PREDICTING STUDENT STRESS AND MENTAL HEALTH USING SUPPORT VECTOR MACHINE AND RANDOM FOREST

Miftahul Farida [1)], Panji Bintoro [*,2)], Ferly Ardhy [3)], Agus Wantoro [4)], Dwi Yana Ayu Andini [5)]

[1)]Informatics Engineering, Faculty of Technology and Informatics, Aisyah University, Indonesia
Jl. A Yani No. 1 A Tambak Rejo, Wonodadi, Gadingrejo, Pringsewu, Lampung, Indonesia, 35372

[2)]Software Engineering, Faculty of Technology and Informatics, Aisyah University, Indonesia
Jl. A Yani No. 1 A Tambak Rejo, Wonodadi, Gadingrejo, Pringsewu, Lampung, Indonesia, 35372

[3)]Informatics Engineering, Faculty of Technology and Informatics, Aisyah University, Indonesia
Jl. A Yani No. 1 A Tambak Rejo, Wonodadi, Gadingrejo, Pringsewu, Lampung, Indonesia, 35372

[4)]Informatics Engineering, Faculty of Technology and Informatics, Aisyah University, Indonesia
Jl. A Yani No. 1 A Tambak Rejo, Wonodadi, Gadingrejo, Pringsewu, Lampung, Indonesia, 35372

[5)]Software Engineering, Faculty of Technology and Informatics, Aisyah University, Indonesia
Jl. A Yani No. 1 A Tambak Rejo, Wonodadi, Gadingrejo, Pringsewu, Lampung, Indonesia, 35372

Email: [2]panjibintoro09@aisyahuniversity.ac.id

Abstract

Student mental health is a critical issue in Indonesia, with more than 30% of students experiencing symptoms of stress and depression due to academic, social, and economic pressures. This study aims to develop a stress classification pipeline based on a Likert-scale survey using the Support Vector Machine (SVM) algorithm. A quantitative cross-sectional method was employed with 293 active university student respondents, utilizing a questionnaire adapted from standardized psychometric instruments measuring academic stress and mental health, with established content validity and high internal consistency (Cronbach's alpha > 0.85). The research stages encompassed data preprocessing, labeling based on median scores, normalization, feature selection, model training, and evaluation. The evaluation results show an accuracy of 89.77%, with precision, recall, and F1-score values consistently ranging between 0.897–0.898. The confusion matrix indicates a balanced classification distribution between the "Stress" and "No Stress" classes. The discussion reveals that dominant factors in stress classification include academic pressure, sleep disturbance, and social support, aligning with established psychological stress theories. This study demonstrates that the SVM model is effective in classifying student stress and that the constructed pipeline adheres to the principles of reproducibility, auditability, and data ethics. The proposed system has the potential to be developed into a practical and responsible stress monitoring tool accessible to educational institutions.

Keywords: *stress classification, mental health prediction, support vector machines, likert surveys, machine learning*

1. Introduction

Student mental health has become an increasingly critical issue in Indonesia, particularly in the context of academic pressure, life transitions, and lack of social support. University students are in a developmental phase that is highly vulnerable to stress, and various studies have shown that more than 30% of students experience symptoms of stress, anxiety, or depression, which directly affect their academic performance and psychosocial well-being [1]. Therefore, early detection of stress has become an urgent necessity within higher education environments.

With the advancement of technology, machine learning–based approaches have increasingly been utilized to detect and predict stress. Research conducted by [2] a systematic review of 49 studies and found that the Support Vector Machine (SVM) algorithm is the most widely used in stress classification, primarily due to its capability to handle multidimensional data and produce stable classification results. Other research [3] developed a stress prediction model based on Artificial Neural Networks (ANN); however, the study did not explicitly integrate demographic and psychosocial variables. Furthermore, [4]emphasized that most mental health prediction models still operate as black-box systems, making them difficult to reproduce and therefore unsuitable for educational institutions that require transparency and auditability. Then, [5] employed SVM for stress classification based on ECG signals, but such an approach is impractical for student survey-

based assessments. In, [6] attempted to combine survey data with Random Forest and SVM, yet the study did not provide a replicable pipeline documentation accessible to non-technical users.

From this review, it can be concluded that although SVM has proven effective in stress classification, no study has yet developed a modular, survey-based Likert pipeline that is replicable, auditable, and practically applicable for educational institutions in Indonesia. Previous studies have tended to focus primarily on model accuracy without considering aspects of data ethics, process documentation, and usability for non-technical users. Furthermore, most studies have not included visualizations of stress distribution or output validation mechanisms that could address critical questions from reviewers or end users.

