Classification of Students' Stress Levels Based on Learning Patterns Using Naïve Bayes Algorithm

Meri Puspita Sari[a], Sucipto[a], M.Najibulloh Muzaki[a]
[a]*Departement of Information Systems, Universitas Nusantara PGRI Kediri, Indonesia*

ARTICLE INFO
Article history:
Received 3 May 2026
Received in revised form
June 8, 2026
Accepted 25 June 2026
Available online 30 June 2026

Keywords:
Naïve Bayes
K-Nearest Neighbors
Classification
Student Stress Level
CRISP-DM

ABSTRACT
Student stress levels are one of the factors that can affect academic and psychological conditions. Learning patterns, daily activities, and academic achievements such as GPA are thought to have a relationship with the level of stress experienced by students. This study aims to classify students' stress levels based on lifestyle factors and academic performance using a *Naïve Bayes* algorithm. The approach used is quantitative with the CRISP-DM framework, which includes business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The research data were obtained from a *Kaggle* dataset and a private student dataset. The analysis process was carried out through preprocessing, data balancing using SMOTE, and splitting the data into training and testing sets. The main model used was *Gaussian Naïve Bayes*, with *K-Nearest Neighbors* (KNN) as a comparison model. The results showed that *Naïve Bayes* achieved the best performance on the private student dataset with an accuracy of 92.59%, while KNN performed better on the *Kaggle* dataset. These findings indicate that model performance is strongly influenced by data characteristics. Therefore, *Naïve Bayes* is considered effective for stress level classification on datasets with a limited size and relatively homogeneous distribution.

1 Introduction

Students often face various academic demands that can trigger academic stress. This condition is still a problem that is found in many university environments. Based on research conducted by Rahmi, as many as 88% of students feel pressured by lecture assignments, while 91% of students experience anxiety related to exams and academic assessments [1]. This high academic pressure can have an impact on decreasing study concentration, academic motivation, and student performance in attending lectures. In addition, the implementation of *hybrid* learning is known to increase the academic burden of students [2]. Meanwhile, the pressure due to technological adaptation also has the potential to increase stress and depression [3].

Learning patterns are an important factor that affects students' stress levels. Habits such as lack of sleep, high learning intensity, and non-academic activities can lead to an imbalance of physical and mental conditions. Research shows that sleep quality has a significant relationship with academic stress. On the other hand, the Student Achievement Index (GPA) also acts as an academic indicator as well as a source of pressure, both in maintaining achievement and managing low achievements [1], [4].

The development of digital technology also increases the potential for distractions in the learning process. Excessive use of social media can decrease productivity and increase psychological stress on students. Therefore, an approach that is able to analyze these various factors in an integrated manner is needed. The *Educational Data Mining* (EDM) approach allows for the analysis of academic data to identify student behavior patterns more comprehensively [5].

One of the methods widely used in EDM is machine learning-based classification.The *Naïve Bayes algorithm* has advantages in simplicity and computational efficiency, and is able to provide stable performance in predicting educational data. In addition, this method has also proven to be effective in predicting the psychological condition of students based on survey data [3], [6]. The selection of *the Naïve Bayes* algorithm in this study is based on its ability to handle data with a simple computational process and provide good performance on datasets with a relatively small amount of data.

Although much research has been done on student stress, the focus of the study still tends to be separate. The research by Widyawati used *a Naïve Bayes* algorithm to predict students' stress levels based on academic characteristics and lecture conditions, while the research by Rahmi examined the relationship between academic stress and student sleep quality [2], [1]. The results of the study show that student stress is influenced by various academic factors and lifestyles. However, research that integrates various aspects of students' lifestyles and activities, such as study time, sleep duration, physical activity, organizational activities, social activities, and academic performance in a single stress level classification model using *a machine learning* approach is still limited. Therefore, this study aims to classify students' stress levels based on lifestyle and academic performance using *the Naïve Bayes* algorithm and evaluate the performance of the model using *accuracy, precision, recall* and *F1-score* metrics. This research is expected to contribute to the development of data-based student behavior analysis [7].

2 Literature Study

2.1 Data Mining and Machine Learning

Data mining is a data analysis process to find hidden patterns that can be used in decision-making. In the field of education, this concept has developed into *Educational Data Mining* (EDM) to analyze student behavior and academic performance [8].

