Application of SMOTE in Multiclass Body Mass Index Classification: A Study on Data Imbalance and Model Performance

Selly Anastassia Amellia Kharis[1*], Melisa Arisanty[2], and Arman Haqqi Anna Zili[3]

[1]Universitas Terbuka, Mathematics Study Program, South Tangerang, Banten, Indonesia, 15437
[2]Universitas Terbuka, Library and Information Science Study Program, South Tangerang, Banten, Indonesia, 15437
[3]Department of Mathematics, Universitas Indonesia, Depok, West Java, Indonesia, 16424

Abstract - The Body Mass Index (BMI) is a commonly utilized measure that calculates body fat by analyzing a person's height and weight. BMI can monitor and explain a person's nutritional status. BMI classification is not always limited to binary classification but can extend to multiclass scenarios. However, a common challenge in BMI classification is the imbalance in data distribution across different classes, where some classes have significantly fewer instances compared to others. This research aims to evaluate the effectiveness of multiclass BMI classification both with and without the application of the Synthetic Minority Over-Sampling Technique (SMOTE). This study divides BMI into five groups using different machine learning algorithms: extremely weak, weak, normal, overweight, obesity, and extreme obesity. The machine learning algorithm utilized in this research include Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Logistic Regression. After applying SMOTE, the F1-score improved significantly across all models, with SVM showing an increase from 82.72% to 93.67% and KNN from 87.02% to 94.95%. Similarly, the overall accuracy improved by up to 7.84% in the SVM model. These results demonstrate that SMOTE effectively enhances the predictive performance of multiclass classification, especially in recognizing underrepresented classes.

Keywords: body mass index, classification, decision tree, K-nearest neighbor, logistic regression, random forest, SMOTE, support vector machine

1. Introduction

Machine learning has evolved into one of the main approaches in data analysis and decision-making across various disciplines, ranging from education and economics to health. Over the past two decades, machine learning has evolved from being a laboratory novelty to a practical technology with extensive commercial applications. Within the field of artificial intelligence (AI), it has become the go-to approach for creating functional software solutions for tasks such as computer vision, speech recognition, natural language processing, robotic control, and numerous other applications [1]. Machine learning makes predictions based on data by creating models and algorithms based on the

[*] Corresponding author: selly@ecampus.ut.ac.id

training data [2]. Machine learning studies training data to identify patters and uses the to predict the given testing data. Machine learning can process large amounts of data in a short time and produce predictions that can be used to make subsequent decisions or policies.

One application of machine learning is classification [3]. Classification is the process in which a model is trained to predict the category or class of input data based on previously labelled historical data. A classification model works by learning patterns from the labelled data, enabling it to make predictions for new data whose class is unknown. In general, the classification process involves dividing data into two main groups: training and testing data. Training data refers to data that already contains labelled categories, which are used to train the model [4]. Algorithms of machine learning will identify the relationship between data features and their existing category labels. Data testing are new data whose category is not yet known, used to evaluates the accuracy of the model in making predictions. In the modern data era, classification has become one of the essential approaches across various disciplines, including the field of health. The use of classification methods enables more accurate decision-making in the diagnosis and evaluation of health conditions.

Data imbalance is a prevalent issue in data classification. This imbalance usually occurs when the sample size in one or more classes is significantly larger than other classes, allowing the machine learning model to focus on learning patterns from the majority class while overlooking the minority class [5]. This imbalance can lead to bias in the model, where predictions tend to be more accurate for the majority class than for the minority class [6]. This condition is of particular concern in classification cases where the model is required to categorize data into more than two classes. The imbalance across multiple categories in multiclass classification cases can cause severe misclassification for minority classes, which would further impair the performance of the model as a whole and produce predictions that are not trustworthy.

