Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Application of Categorical Naive Bayes for Classifying Student Stress Levels Based on Academic and Lifestyle Factors

Gladly Parmonangan[1]*, Muhamad Haviz[2], Yuki Saidi Moses[3], Giatika Chrisnawati[4], Sulaeman Hadi Sukmana[5]

[1,2,3,4,5] *Program Studi Infirmatika, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika,Indonesia*
gladlyparmonangan11@gmail.com[1*], muhamadhaviz250@gmail.com[2], yukimosez354@gmail.com[3], Giatika.gcw@bsi.ac.id[4], sulaeman.sdu@bsi.ac.id[5]

Abstract

Student stress has become an important issue in higher education due to increasing academic demands and lifestyle pressures. Early identification of student stress levels is necessary to support appropriate interventions. This study applies a machine learning approach to classify student stress levels based on academic and lifestyle factors using the Categorical Naive Bayes algorithm. The dataset was obtained from a student stress survey consisting of categorical attributes and stress level labels ranging from 1 to 5. Data preprocessing was conducted to ensure compatibility with the classification model, followed by data splitting into training and testing sets using an 80:20 ratio. Model performance was evaluated using a confusion matrix and standard classification metrics, including accuracy, precision, recall, and F1-score. The experimental results show that the proposed model achieved an accuracy of 0.50, with weighted average values of 0.52 for precision, 0.50 for recall, and 0.50 for F1-score. These results indicate that Categorical Naive Bayes is capable of performing stress level classification with moderate performance on categorical survey data. This study demonstrates the potential of machine learning techniques as a supporting tool for analyzing student stress levels and provides a basis for further model improvement in future research.

Keywords: *Categorical Naive Bayes; CRISP-DM; Machine Learning; Student Stress; Stress Classification*

1. Introduction

Student stress has become a significant concern in higher education due to increasing academic demands, performance expectations, and lifestyle-related pressures. Academic stress has been reported to negatively affect students' mental health, learning outcomes, and overall well-being, highlighting the importance of early identification and proper stress management strategies [1].

Previous studies indicate that student stress is influenced by a combination of academic and lifestyle factors, including workload, examination pressure, sleep quality, and daily habits [1], [2]. As the volume of survey-based data increases, conventional stress assessment methods that rely on manual analysis and self-report evaluation become less efficient and prone to subjectivity.

In recent years, machine learning techniques have been widely applied to analyze psychological and behavioral data due to their ability to identify patterns from large datasets. Several studies have demonstrated that machine learning models can effectively classify student stress levels using survey-based data, showing promising results in stress prediction tasks [2], [3]. However, the performance of such models often depends on data characteristics and algorithm complexity.

Various machine learning algorithms have been explored for student stress classification, including Decision Tree, Artificial Neural Networks, and ensemble-based methods. While these approaches can achieve high accuracy, they often require complex parameter tuning and higher computational costs [3], [4], [5]. Therefore, there is a need for simpler and more interpretable classification models that remain effective when applied to categorical survey data.

Naive Bayes is a probabilistic classification algorithm known for its simplicity, efficiency, and suitability for categorical data analysis. Previous studies have reported that Naive Bayes can perform effectively in classifying psychological and behavioral conditions using categorical survey data [6], [7]. To ensure methodological rigor and reproducibility, this study adopts the Cross-Industry Standard Process

for Data Mining (CRISP-DM) framework as the research methodology, which has been recognized as a structured and systematic approach for machine learning projects [8], [9].

Based on these considerations, this study applies the Categorical Naive Bayes algorithm within the CRISP-DM framework to classify student stress levels based on academic and lifestyle-related factors. Model performance is evaluated using a confusion matrix and standard classification metrics, including accuracy, precision, recall, and F1-score. The findings of this study are expected to contribute to the application of standardized data mining methodologies and simple yet effective machine learning models for student stress analysis.

2. Literature

Student stress has been extensively studied as a critical issue in higher education due to its impact on academic performance and mental health. Anggraini [1] conducted a literature analysis on the factors contributing to academic stress among students and emphasized that prolonged stress can negatively affect students' psychological well-being and learning motivation. These findings highlight the importance of identifying stress levels at an early stage.