Based on these identified gaps, this study aims to develop a modular, transparent, and ethical SVM-based student stress classification pipeline. The proposed pipeline is designed to systematically process Likert-scale survey data, encompassing stages of data preprocessing, labeling, normalization, feature selection, and model evaluation. The novelty of this study lies in the integration of auditability, reproducibility, and anti-overfitting principles at each stage of development, as well as the provision of comprehensive documentation that can be readily utilized by educational institutions or other researchers.

Therefore, the main research question posed in this study is: How effective is the Support Vector Machine (SVM) method in classifying and predicting student stress levels based on survey data processed in a modular and ethical manner? This article presents a novel approach that not only focuses on model performance but also emphasizes process authenticity, transparency, and the practical implementation potential of stress monitoring systems within university environments. In addition to the SVM algorithm, this research also implements the random forest algorithm as a comparison.

2. Methods

This study employs a quantitative approach with a cross-sectional design, which is commonly used in classification studies based on psychometric surveys. The primary objective of this method is to develop a stress classification workflow that is reproducible, auditable, and user-friendly for non-technical users. The workflow is designed to systematically process Likert-scale survey data, encompassing all stages from data collection to classification model evaluation [7].

The stress classification model in this study is grounded in [8] stress theory, which posits that stress arises when individuals perceive that environmental demands exceed their adaptive capacities. In the context of machine learning, the Support Vector Machine (SVM) algorithm was selected due to its ability to generate an optimal hyperplane for binary classification and its effectiveness in handling multidimensional data [9]. SVM has been widely applied in stress classification based on both survey data and physiological signals. Figure 1 shows the flowchart of this research.

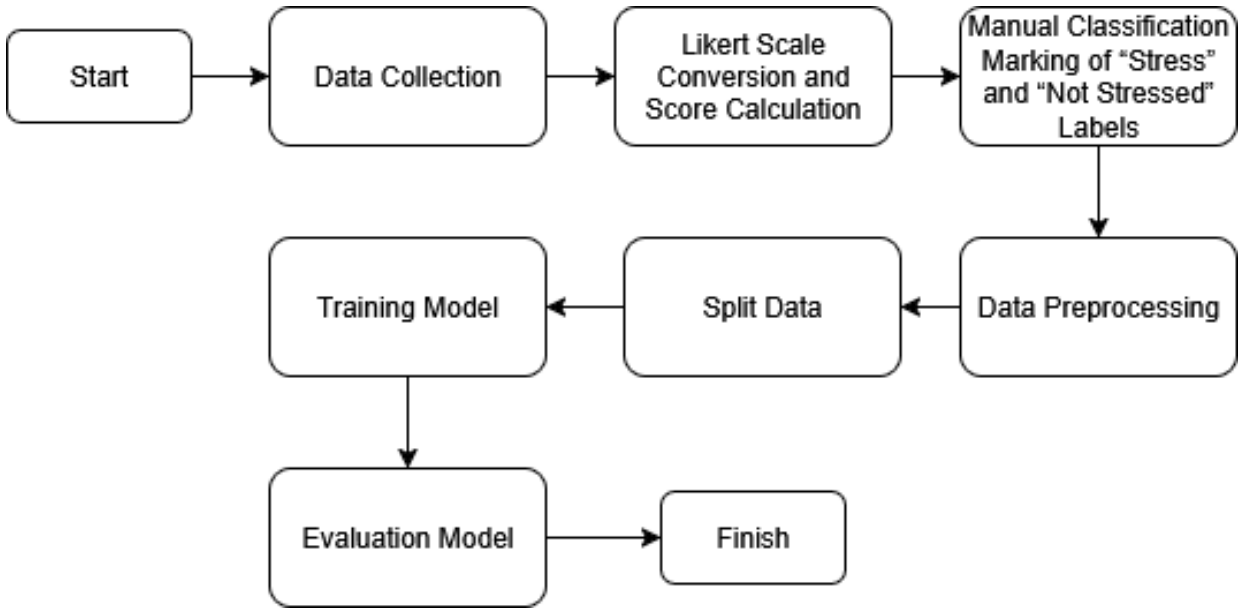

Figure 1. Research Stage

The explanation regarding the research stages is as follows:

a. Data Collection: pimary data were collected through an online questionnaire using a Likert scale (1–5), adapted from standardized psychometric instruments designed to measure academic stress and mental health among university students. The questionnaire consisted of 50 items covering aspects such as workload, time pressure,

sleep disturbances, and social support. Content validity and internal consistency were verified in previous studies, yielding a Cronbach's alpha coefficient greater than 0.85. The respondents comprised 293 active undergraduate students in semesters 2 to 8, selected using a purposive sampling technique.

