JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.7 No.2 May 2025: 175-182
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v7i2.13521

SMOTE Effectiveness and various Machine Learning Algorithms to Predict Self-Esteem Levels of Indonesian Student

Mochammad Anshori[1*], Risqy Siwi Pradini[2], Wahyu Teja Kusuma[3]

[1-3]S1-Informatika, ITSK RS.DR. Soepraoen,
Malang, Indonesia 65147
moanshori@itsk-soepraoen.ac.id; risqypradini@itsk-soepraoen.ac.id;
wtkusuma@itsk-soepraoen.ac.id

*Correspondence: moanshori@itsk-soepraoen.ac.id

Abstract *- Self-esteem plays a crucial role in students' psychological well-being, influencing their academic performance and personal development. Despite its importance, self-esteem is challenging to measure due to its abstract and subjective nature. This study aims to develop a predictive model to classify students' self-esteem levels as high or low using machine learning and tabular data obtained through questionnaires. A dataset comprising 47 student responses, with 19 features consisting of social, emotional, demographic aspects, were analyzed. Five machine learning models were evaluated: Naïve Bayes, Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine (SVM). To address the class imbalance in the dataset, the study applied SMOTE for data balancing and min-max normalization for feature standardization. Model performance was assessed using accuracy and F1-score. The results reveal that SVM, particularly with an RBF kernel, outperformed other models across all scenarios. On raw data, SVM achieved 66% accuracy and an F1-score of 57.3%. After applying SMOTE, the performance improved to 80% accuracy and a 79.9% F1-score. Further enhancement with normalization resulted in the best performance, with SVM achieving 83.33% accuracy and an F1-score of 83.3%. These results demonstrate how well preprocessing methods work to enhance machine learning models for datasets that are unbalanced. The proposed SVM-based model offers promising applications in educational and psychological settings, enabling early interventions to support students' mental health.*

Keywords: *self-esteem; machine learning; psychoinformatics; health-informatics*

I. INTRODUCTION

Mental health has emerged as a pressing concern in contemporary society, particularly in the wake of the global pandemic, which has significantly affected individuals across various demographics. Issues such as loneliness, depression, anxiety disorders, and cognitive decline have been observed to escalate during this period, impacting people's overall well-being and life expectancy (Mirah Yunita et al., 2022; Xie et al., 2022). Among these psychological challenges, self-esteem plays a pivotal role, particularly for students, as it profoundly influences their development, social interactions, and academic performance (Nidia Suryani & Hamidah Rahim, 2022). Self-esteem is defined as a person's assessment of their own value based on their own beliefs, achievements, and self-perceptions (Selfilia Arum Kristanti & Eva, 2022). High self-esteem is associated with positive outcomes, including motivation, resilience, and the ability to form healthy relationships. Conversely, low self-esteem can lead to negative self-perceptions, social anxiety, and diminished productivity, posing significant challenges for individuals and society (Dewi & Ibrahim, 2019; Zhao et al., 2021).

The pandemic-induced transition to remote learning further amplified the challenges associated with low self-esteem, particularly among students. The reduced opportunities for face-to-face social interactions fostered feelings of isolation and disconnection from community life, exacerbating the psychological distress faced by many (Mirah Yunita et al., 2022). Research indicates that self-esteem is not merely an inherent trait but is shaped by various factors, including parental care and family education; teacher-student relationship at school; and peer relationship (Chen & Ma, 2023). Low self-esteem has been linked to critical outcomes such as substance abuse, depression, and, in extreme cases, suicidal tendencies. These findings underscore the importance of identifying and addressing self-esteem issues at an early stage to mitigate their adverse effects on students' academic achievements and overall development.

Despite the recognized importance of self-esteem, its evaluation and prediction remain challenging due to its inherently psychological nature. Traditional approaches to measuring self-esteem often rely on subjective self-report methods, which are susceptible to biases and inaccuracies. Recent advancements in data science and machine learning (ML) offer promising avenues for addressing these limitations(Suhartono et al., 2024). Machine learning has demonstrated significant potential in various domains, including healthcare, by enabling the development of predictive models that leverage large datasets to provide actionable insights (Cutillo et al., 2020). In the context of mental health, ML applications have been explored to predict conditions such as depression, anxiety, and cognitive decline, emphasizing their role in enhancing early diagnosis and intervention strategies (Callahan & Shah, 2017). Applying these advancements to the assessment of self-esteem represents a logical extension of this trend, particularly given the availability of increasingly sophisticated algorithms capable of handling complex, multidimensional data.

