COMPARISON OF SVM, RANDOM FOREST, AND LOGISTIC REGRESSION PERFORMANCE IN STUDENT MENTAL HEALTH SCREENING

[1*]VANNES WIJAYA, [2]NUR RACHMAT

Faculty of Informatics Engineering, University Multi Data Palembang

Jl. Rajawali No.14, 30113 Palembang, Sumatera Selatan

e-mail: [1]vanneswijaya04@mhs.mdp.ac.id, [2]nur.rachmat@mdp.ac.id

[*]Corresponding author

ABSTRACT

Mental health is an essential aspect for university students, as undetected mental health disorders can have a significant impact on students' academic performance and well-being. This study contributes by evaluating Synthetic Minority Oversampling Technique (SMOTE)'s role in improving classification models' performance. Despite the increasing use of machine learning in mental health detection, limited research has addressed the challenges posed by imbalanced datasets, particularly in smaller student populations. This research aims to develop a mental health early detection system based on student data from Multi Data University Palembang using the Mental Health Scale (SKM)-12 mental health measurement. The system aims to remind students' awareness of the importance of mental health. To improve accuracy, this research compares the performance of three models, namely Support Vector Machine, Random Forest, and Logistic Regression, both with and without using SMOTE. The dataset obtained is 78 students, and SKM-12 consists of several groups, namely optimal mental health profile with symbol (+-), maximum mental illness profile with symbol (++), minimum mental illness profile with symbol (--), and minimal mental health profile with symbol (-+). The results of this study using the Logistic Regression method using SMOTE obtained better model performance compared to other methods, with an accuracy of 89.28%, an average class precision of 89.5%, an average class recall of 89.75%, and an average F1 - class score of 88.5%. This research shows that overcoming class imbalance using SMOTE can significantly improve the performance of mental health classification models.

Keywords: *Logistic Regression, Random Forest, Mental Health Scale, SMOTE, Support Vector Machine, Health Screening*

1. INTRODUCTION

Mental health is an important aspect in realizing overall health [1]. Mental health is a stable psychological and emotional state where a person can utilize their cognitive and emotional abilities to fulfill their daily needs and participate in their community [2]. Often, adolescents experience stress, especially at certain moments in their lives. Adolescents are considered vulnerable to mental disorders and thus require more attention as they are the country's assets and the next generation of the nation. Mental health problems in today's modern era can arise due to various pressures in life. Students are an adult age group who often experience pressure and confusion about studies, family, and other aspects of life [3]. College students as a group of ages who experience the transition from adolescence to adulthood, students tend to experience stress, especially those originating from the academic process [4]. The results of several interviews with university resource persons, factors or causes of students experiencing mental health disorders accumulate tasks, family economic factors, final project or theses, and personal or family problems.

Mental health is important for freshmen, undergraduates, and graduating students. Mental health for new college students is very important in order to adapt to the lecture environment. The environment during school and college is certainly very different. College students will find different learning methods compared to their school days. Mental health for college students who are running the lecture process is important in order to complete their academic

DOI: https://doi.org/10.54732/jeecs.v9i2.9

tasks well. For final year students who are in the process of studying, thesis is one of the causes of mental health problems [5].

Previous research aimed to evaluate various machine learning algorithms in the context of depression prediction, utilizing the growing availability of mental health data. The study sought to develop predictive models that could significantly contribute to understanding depression risk and implementing more timely interventions. The methods used in the research included Random Forest, Naïve Bayes, and K-Nearest Neighbors (KNN). The findings revealed that the Random Forest method achieved exceptional performance, with an accuracy of 91%, an F1-score of 91%, and both precision and recall around 91% [6]. A related study discusses the classification of student mental health data, where mental health data is used as input for a model to develop and apply for classifying test data. The study employed the SVM and Naïve Bayes methods, concluding that the SVM algorithm outperformed the Naïve Bayes classifier in classifying students' mental health. The classification results include "Yes," indicating that students require specialized therapy, and "No," indicating no need for such therapy. The SVM method achieved an accuracy of 94.37% [7].

A study on mental health focuses on classifying students into various mental health issue categories, including stress, depression, and anxiety, using machine learning algorithms. The methods employed in this research include Decision Tree, Neural Network, Support Vector Machine, Naïve Bayes, and Logistic Regression. For the stress model, the Decision Tree method achieved the highest accuracy of 84.44%. In the depression model, the Support Vector Machine (SVM) method attained the highest accuracy of 88.15%. Meanwhile, for the anxiety model, Logistic Regression achieved the highest accuracy of 71.85% [8]. Another study aimed to develop a method for predicting MBTI personality types based on textual data. The research utilized the SMOTE technique to address data imbalance issues. Six different machine learning models were individually tested, including Logistic Regression, LSVC (Linear Support Vector Classification), SGD (Stochastic Gradient Descent), Random Forest, XGBoost, and CatBoost. The findings revealed that Logistic Regression was the best-performing model, achieving an average F1-score of 0.8282. Additionally, the use of the SMOTE technique successfully improved model performance, increasing the F1-score to 0.8337 [9].