One approach to *data mining* is *machine learning,* which is a method that allows the system to learn patterns from data and generate predictions automatically. The *supervised learning approach* is widely used in classification assignments, including in the analysis of student stress levels [9], [10]. The application of machine learning in the educational field has been widely utilized to support academic analysis and decision-making processes. Previous research has demonstrated that machine learning techniques can be employed to predict students' learning styles based on their characteristics and learning-related data. These findings indicate that Educational Data Mining (EDM) approaches can generate valuable insights into student learning behavior and support the development of more effective educational strategies [5].

Corresponding Author: meripuspita901@gmail.com

doi: 10.30996/elsains.v8i1.133498

2.2 Classification and Naïve Bayes

Classification is a technique to group data into certain classes based on the attributes they have. This method is widely used in decision support systems because it is able to generate predictions of new data based on patterns learned from previous data. One of the widely used classification algorithms is *Naive Bayes,* which is based on *Bayes' Theorem* assuming that each feature is independent of the others. This algorithm works by calculating the probability of each class based on the observed feature, then assigning the data into the classes that have the highest probability. Because it has a simple, efficient computational process and is able to provide stable performance, *Naïve Bayes is* widely applied in various classification problems, including the analysis of educational data and the prediction of student stress levels [11]. For numerical data, one commonly used variant is *Gaussian Naïve Bayes*, which assumes that the feature values follow a *Gaussian* (normal) distribution. This variant is particularly suitable for datasets containing continuous numerical attributes. [12].

$$P(C_k \mid x) = \frac{P(x|C_k)\cdot P(C_k)}{P(x)} \quad (1)$$

Equation (1) is used to calculate the probability of a data instance x belongs to class C_k. In this equation, $P(C_k|x)$ represents the *posterior* probability, which is the probability of a data instance belonging to class C_k after considering the observed data x. The term $P(x|C_k)$ represents the *likelihood*, or the probability of observing data x given class C_k. Meanwhile, $P(C_k)$ denotes the prior probability of class C_k. Furthermore, $P(x)$ represents the evidence, which is the probability of observing data x and serves as a normalization factor in the probability calculation.

The application of *the Naïve Bayes* algorithm in the classification of students' stress levels has been carried out in previous studies. The research by Susanti used *a Naïve Bayes algorithm* to classify the stress level of Informatics Engineering students in the hybrid lecture system and obtained an accuracy score of 80.00% on the test data. These findings indicate that *Naïve Bayes* is capable of effectively identifying student stress levels based on student characteristics [13].

2.3 Model Evaluation and SMOTE

Classification model evaluation is carried out using metrics such as *accuracy, precision, recall,* and *F1-score* to measure model performance. One of the key metrics is *accuracy* which is calculated as:

In the case of unbalanced data, the *Synthetic Minority Oversampling Technique* (SMOTE) method is used to balance the class distribution. This technique has been proven to improve model performance and reduce bias against the majority of classes [14], [15]. Previous studies have also demonstrated the effectiveness of SMOTE in handling imbalanced datasets. Sulistiyowati and Jajuli integrated SMOTE with the *Naïve Bayes* algorithm and reported that the use of SMOTE improved classification performance by increasing the accuracy and overall classification results compared to the model without SMOTE [16].

2.4 Student Learning Pattern

Learning patterns are the habits of students in regulating learning activities that are influenced by psychological conditions and the academic environment. Unstructured learning patterns can trigger academic stress and decrease learning effectiveness, . In addition, the development of learning technology requires students to be more independent in managing their learning process [17], [18], [19]. Previous research has also shown that students' lifestyles are related to academic stress levels. Research conducted by Rahmi found that most students experience academic stress due to homework and anxiety about exams, which then impacts their sleep quality. The findings suggest that factors such as sleep duration and quality can affect academic conditions as well as overall student well-being [1].

Data analysis approaches in education are also often combined with various methods to improve prediction outcomes. Research shows that the integration of techniques such as *clustering* and other methods can improve the quality of analysis compared to a single method [20]. Several previous studies have applied machine learning approaches to analyze academic conditions and student stress levels. A comparative study between the Decision Tree C4.5 and *K-Nearest Neighbor* algorithms showed that KNN achieved better performance in classifying student stress levels, with an accuracy of 64.17%. The study also identified GPA as one of the factors influencing student stress levels. These findings indicate that academic data have significant potential for developing student stress classification models [21].

3 Methodology

This study uses a quantitative approach with the *Cross-Industry Standard Process for Data Mining* (CRISP-DM) methodology. This methodology was chosen because it provides systematic data analysis stages ranging from problem understanding to model evaluation. CRISP-DM consists of six main stages, namely *business understanding, data understanding, data preparation, modeling, evaluation,* and *deployment,* [22]. This approach allows the analysis process to be carried out in a directional and consistent manner in data-driven research [23].