A growing body of literature supports the use of machine learning for predicting psychological traits, including self-esteem. Prior research has shown that it is feasible to predict self-esteem levels using ML models in combination with physiological data, such as EEG recordings (Buettner et al., 2021). While these approaches offer high accuracy, they often require specialized equipment and expertise, limiting their applicability in resource-constrained settings. Alternatively, the use of tabular data derived from questionnaires provides a more accessible and cost-effective solution. Questionnaires designed to capture relevant features such as social support, emotional well-being, and interpersonal relationships offer a practical means of collecting data for ML-based predictions(Ariyanti & Purwoko, 2023). However, in order to guarantee the robustness and dependability of predictive models, tabular data also poses issues, such as unequal class distributions and fluctuating data formats.

The current study builds upon these insights by employing a structured approach to develop an optimal machine learning model for predicting self-esteem levels among students. The methodology incorporates data pretreatment methods like Synthetic Minority Oversampling Technique (SMOTE) to address class imbalances and normalization to standardize feature ranges. These steps are critical for enhancing the performance of ML models, as they reduce biases introduced by imbalanced datasets and improve the interpretability of results(S. Wang et al., 2021). Several classification algorithms, including Support Vector Machines (SVM), Random Forests (RF), Decision Trees (DT), Naïve Bayes (NB), and Logistic Regression (LR), are evaluated to identify the most effective model for this task. Among these, SVM's robustness against overfitting and capacity to handle nonlinear data have allowed it to perform better in prior research (Anshori & Pangestu, 2024).

A review of relevant literature reveals that preprocessing techniques significantly influence the performance of ML models in psychological research. For instance, prior research highlighted the efficacy of SMOTE in improving classification evaluation metrices for imbalanced dataset (Wu et al., 2022), while Henderi (2021) emphasized the importance of normalization in ensuring consistency across features with varying scales(Henderi, 2021). These findings align with the objectives of the present study, which seeks to leverage these preprocessing methods to enhance the predictive accuracy of ML models. Additionally, cross-validation techniques are employed to ensure the generalizability of the results, addressing potential overfitting issues that may arise in small datasets(Gupta et al., 2021). By systematically comparing the performance of different models across various preprocessing scenarios, this study aims to identify the most reliable approach for predicting self-esteem levels.

While prior research has demonstrated the feasibility of using ML for self-esteem prediction, significant gaps remain in the literature. It is noteworthy that not many studies have examined how data balance and normalization work together to affect ML model performance in this situation. Additionally, most existing research focuses on either physiological data or raw tabular data without adequately addressing preprocessing challenges. This paper aims to close these gaps by offering a thorough examination of how preprocessing methods affect the performance of ML models. The integration of SMOTE and normalization represents a novel contribution, offering valuable insights into best practices for data preparation in psychological research.

This study's main goal is to create a reliable and accurate machine learning model that uses questionnaire-based tabular data to predict students' levels of self-esteem. The study's novelty lies in its systematic application of advanced preprocessing techniques, coupled with an evaluation of multiple ML algorithms, to identify the most effective approach for this task. By addressing the challenges associated with imbalanced and heterogeneous datasets, the study aims to provide a scalable and accessible solution for self-esteem assessment. The findings are expected to contribute to the growing body of literature on ML applications in

psychoinformatics and health informatics, offering practical implications for educators, psychologists, and policymakers. Furthermore, the study underscores the potential of ML in advancing mental health research, highlighting the importance of data-driven approaches in understanding and addressing psychological challenges.

The introduction show what is already known from the previous studies, defines the importance of the study, literature review, and state the research question. In order to understand what is already known from the previous study, the introduction must consist of discussing the relevant journal article (with citation) and summarizing the current understanding of the problem encounter.

II. METHODS

This study aims to develop a predictive machine learning (ML) model for identifying self-esteem levels among students. The research methodology is systematically structured into four key stages as shown in Figure 1. There is dataset collection, data preprocessing, model development, and model evaluation. Each stage is designed to address specific challenges and ensure the robustness and reliability of the final predictive model.