b. Likert Scale Conversion and Score Calculation: the questionnaire responses were converted into a numerical format using a standard mapping: Strongly Agree (SA) = 5, Agree (A) = 4, Neutral (N) = 3, Disagree (D) = 2, and Strongly Disagree (SD) = 1.
c. Manual Classification: the labels "Stressed" and "Not Stressed" were assigned based on the median threshold (193) rather than a fixed value, in order to adjust to the local data distribution and avoid classification bias.
d. Data Preprocessing: the data were cleaned by removing duplicates and missing values. Normalization was performed using Z-score standardization to ensure consistency across feature scales. Feature selection was conducted through variable correlation analysis and Recursive Feature Elimination (RFE) to minimize overfitting.
e. Split data: the dataset was divided into 80% training data and 20% testing data using stratified sampling to maintain balanced class proportions.
f. Training model: The SVM and Random Forest model with a linear kernel was trained using a modular pipeline that included missing value imputation and classification processes.
g. Evaluation Model he model was evaluated using accuracy, precision, recall, and F1-score as performance metrics.

3. Results and Discussion

This study involved 293 active university students from various semesters as respondents. Data were collected through a Likert-scale questionnaire consisting of 50 items related to academic stress, anxiety, and social support. The data were processed using a Python-based pipeline in Google Colaboratory, with the Support Vector Machine (SVM) and Random Forest (RF) algorithm serving as the primary classification method [10]. The dataset division includes 80% training data and 20% testing data.

The classification was performed using a median stress score threshold of 193. Respondents with total scores above the threshold were categorized as "Stressed", while those with scores below or equal to the threshold were classified as "Not Stressed." The classification results indicated a relatively balanced distribution between the two categories.

Table 1. Distribution of student stress based on label

Category	Number of Respondents	Percentage
Stress	147	50.2%
Not Stress	146	49.8%
Total	293	100%

After removing the identity columns, the data were converted from text to numeric format using a standard Likert scale mapping. The total stress score was calculated for each respondent, and classification was performed based on the median threshold value (193). Respondents with scores ≥193 were labeled as "Stressed," while the remaining respondents were labeled as "Not Stressed".

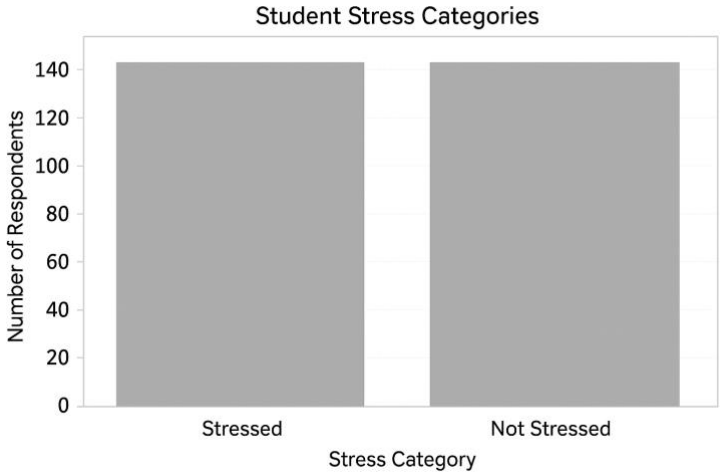

Figure 2. Stress distribution based on median

Figure 2 shows a bar chart titled “Student Stress Categories.” The horizontal axis represents two stress categories Stressed and Not Stressed while the vertical axis shows the number of respondents. Both bars are nearly equal in height, each with approximately 140 respondents, of which 50.2% were classified as "Stressed" and 49.8% as "Not Stressed," indicating that the number of students who felt stressed was nearly equal to the number of students who did not feel stressed. This indicates a balanced distribution of stress levels among the students surveyed, with no dominant category.