With the growing availability of survey-based data, machine learning techniques have increasingly been used to analyze and predict student stress levels. Fadhila et al. [2] demonstrated that machine learning methods could be applied to detect stress levels based on sleep quality, indicating that behavioral and lifestyle factors play a significant role in stress prediction. Similarly, Dzakiyyah and Mahdiana [10] applied machine learning models to predict academic stress levels and reported that data-driven approaches could support early stress identification among students.

Several machine learning algorithms have been explored in student stress classification tasks. Harahap [4] applied Artificial Neural Networks to classify student stress levels and achieved promising performance, showing the ability of complex models to capture non-linear relationships in stress-related data. In addition, ensemble-based methods such as Random Forest have been used to improve classification accuracy in stress prediction, although these methods often require extensive parameter tuning and higher computational resources [5].

Naive Bayes has been widely used as a probabilistic classification algorithm due to its simplicity and computational efficiency. Amelia et al. [6] applied Naive Bayes to classify sleep disorders and demonstrated that the algorithm performs well when handling categorical health-related data. Furthermore, Susanti [7] reported that Naive Bayes is effective in classifying student stress levels using categorical survey data, making it suitable for stress-related classification tasks.

In terms of methodological frameworks, standardized data mining processes are essential to ensure consistency and reproducibility in machine learning research. The Cross-Industry Standard Process for Data Mining (CRISP-DM) has been recognized as a structured framework that guides machine learning projects from problem definition to evaluation [8], [9]. Recent studies confirm that CRISP-DM remains relevant and applicable in modern data science and machine learning applications.

Based on previous studies, it can be concluded that machine learning provides a promising approach for classifying student stress levels. However, there is still a research gap in studies that combine simple and interpretable classification algorithms with standardized data mining frameworks. Therefore, this study adopts the CRISP-DM framework and applies the Categorical Naive Bayes algorithm to classify student stress levels based on academic and lifestyle-related factors, with evaluation conducted using standard classification metrics.

3. Methodology

3.1. Research Framework

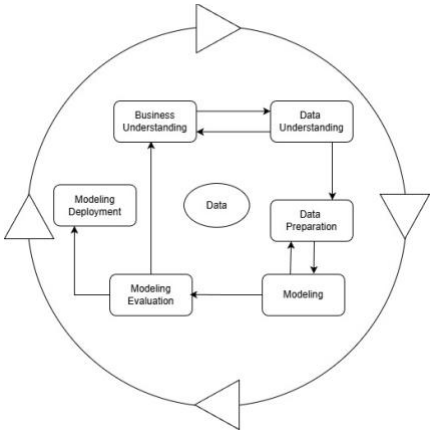

Fig .1: Research Methodology Based on the CRISP-DM Framework (adapted from[8], [9])

Figure 1 presents the overall research methodology employed in this study, starting from data collection to model evaluation.

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework as the research methodology. CRISP-DM provides a structured and systematic approach for machine learning research, ensuring consistency from problem understanding to model evaluation. As illustrated in **Figure 1**, this study applies the CRISP-DM stages of business understanding, data understanding, data preparation, modeling, and evaluation. The deployment phase is not included, as the focus of this research is on classification performance analysis rather than system implementation.

Table 1: Dataset Description

Description	Value
Data source	Student stress survey
Number of instances	520
Number of attributes	5
Attribute type	Categorical
Stress level labels	1-5

Table 1 The dataset consists entirely of categorical attributes representing academic and lifestyle-related factors, with stress level labels ranging from 1 to 5. The dataset characteristics were obtained directly from the original dataset before the data were divided into training and testing sets.

3.2. Business Understanding

The business understanding phase focuses on identifying the research problem and defining the study objectives. In this study, the main problem addressed is the increasing level of academic stress among university students. The objective of the research is to classify student stress levels based on academic and lifestyle-related factors using a machine learning approach.

3.3. Data Understanding

The data used in this study were obtained from the Student Stress Factors dataset. The dataset consists of 520 student records with four categorical attributes, namely academic pressure, sleep quality, headache frequency, and study habits. The target variable represents student stress levels measured on a Likert scale ranging from 1 to 5.