Another study examined the impact of emotions and mental health on students' cumulative grade point average (CGPA) using machine learning algorithms. To address data imbalance, the research employed the Synthetic Minority Oversampling Technique (SMOTE). The methods used in this study included Logistic Regression, Decision Tree, Random Forest, SVC, XGBoost, KNN, Voting Classifier, and Stacking Classifier. The results indicated that the Logistic Regression model achieved an accuracy of 86.55%, while the Random Forest model achieved a slightly higher accuracy of 87.62% [10]. Previous studies have demonstrated that Support Vector Machine (SVM), Random Forest, and Logistic Regression methods can achieve good accuracy in various classification cases. Additionally, the use of the Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance has proven effective in improving model accuracy. Therefore, this study employs Logistic Regression, SVM, Random Forest, and SMOTE to address class imbalance, aiming to achieve optimal model performance in predicting students' mental health.

In this study using the SKM-12 mental health measurement tool [11]. SKM-12 is the result of modified questions from the Mental Health Inventory measuring instrument [12]. It measures mental health from positive aspects (positive emotions, love, life satisfaction) and negative aspects (anxiety, depression, and loss of control) [13]. Then it was refined again by simplifying the number of items question. In each aspect (positive and negative), 12 items question were reduced to 6 for each aspect, so that there were 12 items question [14]. The positive and negative aspects are referred to as psychological well-being and psychological distress. Mental health data can be classified based on highs and lows of psychological well-being and psychological distress. The classified data is put into four separate groups. First, the optimal mental health profile (+-) indicates high psychological well-being and low psychological distress. Second, the maximum mental illness profile (++) indicates high psychological well-being and high psychological distress. Third, the minimum mental illness profile (--) indicates low psychological well-being and low psychological distress. Finally, the minimal mental health profile (-+) indicates low psychological well-being and high psychological distress [11].

This study used the SKM-12 mental health measurement tool to evaluate the mental health of university students. This study aims to compare several machine learning models, namely Support Vector Machine (SVM), Random Forest, and Logistic Regression, both with and without SMOTE method. Model performance evaluation was conducted using Confusion Matrix to determine the best model in predicting students' mental health. The results show that the Logistic Regression model with the SMOTE imbalance method provides the best performance in predicting student mental health, based on evaluation using Confusion Matrix with an accuracy of 89.28%, an average class precision of 89.5%, an average class recall of 89.75%, and an average F1 - class score of 88.5%.

2. RESEARCH METHODOLOGY

Before the research is carried out, first conduct a theoretical review and literature study of publication and research manuscripts, in order to understand the methods and steps in the research [15].

2.1 Research flow

This research flow consists of data collection, preprocessing data, data splitting, implementation SMOTE method, or not using the SMOTE method, further implementation algorithm (SVM, Random Forest, Logistic Regression), classification evaluation, and result. At the data collection stage where data is collected using the SKM-12 questionnaire in the Multi Data University Palembang environment. Then in the preprocessing data stage, the answer data is converted into likert scale values to facilitate calculation and produce mental health profile classes. The answer values include very often 5, often 4, sometimes 3, rarely 2, and never 1. The four classes consist of an optimal mental health profile with a symbol (+-), a maximum mental illness profile with a symbol (++), a minimum mental illness profile with a symbol (--), and a minimal mental health profile with a symbol (-+).

After the preprocessing data stage, the next stage is the data splitting stage, where the data will be divided into 2, namely 70% data for training and 30% data for testing. For the next stage, the author tests for the first stage not using the SMOTE method, which is directly at the stage of implementing the three methods, namely SVM, Random Forest, and Logistic Regression. The second stage the author tests by using the SMOTE method to balance the class, after using the SMOTE method, then the next stage is to implement the three methods, namely SVM, Random Forest, and Logistic Regression.

For the first stage, not using the SMOTE method resulted in 3 models, namely SVM, Random Forest, and Logistic Regression. For the second stage using the SMOTE method produces 3 models namely SVM, Random Forest, and Logistic Regression. So that it produces 6 models, namely, SVM not SMOTE, Random Forest not SMOTE, Logistic Regression not SMOTE, SVM with SMOTE, Random Forest with SMOTE, and Logistic Regression with SMOTE. The next stage is classification evaluation, where the six models are tested with testing data to produce TP, TN, FP, and FN which form a confusion matrix. At the result stage, the TP, TN, FP, and FN results of the six models will be calculated and produce accuracy, precision, recall, and f1-score. For the research flow can be seen in the research flow chart (Figure 1).

2.2 Support Vector Machine (SVM)

One of the statistical methods that can be used for classification is Support Vector Machine (SVM). SVM is a technique that aims to find a hyperplane to separate two data sets from two different classes [16]. Support Vector Machine works by separating classes of data using an algorithm to find the optimal hyperplane in the input space. The best method to find the hyperplane that separates two classes is to measure the margin of the hyperplane and find the maximum point of the margin [17].