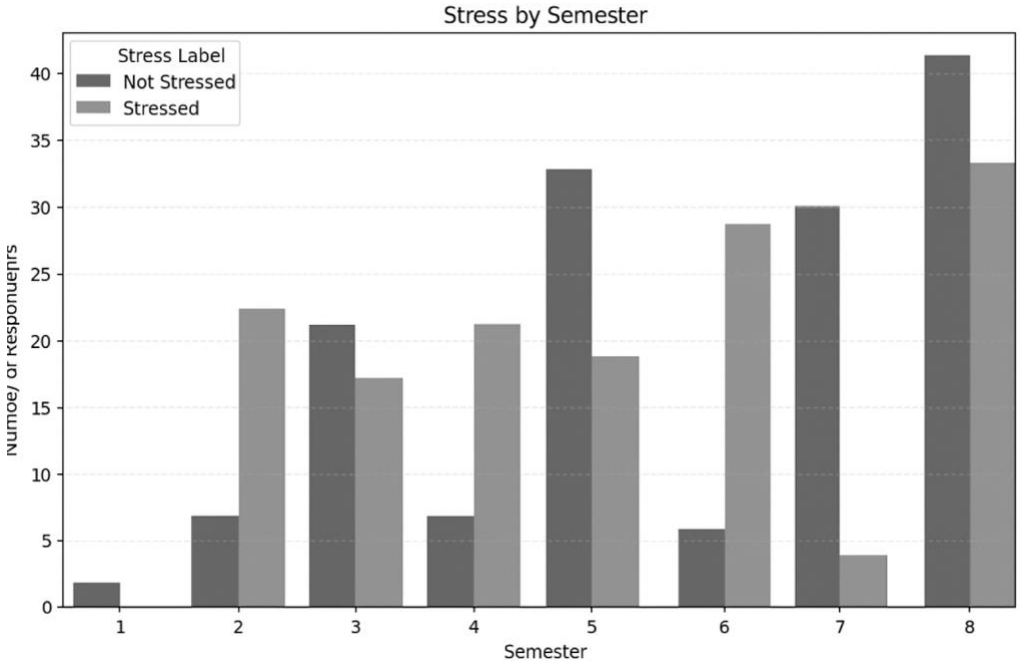

Figure3. Stress distribution by semester

Figure 3 shows a bar chart titled “Stress by Semester,” which illustrates the distribution of student stress levels across semesters 1 to 8. The vertical axis shows the number of respondents, while the horizontal axis represents the semester. Two categories are compared in each semester: Stressed and Not Stressed, as indicated in the legend. Overall, the chart shows that stress levels vary across semesters. In the earlier semesters (1–2), the number of stressed students is relatively low compared to those not stressed. As semesters progress, particularly from semesters 3 to 6, the number of stressed students increases and in some semesters exceeds the number of students who are not stressed. In the later semesters (7–8), both categories show higher respondent counts, with stress levels remaining prominent. This pattern suggests that academic stress tends to increase in higher semesters, possibly due to greater academic workload, projects, or preparation for graduation.

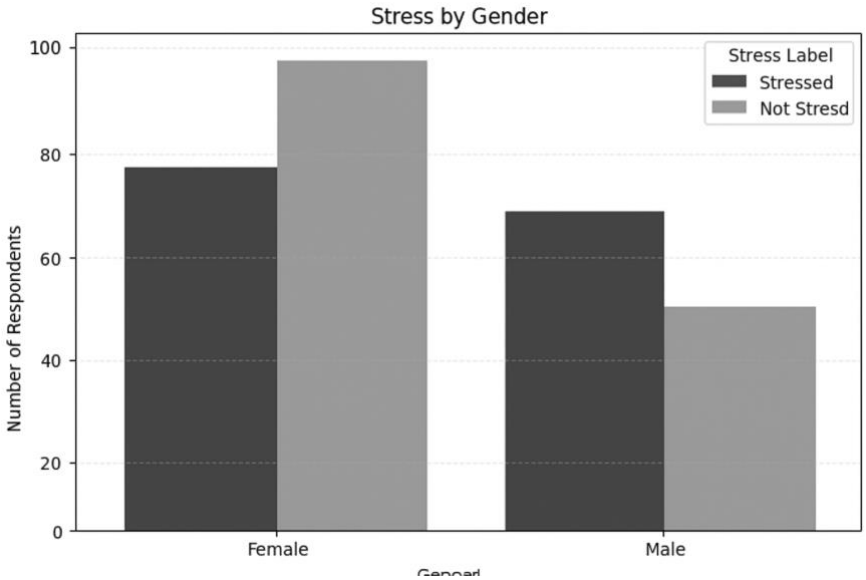

Figure 4. Stress distribution by gender

Figure 4 shows shows a bar chart titled “Stress by Gender,” which compares the number of respondents experiencing stress and not experiencing stress based on gender. The horizontal axis represents gender (Female and Male), while the vertical axis indicates the number of respondents. Two categories are shown for each gender: Stressed and Not Stressed. The chart indicates that among female respondents, the number of students who are not stressed is higher than those who are stressed. In contrast, among male respondents, the number of stressed students exceeds those who are not stressed. Overall, the figure suggests differences in stress distribution between genders, with female students showing a higher proportion of non-stressed respondents, while male students exhibit a relatively higher level of stress.

The input data for SVM and Random Forest is then used to predict the x_train data by initializing the SVM and Random Forest parameters used to predict student stress and mental health. The SVM models are fitted, as shown in Figures 5.