3.4. Data Preparation

In the data preparation phase, the dataset was examined to ensure data quality. No missing values or duplicated records were found. All categorical attributes were transformed into numerical representations using encoding techniques to enable processing by the classification algorithm. The dataset was then divided into training and testing sets with a ratio of 80:20

Table 2: Stress Level Categorization

Original Stress Scale	Stress Category
1	Low
2	Low
3	Medium
4	High
5	High

3.5. Modeling

The modeling phase involved the implementation of the Categorical Naive Bayes algorithm to classify student stress levels. This algorithm is suitable for categorical data and operates based on probabilistic principles. The model was trained using the training dataset and subsequently tested using the testing dataset to generate classification results.

3.6. Evaluation

Model performance was evaluated using a confusion matrix and standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics were used to assess the effectiveness of the Categorical Naive Bayes model in classifying student stress levels. The CRISP-DM process in this study was applied up to the evaluation stage, as the research focuses on performance analysis rather than system deployment.

4. Results and Discussion

4.1. Results

This section presents the results of the Categorical Naive Bayes classification model applied to the *Student Stress Factors* dataset. The results are organized according to the evaluation stage of the CRISP-DM framework.

4.1.1. Data Processing Results

The dataset consisted of 520 student records with categorical attributes related to academic pressure, sleep quality, headache frequency, and study habits. Data inspection showed no missing values or duplicate records. All categorical attributes were encoded to numerical values to enable processing by the classification algorithm. The dataset was divided into training and testing sets with a ratio of 80:20.

4.1.2. Model Training Results

The Categorical Naive Bayes model was trained using the training dataset. The model learned the probabilistic relationships between categorical features and stress level classes. The trained model was then applied to the testing dataset to generate stress level predictions.

4.1.3. Model Training Results

Model performance was evaluated using a confusion matrix and standard classification metrics, including accuracy, precision, recall, and F1-score. The evaluation results indicate that the model achieved an overall accuracy of 50%.

Table X presents the confusion matrix of the classification results, while Table Y summarizes the evaluation metrics obtained from the testing data.

Table 3: Confusion Matrix of Stress Level Classification.

Actual \ Predicted	1	2	3	4	5
1	12	0	6	3	0
2	7	12	5	1	0
3	1	3	12	3	0
4	2	10	3	4	1
5	4	0	2	1	12

Table 4: Model Evaluation Metrics

Metric	Value
Accuracy	0.50
Precision	0.52
Recall	0.50
F1-Score	0.50

4.2. Discussion

This section discusses the classification results obtained in this study and interprets the model performance in relation to previous research.

4.2.1. Interpretation of Classification Performance

The results show that the Categorical Naive Bayes model performs better in classifying extreme stress levels compared to intermediate categories. Most misclassifications occur between adjacent stress categories, which may be attributed to similarities in stress-related patterns among neighboring classes.

4.2.2. Comparison with Previous Studies

The findings of this study are consistent with previous research indicating that machine learning approaches can be applied to classify student stress levels using survey-based data [3], [4]. Although the achieved accuracy is moderate, the results demonstrate that Categorical Naive Bayes can serve as a baseline classification method for categorical stress data.

4.3. Model Evaluation

To quantitatively assess model performance, standard classification metrics were used, including accuracy, precision, recall, and F1-score. The overall accuracy achieved by the proposed model is **0.50**, indicating that the model is able to correctly classify approximately half of the test instances.

The weighted average precision, recall, and F1-score are **0.52**, **0.50**, and **0.50**, respectively. The use of weighted average metrics accounts for differences in class distribution and provides a more representative evaluation of multi-class classification performance. These results suggest that the Categorical Naive Bayes model achieves moderate classification performance when applied to categorical survey data.

4.4. Discussion

The results indicate that the Categorical Naive Bayes algorithm is capable of handling categorical attributes in student stress classification tasks with reasonable performance. The observed misclassifications are primarily found between neighboring stress levels, which suggests similarities in academic and lifestyle-related factors across these categories.

Compared to previous studies that applied more complex classification models, the performance achieved in this study is competitive considering the simplicity and efficiency of the Categorical Naive Bayes algorithm. The findings support the suitability of probabilistic models for analyzing survey-based stress data, particularly when interpretability and computational efficiency are prioritized.

Overall, the results demonstrate that machine learning techniques can support the analysis of student stress levels and provide insights into stress patterns among students. The proposed approach serves as a baseline for further research and can be enhanced through feature selection, parameter tuning, or the integration of additional data sources in future studies. The evaluation results indicate that the modeling and evaluation stages of the CRISP-DM framework were successfully implemented using a simple and interpretable classification algorithm.