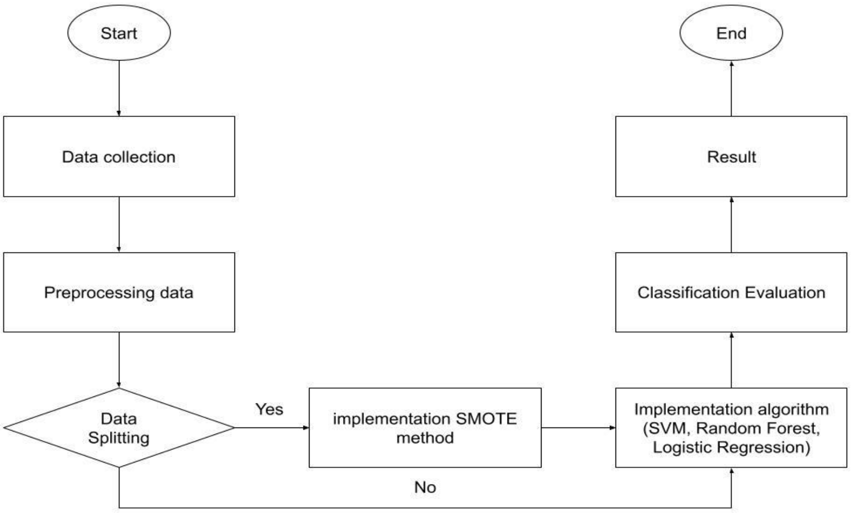
Start
End
Data collection
Result
Preprocessing data
Classification Evaluation
Data Splitting
Yes
implementation SMOTE method
Implementation algorithm (SVM, Random Forest, Logistic Regression)
No

Figure 1. Research flow chart

The main problem in SVM is to find a hyperplane, expressed by the equation <w, x> + b = 0, to separate the data x_j consisting of two classes, namely *yi = {+1, -1},* with a maximum margin. This margin refers to the distance between the hyperplane and the data from each class. The hyperplane is then used as the decision function *f(x)* in solving the two-class classification problem. The following is the formula for *f(x)* in equation 1 [18].

$$f(\emptyset(x)) = \ sign(w.\emptyset(x) + b) = sign(\sum_{i=1}^{N} \alpha_i y_i \emptyset(x_i)^T.\emptyset(x) + b) \tag{1}$$

Description:
w : weight
x : input variable value
b : bias

The formula is used to calculate the prediction results. Hyperplane can be uniquely determined based on the values of *w* and *b* obtained. The data x_i which is a subset of the training data that is on the margin, is called the support vector [18].

2.3 Randon Forest

Random Forest applies a straightforward analysis method to select nodes for constructing the root node, internal nodes, and leaf nodes using the same attributes and information, regardless of the criteria applied. This method achieves high accuracy [19]. The working mechanism of Random Forest involves combining multiple Decision Trees to achieve stable and accurate predictions. Random Forest consists of a collection of Decision Trees trained using the bagging method.

The random forest method is an evolution of the CART method, using boostrap aggregating (bagging) and random feature selection [20]. Random Forest is a collection of Decision Trees trained with bagging methods to produce stable and accurate predictions. The Decision Tree algorithm includes several variants, such as ID3, which utilizes entropy, and CART, which relies on the Gini index. The following is the impurity value formula in the CART algorithm in equation 2 and the Gini index value is represented in equation 3 [21].

.

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2 \tag{2}$$

Description:
P_i : the probability value of a tuple value D in a class
m : number of class labels

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{3}$$

The Gini index evaluates binary splits for each attribute. To assess a binary split, it calculates the weighted sum of the impurities for each resulting partition. For instance, if a binary split divides partition D into D1 and D2, the Gini index value for D based on that split can then be determined [21]. Regression analysis is a method aimed at understanding the effect of one variable on another.

2.4 Logistic Regression

Logistic Regression is a supervised method in machine learning, used to evaluate data and explain the relationship between one or more prediction variables and one response variable. The value of the Logistic Regression response variable ranges between 0 and 1, with a value cutoff of 0.5 [22]. The following represents the simple linear regression model as shown in equation 4 [23].

$$Y = \ \beta_0 + \beta_1 X + \mathcal{E} \tag{4}$$

Description:
Y : the dependent variable (predicted value).
X : independent variable
β_0 : constant
β_1 : regression coefficient (increase or decrease value)

ε : random error

Because this study uses more than 2 mental health profiles, the softmax function can be used for multi-class logistic regression classification. The softmax function is used to calculate probabilities from output results, with the highest probability value from the output layer taken as the prediction result. Softmax computes the probability distribution from a vector of real numbers. It produces outputs ranging between 0 and 1, with the total probabilities summing to 1. The following is the softmax equation in equation 5 [24].