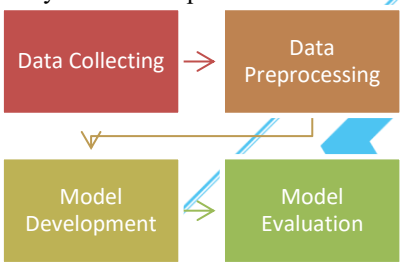

Figure 1 Methodology

2.1 Dataset Collecting

The data set used in this study was obtained through a questionnaire distributed to students aged 16 to 30 years. The questionnaire included 19 items, with two questions addressing demographic information (age and gender) and 17 questions related to self-esteem factors. These factors were derived from previous studies and included variables such as social relations, psychological well-being, social support, and emotional regulation (Ariyanti & Purwoko, 2023) . Each item was designed to capture relevant features influencing self-esteem levels. The collected dataset comprised 47 records, with 64% of the responses indicating high self-esteem and 36% indicating low self-esteem. This initial class imbalance necessitated the use of preprocessing techniques to ensure equitable representation of both classes in the analysis.

2.2 Data Preproccessing

Variables, such as gender and binary responses, were transformed into numerical values using ordinal encoding. This conversion maintained the dataset's dimensional integrity without increasing computational complexity (Trang & Nguyen, 2022). To rectify the class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was utilized.

$$P_{ij} = x_i + random(0,1) \times (x_{ij} - x_i) \quad (1)$$

Equation (1) is SMOTE formula to produce new data. SMOTE interpolates between existing data samples to create synthetic data points for the minority class, resulting in a balanced dataset. After applying SMOTE, the dataset included an equal distribution of high and low self-esteem records (S. Wang et al., 2021). To ensure consistency across variables with differing scales, min-max normalization was applied to all numerical features.

$$x' = \frac{x - min(x)}{max(x) - min(x)} \quad (2)$$

Equation (2) generates a new x' as converted data according to each feature's lowest and maximum values. The data values were scaled using this method to fall between 0 and 1, thereby reducing biases introduced by large data ranges (Henderi, 2021).

2.3 Model Development

Five widely used supervised ML algorithms were selected for this study: Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM). Each algorithm was implemented to classify the self-esteem levels into high and low categories. Three scenarios were tested: using raw data, SMOTE-balanced data, and SMOTE-balanced data with normalization.

SVM was chosen for its ability to handle nonlinear data using the radial basis function (RBF) kernel. This kernel provides effective classification by creating optimal hyperplanes between data points(Anshori, Mahmudy, et al., 2019; Anshori & Pangestu, 2024; Shiddiqi et al., 2025). SVM also offering the capability to handle high-dimensional data (Saraswati et al., 2024). DT offers interpretable classification models, while RF enhances performance by aggregating multiple decision trees and using majority voting for final predictions (Anshori et al., 2023; Charbuty & Abdulazeez, 2021). The benefit of RF is its ability to filter out noise in the data and work with vast and diverse volumes of data. The fact that it typically yields positive outcomes is another benefit.

$$\boldsymbol{Entropy}\ (\boldsymbol{S}) = -\sum_{x \in X} \boldsymbol{p}(\boldsymbol{x}) \boldsymbol{\log}\ (\boldsymbol{p}(\boldsymbol{x})) \quad (3)$$

$$Gain(S, A) = \sum_{v \in V(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (4)$$

Equation (3) and Equation (4) is respectively formula for entropy and information gain to produce tree classifier model. Where $\boldsymbol{X}$ is data and $\boldsymbol{p}(\boldsymbol{x})$ is probability density of $\boldsymbol{X}$. Entropy is equation to

measure uncertainty of random variabel. Entropy is used to calculate heterogenitas and information gain to create branch of its decision.

NB applies the Bayes theorem to estimate the probability of class membership based on observed features (Farida et al., 2023; Wicahyo et al., 2021). This method's simplicity makes it efficient for small datasets (Anshori, Mar'i, et al., 2019).

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)} \quad (5)$$

Equation (5) is the formula of NB, where $\boldsymbol{P(X|H)}$ is likelihood, $\boldsymbol{P(H)}$ is the hypothesis likelihood, $\boldsymbol{P(x)}$ is probability predictor, and $\boldsymbol{P(H|X)}$ is the probability of hypothesis H based on X condition.