Classification is usually performed only for binary classes, which consist of two labels. However, in real world cases, labelling often does not just split into two labels. One case that requires labelling with more than two classes is the Body Mass Index (BMI). BMI is an indicator widely used to categorize a person’s body status into several classes, such as underweight, normal, overweight, and obesity based on height and weight [7]. Cases of classification with more than two classes are referred to as multiclass classification [8]. In multiclass classification, class imbalance can exacerbate the problem, as the model must deal with highly varied distribution among several classes. In BMI classification, it is often found that there is data imbalance between classes, especially since the amount of data in categories like obesity tends to be lower compared to the normal category. This imbalance can lead to less accurate prediction models for minority classes, resulting in suboptimal overall model performance. Therefore, an approach is needed that can address this imbalance issue in order to improve the quality of predictions across all BMI categories.

One promising approach to addressing class imbalance is the Synthetic Minority Over Sampling Technique (SMOTE) [9]. SMOTE works by synthesizing new samples in the minority class, thereby enriching the amount of data in the underrepresented class without altering the overall original data distribution [10]. Unlike Random Oversampling (ROS), which duplicates existing samples and risks overfitting [11], SMOTE creates synthetic instance by interpolating between minority class samples, resulting in better generalization. Compared to ADASYN, which may focus excessively on difficult-to-learn examples, potentially introducing noise, SMOTE generates a more balanced synthetic distribution. SMOTE technique has proven effective in improving model performance in binary classification, but its application in multiclass classification, especially in the context of BMI prediction, has yet to be explored in depth. The study by Kharis & Rustam [12] explored multiclass classification in brain cancer using Support Vector Machine and Feature Selection based on kernel functions. Another study was conducted by Guleria et al. [13], which employed various machine learning techniques such as decision trees, random forests, and ANN to predict hypothyroidism. The novelty of this research lies in two main aspects. First, this research comprehensively explores the use of SMOTE in multiclass classification scenarios, which is still rarely discussed in the literature, especially in the context of Health and BMI prediction. Second, this study not only focuses on improving overall accuracy but also analyzes in detail how SMOTE impacts model performance in each BMI class, both in majority and minority classes. This research also examines how the use of SMOTE can affect other evaluation metrics such as precision, recall, and F1-Score in multiclass scenarios, which is crucial in ensuring balanced prediction quality across all BMI categories.

2. Research Methods

2.1 Data

The data used in this research was obtained from the Kaggle website, which contains 500 individual samples with three features: Gender (Male/Female), Height (cm), and Weight (kg). In this study, model evaluation was conducted using 5-fold cross validation to ensure the robustness and reliability of the results.

Table 1. Descriptive Statistics of BMI Data

Description	Height	Weight
Sampel	500	500
Mean	500	500
Standard Deviation	16.375	32.383
Minimal	140	50
Q1	156	80
Median(Q2)	170.50	106
Q3	184	136
Maximum	199	160

The distribution of the number of female and male in the BMI data is shown in Figure 1. In the BMI data, there are 255 females and 245 males. Furthermore, the data includes classes consisting of Extremely Weak (0), Weak (1), Normal (2), Overweight (3), Obesity (4), dan Extreme Obesity (5). The number of data points in each class is shown in Table 2.

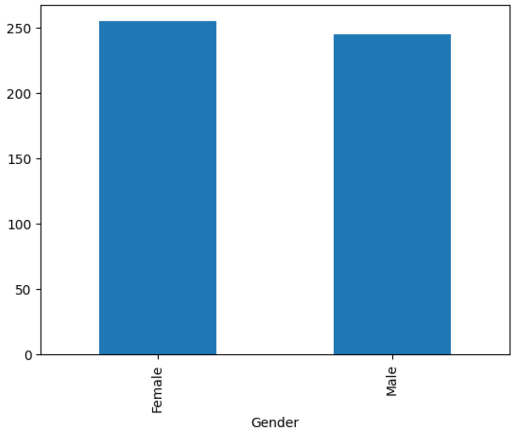

Fig. 1. Distribution of the number of female and male in the BMI data

Table 2. Details of data for each class

Class	Number of data
0	13
1	22
2	68
3	69
4	130
5	198

2.2 Preprocessing Data

After the initial exploration process, preprocessing is carried out to prepare the data for the multiclass classification stage. The categorical feature in the dataset, namely Gender has two categories: Male and Female. In this study, the Gender feature is transformed into Males with a value of 0 and Female with a value of 1. The numerical features in the dataset consist of height and weight, which have different scales. Both features are normalized so that there are no scale differences affecting the performance of the classification model. Normalization in this research is performed using the Z-Score normalization method, as shown in equation (1), where Z represent the normalized value, X is the original value of the feature, μ is the mean of the feature, and σ is the standard deviation of the feature.