$$softmax(xi) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (5)$$

2.5 Synthetic Minority Over-Sampling Technique (SMOTE)

To overcome class imbalance in dataset, the synthetic minority over-sampling technique (SMOTE) is popular [25]. SMOTE is a technique that equalizes the dataset by artificially generating new instances for the minority class, helping to achieve a balanced dataset [26]. SMOTE (Synthetic Minority Oversampling Technique) generates synthetic data for the minority class without simply duplicating existing samples, helping to address the challenge of overfitting. The process begins by sequentially selecting each minority class sample as the base for generating additional synthetic samples. This process is repeated n times. Finally, linear interpolation is applied between the base sample and the selected neighbors to create n new synthetic samples [27].

2.6 Mental Health Scale (SKM-12)

The Mental Health Scale (SKM-12) is a mental health measurement tool consisting of 12 questions containing 6 questions for aspects of psychological well-being and 6 questions for aspects of psychological distress. From these two aspects, it produces 4 group categories, namely profile optimal mental health (+-), profile maximum mental illness (++), profile minimum mental illness (--), and profile minimal mental health (-+). Table 1 shows the list of questions SKM-12 on each aspect.

2.7 Data collection

The first stage of this research is data collection. In SKM-12 there are 12 questions consisting of 6 psychological well-being questions and 6 psychological distress questions, which are then made into a questionnaire in the form of a Google Form so that it can be accessed easily. For the distribution of questionnaires in the Multi Data Palembang University environment that can access the Google Form web link. For the answer to each questionnaire question using a Likert scale consisting of 5, namely very often, often, sometimes, rarely, and never [11]. The results of data collection from May to June 2024, with different semester levels and majors, obtained 78 respondents. Following Table 2 are 10 sample data that have been created in csv format, where questions 1 to 6 are psychological wellbeing questions and questions 7 to 12 are psychological distress questions.

Tabel 1. Question SKM-12

No	**Question**	
	Psychological wellbeing	**Psychological distress**
1	Daily life is full of interesting things	Finding yourself as a confused or frustrated person
2	You generally enjoy what you do	Feeling like a tired person or feeling helpless
3	Feel comfortable communicating with your friends	Feeling at the lowest point
4	Feeling valuable because of your friend's treatment	Taking time to enjoy the feeling of despair
5	Feeling happy in living this life	Feeling a loss of control over thoughts, feelings, and behavior
6	Enjoying what happens in this life	Feeling like you have nothing to look forward to in the future

Tabel 2. Sample Data SKM-12

Variable	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8	Data 9	Data 10
question 1	2	5	4	2	3	3	3	3	4	4
question 2	3	4	4	2	4	3	4	5	4	5
question 3	4	4	4	3	5	4	3	5	4	4
question 4	4	5	4	2	4	3	2	3	3	4
question 5	3	5	4	4	5	3	3	4	2	4
question 6	3	4	4	4	5	3	4	5	5	4
question 7	3	5	3	2	3	3	3	4	2	3
question 8	3	3	3	2	4	3	2	4	3	2
question 9	3	4	2	1	2	2	2	4	2	2
question 10	3	1	3	1	1	1	2	5	1	2
question 11	3	4	3	3	3	1	3	4	2	2
question 12	3	2	2	2	4	1	2	2	1	1
mental health profile	minimal mental health	maximum mental illness	optimal mental health	minimum mental illness	optimal mental health	minimum mental illness	minimum mental illness	maximum mental illness	optimal mental health	optimal mental health

2.8 Preprocessing data

The results of the questionnaire were then processed to categorize into 4 mental health profile classes using the SKM-12 calculation method [11]. The answers were converted into likert scale values for easy calculation and to produce mental health profile classes. Answer values include very often 5, often 4, sometimes 3, rarely 2, and never 1. The 4 classes consist of optimal mental health profile with symbol (+-), maximum mental illness profile with symbol (++), minimum mental illness profile with symbol (--), and minimum mental health profile with symbol (-+) [11]. The result of dataset management contains 13 columns and 78 data, 12 columns of questions and mental health profiles. The mental health profile consists of 23 optimal mental health (+-), 16 maximum mental illness (++), 23 minimum mental illness (--), 16 minimal mental illness (-+). The dataset is stored in csv format which is then stored in Google Drive to make it easier to integrate into Google Colaboratory.

2.9 Data splitting

This study compares several models with different machine learning methods to get the best performing model. For the comparison of machine learning methods using SMOTE and without SMOTE. In this study, we divided the training data as much as 70% and 30% testing data.

2.10 Implementation SMOTE method

This study uses the SMOTE method or without the SMOTE method to see whether the algorithm's evaluation performance increases or decreases. The SMOTE method addresses class imbalance by increasing the number of samples in the minority class by generating new synthetic examples. This helps the machine learning model to pay more attention to the minority class, reducing the tendency to ignore the minority class. In this research for the first stage using SMOTE, the results obtained after doing SMOTE are from 78 data to 92 data. Where the data for the SKM-12 mental health profile becomes balanced, namely 23 data from 4 classes.

