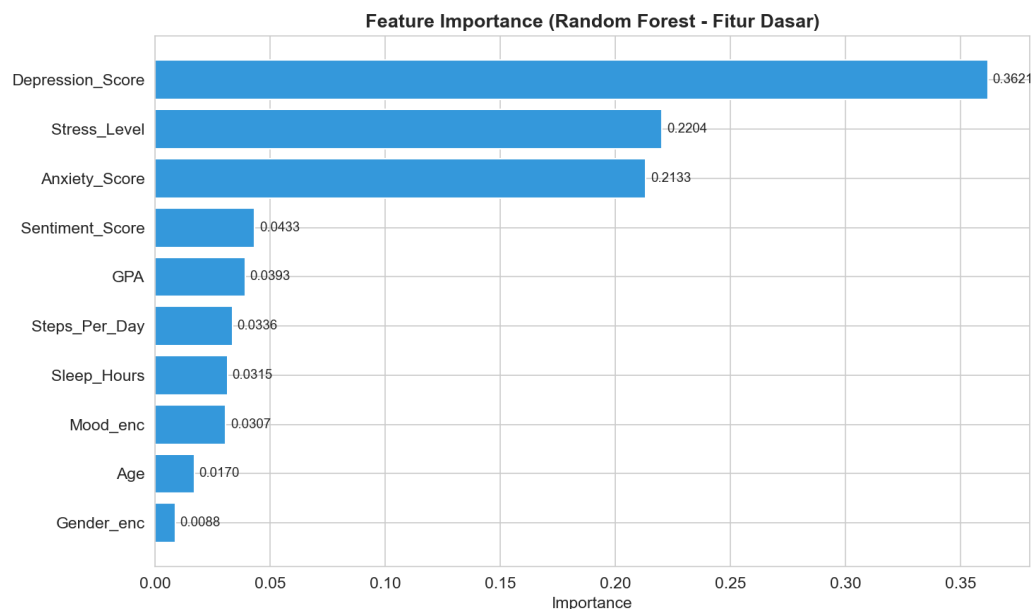
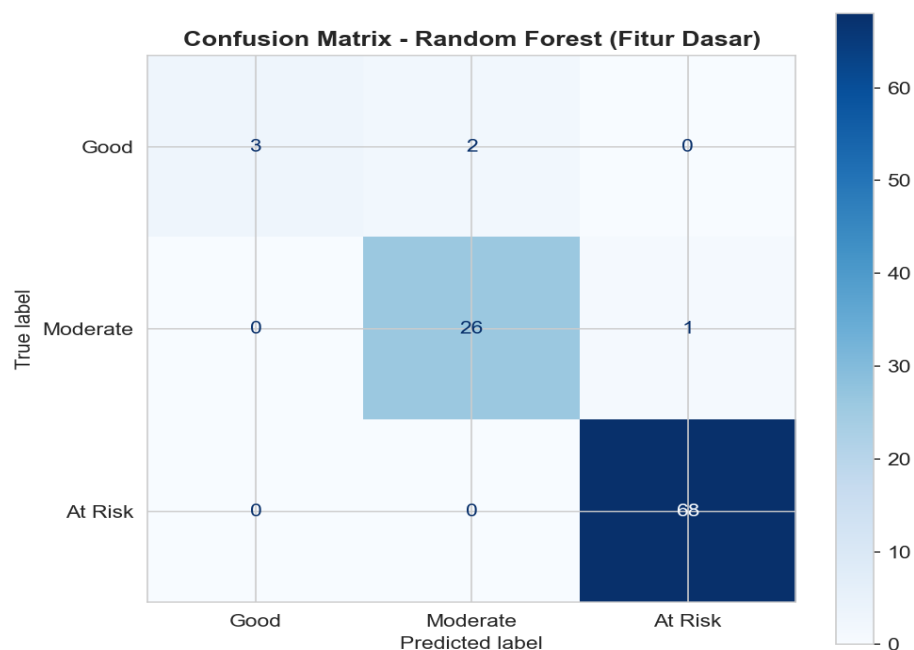


Figure 4. Grafik Feature Importances

As seen in Figure 4, the modeling results conducted with the Random Forest algorithm show that the key variable is the Depression\_Score variable, with an importance level of

0.3621, which means that depression will most determine the category of mental health risk. The Stress\_Level variable ranks second with an importance level of 0.2204, and the third is the Anxiety\_Score with an importance level of 0.2133. From these three variables, it can be concluded that the internal psychological factors of the students are the main factors considered in this case. In addition, several other variables have a lower level of importance compared to those mentioned. These variables include Sentiment\_Score with an importance of 0.0433, GPA 0.0393, Steps\_Per\_Day 0.0336, and finally Sleep\_Hours 0.0315. Mood\_enc is another important variable with a significance level of 0.0307. However, its importance is still lower compared to the key variables. The importance of the Age variable is 0.0170, and the least important variable in the model is Gender\_enc, which has an importance of only 0.0088. Therefore, based on these results, it can be concluded that variables related to psychological disorders such as depression, stress, and anxiety have a much greater impact compared to demographic variables or those related to lifestyle.