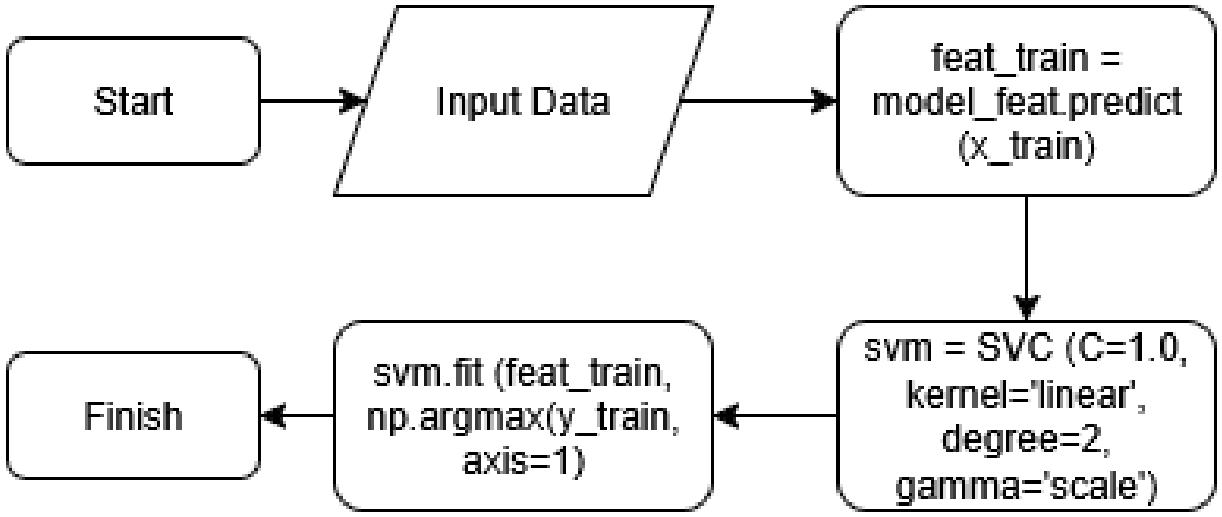

Figure 5. SVM algorithm flowchart

The flowchart illustrates the workflow of a Support Vector Machine (SVM) based classification process. The procedure begins with the input of training data, which consists of feature vectors x_train and their corresponding labels y_train. Next, a predictive model is applied to the training data to generate intermediate feature representations, denoted as feat_train = model_feat.predict(X_train). These extracted features are then used as inputs to the SVM classifier. The SVM is configured with specific hyperparameters, namely a regularization parameter (C = 1.0), a linear kernel, degree set to 2, and gamma scaled automatically. After the SVM model is trained using the feature data and class labels where class labels may be determined using the maximum value across the target vector via np.argmax(y_train, axis=1) the classifier learns an optimal decision boundary that separates the classes. Once training is completed, the process ends, resulting in a trained SVM model ready for classification or further evaluation. Next, the Random Forest models are fitted, as shown in Figures 6.