2.11 Implementation algorithm

in this study compared the SVM, Random Forest, and Logistic Regression algorithms. in this study using Google Colaboratory to simplify the operation of the algorithm so as to produce several models. the resulting models include SVM with SMOTE, SVM without SMOTE, Random Forest with SMOTE, Random Forest without SMOTE, Logistic Regression with SMOTE, and Logistic Regression without SMOTE.

2.12 Classification evaluation and confusion matrix

After implementing the algorithm, classification evaluation will be conducted, where the model will be tested with test data. The 4 mental health classes will produce TP, TN, FP, and FN which will form a confusion matrix. Confusion matrix will get the results of model accuracy, precision of each class, recall of each class, and F1 - score of each class.

Table 3. Confusion Matrix

	Positive Prediction	Negative Positive
Actual Positive	TP	FN
Actual Negative	FP	TN

Confusion Matrix is information about the actual classification results that can be predicted by a classification system. Accuracy is the determination of the system in performing the classification process correctly. Precision is the ratio of the number of relevant documents to the total number of documents found in the classification system. Recall is the ratio of the number of documents recovered by the classification system to the total number of relevant documents. F-measure is a popular evaluation metric for dealing with class imbalance problems [28].

Description:

TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative

Here are the equations for calculating the confusion matrix algorithm in equations 6, 7, 8, and 9:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (8)$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \quad (9)$$

3. RESULTS AND DISCUSSIONS

In this research, the algorithm implementation stage consists of SVM, Random Forest, and Logistic Regression algorithms. This stage is carried out in 2 stages, namely the test results without SMOTE and using SMOTE. In these 2 stages, 6 models were produced. The model will then be tested with testing data that has been divided previously with a ratio of 30% of the data. resulting in TP, TP, TN, FP, and FN which will evaluate the model using confusion matrix. The following are the steps in implementing the algorithm.

3.1 Result implementation and classification evaluation algorithm not SMOTE

At this stage of the research, after data splitting is done, the training data will be trained on the algorithm, then the classification evaluation will be carried out with the testing data to obtain model performance results. The following Table 4 displays the evaluation of the algorithm model without SMOTE.

3.2 Result implementation and classification evaluation algorithm with SMOTE

At this stage of the research, after data splitting is done, training data and testing data are SMOTE to increase the minority class. After the data is SMOTE, the training data will be trained on the algorithm, then the classification evaluation will be carried out with the testing data to obtain model performance results. The following Table 5 displays the evaluation of the algorithm model with SMOTE.

Table 4. Classification evaluation algorithm not SMOTE.

Algorithm	Mental Health Class	Accuracy	Precision	Recall	F1 - score
SVM	profile optimal mental health (+-)	87.5%	86%	100%	92%
	profile maximum mental illness (++)		100%	83%	91%
	profile minimum mental illness (--)		80%	67%	73%
	profile minimal mental health (-+)		86%	100%	92%
Random Forest	profile optimal mental health (+-)	70.83%	83%	83%	83%
	profile maximum mental illness (++)		67%	67%	67%
	profile minimum mental illness (--)		60%	50%	55%
	profile minimal mental health (-+)		71%	83%	77%
Logistic Regression	profile optimal mental health (+-)	83%	83%	83%	83%
	profile maximum mental illness (++)		100%	83%	91%
	profile minimum mental illness (--)		67%	67%	67%
	profile minimal mental health (-+)		86%	100%	92%

Table 5. Classification evaluation algorithm with SMOTE.

Algorithm	Mental Health Class	Accuracy	Precision	Recall	F1 - score
SVM	profile optimal mental health (+-)	85.71%	100%	89%	94%
	profile maximum mental illness (++)		70%	100%	82%
	profile minimum mental illness (--)		75%	75%	75%
	profile minimal mental health (-+)		100%	75%	86%
Random Forest	profile optimal mental health (+-)	85.71%	100%	89%	94%
	profile maximum mental illness (++)		75%	86%	80%
	profile minimum mental illness (--)		75%	75%	75%
	profile minimal mental health (-+)		88%	88%	88%
Logistic Regression	profile optimal mental health (+-)	89.28%	100%	100%	100%
	profile maximum mental illness (++)		100%	71%	83%
	profile minimum mental illness (--)		80%	100%	89%
	profile minimal mental health (-+)		78%	88%	82%