$$Z = \frac{X-\mu}{\sigma} \tag{1}$$

Each feature value will represent how far that value is from the average in terms of standard deviations. Normalization has the advantage of standardizing the scale while also preserving information about the data distribution around the mean. This is very important, especially if the features have significant variability, such as height and weight in this dataset.

2.3 Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is a technique used to address the problem of data imbalance by increasing the number of samples in the minority class by creating new synthetic data based on existing samples. SMOTE generates new samples by using a certain distance between two or more existing samples in the feature space of the minority class. In this study, SMOTE was used to balance the imbalanced training data across various classification classes. Based on Table 2, it can be seen that the Extreme Weak and Extreme Obesity classes have a very different number of samples. The data for the Extreme Weak class consists of 13 individuals, while the Extreme Obesity class has 198 individuals. SMOTE is applied only to the training data to make the class distribution more balanced without affecting the test data. Adding data to balance classification classes allows the model to learn better from each class and prevents bias towards the majority class. Synthetic samples are generated based on the original features by randomly selecting minority samples and creating new samples in the vector space that lie between two or more existing samples. By balancing the class distribution, it is hoped that the classification model will have a better ability to identify and predict the minority class that was previously underrepresented. A model trained with more balanced data can produce more accurate predictions across all classes.

After applying SMOTE to the training data, the classification model is trained using various machine learning algorithms. The model was then tested using test data that did not utilize SMOTE, ensuring that the evaluation of the model's performance remained realistic and was not influenced by the synthetic samples generated by SMOTE. The evaluation metrics employed in this research include accuracy, precision, recall, and F1-Score. Accuracy is the ratio of the number of correct predictions (both true positives and true negatives) to the total number of predictions made by the model [14]. Precision is an evaluation metric that measures how accurately the model predicts the positive class [15]. Recall, commonly referred to as Sensitivity, is an evaluation metric that measures how well the model identifies all positive examples in the data [16]. F1-Score is an evaluation metric that measures the balance between precision and recall [17]. These metrics will provide a clearer picture of how the model performs in handling class imbalance after the application of SMOTE.

2.4 Multiclass Classification

This research uses multiclass classification with various machine learning models. The models used are Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression. The selection of this model is based on its ability to handle multiclass data and its popularity in classification tasks. Each model has a different approach to processing data, allowing for comparative analysis to find the best model.

A decision tree is a prediction model based on a decision tree that divides data into subsets based on decision rules generated from existing features. At each node, data is split based on the feature that provides the most information regarding class separation. Decision trees have the advantage of providing easily interpretable results, as this model works by forming a clear decision path from the root to the leaf node, where each leaf represents a class. However, decision trees tend to overfit, especially on small or varied datasets.

Random Forest is an advancement of Decision Tree that employs an ensemble learning approach. This model builds many decision trees (a forest) from random subsets of data, and then the prediction results from each tree are combined through a voting mechanism (majority voting). In this way, Random Forest addresses the weaknesses of Decision Trees, which are prone to overfitting and are more stable against variations in data. Random Forest is very effective in handling large and complex datasets, and it performs well in multiclass classification.

Support Vector Machine is one of the supervised learning models, working by finding the optimal hyperplane that can separate data into different classes. In multiclass classification, SVM uses approaches like one-vs-one or one-vs-rest to transform the multiclass problem into a series of binary classification problems. SVM is very useful for high-dimensional data. SVM can become less efficient when the dataset is too large or not linearly separable, which requires kernel adjustments.