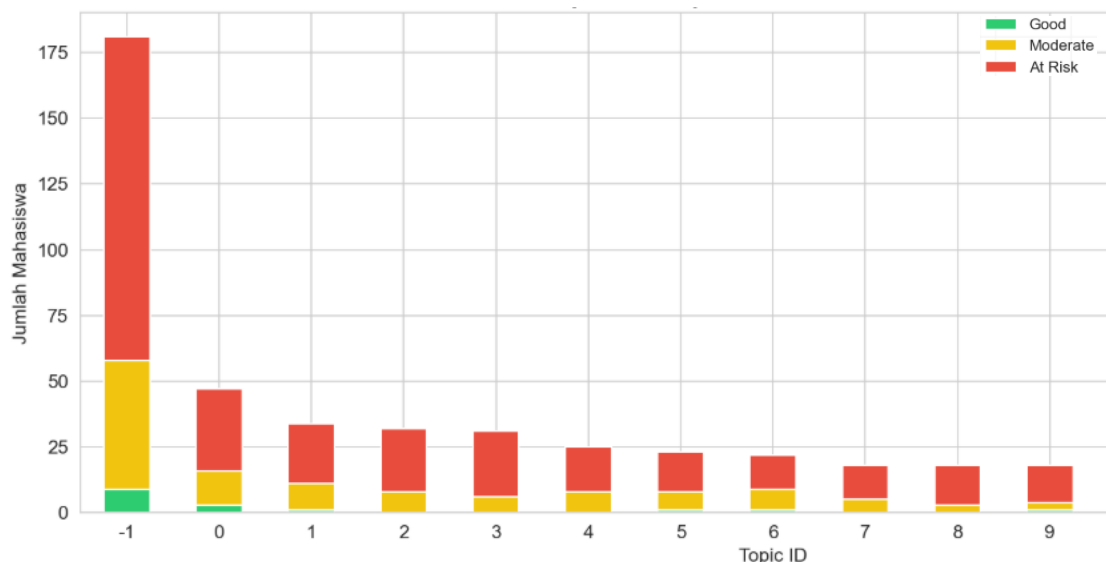


**Figure 5.** Confusion Matrix Prediction Results

The evaluation results using the confusion matrix in Figure 5 show that the model correctly predicted 3 cases but incorrectly predicted 2 cases as Moderate. Conversely, the Moderate classification had 26 cases correctly predicted while 1 case was incorrectly predicted as At Risk. The At Risk classification showed perfect results with 68 cases correctly predicted without any errors. Overall, out of 100 test data, there were 97 correct predictions and 3 incorrect predictions, resulting in an accuracy rate of 97%

### 3.4 Text Data Analysis using BERTopic

The data analysis phase involved analyzing text data in the form of the “Daily Reflection” variable, which captures students’ psychological states, using the BERTopic method to identify the main topics related to the issues students face and to explore the information and patterns found in the text data, thereby providing a deeper understanding of students’ mental health.



**Figure 6.** Distribution of BERTopic Topics by Mental Health Status

The visualization in Figure 6 shows the relationship between text topics extracted using BERTopic and the distribution of students’ mental health statuses (Good, Moderate, and At Risk). From the graph, it is clear that the largest group of respondents falls under Topic ID -1, which represents outliers topics that the HDBSCAN algorithm could not classify into specific clusters. Meanwhile, for the cluster-based topics (Topic IDs 0 through 9), the distribution tends to be more even, with the number of respondents ranging from 15 to 50 students per topic. Topic ID 0, represented by the keyword “politics,” has a relatively high number of students compared to other specific topics, though the pattern remains dominated by the “At Risk” and “Moderate” groups. Overall, the At Risk group dominates almost all of the topic clusters formed. However, distribution analysis shows that Topic ID 0 and Topic -1 (outlier) are the clusters most populated by the At Risk and Moderate groups. This indicates that mental health issues faced by students are more diverse among those in the At Risk and Moderate groups. On the other hand, the Good group only appears very minimally and sporadically, mainly in Topic -1, Topic 0, and Topic 1. The results of the analysis using BERTopic identified the following topics:



**Figure 7.** Results of the BERTopic text data analysis

The analysis results show topics using the BERTopic approach, which involves dividing the text data into ten topics with words that carry certain weights within those topics. For example, in Topic 0, words like republican (0.06), democrat (0.045), and congress

(0.035) represent themes related to politics. On the other hand, words in Topic 1 include investment (0.045), hand (0.044), and store (0.041), indicating themes related to the economy or business transactions. Words like sport (0.06), shoulder (0.05), and threat (0.048) represent Topic 2 and indicate physical actions or the competitive nature of the events. Additionally, words like fish (0.07), energy (0.05), and religious (0.048) dominate Topic 3 and can be associated with the environment or spirituality. Topic 4 has words like culture (0.05), improve (0.047), and fine (0.042). Care, laugh, and describe have been rated most frequently in Topic 5, indicating the emotionality and uniqueness of that topic. Meanwhile, for Topic 6, several terms including lawyer (0.07), case (0.06), and attorney (0.05) appear, showing a legal character. Topic 7 features the most frequently used words, namely dog (0.10), coach (0.095), and scientist (0.07). The last three words can define topics related to professional activities. High scores for the words middle (0.09), career (0.08), and sort (0.06) describe the topic of personal growth careers. Finally, Topic 9 is marked by high scores of words such as kid (0.09), daughter (0.08), and industry (0.06), indicating a topic about family or social life. Thus, it should be noted that these topics display varying frequencies of word assessments. This shows that there is a lot of diversity in the topics discussed by the analyzed text data, which includes topics of a psychological, social, and even everyday activity nature.

### 3.5 Model Evaluation

In the model evaluation stage, the performance of the Random Forest algorithm in predicting the risk of mental health disorders among students is measured.

**Table 3.** Random Forest Evaluation Results

|              | Precision | recall | F1-score | support |
|--------------|-----------|--------|----------|---------|
| Good         | 1.00      | 0.60   | 0.75     | 5       |
| Moderate     | 0.93      | 0.96   | 0.95     | 27      |
| At Risk      | 0.99      | 1.00   | 0.99     | 68      |
| Accuracy     | -         | -      | 0.97     | 100     |
| Macro avg    | 0.97      | 0.85   | 0.90     | 100     |
| Weighted avg | 0.97      | 0.97   | 0.97     | 100     |

As seen in the evaluation table, the mental health risk classification model shows results for all categories, with performance analysis varying but remaining optimal. At Risk is the category with the highest number of samples (68 samples). For this specific category, the model made predictions with optimal Recall (1.00) and Precision (0.99). This can be seen in the Confusion Matrix where no samples from the At Risk category were misclassified into other categories. For the Moderate category, Precision is equal to 0.93, while Recall is equal to 0.96, with 26 out of 27 data points correctly predicted by the algorithm. The Good category also has optimal Precision (1.00), but lower Recall (0.60) due to two predictions incorrectly classified as Moderate. The low recall value is primarily due to the very small data size (5 respondents). However, with the help of the F1-Score

metric, which is equal to 0.90 (Macro Avg), it can be said that the model works well even when under pressure. In conclusion, this model is very successful in predicting medium to high risk, but there is still room for improvement regarding the low-risk category.

#### 4. CONCLUSION

Based on the results of the study, the Random Forest algorithm was able to predict mental health risk levels in tabular data with very high performance, achieving an accuracy of 97% and a macro F1-score of 0.90. Based on the Feature Importance analysis, internal psychological factors such as depression score (0.3621), stress level (0.2204), and anxiety score (0.2133) proved to be the most dominant variables influencing the model's classification decisions. In contrast, lifestyle and demographic variables such as sleep duration, daily step count, age, and gender made a significantly lower contribution. Meanwhile, text modeling using BERTopic successfully extracted ten main themes from students' daily reflections, covering various dimensions ranging from academics and careers to social interactions. However, the discovery of a dominant outlier (Topic -1) indicates that the students' text data exhibits very high variability and is not yet fully tied to students' psychological issues.

#### 5. RECOMMENDATION

From the results of the research conducted, the performance of the built classification model is very good, but there are several limitations that should be addressed for future research in this field. The most prominent limitation is related to the dataset imbalance, as the number of people in the low-risk category (good) is only 22 or 4.4% of the total sample size. As a result, the model's recall rate in predicting this category is quite low (0.60). This issue arises due to the tendency to recognize patterns in the majority class. Additionally, although the BERTopic algorithm was successfully applied to the text data, the data is still too diverse to solely reflect topics related to mental health, as it touches on various themes such as politics and economics, thus requiring additional effort in data preparation. Therefore, it is highly recommended that future research conduct balanced sampling across all categories.