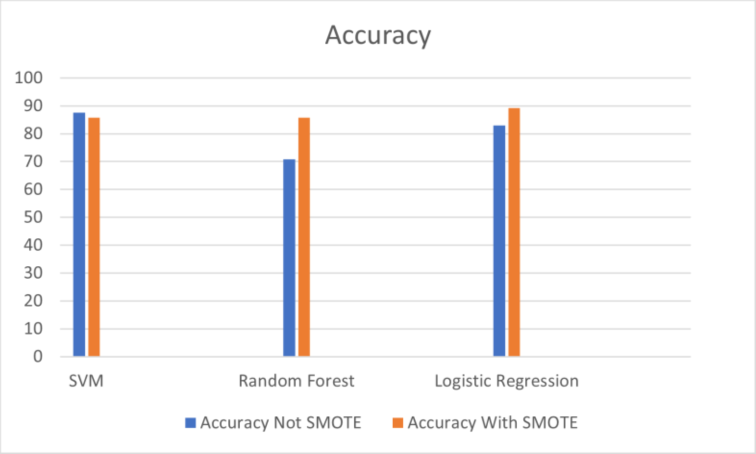

Figure 2. Accuracy comparison chart

3.3 Discussions

This discussion compares the classification evaluation results, in the form of a chart. The following Figure 2 shows the accuracy comparison chart.

The accuracy of the logistic regression method has increased with SMOTE, the accuracy obtained before using SMOTE is 83% accuracy by using SMOTE to 89.28%. The following Table 6 is a comparison of class averages for precession, recall, and f1 - score using SMOTE and without SMOTE.

The precision of Logistic Regression increases with SMOTE, and has a higher average than other methods with a precision of 89.5%. The following Figure 3 shows the precision comparison chart. The recall of Logistic Regression increases with SMOTE, and has a higher average than other methods with a precision of 89.75%. The following Figure 4 shows the recall comparison chart.

The F1 - score of Logistic Regression increases with SMOTE, and has a higher average than other methods with a precision of 88.5%. The following Figure 5 shows the F1 - score comparison chart.

Table 6. Comparison of class averages for precession, recall, f-1 score using SMOTE and without SMOTE

Algorithm	Precision		Recall		F1 - score	
	Not SMOTE	With SMOTE	Not SMOTE	With SMOTE	Not SMOTE	With SMOTE
SVM	88%	86.25%	87.5%	84.75%	87%	84.25%
Random Forest	70.25%	84.5%	70.75%	84.5%	70.5%	84.25%
Logistic Regression	84%	89.5%	83.25%	89.75%	83.25%	88.5%