Figure 1 shows the stages of research starting from understanding the problem to the implementation of the model. Each stage is interrelated in building a student stress level classification model.

3.1 Business Understanding

This stage is carried out by formulating the research objectives, namely classifying students' stress levels based on learning patterns and GPAs. The variables used include sleep hours, study time, organizational activities, physical activity, socializing time, and GPA as the main indicators.

3.2 Data Understanding

The data understanding stage includes exploring datasets obtained from *Kaggle* Mulyasyira and *private student* questionnaire data [24]. The Kaggle *dataset* is used as the primary dataset in the training and testing process of the model because it has a larger amount of data and variables that suit the research objectives. Meanwhile, the private dataset was used as validation data to test the model's ability to classify stress levels in student data obtained directly from real conditions.

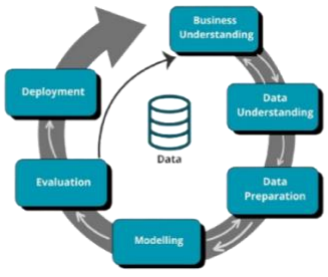

Figure 1. CRISP-DM Research Flow

Corresponding Author: meripuspita901@gmail.com
doi: 10.30996/elsains.v8i1.133498

This process includes analyzing data structure, variable types, identifying *missing values,* and distributing data to understand the characteristics of each dataset [23].

The private dataset consisted of 133 student responses collected from Universitas Nusantara PGRI Kediri through an online questionnaire distributed over a two-week period, from the last week of February to the first week of March 2026. The respondents came from various study programs, with most of them being eighth-semester students. The questionnaire was designed based on the variables used in this study, including study time, sleep duration, organizational activities, physical activities, social activities, GPA, and self-reported stress levels. Stress level labels were obtained directly from respondents by selecting one of the predefined categories in the questionnaire. Participation was voluntary, and respondents' identities were kept confidential to ensure data privacy.

3.3 Data Preparation

The data preparation stage consists of several steps. First, the dataset is explored through *Exploratory Data Analysis* (EDA) to understand its characteristics and identify potential issues. Next, the feature variables (x) and target variable (y) are determined, followed by encoding the target labels into numerical values. The dataset is then divided into training and testing data using the *train-test split* technique. To address class imbalance, the *Synthetic Minority Oversampling Technique* (SMOTE) is applied only to the training data. After that, feature scaling is performed by fitting the scaler on the SMOTE-resampled training data and transforming both the training and testing data. This approach helps prevent data leakage and ensures that the evaluation results are obtained from unseen test data[14].

3.4 Modeling

The modeling stage uses the *Gaussian Naïve Bayes* algorithm as the main classification method. *Gaussian Naïve Bayes* was selected because the variables used in this study, including study time, sleep duration, organizational activities, physical activity, social activities, and GPA, are predominantly numerical and can be effectively modeled using a *Gaussian* probability distribution. In addition, this algorithm offers fast training, computational efficiency, and simple interpretation. The selection of *Gaussian Naïve Bayes* was also motivated by the objective of evaluating its performance on datasets of different sizes, namely the *Kaggle* dataset and the private questionnaire dataset. Furthermore, *Naïve Bayes* has been widely applied in educational data classification and student stress analysis. To provide an objective performance comparison, the *K-Nearest Neighbors* (KNN) algorithm was also employed as a benchmark model.

For the comparison model, *K-Nearest Neighbors* (KNN) was implemented using $K = 5$ with the Euclidean distance metric. Feature scaling was performed using *StandardScaler* before model training to ensure that all numerical features contributed proportionally to the distance calculation. No *hyperparameter tuning* was performed, and the default parameter settings were used to provide a straightforward baseline comparison with the *Gaussian Naïve Bayes* model.

3.5 Evaluation

The evaluation stage was carried out using *a confusion matrix* with accuracy, precision, recall, *and F1-score metrics*.One of the metrics used is *accuracy* formulated in equation (2).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Equation (2) is used to calculate the accuracy of the classification model. In this equation, TP *(True Positive)* represents the number of positive instances correctly classified as positive, TN *(True Negative)* represents the number of negative instances correctly classified as negative, FP *(False Positive)* represents the number of negative instances incorrectly classified as positive, and FN *(False Negative)* represents the number of positive instances incorrectly classified as negative. The accuracy value indicates the proportion of crorrectly classified instances compared to the total number of instances in the test dataset.