KNN is an algorithm that works by finding the nearest data samples to a given point and making predictions based on the majority class among those neighbors. The class of a new sample is determined based on the majority class of its nearest neighbor samples (number of k). KNN is a simple yet effective model, especially for small datasets that do not require assumptions about data distribution. KNN is highly influenced by the size of the chosen neighbor sample and can become inefficient on large datasets because it has to calculate the distance between the new sample and all existing samples.

Logistic Regression is a classification model that utilizes the logistic function to predict the probability of a sample belonging to a particular class. While it is typically used for binary classification, it can be extended to multiclass classification through methods like one-vs-rest or softmax regression. Logistic regression provides interpretable results because the coefficients of each feature indicate their influence on the probability of a certain class. This model works well in the data is linear.

3. Results and Discussion

This research includes the implementation of two separate experiments to assess the effect of addressing class imbalance in the dataset. In the first experiment, the machine learning model was trained and evaluated without applying any oversampling techniques, while in the second experiment, SMOTE was used to balance the dataset by generating synthetic examples for the minority class. Each type of machine learning model was tested under both conditions to analyze the impact of SMOTE on model performance. To ensure robustness and reduce variance in the evaluation, experiments were conducted using 5-fold cross validation, which divided the dataset into five parts, with each part used for training data.

Table 3. Details of the sample for each class after SMOTE

Class	Number of data
0	195
1	195
2	195
3	195
4	195
5	195

The details of the amount of data in each class after the minority data was generated using SMOTE are shown in Table 3. As shown in Table 3, each class has the same amount of data, which is 195 data after the minority class was generated using the SMOTE method. The results of the experiment show that the application of SMOTE successfully improved the overall performance of the model across various machine learning algorithms. This improvement can be seen from the increase in evaluation metrics such as Accuracy, Precision, Recall, and F1-Score for each model after SMOTE compared to before SMOTE. The results of the multiclass classification using the decision tree are shown in Table 4. The application of SMOTE on the Decision Tree model has increased the accuracy from 0.8201 to 0.8528 and improved the F1-Score from 0.8319 to 0.8499. Figure 2 show the comparison before and after SMOTE using Decision Tree.

Table 4. Multiclass Classification Results using Decision Tree

Description	Before SMOTE	After SMOTE
Accuracy	0.8201	0.8528
Precision	0.8384	0.8526
Recall	0.8262	0.8446
F1-Score	0.8319	0.8499