LR models the relationship between features and class membership using a sigmoid function, providing probabilities for each class(Maulud & Abdulazeez, 2020). The advantage of LR is that it does not require any hyperparameters, making it easy to implement (Nusinovici et al., 2020).

$$\hat{y} = E(y|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \quad (6)$$

The mathematical function of LR is based on linear regression combined with the sigmoid function as a classification determinant. Equation (6) shows the LR equation where $\hat{\boldsymbol{y}}$ is a prediction, β is coefficient of the regression, and $\boldsymbol{x}$ is data. This equation is known as the logistic-probability function, or logit.

2.4 Model Evaluation

F1-score metrics and accuracy were used to assess the model's performance.

$$Accuracy = \frac{all\ true\ predicted}{all\ datas} \quad (7)$$

$$F1\ score = \frac{TP}{TP + 0.5\,(FP + FN)} \quad (8)$$

Equation (7) and (8) is used to calculate accuracy and F1 Score. TP (true positive) is the quantity of correctly data that is predicted, FP (false positive) is the quantity of low self-esteem data identified as high, and FN (false negative) is the quantity of high self-esteem data identified as low. The fraction of properly identified instances is measured by accuracy; however, the F1-score offers a more robust assessment for imbalanced datasets by providing a harmonic mean of precision and recall.(Douzas et al., 2019). So, F1 score can be used to indicate high sensitivity and precision of the model (W. Wang & Sun, 2021).

A 10-fold cross-validation approach was adopted to ensure the generalizability of the models. This method aids in evaluating the model's functionality and figuring out how well it can forecast the future. (Anshori & Haris, 2022; Sugihdharma & Bachtiar, 2022).

Referring Figure 2 above, in each iteration, the dataset was split into ten subsets, nine of which were utilized for training and one for testing. The average performance across all folds was reported as the final evaluation metric (Anshori & Haris, 2022).

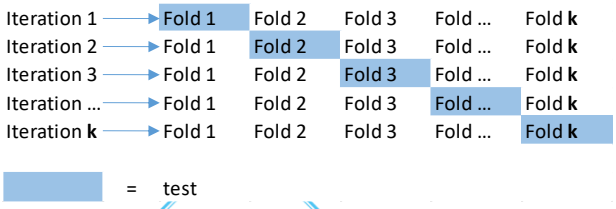

Figure 2 Illustration of k=10-fold cross validation

2.5 Analysis Scenarios

To find the best preprocessing and modeling strategy, three situations were assessed: First, the original, unbalanced dataset was used to train and test the models; second, the dataset was subjected to SMOTE oversampling to balance the class distributions; and third, the balanced and normalized dataset was used to test the models and evaluate the effect of standardization on performance.

III. RESULT AND DISCUSSION

This work aimed to create a machine learning model that could accurately predict self-esteem levels among students. The results were analyzed in the context of three preprocessing scenarios—raw data, SMOTE-balanced data, and SMOTE with normalization. The performance of the models was evaluated based on accuracy and F1-score metrics, with a particular focus on identifying the most effective preprocessing and classification methods.

3.1 Dataset Characteristics

Table 1 displayed the dataset's characteristics, which are described below. There are 20 features in the dataset, including one dependent feature and 19 independent variables. Numerical and categorical data are the two categories of data. Ordinal encoding will be used to convert categorical kinds into numerical. It will be beneficial. Nearly every attribute has a data range between 1 and 5. It is the result of the questionnaire's Likert scale. Gender is classified as either male or female. psychological strain between yes and no value and self-presentation. The class has low and high values and is categorical.

According to Figure 3, the dataset consisted of 47 records, with an initial imbalance of 64% high self-esteem and 36% low self-esteem. After applying SMOTE, the dataset was balanced, with equal representation of both classes. Table 2 of the study outlines the features and their respective data types,

while Figures 3 illustrates the class distributions before and after SMOTE application. The preprocessing steps, including normalization using min-max scaling, ensure consistency across variables, improving model performance on downstream tasks.