Another limitation of this study is the use of a synthetic dataset. The dataset used in this research consists of 500 student respondents and was obtained from Kaggle. Although synthetic data can be useful for developing and testing machine learning models, it may not fully represent the real conditions experienced by students. Real students may have different backgrounds, psychological conditions, academic environments, and ways of expressing their feelings. Therefore, the good performance of the Random Forest model in this study, with an accuracy of 97% and a macro F1-score of 0.90, should be interpreted carefully. The use of synthetic data can also affect the generalization of the findings. A model that performs well on synthetic data does not necessarily have the same performance on real data. Therefore, future research should validate the model using larger and more diverse real-world datasets from different universities, regions, and student backgrounds. This would help determine whether the identified patterns and important variables, such as depression, stress, and anxiety, remain consistent in real student populations.

## 6. REFERENCES

- Figueira, G. M., Demarchi, M. E., Casselli, D. D. N., Silva, E. de S. M. e, & Souza, J. C. (2020). Fatores de risco associados ao desenvolvimento de transtornos mentais em estudantes universitários. *Research, Society and Development*, 9(9), e432997454. <https://doi.org/10.33448/rsd-v9i9.7454>
- Gao, X., & Sazara, C. (2023). *Discovering Mental Health Research Topics with Topic Modeling*. <http://arxiv.org/abs/2308.13569>
- George, L., & Sumathy, P. (2023). An integrated clustering and BERT framework for improved topic modeling. *International Journal of Information Technology (Singapore)*, 15(4), 2187–2195. <https://doi.org/10.1007/s41870-023-01268-w>
- Liao, S., Zhang, Q., & Gan, R. (2021). Construction of real-time mental health early warning system based on machine learning. *Journal of Physics: Conference Series*, 1812(1). <https://doi.org/10.1088/1742-6596/1812/1/012032>
- Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. In *SpringerPlus* (Vol. 5, Number 1). SpringerOpen. <https://doi.org/10.1186/s40064-016-3252-8>
- Mersha, M. A., yigezu, M. G., & Kalita, J. (2024). *Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms*. <http://arxiv.org/abs/2410.00134>
- Mohana, R. M., Reddy, C. K. K., Anisha, P. R., & Murthy, B. V. R. (2021). WITHDRAWN: Random forest algorithms for the classification of tree-based ensemble. *Materials Today: Proceedings*. <https://doi.org/10.1016/j.matpr.2021.01.788>
- Russ, T. C., Woelbert, E., Davis, K. A. S., Hafferty, J. D., Ibrahim, Z., Inkster, B., John, A., Lee, W., Maxwell, M., McIntosh, A. M., Stewart, R., Anderson, M., Aylett, K., Bourke, S., Burhouse, A., Callard, F., Chapman, K., Cowley, M., Cusack, J., ... Zammit, S. (2019). How data science can advance mental health research. *Nature Human Behaviour*, 3(1), 24–32. <https://doi.org/10.1038/s41562-018-0470-9>
- Sakthi Vel, S. (2021). Pre-Processing techniques of Text Mining using Computational Linguistics and Python Libraries. *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*, 879–884. <https://doi.org/10.1109/ICAIS50930.2021.9395924>
- Saraswat, P., & Raj, S. (2022). DATA PRE-PROCESSING TECHNIQUES IN DATA MINING: A REVIEW. *International Journal of Innovative Research in Computer Science & Technology*, 122–125. <https://doi.org/10.55524/ijircst.2022.10.1.22>
- Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7, 44883–44893. <https://doi.org/10.1109/ACCESS.2019.2909180>
- Tajima, K., Ichijo, N., Nakahara, Y., & Matsushima, T. (2024). *An Algorithmic Framework for Constructing Multiple Decision Trees by Evaluating Their Combination Performance Throughout the Construction Process*. <http://arxiv.org/abs/2402.06452>
- Tiwari, V., Garg, B., & Sharma, U. P. (2020). Significant Impact of Improved Machine Learning Algorithm in The Processes of Large Data Sets. *International Journal of Scientific Research*

- in Computer Science, Engineering and Information Technology*, 458–467.  
<https://doi.org/10.32628/cseit206133>
- Xu, Bing., & Mou, Kefen. (2020). *Proceedings of 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC 2020) : June 12-14, 2020, Chongqing, China*. IEEE Press.
- Zhang, X., & milios, E. (2023). *MPTopic: Improving topic modeling via Masked Permuted pre-training*. <http://arxiv.org/abs/2309.01015>
- Ziya (2025). Student Mental Health Survey [Dataset]. Kaggle.  
<https://www.kaggle.com/datasets/ziya07/student-mental-health-and-resilience-dataset>