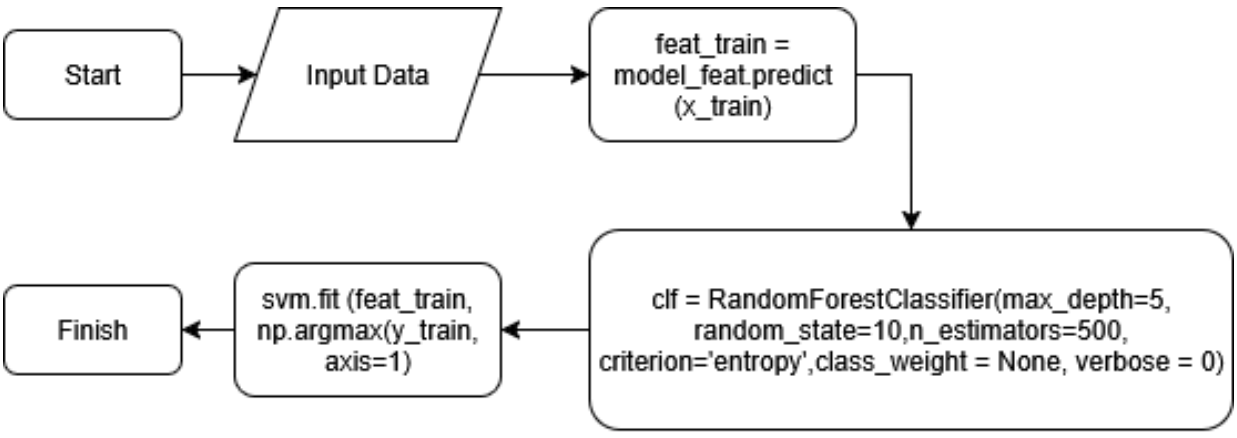

Figure 6. Random Forest algorithm flowchart

The flowchart describes the process of building a classification model using the Random Forest (RF) algorithm. The procedure starts with the input of training data, which includes the feature matrix x_train and the corresponding target labels y_train. The input data are then processed by a feature extraction model to generate transformed features, expressed as feat_train = model_feat.predict(X_train). These extracted features serve as the input for the Random Forest classifier. The RF model is initialized with predefined hyperparameters, including a maximum tree depth of 5, a fixed random state of 10 to ensure reproducibility, 500 decision trees (n_estimators = 500), and the entropy criterion to measure the quality of splits, with no class weighting and verbosity disabled. After initialization, the classifier is trained using the extracted features and class labels, where the labels are obtained by selecting the class with the highest probability via np.argmax(y_train, axis=1). Through this training process, the Random Forest learns an ensemble of decision trees to improve classification robustness and accuracy. Once training is completed, the process ends, resulting in a trained Random Forest model ready for prediction or evaluation.

The SVM model was trained using data normalized with selected features. Evaluation was performed on 88 test data samples. Table 2 shows the evaluation results of the SVM model.

Table 2. Evaluation model SVM and RF

Class	SVM				Random Forest			
	Precision	Recall	F1-score	Accuray	Precision	Recall	F1-score	Accuray
No Stress	0.90	0.93	0.92	0.92	0.92	0.80	0.86	0.86
Stress	0.93	0.90.	0.91		0.82	0.93	0.87	

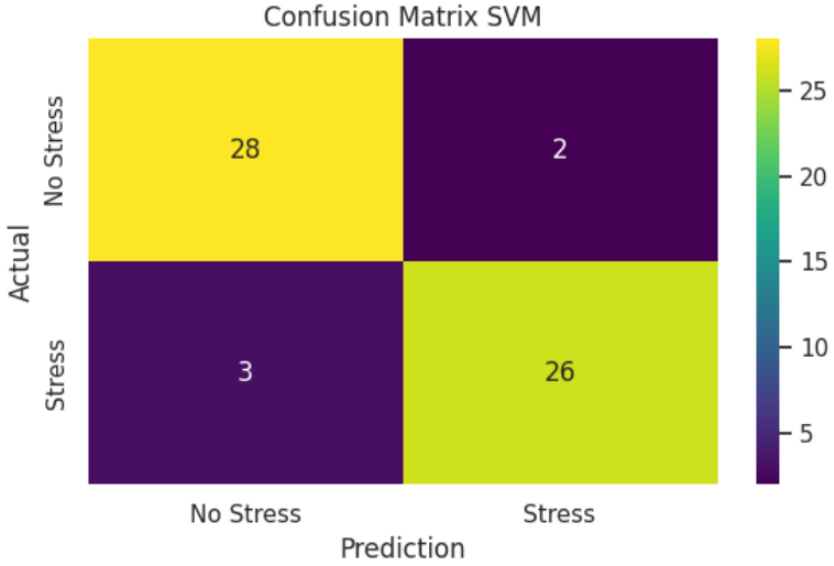

Figure 7. Confusion matrix SVM model

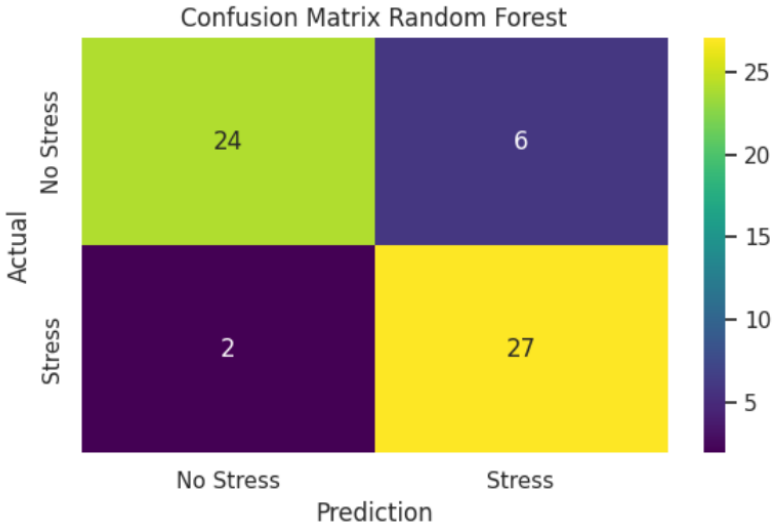

Figure 8. Confusion matrix SVM model

Figure 7 shows the confusion matrix of the SVM model. The model successfully classified 28 respondents as "Not Stressed" and 26 respondents as "Stressed" accurately, with only 5 misclassifications. This indicates that the model is not only accurate but also balanced in handling class distributions. Then, Figure 8 shows the confusion matrix of the Random Forest model. The model successfully classified 24 respondents as "Not Stressed" and 26 respondents as "Stressed" accurately, with only 8 misclassifications. This indicates that the model is not only accurate but also balanced in handling class distributions.