Table 1. Dataset details

Feature	Data type	Range
Age	Numerical	16 - 30
Gender	Categorical	M, F
Social relation	Numerical	1 – 5
Ability	Numerical	1 – 5
Psychological well-being	Numerical	1 – 5
Positive emotion	Numerical	1 – 5
Social media usage	Numerical	1 – 5
Satisfaction in living life	Numerical	1 – 5
self-presentation	Categorical	Y, N
Feelings of Shame	Numerical	1 – 5
Relationships in friendship	Numerical	1 – 5
Childhood	Numerical	1 – 5
Psychological pressure	Categorical	Y, N
Sosial support	Numerical	1 – 5
Participation in sports	Numerical	1 – 5
Interpersonal relation	Numerical	1 – 5
Management of negative emotions	Numerical	1 – 5
Control over the events experienced	Numerical	1 – 5
Emptiness	Numerical	1 – 5
Class	Categorical	L, H

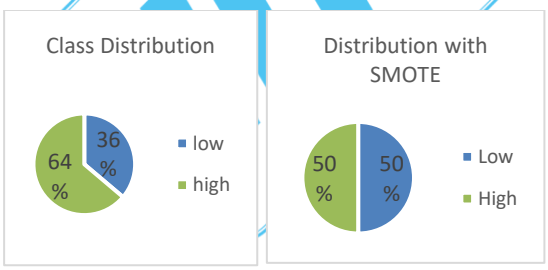

Figure 3. Dataset class distribution before and after SMOTE applied

3.2 Model Evaluation on Raw Data and SMOTE Implemented

Initial tests using the raw, unbalanced dataset revealed suboptimal performance across all machine learning models.

As shown in Table 2, the Support Vector Machine (SVM) achieved the highest accuracy (66%) and an F1-score of 57.3%. Other models, including Random Forest (RF) and Logistic Regression (LR), demonstrated comparable performance with F1-scores ranging from 57.6% to 58.2%, while Decision Tree (DT) performed the worst with an F1-score of 47.6%. These results indicate that imbalanced data significantly impacts the predictive accuracy of machine learning algorithms, necessitating effective preprocessing methods to address class imbalance.

Table 2 illustrates the evaluation metrics for this scenario. SVM outperformed other models with an accuracy of 80% and an F1-score of 79.9%. Random Forest and Logistic Regression followed, with F1-scores of 73.3% and 71.7%, respectively. These findings confirm the effectiveness of SMOTE in mitigating the effects of class imbalance by generating synthetic samples for the minority class. However, while all models exhibited enhanced performance, SVM consistently demonstrated superior results due to its capability to handle complex and high-dimensional data(Anshori & Pangestu, 2024).

Table 2 Evaluation result between raw data and oversampling with SMOTE

Method	Raw Data		SMOTE	
	Acc	F1-score	Acc	F1-score
NB	0,574	0,582	0,683	0,681
DT	0,468	0,476	0,7	0,7
RF	0,574	0,582	0,733	0,733
LR	0,596	0,576	0,717	0,717
SVM	0,66	0,573	0,8	0,799

3.3 Model Evaluation with SMOTE and Normalization

The third scenario involved applying both SMOTE and min-max normalization.

Table 3 Evaluation result with SMOTE + normalization

Method	Acc	F1-score
NB	0,683	0,681
DT	0,683	0,683
RF	0,7	0,697
LR	0,667	0,665
SVM	0,833	0,833

As detailed in Table 3, this preprocessing combination resulted in further improvements for SVM, achieving an accuracy and F1-score of

83.33%. Random Forest and Logistic Regression showed modest gains, with F1-scores of 69.7% and 66.5%, respectively. In contrast, Naïve Bayes (NB) and Decision Tree exhibited no significant improvements compared to the SMOTE-only scenario. These results align with prior studies emphasizing the importance of normalization in reducing the impact of variable scaling on machine learning models [18].

3.5 Comparative Analysis of Models

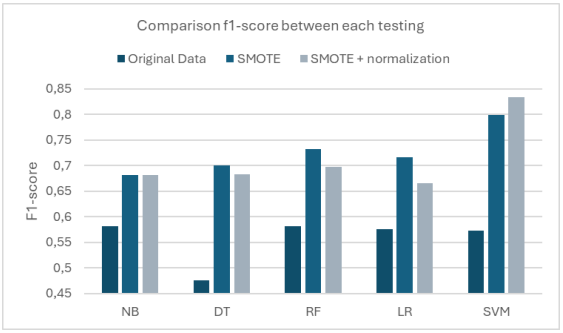

Figure 4 Comparison F1-score each experiment

Figure 4 compares the F1-scores of all models across the three scenarios. The results clearly indicate that SVM consistently outperforms other models, particularly after SMOTE and normalization. While Decision Tree and Naïve Bayes struggled to achieve competitive results, Random Forest and Logistic Regression demonstrated moderate performance gains.