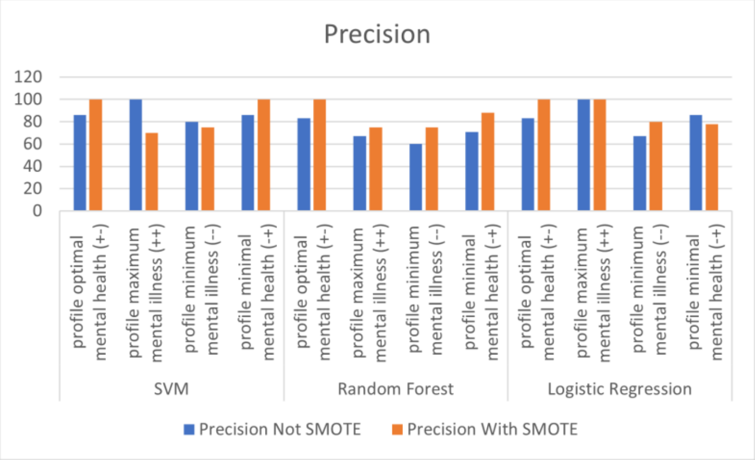

Figure 3. Precision comparison chart

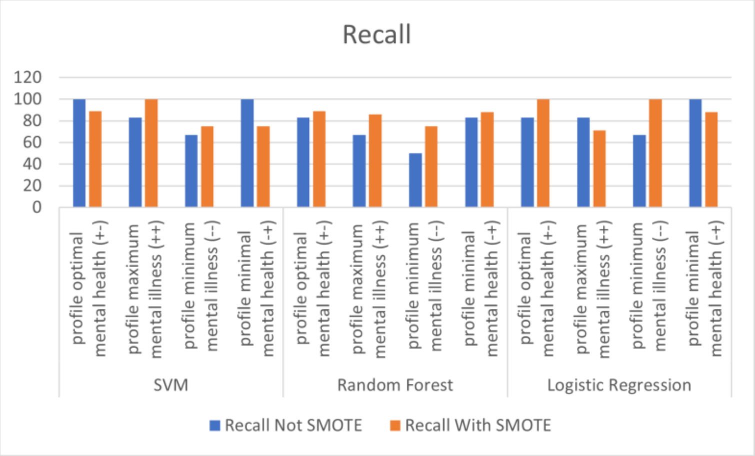

Figure 4. Recall comparison chart

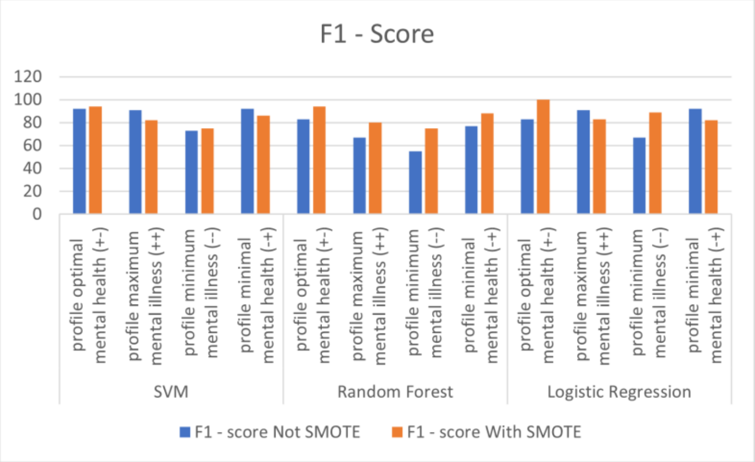

Figure 5. F1 - score comparison chart

4. CONCLUSION

The results of this study show that the use of SMOTE improves the performance of the model in Classification Evaluation. Logistic Regression method without SMOTE is lower than SVM method without SMOTE, but the results show that the Logistic Regression method has improved accuracy, precision, recall, and f1-score after using SMOTE and higher performance than SVM method without SMOTE. The results show that the Logistic Regression method with SMOTE has an accuracy of 89.28%, an average class precision of 89.5%, an average class recall of 89.75%, and an average class f1 - score of 88.5%. So that in further research the model can be applied to mental health screening systems, or for further research try to increase data and also use different methods to increase the accuracy of the model.

REFERENCES

[1] I. A. Ridlo, (2020), “Pandemi COVID-19 dan Tantangan Kebijakan Kesehatan Mental di Indonesia,” *INSAN Jurnal Psikologi dan Kesehatan Mental*, vol. 5, no. 2, pp. 155–164, doi:10.20473/jpkm.v5i12020.155-164.

[2] Z. Zulkarnain and S. Fatimah, (2019), “Kesehatan dan Mental dan Kebahagiaan: Tinjauan Psikologi Islam,” *Mawa’Izh: Jurnal Dakwah Dan Pengembangan Sosial Kemanusiaan*, vol. 10, no. 1, pp. 18–38, doi:10.32923/maw.v10i1.715.

[3] Y. Huang, S. Li, B. Lin, S. Ma, J. Guo, and C. Wang, (2022), “Early Detection of College Students’ Psychological Problems Based on Decision Tree Model,” *Frontiers in Psychology*, vol. 13, pp. 1–10, doi:10.3389/fpsyg.2022.946998.

[4] A. H. Azizah, S. Warsini, and K. P. Yuliandari, (2023), “Hubungan Stres Akademik dengan Kecenderungan Depresi Mahasiswa Ilmu Keperawatan Universitas Gadjah Mada pada Masa Transisi Pandemi COVID-19,” *Jurnal Keperawatan Klinis dan Komunitas (Clinical and Community Nursing Journal)*, vol. 7, no. 2, p. 114, doi:10.22146/jkkk.84827.

[5] M. K. Sari and E. A. Susmiatin, (2023), “Deteksi Dini Kesehatan Mental Emosional pada Mahasiswa,” *Jurnal Ilmiah STIKES Yarsi Mataram*, vol. 13, no. 1, pp. 10–17, doi:10.57267/jisym.v13i1.226.

[6] M. Rijal, F. Aziz, and S. Abasa, (2024), “Prediksi Depresi : Inovasi Terkini Dalam Kesehatan Mental Melalui Metode Machine Learning Depression Prediction : Recent Innovations in Mental Health Journal Pharmacy and Application,” *Journal Pharmacy and Application of Computer Sciences*, vol. 2, no. 1, pp. 9–14.

[7] H. D. Putra, L. Khairani, and D. Hastari, (2023), “Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Data Kesehatan Mental Mahasiswa,” in *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat*, pp. 120–125.

[8] S. Mutalib, N. S. M. Shafiee, and S. Abdul-Rahman, (2021), “Mental Health Prediction Models Using Machine Learning in Higher Education Institution,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 5, pp. 1782–1792, doi:10.17762/turcomat.v12i5.2181.

[9] G. Ryan, P. Katarina, and D. Suhartono, (2023), “MBTI Personality Prediction Using Machine Learning and SMOTE for Balancing Data Based on Statement Sentences,” *Information (Switzerland)*, vol. 14, no. 4, doi:10.3390/info14040217.

[10] F. M. Basysyar, G. Dwilestari, and A. I. Purnamasari, (2024), “Analysis Student Emotions And Mental Health on Cumulative GPA Using Machine Learning and Smote,” *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 2, pp. 361–368, doi:10.33480/jitk.v10i2.5967.ANALYSIS.