3.6 Deployment

The last stage is *deployment,* the model that has been built is applied to real data to test its usefulness in predicting student stress levels based on questionnaire data. The prediction results are then compared with patterns in the main dataset to see the consistency and ability of the model to generalize to the new data. This stage is important to ensure that the model is not only accurate on training data, but also remains stable when used in different populations, in accordance with the deployment role in CRISP-DM that emphasizes the implementation of results or the presentation of final reports [25].

In addition, the results obtained are also interpreted to identify learning pattern factors that affect student stress levels, so that they can be used as a basis for decision-making by the campus, as well as open up opportunities for the development of data-based systems that are more applicable in the academic environment.

4 Discussion

4.1 Dataset Description

This research uses two dataset sources, namely *the Kaggle* dataset consisting of 2000 data and *the Private* dataset of 133 datasets [24]. Both datasets contain variables that represent student activity patterns such as study time, sleep time, organizational activities, physical activity, social time, and GPA, with stress level as the target variable. *The Kaggle* dataset is used as the main data in the modeling process, while the student dataset is used as validation data to test the model's consistency on real data.

Based on Table 1, the research dataset consists of 7 variables that represent student learning patterns and activities. Six variables are in the form of numerical data, namely, sleep hours, learning time, SME activities, physical activity, socialization time and GPA, while stress level is a categorical variable used as a target. Overall, the dataset consist of 6 independent variables and 1 dependent variable which the basis for classification process.

Table 1. Research Features

Variable	**Features**	
	Description	**Data Type**
Jam Tidur	Average student sleep hours per day	Numerical
Waktu Belajar	Length of time students study per day	Numerical
Waktu UKM	Duration of student organization activities	Numerical
Aktivitas Fisik	Long Duration of Physical Activity of Students	Numerical
Waktu Bersosialisasi	A Long Time to Socialize with Students	Numerical
IPK	Student's Cumulative Achievement Index	Numerical
Tingkat Stres	Student Stress Level Categories	Categorical

Corresponding Author: meripuspita901@gmail.com

doi: 10.30996/elsains.v8i1.133498

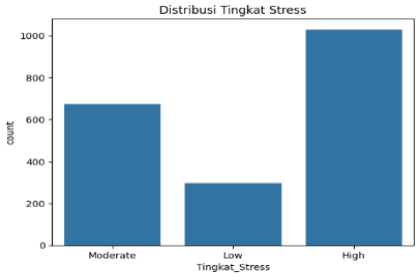

Figure 2. Kaggle Dataset Target Distribution

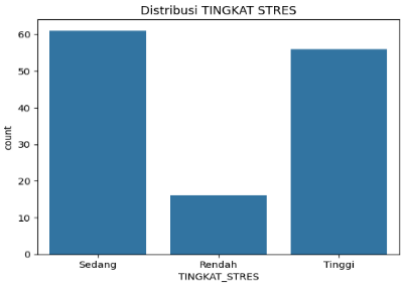

Figure 3. Private Dataset Target Distribution

4.2 *Distribution of Stress Levels*

Figure 2 shows the distribution of stress levels in the dataset indicating that the amount of data in each category is unbalanced. In *the Kaggle dataset,* the *High category* has the most dominant number, followed by *Moderate,* while *Low* is the category with the least number [24]. This condition identifies the existence of class imbalances *(imbalanced data)* that have the potential to affect the performance of the classification model. Therefore, special handling techniques are needed so that the model is not biased against the majority class.

Referring to Figure 3, it can be seen that the distribution of stress level data is uneven. The *Moderate* category dominates the amount of data, followed by *the High* category with a relatively small difference. On the other hand, *the Low* category has the least amount. This condition indicates a class imbalance, although the difference is still within reasonable limits and not too extreme.

4.3 *Data Pre-processing*

The *pre-processing* stage is carried out to ensure the data is in optimal condition before the modeling process. The categorical data on the target variable is first converted into numerical form using *the label encoding technique.* Furthermore, the dataset is divided into training data and test data with an 80:20 ratio. To overcome class imbalances, the SMOTE method is applied which results in a more balanced distribution of data in each category. In addition, standardization is carried out using *the StandardScaler* to equalize the scale between numerical features. This process aims to improve the stability and performance of the model in classifying.

Table 2 shows that prior to the implementation of SMOTE, the distribution of data in each dataset was unbalanced. After SMOTE is carried out, the amount of data in each class becomes the same so that the data distribution becomes more balanced and not biased towards certain classes.