This study successfully developed a modular, transparent, and reproducible student stress classification pipeline based on the Support Vector Machine (SVM) algorithm. The pipeline was designed to systematically process Likert-scale survey data through stages including preprocessing, labeling, normalization, feature selection, and model evaluation. All stages were conducted with comprehensive documentation and audit trails, enabling replication by other researchers.

The balanced distribution of stress labels (50.2% “Stressed” and 49.8% “Not Stressed”) indicates that the data are not dominated by a single class, allowing the model to be trained fairly without bias toward either category. The use of the median threshold (193) as the classification cutoff represents an adaptive approach that accounts for the actual data distribution rather than assuming normality. This approach has been applied in psychometric studies to prevent classification bias caused by outliers.

Table 2 shown which presents a comparison of the performance evaluation of two classification models, Support Vector Machine (SVM) and Random Forest (RF), in predicting student stress levels. The evaluation is based on four metrics: precision, recall, F1-score, and accuracy, reported for two classes: No Stress and Stress. For the SVM model, the results indicate strong and balanced performance across both classes. The No Stress class achieves a precision of 0.90, recall of 0.93, and F1-score of 0.92, while the Stress class records a precision of 0.93, recall of 0.90, and F1-score of 0.91. The overall accuracy of the SVM model is 0.92, showing high reliability in distinguishing between stressed and non-stressed students. In comparison, the Random Forest model shows slightly lower overall performance. For the No Stress class, it achieves a precision of 0.92, recall of 0.80, and F1-score of 0.86. For the Stress class, the precision is 0.82, recall is 0.93, and the F1-score is 0.87. The overall accuracy of the Random Forest model is 0.86. Overall, the table indicates that the SVM model outperforms the Random Forest model, particularly in terms of overall accuracy and balanced performance across both classes.

The features that contributed most significantly to stress classification were related to academic pressure, sleep disturbances, and social support. These findings align with [8] stress theory, which posits that stress arises when environmental demands exceed an individual’s coping capacity. Previous studies have similarly shown that workload, disrupted rest patterns, and a lack of emotional support are major risk factors for stress among university students.

The distribution of stress across semesters indicates that academic pressure is experienced not only by final-year students but also by those in the early and middle semesters [11]. This suggests that stress is systemic and not confined to the final stages of study. Meanwhile, the distribution of stress by gender shows that female students experience slightly higher levels of stress; however, the difference is not statistically significant. Psychosocial and academic factors appear to play a more dominant role than demographic factors in influencing stress levels [12].

The developed pipeline not only produces a predictive model but also provides a documentation system and audit log that enable validation by other researchers. With its modular and transparent structure, this pipeline can be further developed into a real-time stress monitoring system within university environments.

To enhance the analytical depth of this study, inferential statistical tests were conducted to examine whether the observed differences in stress distribution across gender and academic semester were statistically significant. Given that stress labels (“Stressed” and “Not Stressed”) and demographic variables are categorical, a chi-square (χ^2) test of independence was applied. The chi-square test was used to assess the association between gender and stress category. The results indicate that there was no statistically significant association between gender and stress status ($p > 0.05$). Although descriptive results showed slight differences in stress distribution between male and female students, these differences were not sufficient to conclude that gender independently influenced stress classification outcomes. This suggests that psychosocial and academic factors play a more dominant role than gender alone in determining student stress levels. A chi-square test was also conducted to examine the relationship between academic semester and stress category. The results showed no statistically significant difference in stress distribution across semesters ($p > 0.05$). While descriptive analysis suggested an increasing trend of stress in middle to higher semesters, particularly during semesters associated with heavier academic workloads, the inferential analysis indicates that stress is experienced relatively uniformly across academic stages. The absence of statistically significant differences across gender and semester supports the robustness of the proposed stress classification pipeline, as the model performance does not appear to be biased toward specific demographic groups. These findings reinforce the notion that student stress is a multidimensional phenomenon influenced primarily by academic pressure, sleep disturbance, and social support rather than demographic attributes alone. This aligns with the recommendations of [2] and [4] who emphasize the importance of reproducibility and auditability in machine learning–based mental health classification systems.

Overall, this study demonstrates that the Support Vector Machine (SVM) method is effective in classifying student stress, and that the developed pipeline adheres to the principles of data ethics, transparency, and practicality. This system

has the potential to be implemented by educational institutions as a tool for early detection and intervention of student stress risks.