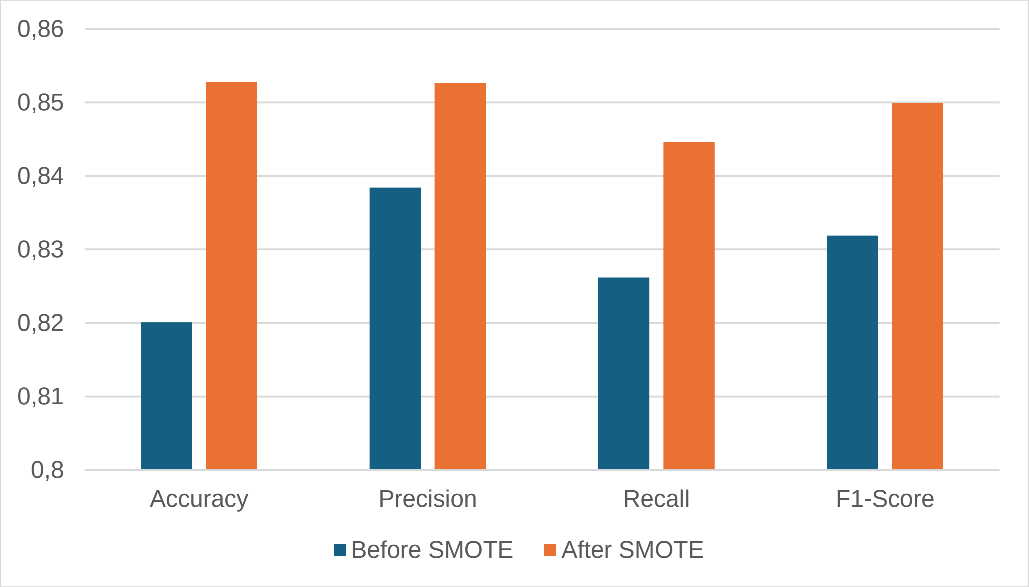

Fig. 2. Comparison before and after SMOTE using Decision Tree

Furthermore, a comparison is made with the results of multiclass classification using Random Forest. The classification results are shown in Table 5. Based on Table 5, it can be seen that there is an increase in the F1-Score, but there is a decrease in accuracy. This happens because accuracy only takes into account the number of correct predictions, both for the majority and minority classes, without considering the distribution of data between classes. In the context of classification, accuracy is often regarded as the primary metric for evaluating model performance. However, accuracy is not a reliable metric when the data used is imbalanced, as it can provide a misleading picture of the model's ability to classify data as a whole [3]. Therefore, in situations where minority classes are important, choosing the F1-Score metric is better to ensure that the model can predict all classes effectively. Figure 3 show the comparison before and after SMOTE using Random Forest.

Table 5. Multiclass Classification Results using Random Forest

Description	Before SMOTE	After SMOTE
Accuracy	0.8691	0.8610
Precision	0.8764	0.8680
Recall	0.8610	0.8671
F1-Score	0.8601	0.8714

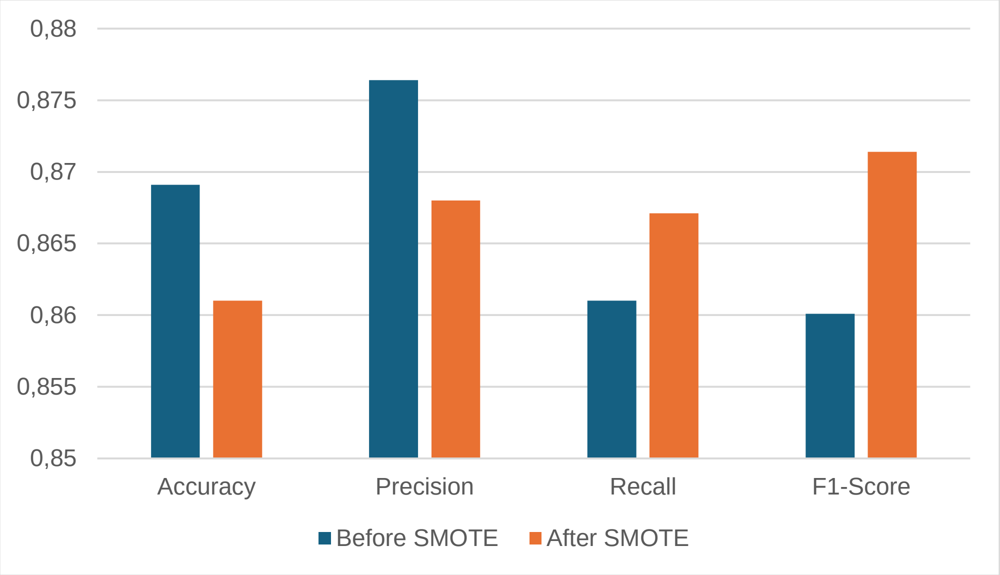

Fig. 3. Comparison before and after SMOTE using Random Forest

The increase in F1-Score from 0.8601 to 0.8714 indicates that the model has become more balanced in identifying minority and majority classes after the application of SMOTE. The F1-Score is the harmonic mean of precision and recall, so this improvement indicates that the model is better at identifying the minority class (increased recall) while also being more accurate in its predictions (increased precision). The increase in F1-Score is more significant than a slight decrease in accuracy because the F1-Score is more relevant in the context of class imbalance.

The next classification was carried out using SVM. The test results are shown in Table 6. The application of SMOTE using the SVM model in Table 6 has increased the accuracy from 86.30% to 93.68%, and precision, recall, and F1-Score have also significantly improved. The F1-Score increased from 82.72% to 93.67%. This indicates that SMOTE helps the model better detect the minority class and reduces prediction errors. This significant improvement indicates that by balancing the dataset, the SVM model is able to produce more accurate and balanced predictions across all classes, including the underrepresented classes. Figure 4 show the comparison before and after SMOTE using SVM.