4.4 *Modeling Results*

The modeling stage was carried out using *the Naïve Bayes* algorithm as the main model in this study. The selection of this method is based on its ability to process numerical data as well as good computational efficiency. For comparison, *the K-Nearest Neighbors* (KNN) algorithm is used to provide an objective picture of the model's performance. The implementation of KNN on *the Kaggle* dataset refers to previous research by Mulyasyira. Thus, the main focus of the analysis remains on *the performance of Naïve Bayes,* while the KNN is used as a *comparative baseline* [24].

Table 2. *Before-After* SMOTE

Dataset	SMOTE		
	Categories	***Before***	***After***
Kaggle	*High*	823	823
	Low	238	823
	Moderate	539	823
Private	*Rendah*	13	48
	Sedang	48	48
	Tinggi	45	48

Table 3. Comparison of Accuracy Between Models

Dataset	Model Accuracy	
	Naïve Bayes(%)	**KNN(%)**
Kaggle	82.75	87.75
Private	92.59	88.89

Table 4.*Classification Report* Dataset *Private*

Model	*Classification Report*			
	Class	***Precision***	***Recall***	***F1-Score***
Naïve Bayes	*Rendah*	0.75	1.00	0.86
	Sedang	1.00	0.85	0.92
	Tinggi	0.92	1.00	0.96
KNN	*Rendah*	0.50	1.00	0.67
	Sedang	1.00	0.77	0.87
	Tinggi	1.00	1.00	1.00

Based on Table 3, The test results showed that both models performed quite well in classifying students' stress levels. In the *Kaggle* dataset, the KNN algorithm obtained an accuracy of 87.75% higher than *Naïve Bayes* which reached 82.75%. Meanwhile, in the Private dataset, *Naïve Bayes* showed superior performance with an accuracy of 92.59% compared to KNN of 88.89%. These differences in results suggest that the performance of the model can be affected by the characteristics and distribution of the data used. In general, both algorithms are able to provide good classification results, but *Naïve Bayes* is more optimal on the main research dataset.

4.5 *Model Evaluation*

Model evaluation was carried out using *accuracy, precision, recall,* and *F1-score* metrics to measure classification performance. The test results showed that both algorithms were able to provide good performance on each dataset. In *the Kaggle* dataset [24], KNN showed higher accuracy than *Naïve Bayes,* while in the Private dataset, *Naïve Bayes* had superior performance. This shows that the performance of the model is influenced by the characteristics of the data used.

Based on Table 4, the *Naïve Bayes* model shows consistent performance with high *precision, recall,* and *F1-score* values across all classes. A *recall* value of 1.00 in some classes indicates that the model is able to identify data very well without many errors. Meanwhile, KNN still shows weaknesses in certain classes, especially in lower *precision* values, so that overall *Naive Bayes* is more optimal in student datasets.

Corresponding Author: meripuspita901@gmail.com
doi: 10.30996/elsains.v8i1.133498

4.6 Confusion Matrix

The results *of the confusion matrix* show that the *Naïve Bayes* model is able to classify most of the data correctly, which is indicated by the dominance of values on the diagonal of the grid. Misclassification is still defined in some classes, especially in categories with adjacent distributions. However, overall, the error rate is relatively small so the model can be said to have good classification ability in predicting student stress levels.

Based on Figure 4, the *confusion matrix* shows that *the Naïve Bayes* model is able to classify most of the data correctly, which is indicated by the dominance of values on the diagonal of the matrix. The number of correct predictions in each class looks quite high, namely 3 data in class 0, and 11 data in class 1 and class 2. However, there are still misclassifications in class 1, where each 1 data is predicted as class 0 and class 2. Meanwhile, in grade 0 and grade 2, no prediction errors were found.

4.7 Model Comparison

The performance comparison in Table 5, The performance of the classification algorithm shows the differences affected by the characteristics of each dataset. In the Kaggle dataset, the K-Nearest Neighbors (KNN) algorithm obtained higher evaluation scores than Naïve Bayes on metrics, accuracy, precision, recall, and fi-score. These results show that KNN is more effective in recognizing data patterns that provide fairly good classification results and can be used as an alternative to a comparative model. Meanwhile, in the Private dataset, Naïve Bayes showed better performance than KNN. This advantage is associated with higher accuracy values and F1-score, which indicate the model's ability to produce more consistent predictions in each class. Although the KNN has a slightly higher precision value, the lower recall and F1-score values indicate that the model has not been able to identify all categories optimally.

The difference in results in these two datasets confirms that the effectiveness of the algorithm is highly dependent on the structure and distribution of the data used. Therefore, Naïve Bayes is considered more suitable for the Private dataset, while KNN is more optimal when applied to the Kaggle dataset.