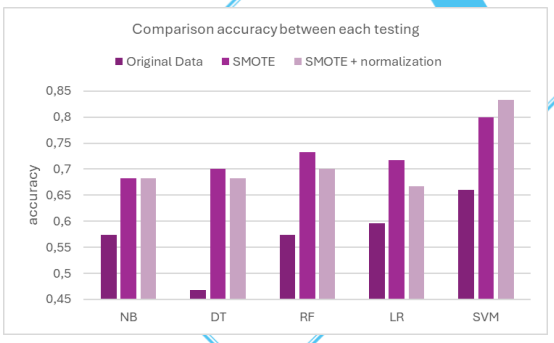

Figure 5 Comparison accuracy in each experiment

Figure 5 further corroborates these findings, showing that SVM achieved the highest accuracy in all scenarios, underscoring its robustness and adaptability to varying data preprocessing techniques.

The comparative performance of SVM across all three scenarios is graphically summarized in Figure 6. SVM consistently outperformed other models due to its ability to construct optimal hyperplanes for data separation. The use of the radial basis function (RBF) kernel proved instrumental in handling non-linear relationships within the dataset. These findings align with existing literature, which recognizes SVM as a robust classifier for small and imbalanced datasets. The study highlights the critical role of data preprocessing in enhancing the performance of machine learning models. SMOTE effectively addressed class imbalance, improving the representation of minority class samples and reducing bias. Normalization further standardized feature ranges, enhancing the interpretability and comparability of data for algorithms reliant on distance metrics. While Random Forest and Logistic Regression showed reasonable performance, their reliance on simpler decision boundaries and linear relationships limited their effectiveness compared to SVM. Decision Tree, in particular, struggled with overfitting, as evidenced by its lower F1-scores. Naïve Bayes, despite its computational efficiency, exhibited limited adaptability to the dataset's complexity, resulting in subpar performance.

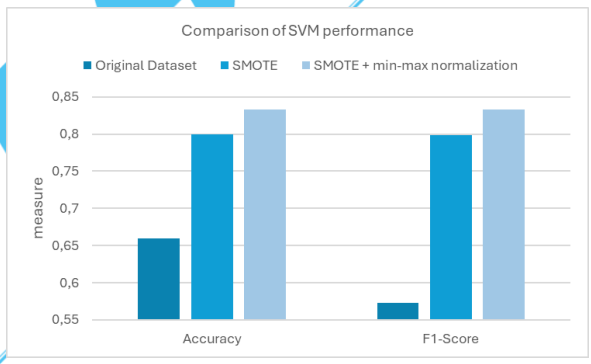

Figure 6 Comparison SVM performance

IV. CONSLUSION

This study presents a systematic approach to predicting self-esteem levels in students using various machine learning, focusing on improving classification accuracy through data preprocessing techniques. By analyzing self-esteem as a critical psychological factor influencing academic performance and personal growth, the research highlights its significance for early detection and intervention. Utilizing a dataset derived from questionnaires, the study applied Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM) models to classify students into high and low self-esteem categories. Results from raw data without SMOTE indicate that SVM performs better than any other machine learning models. The F1-score reaches 0.573 and the accuracy reaches 0.66. Subsequent comparison analysis demonstrates that SVM consistently outperforms alternative models, achieving the maximum accuracy = 0.833 and F1-

score = 0.833 when the dataset is treated using min-max normalisation for standardisation and SMOTE for data balance. These results underline the efficacy of combining preprocessing techniques with advanced machine learning algorithms to address issues such as class imbalance, which are common in psychological datasets. The superior performance of SVM with an RBF kernel reaffirms its suitability for handling nonlinear data, making it a valuable tool for psychoinformatics applications. This research contributes to the growing body of knowledge in machine learning applications for mental health by providing a framework for classifying self-esteem using accessible tabular data rather than more complex EEG data. The findings also emphasize the importance of preprocessing techniques in enhancing model performance, particularly in scenarios involving limited and imbalanced datasets. The implications of this study are twofold. Practically, it offers a predictive tool that could support early intervention strategies for students with low self-esteem. Theoretically, it lays a foundation for future research in applying machine learning to psychological assessments. Further investigations could explore larger datasets, alternative features, and optimized hyperparameters to refine the proposed model and validate its applicability in diverse educational settings.