Table 6. Multiclass Classification Results using SVM

Description	Before SMOTE	After SMOTE
Accuracy	0.8630	0.9368
Precision	0.8388	0.9412
Recall	0.8316	0.9368
F1-Score	0.8272	0.9367

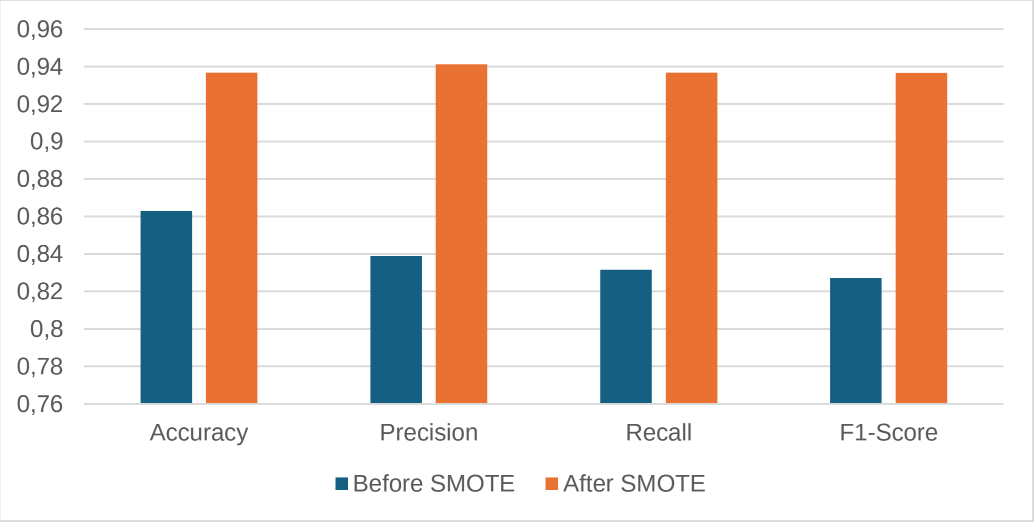

Fig. 4. Comparison before and after SMOTE using SVM

The KNN model also showed similar results, with an accuracy increase from 87.12% to 94.96%, as well as improvements in other metrics. The F1-Score also rose from 87.02% to 94.95%. This indicates that the application of SMOTE has a positive contribution to the model's performance in handling the minority class. Table 7 shows the detailed results of the multiclass classification.

Table 7. Multiclass Classification Results using KNN

Description	Before SMOTE	After SMOTE
Accuracy	0.8712	0.9496
Precision	0.8731	0.9515
Recall	0.8712	0.9496
F1-Score	0.8702	0.9495

Figure 5 show the comparison before and after SMOTE using KNN.

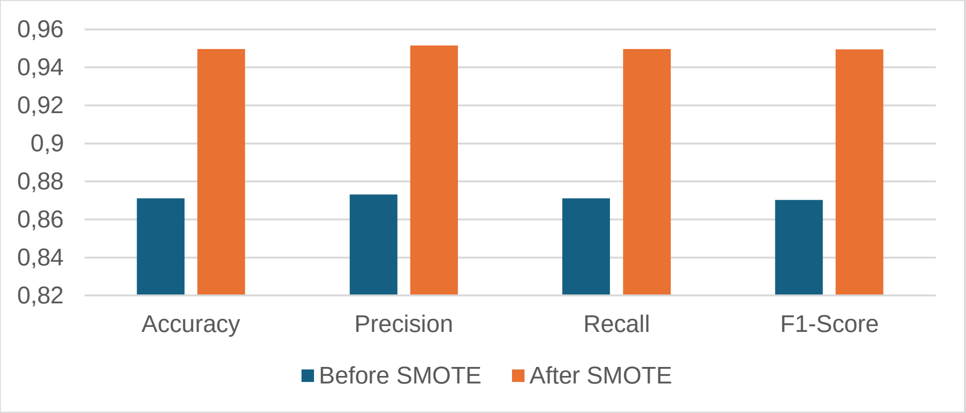

Fig. 5. Comparison before and after SMOTE using KNN

The Logistic Regression model also demonstrated an increase in accuracy from 83.85% to 86.51% after the application of SMOTE, with a significant improvement in precision and F1-Score, indicating that the model became better at identifying the minority class and reducing prediction errors. The F1-Score increased by 5.28% from before SMOTE was applied, rising from 81.56% to 86.84%. Table 8 shows the results of multiclass classification using logistic regression. Figure 6 show the comparison before and after SMOTE using KNN.