The performance comparison in Table 5, The performance of the classification algorithm shows the differences affected by the characteristics of each dataset. In *the Kaggle* dataset, *the K-Nearest Neighbors* (KNN) algorithm obtained higher evaluation

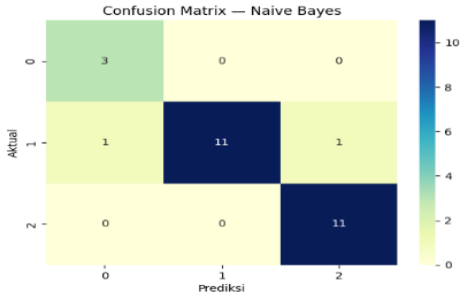

Figure 4. *Matrix Naïve Bayes Confusion*

Table 5. Comparison of Model Performance on the Two Datasets

Dataset	Model Performance				
	Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Kaggle	*Naïve Bayes*	0.83	0.84	0.83	0.83
	KNN	0.88	0.89	0.88	0.88
Private	*Naïve Bayes*	0.93	0.94	0.93	0.93
	KNN	0.89	0.94	0.89	0.90

TIDUR	BELAJAR	UKM	OLAHRAGA	SOSIAL	IPK	Aktual	Pred_NB	Pred_KNN
6	3	5	2	3	3.00	Sedang	Sedang	Sedang
6	3	4	3	4	3.69	Sedang	Sedang	Sedang
5	2	0	2	3	3.70	Sedang	Sedang	Rendah
5	2	6	2	5	3.82	Tinggi	Tinggi	Tinggi
5	2	5	2	4	3.90	Tinggi	Tinggi	Tinggi
6	4	0	2	1	3.89	Sedang	Rendah	Rendah
5	3	3	3	3	3.00	Sedang	Sedang	Sedang
5	3	4	4	3	3.69	Sedang	Sedang	Sedang
7	4	0	4	2	2.85	Rendah	Rendah	Rendah
5	2	0	2	4	3.50	Sedang	Sedang	Rendah
5	2	5	2	5	3.27	Tinggi	Tinggi	Tinggi
7	4	0	4	2	2.65	Rendah	Rendah	Rendah
7	3	2	4	2	2.78	Sedang	Sedang	Sedang
5	2	4	2	4	3.89	Tinggi	Tinggi	Tinggi
5	2	4	2	4	3.38	Sedang	Tinggi	Sedang

Figure 5. Private Dataset Prediction Results Table

scores than *Naïve Bayes* on metrics, *accuracy*, *precision, recall*, and *fi-score*. These results show that KNN is more effective in recognizing data patterns that provide fairly good classification results and can be used as an alternative to a comparative model.

Meanwhile, in the Private dataset*, Naïve Bayes* showed better performance than KNN. This advantage is associated with higher accuracy values and *F1-score*, which indicate the model's ability to produce more consistent predictions in each class. Although the KNN has *a slightly higher* precision value, the *lower recall* and *F1-score* values indicate that the model has not been able to identify all categories optimally. The difference in results in these two datasets confirms that the effectiveness of the algorithm is highly dependent on the structure and distribution of the data used. Therefore, *Naïve Bayes* is considered more suitable for *the Private* dataset, while KNN is more optimal when applied to *the Kaggle dataset.*

4.8 Prediction Results

As shown in Figure 5, the two models exhibit different levels of consistency in class prediction. Overall, most data points were classified correctly, as indicated by the agreement between actual and predicted values. However, Naïve Bayes demonstrates greater stability, producing predictions that align well with the actual data across categories. This suggests that its probabilistic approach effectively captures the underlying data distribution, even with a limited dataset.

In contrast, the K-Nearest Neighbours (KNN) model shows several inconsistencies, particularly in the *moderate* category, which is frequently misclassified as low. This indicates that KNN is sensitive to data distribution and sample size, leading to reduced performance when the data is limited or unevenly distributed. Furthermore, its distance-based mechanism may introduce bias when feature characteristics overlap between classes.

These findings highlight that each algorithm's performance is shaped by the characteristics of the dataset. In this study, Naïve Bayes produced more consistent and reliable results than KNN, making it a more suitable approach for private datasets, particularly those with a small sample size and relatively simple feature distribution.

5 Conclusion

This study shows that both *Naïve Bayes* and KNN can be used to classify students' stress levels. KNN achieved better performance on the *Kaggle* dataset, while *Gaussian Naïve Bayes* produced more optimal results on the private student dataset with an accuracy of 92.59%. These results indicate that model performance is influenced by the characteristics and size of the dataset. Therefore, *Gaussian Naïve Bayes* can be considered effective for classifying student stress levels on datasets with relatively limited and homogeneous data.