REFERENCES

Anshori, M., & Haris, M. S. (2022). Predicting Heart Disease using Logistic Regression. Knowledge Engineering and Data Science, 5(2), 188. https://doi.org/10.17977/um018v5i22022p188-196

Anshori, M., Mahmudy, F., & Supianto, A. A. (2019). Preprocessing Approach for Tuberculosis DNA Classification using Support Vector Machines (SVM). Journal of Information Technology and Computer Science, 4(3), 233–240. https://doi.org/https://doi.org/10.25126/jitecs.201943113

Anshori, M., Mar'i, F., & Bachtiar, F. A. (2019). Comparison of Machine Learning Methods for Android Malicious Software Classification based on System Call. Proceedings of 2019 4th International Conference on Sustainable Information Engineering and Technology, SIET 2019, 343–348. https://doi.org/10.1109/SIET48054.2019.8985998

Anshori, M., & Pangestu, G. (2024). Support vector model to predict smartphone addiction in early adolescents. AIP Conference Proceedings, 2927(1). https://doi.org/10.1063/5.0192301/3279174

Anshori, M., Rikatsih, N., Haris, M. S., Kesehatan, T., Rs, I., & Kesdam, S. (2023). PREDIKSI PASIEN DENGAN PENYAKIT KARDIOVASKULAR MENGGUNAKAN RANDOM FOREST. TEKTRIKA, 7(2), 58–64.

Ariyanti, V., & Purwoko, B. (2023). Faktor-Faktor yang Memengaruhi Self-Esteem Remaja: Literature Review. Teraputik: Jurnal Bimbingan Dan Konseling, 6(3), 362–368. https://doi.org/10.26539/teraputik.631389

Buettner, R., Sauter, D., Eckert, I., & Baumgartl, H. (2021). Classifying High and Low Self-Esteem using a Novel Machine Learning Method based on EEG Data. PACIS 2021 Proceedings.

Callahan, A., & Shah, N. H. (2017). Machine Learning in Healthcare. In Key Advances in Clinical Informatics: Transforming Health Care through Health Information Technology (pp. 279–291). Elsevier Inc. https://doi.org/10.1016/B978-0-12-809523-2.00019-4

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20–28. https://doi.org/10.38094/jastt20165

Chen, X., & Ma, R. (2023). Adolescents' Self-Esteem: The Influence Factors and Solutions. Journal of Education, Humanities and Social Sciences, 8, 1562–1566. https://doi.org/10.54097/ehss.v8i.4520

Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., Mandl, K. D., Beck, T., Collier, E., Colvis, C., Gersing, K., Gordon, V., Jensen, R., Shabestari, B., & Southall, N. (2020). Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. Npj Digital Medicine, 3(1), 1–5. https://doi.org/10.1038/s41746-020-0254-2

Dewi, C. G., & Ibrahim, Y. (2019). Hubungan Self-Esteem (Harga Diri) dengan Perilaku Narsisme Pengguna Media Sosial Instagram pada Siswa SMA. Jurnal Neo Konseling, 1(2), 2019. https://doi.org/10.24036/0099kons2019

Douzas, G., Bacao, F., Fonseca, J., & Khudinyan, M. (2019). Imbalanced learning in land cover classification: Improving minority classes' prediction accuracy using the geometric SMOTE algorithm. Remote Sensing, 11(24). https://doi.org/10.3390/rs11243040

Farida, Y., Ulinnuha, N., Sari, S. K., & Desinaini, L. N. (2023). Comparing Support Vector Machine and Naïve Bayes Methods with A Selection of Fast Correlation Based Filter

Features in Detecting Parkinson's Disease. Lontar Komputer : Jurnal Ilmiah Teknologi Informasi, 14(2), 80. https://doi.org/10.24843/lkjiti.2023.v14.i02.p02

Gupta, V. K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. Big Data Mining and Analytics, 4(2), 116–123. https://doi.org/10.26599/BDMA.2020.9020016

Henderi, H. (2021). Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. International Journal of Informatics and Information Systems, 4(1), 13–20. https://doi.org/10.47738/ijiis.v4i1.73

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends, 1(4), 140–147. https://doi.org/10.38094/jastt1457

Mirah Yunita, M., Isabel, K., Ernest Keziah, B., Cristina Natasya, M., Chandra Wijaya, S., & Studi Psikologi, P. (2022). Self-Esteem Dan Kesepian Pada Mahasiswa Selama Masa Pandemi. Jurnal Psikologi Malahayati, 4(2), 114–128.