[11] R. Aziz, R. Mangestuti, Y. Sholichatun, I. T. Rahayu, E. K. Purwaningtyas, and E. N. Wahyuni, (2022), “Model Pengukuran Kesehatan Mental pada Mahasiswa di Perguruan Tinggi Islam,” *Journal of Islamic and Contemporary Psychology (JICOP)*, vol. 1, no. 2, pp. 83–94, doi:10.25299/jicop.v1i2.8251.

[12] C. T. Veit and J. E. Ware, (1983), “The structure of psychological distress and well-being in general populations,” *Journal of Consulting and Clinical Psychology*, vol. 51, no. 5, pp. 730–742, doi:10.1037/0022-006X.51.5.730.

[13] R. Aziz, (2015), “Aplikasi Model RASCH dalam Pengujian Alat Ukur Kesehatan Mental di Tempat Kerja,” *Psikoislamika : Jurnal Psikologi dan Psikologi Islam*, vol. 12, no. 2, p. 29, doi:10.18860/psi.v12i2.6402.

[14] R. Aziz and Zamroni, (2020), “Analisis Faktor Konfirmatori Terhadap Alat Ukur Kesehatan Mental Berdasarkan Teori Dual Model,” *Psikoislamika : Jurnal Psikologi dan Psikologi Islam*, vol. 16, no. 2, p. 1, doi:10.18860/psi.v16i2.8199.

[15] V. Wijaya, M. Fadli, Y. A. Dharma, and M. R. Pribadi, (2022), “Pengembangan UI/UX pada aplikasi Go-Print Dengan menggunakan metode design thinking,” *MDP Student Conference (MSC)*, vol. 1, no. 1, pp. 298–305.

[16] D. I. Pushpita Anna Octaviani, Yuciana Wilandari, (2014), “Penerapan Metode SVM Pada Data Akreditasi Sekolah Dasar Di Kabupaten Magelang,” *Jurnal Gaussian*, vol. 3, no. 8, pp. 811–820.

[17] M. Athoillah, M. I. Irawan, and M. Imah, (2015), “Support Vector Machine Untuk Image Retrieval,” *Seminar Nasional Matematika dan Pendidikan Matematika*, no. 978, pp. 279–287.

[18] R. R. Fiska, (2017), “Penerapan Teknik Data Mining dengan Metode Support Vector Machine (SVM) untuk Memprediksi Siswa yang Berpeluang Drop Out (Studi Kasus di SMKN 1 Sutera),” *Sains dan Teknologi*

Informasi (SATIN), vol. 3, no. 1, pp. 15–23.

[19] F. Akbar and Rahmaddeni, (2022), “Komparasi Algoritma Machine Learning untuk Memprediksi Penyakit Alzheimer,” *Jurnal Komputer Terapan*, vol. 8, no. 2, pp. 236–245.

[20] N. K. Dewi, S. Y. Mulyadi, and U. D. Syafitri, (2012), “Penerapan Metode Random Forest Dalam Driver Analysis,” *Forum Statistika Dan Komputasi*, vol. 16, no. 1, pp. 35–43.

[21] G. A. M. Ashfania, T. Prahasto, A. Widodo, and T. Warsokusumo, (2023), “Penggunaan Algoritma Random Forest untuk Klasifikasi berbasis Kinerja Efisiensi Energi pada Sistem Pembangkit Daya,” *Rotasi*, vol. 24, no. 3, pp. 14–21.

[22] S. A. Assaidi and F. Amin, (2022), “Analisis Sentimen Evaluasi Pembelajaran Tatap Muka 100 Persen pada Pengguna Twitter menggunakan Metode Logistic Regression,” *Jurnal Pendidikan Tambusai*, vol. 6, no. 2, pp. 13217–13227, doi:doi.org/10.31004/jptam.v6i2.4543.

[23] Y. Tampil, H. Komaliq, and Y. Langi, (2017), “Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado,” *d’CARTESIAN*, vol. 6, no. 2, p. 56, doi:10.35799/dc.6.2.2017.17023.

[24] A. Maulvi Inayat, (2021), “Analisis Sentimen Berdasarkan Aspek Menggunakan Elman Recurrent Neural Network,” *Thesis, Universitas Komputer Indonesia.*, pp. 9–30.

[25] R. Siringoringo, (2018), “Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan K-Nearest Neighbor,” *Jurnal ISD*, vol. 3, no. 1, pp. 44–49.

[26] M. Fadli, V. Wijaya, M. R. Pribadi, and W. Widhiarso, (2023), “Effect of TF-IDF Extraction and Application of SMOTE on Model Performance in Detecting Spam Email,” in *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pp. 637–641.

[27] R. Ridwan, E. H. Hermaliani, and M. Ernawati, (2024), “Penerapan: Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada Klasifikasi Ujaran Kebencian,” *Computer Science (CO-SCIENCE)*, vol. 4, no. 1, pp. 80–88.

[28] A. N. Kasanah, M. Muladi, and U. Pujianto, (2019), “Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 2, pp. 196–201, doi:https://doi.org/10.29207/resti.v3i2.945.