Corresponding Author: meripuspita901@gmail.com

doi: 10.30996/elsains.v8i1.133498

Author Contribution

Conceptualization, A1, A2; methodology and formal analysis, A1; validation, A1, A2, A3; writing-original draft, A1; writing-review, editing, and visualization, A2, A3.

Reference

[1] A. Rahmi, N. S. Hasibuan, N. G. Purba, S. R. Manurung, and R. Thohiri, "Hubungan Stres Akademik dan Kualitas Tidur pada Mahasiswa," *Jurnal Pengabdian Masyarakat dan Riset Pendidikan*, vol. 3, no. 4, pp. 2017–2024, May 2025, doi: 10.31004/jerkin.v3i4.697.

[2] G. N. Widyawati, M. Khasanah, M. Muttaqin, E. Rasywir, and A. Feranika, "Prediksi Tingkat Stress Pada Mahasiswa Universitas Dinamika Bangsa Jambi Dalam Melakukan Perkuliahan Metode Hybrid Menggunakan Algoritma Naive Bayes," *Jurnal Manajemen Teknologi dan Sistem Informasi (JMS)*, vol. 1, no. 2, 2022, doi: doi.org/10.33998/jms.2022.2.1.44.

[3] W. Philuek, "Using machine learning techniques to study of stress, depression, and academic performance of grade 12th students," *Asian Journal of Education and Training*, vol. 10, no. 3, pp. 146–153, Aug. 2024, doi: 10.20448/edu.v10i3.5885.

[4] A. Sastra, K. Sabri, B. Yanto, F. Asmen, I. Inal, and A. Crisdianto, "Analisis Pengaruh Gaya Hidup Mahasiswa terhadap Indeks Prestasi Kumulatif (IPK) Menggunakan Model Regresi," *KERNEL: Jurnal Riset Inovasi Bidang Informatika dan Pendidikan Informatika*, vol. 4, no. 2, pp. 89–101, Dec. 2023, doi: 10.31284/j.kernel.2023.v4i2.7384.

[5] D. Anggraini, "DATA MINING PENDIDIKAN: PREDIKSI GAYA BELAJAR MAHASISWA TEKNIK MENGGUNAKAN MACHINE LEARNING," vol. 11, no. 3, pp. 563–572, 2025, doi: 10.25126/jtiik.2025129190.

[6] R. P. Sonjaya, A. R. Gintara, L. S. Riza, M. Nursalman, E. Nugraha, and D. Wahyudin, "Predicting Student Depression Using the Naive Bayes Model on the Student Depression Dataset from Kaggle," *JENTIK : Jurnal Pendidikan Teknologi Informasi dan Komunikasi*, vol. 4, no. 1, pp. 9–20, Jun. 2025, doi: 10.58723/jentik.v4i1.448.

[7] S. Sucipto *et al.*, "Pelatihan Tim Pengembang Akademik SMAN 6 Kediri dalam Menggunakan E-Learning Berbasis Moodle," *Jurnal Pengabdian Kepada Masyarakat*, vol. 3, no. 1, pp. 105–115, Nov. 2023, doi: 10.55506/arch.v3i1.80.

[8] Y. Apridiansyah, N. David, M. Veronika, E. D. Putra, U. Muhammadiyah, and Y. A. Id;, "Prediksi Kelulusan Mahasiswa Fakultas Teknik Informatika Universitas Muhammadiyah Bengkulu Menggunakan Metode Naive Bayes," *JSAI : Journal Scientific and Applied Informatics*, vol. 4, no. 2, 2021, doi: 10.36085.

[9] A. Mulyoto, *NAÏVE BAYES PADA GOOGLE COLABS PENERBIT CV. EUREKA MEDIA AKSARA*. Eureka Media Aksara, 2021.

[10] A. A. Permana *et al.*, *MACHINE LEARNING*. PT GLOBAL EKSEKUTIF TEKNOLOGI, 2023. [Online]. Available: www.globaleksekutifteknologi.co.id

[11] Amna *et al.*, *DATA MINING*, Pertama. PT GLOBAL EKSEKUTIF TEKNOLOGI, 2023. [Online]. Available: www.globaleksekutifteknologi.co.id

[12] D. Berrar, "Bayes' theorem and naive bayes classifier," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, Elsevier, 2018, pp. 403–412. doi: 10.1016/B978-0-12-809633-8.20473-1.