Table 8. Multiclass Classification Results using Logistic Regression

Description	Before SMOTE	After SMOTE
Accuracy	0.8385	0.8651
Precision	0.8212	0.8810
Recall	0.8385	0.8651
F1-Score	0.8156	0.8684

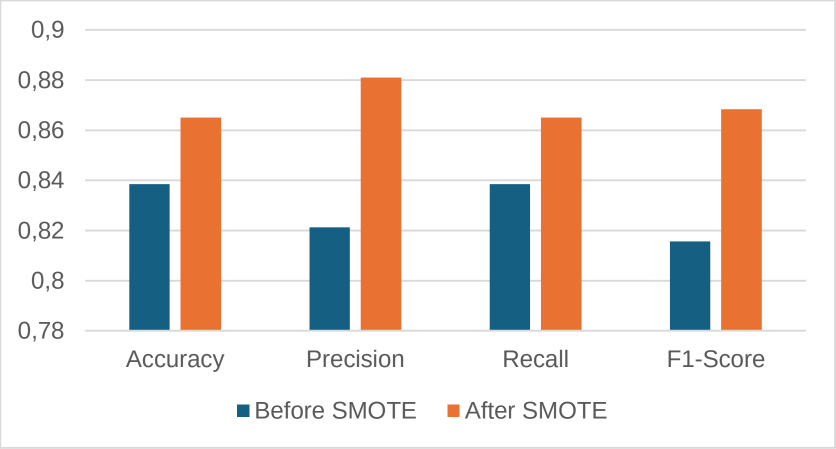

Fig. 6. Comparison before and after SMOTE using Logistic Regression

The results of this research indicate that the application of SMOTE has a significant impact on improving the performance of machine learning models, particularly in addressing class imbalance issues. In the experiments conducted, it was evident that SMOTE not only enhanced the model's ability to detect the minority class but also assisted the model in reducing the bias that tends to occur towards the majority class. This performance improvement can be seen through all the evaluation metrics used, namely accuracy, precision, recall, and F1-Score. Before the implementation of SMOTE, the models tested tended to have suboptimal performance in detecting the minority class, as the models were more focused on the majority class, which had a larger amount of data. However, after SMOTE was applied, the models became more balanced in recognizing patterns in the minority class. After SMOTE was applied, the extremely weak and weak showed substantial improvement in recall, with gains of over 20% in the SVM and KNN models. Performance metrics for the Normal and Overweight classes relatively stable before and after SMOTE, demonstrating that synthetic oversampling did not distort or harm well-represented classes. Overall, SMOTE significantly reduced the disparity in prediction quality across classes.

4. Conclusion

Research results show that the application of the SMOTE technique improves machine learning models' performance in handling class imbalance issues in datasets, including multiclass problems. SMOTE successfully enhances the model's ability to detect minority classes and reduces the bias that tends to occur towards majority classes in various machine learning models used. The machine learning models employed include Decision Tree, Random Forest, SVM, KNN, and Logistic Regression. The evaluation metrics used are Accuracy, Precision, Recall, and F1-Score. The experimental findings demonstrate that the model used following SMOTE performs better in identifying patterns in the minority class. The application of SMOTE to the Decision Tree model increased accuracy from 82.01% to 85.28% and F1-Score from 83.19% to 84.99%. The application of SMOTE to the Random Forest, although accuracy slightly decreased, F1-Score increased from 86.01% to 87.14%, indicating that the model became more balanced in identifying minority and majority classes. The application of SMOTE to the SVM and KNN models showed very significant improvements, with substantial improvements, with substantial increases in accuracy and F1-Score.