5. Conclusion

This study concludes that the Categorical Naive Bayes algorithm, when applied within the CRISP-DM framework, is suitable as a baseline approach for classifying student stress levels based on academic and lifestyle factors. The experimental results demonstrate that the proposed model is capable of performing multi-class stress level classification with moderate performance. The evaluation results show an accuracy of 0.50, with weighted average precision, recall, and F1-score values of 0.52, 0.50, and 0.50, respectively.

The findings indicate that Categorical Naive Bayes can handle categorical attributes commonly found in student stress survey data while maintaining computational simplicity and interpretability. Misclassifications observed between adjacent stress levels reflect the overlapping characteristics of stress-related factors in real-world survey responses.

Although the achieved performance is moderate, the proposed approach provides a meaningful reference point for student stress analysis using machine learning techniques. Future research may focus on improving classification performance through feature selection, parameter optimization, class balancing strategies, or the integration of additional data sources to enhance model robustness and accuracy.

Acknowledgement

The authors would like to express their sincere gratitude to all parties who contributed to the completion of this research. Special appreciation is extended to the lecturer of Artificial Intelligence and Machine Learning for valuable guidance, constructive feedback, and continuous support throughout the research process. The authors also thank colleagues and peers for insightful discussions and collaborative support during the development of this study. Furthermore, heartfelt appreciation is conveyed to family and friends for their encouragement and moral support. All contributions and assistance provided are deeply appreciated.

References

[1] A. R. Anggraini, "Analisis literatur tentang faktor penyebab dan dampak stres akademik pada mahasiswa," *Maliki Interdisciplinary Journal (MIJ)*, vol. 3, pp. 318–324, 2025.
[2] A. Van Fadhila, J. A. Azzahra, K. Rizki, T. Zulkarnain, and N. D. Lathifah, "Implementasi metode machine learning untuk mendeteksi tingkat stres manusia berdasarkan kualitas tidur," in *Proceedings of SENAMIKA*, pp. 130–143, 2023.
[3] K. Yeler and S. Sürücü, "Classification of student stress levels using machine learning methods," pp. 210–214, 2025.
[4] S. B. Harahap and Y. Yamasari, "Klasifikasi tingkat stres mahasiswa menggunakan RMSProp untuk arsitektur artificial neural network," *Journal of Informatics and Computer Science (JINACS)*, vol. 5, no. 4, pp. 560–567, 2024, doi: 10.26740/jinacs.v5n04.p560-567.
[5] R. Nur Listiawan Dhito Eka Santoso, "Optimization of stress classification among students using random forest algorithm," *SITEKNIK Sistem Informasi, Teknik dan Teknologi Terapan*, vol. 2, no. 2, pp. 76–87, 2025, doi: 10.5281/zenodo.15130385.
[6] M. M. Amelia, B. M. Fazrin, Y. Y. Panjaitan, M. D. Kurniawan, and N. Khasanah, "Implementasi Naive Bayes untuk klasifikasi gangguan tidur," *Indonesian Journal of Computer Science*, vol. 4, no. 1, pp. 53–60, 2025, doi: 10.31294/t18ryp42.
[7] L. Susanti, "Klasifikasi tingkat stres pada mahasiswa teknik informatika dalam perkuliahan metode hybrid menggunakan algoritma Naive Bayes," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 8, no. 3, pp. 243–248, 2024.
[8] A. M. Shimaoka, R. C. Ferreira, and A. Goldman, "The evolution of CRISP-DM for data science: Methods, processes and frameworks," *SBC Reviews on Computer Science*, vol. 4, no. 1, pp. 28–43, 2024, doi: 10.5753/reviews.2024.3757.
[9] R. A. Casonatto, T. D. P. G. Souza, and A. M. Mariano, "Quality and risk management in data mining: A CRISP-DM perspective," *Procedia Computer Science*, vol. 242, pp. 161–168, 2024, doi: 10.1016/j.procs.2024.08.257.
[10] S. G. Dzakiyyah and D. Mahdiana, "Prediction of student academic stress levels using the decision tree algorithm and particle swarm optimization," *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, vol. 8, no. 3, pp. 513–525, 2025.