[13] L. Susanti, "STRING (Satuan Tulisan Riset dan Inovasi Teknologi) KLASIFIKASI TINGKAT STRESS PADA MAHASISWA TEKNIK INFORMATIKA DALAM MELAKUKAN PERKULIAHAN METODE HYBRID MENGGUNAKAN ALGORITMA NAIVE BAYES," Apr. 2024. doi: 10.30998/string.v8i3.17096.

[14] S. Sofyan and A. Prasetyo, "Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang Pada Tingkat Pendapatan Pekerja Informal Di Provinsi D.I. Yogyakarta Tahun 2019," 2021.

[15] A. Ramadhani, R. Aryanti, and S. Agustiani, "Optimasi Model Machine Learning Menggunakan Teknik SMOTE pada Analisis Sentimen Pengguna RedBus," *Journal of Artificial Intelligence and Technology Information (JAITI)*, vol. 4, no. 1, pp. 1–12, Mar. 2026, doi: 10.58602/JAITI.V4I1.182.

[16] N. Sulistiyowati and M. Jajuli, "INTEGRASI NAIVE BAYES DENGAN TEKNIK SAMPLING SMOTE UNTUK MENANGANI DATA TIDAK SEIMBANG," *NUANSA INFORMATIKA*, vol. 14, no. 1, pp. 34–37, Jan. 2020, doi: 10.25134/NUANSA.V14I1.2411.

[17] H. Herman *et al.*, *Psikologi Belajar dan Pembelajaran*, 1st ed., no. 2. PT GLOBAL EKSEKUTIF TEKNOLOGI, 2022. doi: 10.24198/JPPM.V2I2.13540.

[18] E. Harianto *et al.*, *Psikologi Pendidikan*. CV. Zenius Publisher, 2024.

[19] R. Susanto, E. S. Nasution, and Yulhendri, *PSIKOLOGI PENDIDIKAN Pendidikan di Era Revolusi Industri 4.0*, 1st ed. Malang: PT. Literasi Nusantara Abadi Grup, 2024. Accessed: Jan. 03, 2026. [Online]. Available: www.penerbitlitnus.co.id

[20] M. N. Muzaki, "PENERAPAN TRIANGULAR KERNEL NEAREST NEIGHBOR SEBAGAI METODE CLUSTERING DASAR PADA METODE BAGGING," *Jurnal Teknologi dan Riset Terapan (JATRA)*, vol. 1, no. 1, pp. 9–13, Jun. 2019, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JATRA,https://jurnal.polibatam.ac.id/index.php/JATRA

[21] N. Renaningtias, Gushevinalti, T. Eka Putri, E. Putri Purwandari, and Y. Stainly Ritonga, "Studi Komparasi Algoritma Decision Tree C4.5 dan K-Nearest Neighbor pada Klasifikasi Masa Studi dan Tingkat Stres Mahasiswa," vol. 13, pp. 1176–1785, Dec. 2024, doi: 10.35889/jutisi.v13i3.2272.

[22] A. Lathifah, "Cross-Industry Standard Process For Data Mining (CRISP-DM) Untuk Menemukan Pola Asosiasi Pada Data Tracer Study Lulusan Perguruan Tinggi Keagamaan Islam (Studi Kasus: UIN Syarif Hidayatullah Jakarta)," Universitas Islam Negeri Syarif Hidayatullah Jakarta, Jakarta, 2023.

[23] Y. Suhanda, I. Kurniati, and S. Norma, "Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik," *Jurnal Teknologi Informatika dan Komputer*, vol. 6, no. 2, pp. 12–20, Sep. 2020, doi: 10.37012/JTIK.V6I2.299.

[24] Mulyasyira, "Pengaruh Gaya Hidup Mahasiswa terhadap IPK," 2023, *Kaggle*. Accessed: Apr. 30, 2026. [Online]. Available: https://www.kaggle.com/code/mulyasyira/pengaruh-gaya-hidup-mahasiswa-terhadap-ipk/notebook

[25] A. F. Hakim, Y. Irawan, and R. R. Setiawan, "Implementation of the Crisp-Dm Methodology and Naive Bayes Algorithm on A Raw Material Requirement Prediction System to Reduce Food Waste (Case Study: Adamsafee Bakery, Resto, & Cafe)," *Jurnal Teknologi Informasi dan Pendidikan*, vol. 18, no. 2, pp. 968–983, Aug. 2025, doi: 10.24036/JTIP.V18I2.990.

Corresponding Author: meripuspita901@gmail.com
doi: 10.30996/elsains.v8i1.133498