References

[1] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science (1979)*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aac4520.

[2] B. Mahesh, “Machine Learning Algorithms - A Review,” *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, Jan. 2020, doi: 10.21275/art20203995.

[3] S. A. A. Kharis and A. H. A. Zili, “Predicting life expectancy of lung cancer patients after thoracic surgery using SMOTE and machine learning approaches,” *Jurnal Natural*, vol. 23, no. 3, pp. 152–161, 2023, doi: 10.24815/jn.v23i3.29144.

[4] M. Alawida, S. Mejri, A. Mehmood, B. Chikhaoui, and O. Isaac Abiodun, “A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity,” Aug. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/info14080462.

[5] S. A. A. Kharis, G. F. Hertono, S. R. Irawan, E. Wahyuningrum, and Yumiati, “Students’ success prediction based on the Fuzzy K-Nearest Neighbor method in Universitas Terbuka,” *Education Technology in the New Normal: Now and Beyond*, pp. 212–218, Jun. 2023, doi: 10.1201/9781003353423-22.

[6] R. Blagus and L. Lusa, “Class prediction for high-dimensional class-imbalanced data.,” *BMC Bioinformatics*, vol. 11, p. 523, 2010, doi: 10.1186/1471-2105-11-523.

[7] D. Mohajan and H. K. Mohajan, “Body Mass Index (BMI) is a Popular Anthropometric Tool to Measure Obesity Among Adults,” *Journal of Innovations in Medical Research*, vol. 2, no. 4, pp. 25–33, Apr. 2023, doi: 10.56397/jimr/2023.04.06.

[8] C. Iwendi, S. Khan, J. H. Anajemba, M. Mittal, M. Alenezi, and M. Alazab, “The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems,” *Sensors (Switzerland)*, vol. 20, no. 9, May 2020, doi: 10.3390/s20092559.

[9] S. Gholampour, “Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable,” *Mach Learn Knowl Extr*, vol. 6, no. 2, pp. 827–841, Jun. 2024, doi: 10.3390/make6020039.

[10] C. W. Chiu and L. L. Minku, “Smoclust: synthetic minority oversampling based on stream clustering for evolving data streams,” *Mach Learn*, vol. 113, no. 7, pp. 4671–4721, Jul. 2024, doi: 10.1007/s10994-023-06420-y.

[11] T. Wongvorachan, S. He, and O. Bulut, “A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining,” *Information (Switzerland)*, vol. 14, no. 1, Jan. 2023, doi: 10.3390/info14010054.

[12] Z. Rustam and S. A. A. Kharis, “Multiclass classification on brain cancer with multiple support vector machine and feature selection based on kernel function,” in *AIP Conference Proceedings*, American Institute of Physics Inc., Oct. 2018. doi: 10.1063/1.5064230.

[13] K. Guleria, S. Sharma, S. Kumar, and S. Tiwari, “Early prediction of hypothyroidism and multiclass classification using predictive machine learning and deep learning,” *Measurement: Sensors*, vol. 24, Dec. 2022, doi: 10.1016/j.measen.2022.100482.

[14] J. Wu and C. Hicks, “Breast cancer type classification using machine learning,” *J Pers Med*, vol. 11, no. 2, pp. 1–12, Feb. 2021, doi: 10.3390/jpm11020061.

[15] U. Erkan, “A precise and stable machine learning algorithm: eigenvalue classification (EigenClass),” *Neural Comput Appl*, vol. 33, no. 10, pp. 5381–5392, May 2021, doi: 10.1007/s00521-020-05343-2.

[16] J. Miao and W. Zhu, “Precision-Recall Curve (PRC) Classification Trees,” *ArXiv*, pp. 1–19, Nov. 2020, doi: 10.1007/s12065-021-00565-2.

[17] R. O. Alabi *et al.*, “Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer,” *Int J Med Inform*, vol. 136, Apr. 2020, doi: 10.1016/j.ijmedinf.2019.104068.