4. Conclusions and Suggestions

(SVM) algorithm. Given the high prevalence of stress among Indonesian university students and the limitations of previous studies in comprehensively integrating academic, psychosocial, and demographic factors, this research offers a novel and contextually adaptive approach. Using a quantitative cross-sectional method and Likert-scale survey data from 293 respondents, the workflow includes data conversion, score calculation, median-based labeling, normalization, feature selection, and model evaluation. The classification results indicate a balanced stress distribution across classes, and the developed SVM model achieved an accuracy of 89.77%, with consistent precision, recall, and F1-scores. The primary factors contributing to stress classification include academic pressure, sleep disturbances, and lack of social support. Stress distribution did not show bias by semester or gender but was influenced by students' psychosocial conditions. This study addresses the research question and provides a significant contribution to the field of student mental health classification using machine learning.

Bibliography

[1] A. Hapsari, A. S. Nursuwanda, H. Zuhriyah, and D. J. Vresdian, "Klasifikasi Kesehatan Mental Mahasiswa Model TMAS dengan Algoritma Decision Tree, Logistic Regression, dan Random Forest," *INTEK J. Inform. dan Teknol. Inf.*, vol. 7, no. 2, pp. 55–64, 2024, doi: 10.37729/intek.v7i2.5690.

[2] M. Razavi *et al.*, "Machine Learning, Deep Learning, and Data Preprocessing Techniques for Detecting, Predicting, and Monitoring Stress and Stress-Related Mental Disorders: Scoping Review," *JMIR Ment. Heal.*, vol. 11, pp. 1–28, 2024, doi: 10.2196/53714.

[3] S. Ghosh *et al.*, "Predicting Stress among Students via Psychometric Assessments and Machine Learning," *ACM Int. Conf. Proceeding Ser.*, no. May, pp. 662–669, 2024, doi: 10.1145/3652037.3663949.

[4] U. Madububambachu, A. Ukpebor, and U. Ihezue, "Machine Learning Techniques to Predict Mental Health Diagnoses: A Systematic Literature Review," *Clin. Pract. Epidemiol. Ment. Heal.*, vol. 20, no. 1, pp. 1–16, 2024, doi: 10.2174/0117450179315688240607052117.

[5] M. Kang, S. Shin, G. Zhang, J. Jung, and Y. T. Kim, "Mental stress classification based on a support vector machine and naive bayes using electrocardiogram signals," *Sensors*, vol. 21, no. 23, 2021, doi: 10.3390/s21237916.

[6] A. Singh, K. Singh, A. Kumar, A. Shrivastava, and S. Kumar, "Machine Learning Algorithms for Detecting Mental Stress in College Students," *2024 IEEE 9th Int. Conf. Converg. Technol. I2CT 2024*, 2024, doi: 10.1109/I2CT61223.2024.10544243.

[7] S. B. Harahap and Y. Yamasari, "Klasifikasi Tingkat Stres Mahasiswa Menggunakan RMSProp untuk Arsitektur Artificial Neural Network," *J. Informatics Comput. Sci.*, vol. 5, no. 04, pp. 560–567, 2024, doi: 10.26740/jinacs.v5n04.p560-567.

[8] L. Susanti, "Klasifikasi Tingkat Stress pada Mahasiswa Teknik Informatika dalam Melakukan Perkuliahan Metode Hybrid Menggunakan Algoritma Naive Bayes," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 8, no. 3, p. 243, 2024, doi: 10.30998/string.v8i3.17096.

[9] M. Fadhilla, R. Wandri, A. Hanafiah, P. R. Setiawan, Y. Arta, and S. Daulay, "Analisis Performa Algoritma Machine Learning Untuk Identifikasi Depresi Pada Mahasiswa," *J. Informatics Manag. Inf. Technol.*, vol. 5, no. 1, pp. 40–47, 2025, doi: 10.47065/jimat.v5i1.473.

[10] M. F. Alamsyah and A. Wijaya, "Perbandingan Metode KNN dan Naïve Bayes dalam Deteksi Tingkat Stres Berdasarkan Ekspresi Wajah," *J. Inform. J. Pengemb. IT*, vol. 10, no. 2, pp. 359–369, 2025, doi: 10.30591/jpit.v10i2.8513.

[11] S. D. Amalia, M. A. Barata, and P. E. Yuwita, "Optimization of Random Forest Algorithm with Backward Elimination Method in Classification of Academic Stress Levels," *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 633–641, 2025, doi: 10.30871/jaic.v9i3.9280.

[12] V. Oktaviani, N. Rosmawarni, and M. P. Muslim, "Perbandingan Kinerja Random Forest Dan Smote Random Forest Dalam Mendeteksi Dan Mengukur Tingkat Stres Pada Mahasiswa Tingkat Akhir," *Inform. J. Ilmu Komput.*, vol. 20, no. 1, pp. 43–49, 2024, doi: 10.52958/iftk.v20i1.9158.