Nidia Suryani, & Hamidah Rahim. (2022). Korelasi Self Esteem Dengan Tingkah Laku Sosial Serta Implikasinya Pada SD Muhammadiyah IV Padang. Jurnal Riset Madrasah Ibtidaiyah (JURMIA), 2(2), 237–246. https://doi.org/10.32665/jurmia.v2i2.511

Nusinovici, S., Tham, Y. C., Chak Yan, M. Y., Wei Ting, D. S., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of Clinical Epidemiology, 122, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

Saraswati, N. W. S., Dewi, D. A. P. R., & Pirozmand, P. (2024). Comparative Analysis of SVM and CNN for Pneumonia Detection in Chest X-Ray. Lontar Komputer : Jurnal Ilmiah Teknologi Informasi, 15(1), 38. https://doi.org/10.24843/lkjiti.2024.v15.i01.p04

Selfilia Arum Kristanti, & Eva, N. (2022). Self-esteem dan Self-disclosure Generasi Z Pengguna Instagram. Jurnal Penelitian Psikologi, 13(1), 10–20. https://doi.org/10.29080/jpp.v13i1.697

Shiddiqi, H. A., Setiawan, K. E., & Fredyan, R. (2025). Leveraging Support Vector Machines and Ensemble Learning for Early Diabetes Risk Assessment : A Comparative Study. 7(1), 1–6. https://doi.org/10.21512/emacsjournal.v6

Sugihdharma, J. A., & Bachtiar, F. A. (2022). Myers-Briggs Type Indicator Personality Model Classification in English Text using Convolutional Neural Network Method. Jurnal Ilmu Komputer Dan Informasi, 15(2), 93–103. https://doi.org/10.21609/jiki.v15i2.1052

Suhartono, D., Ciputri, M. M., & Susilo, S. (2024). Machine Learning for Predicting Personality using Facebook-Based Posts. Engineering, MAthematics and Computer Science Journal (EMACS), 6(1), 1–6. https://doi.org/10.21512/emacsjournal.v6i1.10748

Trang, K., & Nguyen, A. H. (2022). A Comparative Study of Machine Learning-based Approach for Network Traffic Classification. Knowledge Engineering and Data Science, 4(2), 128. https://doi.org/10.17977/um018v4i22021p128-137

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. Scientific Reports, 11(1), 1–11. https://doi.org/10.1038/s41598-021-03430-5

Wang, W., & Sun, D. (2021). The improved AdaBoost algorithms for imbalanced data classification. Information Sciences, 563, 358–374. https://doi.org/10.1016/j.ins.2021.03.042

Wicahyo, A., Pudoli, A., & Kusumaningsih, D. (2021). Penggunaan Algoritma Naive Bayes dalam klasifikasi Pengaruh Pencemaran Udara. Jurnal ICT : Information Communication & Technology, 20(1), 103–108. https://ejournal.ikmi.ac.id/index.php/jict-ikmi/article/view/332

Wu, T., Fan, H., Zhu, H., You, C., Zhou, H., & Huang, X. (2022). Intrusion detection system combined enhanced random forest with SMOTE algorithm. Eurasip Journal on Advances in Signal Processing, 2022(1). https://doi.org/10.1186/s13634-022-00871-6

Xie, Y., Xu, E., & Al-Aly, Z. (2022). Risks of mental health outcomes in people with covid-19: Cohort study. The BMJ, 376, 1–13. https://doi.org/10.1136/bmj-2021-068993

Zhao, Y., Zheng, Z., Pan, C., & Zhou, L. (2021). Self-Esteem and Academic Engagement Among Adolescents: A Moderated Mediation Model. Frontiers in Psychology, 12(June). https://doi.org/10.3389/fpsyg.2021.690828