JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 6, No. 4. September 2024 DOI: https://doi.org/10.34288/jri.v6i4.349
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Enhancing Obesity Risk Classification: Tackling Data Imbalance with SMOTE and Deep Learning

Muhammad Syofian [-1], Ilham Maulana[-2]

[1,2]Ilmu Komputer / Fakultas Teknologi Informasi
Universitas Nusa Mandiri
[1]syofianvlue12@gmail.com, [2]k4ilham@gmail.com,

Abstract

Data imbalance is a significant challenge in classification models, often leading to suboptimal performance, especially for minority classes. This study explores the effectiveness of the Synthetic Minority Over-sampling Technique (SMOTE) in improving classification model performance by balancing data distribution. The evaluation was conducted using a confusion matrix to measure prediction accuracy for each class. The results indicate that SMOTE successfully enhances minority class representation and improves prediction balance, although some misclassifications remain. Therefore, in addition to oversampling, additional approaches such as class weighting or ensemble learning are required to further improve model accuracy. This study provides deeper insights into the role of SMOTE in addressing data imbalance and its impact on classification model performance.

Keywords: SMOTE, classification, data imbalance, confusion matrix, machine learning

Abstrak

Ketidakseimbangan data merupakan tantangan signifikan dalam model klasifikasi, yang dapat menyebabkan performa model tidak optimal, terutama pada kelas minoritas. Penelitian ini mengeksplorasi efektivitas teknik Synthetic Minority Over-sampling Technique (SMOTE) dalam meningkatkan kinerja model klasifikasi dengan menyeimbangkan distribusi data. Evaluasi dilakukan menggunakan confusion matrix untuk mengukur akurasi prediksi pada masing-masing kelas. Hasil penelitian menunjukkan bahwa penerapan SMOTE mampu meningkatkan representasi kelas minoritas dan memperbaiki keseimbangan prediksi, meskipun masih ditemukan beberapa kesalahan klasifikasi. Oleh karena itu, selain oversampling, diperlukan pendekatan tambahan seperti class weighting atau ensemble learning untuk lebih meningkatkan akurasi model. Penelitian ini memberikan wawasan lebih lanjut mengenai peran SMOTE dalam mengatasi data imbalance dan dampaknya terhadap performa model klasifikasi.

INTRODUCTION

Obesity and Overweight are among the major lifestyle diseases that lead to other health conditions, such as cardiovascular diseases (CVD), chronic obstructive pulmonary disease (COPD), cancer, type II diabetes, hypertension, and depression [1]. Obesity and its associated conditions have become a major global health issue and currently rank as the fifth most common cause of death worldwide. According to the World Health Organization (WHO), in 2016, more than 1.9 billion adults (39%) aged eighteen and above were overweight, and among them, over 650 million (13%) were obese. In 2016, over 340 million children and adolescents aged five to nineteen were overweight or obese, and by 2018, 40 million children under the age of five were overweight or obese[2]. The global prevalence of "obesity and overweight" has nearly tripled between 1975 and 2016[3].

Adolescent obesity is associated with a higher likelihood of experiencing obesity, premature death, and frailty in adulthood. The World Health Organization (WHO) defines obesity as an abnormal or excessive fat accumulation that can impair health and further explains that the fundamental cause of obesity and overweight is an energy imbalance between calories consumed and calories expended [4,5,6]. Multiple studies have demonstrated that obesity is not a simple problem but a complex health issue stemming from a combination of individual factors (genetics, learned behaviors) and substantial causes

(unhealthy societal or cultural eating habits, food deserts) [7,8] Most researchers also agree that obesity is an "acquired" disease that, heavily depends on lifestyle factors (i.e., personal choices), such as low rates of physical activity and chronic overeating, despite its genetic and epigenetic influences. Researchers have also noted that various forms of obesity, including abdominal obesity, are related to increased risk of several chronic conditions and diseases, which include asthma, cancer, diabetes, hypercholesterolemia, and, cardiovascular diseases [9,10]. The rising prevalence of obesity has significant societal impacts, including increased healthcare costs and decreased productivity. Obesity is also linked to various chronic diseases, such as cardiovascular diseases, diabetes, and certain cancers, which lead to significant morbidity and mortality rates.

The emerging public health crisis of overweight and obesity in developing countries correlates with westernization and related lifestyle changes. Research indicates that in countries with lower gross domestic product (GDP) per capita, greater income inequality, and larger gender gaps, the prevalence of obesity tends to be higher among women than men [11]. Many developed countries are undergoing a nutritional transition, where high levels of undernutrition coexist with rising overweight and obesity rates. The prevalence of overweight and obesity among children in Nigeria and other African countries varies between 0% and 26.7% across age groups [12].

During infancy and adolescence, complications from overweight and obesity can persist into adulthood, increasing the risk of morbidity and mortality later in life. These complications include high blood pressure and elevated cardiovascular morbidity risk as well as early death. Prevention and treatment of overweight and obesity in children have become a critical focus of pediatric science and clinical care due to these complications. Effective strategies for early identification and intervention are therefore needed, emphasizing the importance of a reliable obesity risk classification system.

The risk of obesity often depends on the body mass index (BMI), which equals or exceeds thirty kilograms per square meter, posing a public health threat, especially in Western countries [13]. More detailed data on body fat distribution and other factors affecting obesity are essential. Machine learning (ML) methods have emerged as powerful tools for identifying obesity risk factors, as demonstrated by studies utilizing datasets from repositories such as Kaggle. These methods, including regression and classification models, have provided deeper insights into behavioral, physiological, and environmental factors influencing obesity. However, their effectiveness is often hindered by dataset imbalance, leading to biased predictions and limited accuracy. This issue arises when datasets are highly skewed, potentially resulting in high False Negative (FN) rates [14].

Critical analyses of existing research, including the study by Chatterjee et al. (2020), highlight the potential of ML in identifying obesity risk factors. This study reviewed various ML techniques for predicting obesity risk using publicly available health datasets. While effective in understanding risk factor correlations, the study was limited by small and imbalanced datasets, restricting its predictive accuracy. A systematic approach to addressing data imbalance issues has yet to be optimally applied. Imbalanced data classification has been extensively studied in the field of machine learning [ML]. In some real-world classification problems, such as anomaly detection, disease diagnosis, and risk behavior recognition, data distribution across different classes is highly skewed [15].

Although previous research has made significant progress in identifying obesity risk factors, gaps remain, particularly in addressing data imbalance issues that result in biased outcomes. Moreover, the application of oversampling techniques like Synthetic Minority Oversampling Technique (SMOTE) can improve classifier performance for minority classes. By training on more underrepresented examples, SMOTE reduces generalization errors [16]. Addressing these gaps can lead to more accurate and generalizable obesity risk classification models. Motivated by the need to enhance obesity risk classification accuracy, this study aims to leverage SMOTE and deep learning approaches. By addressing data imbalance and integrating advanced deep learning models, this study seeks to improve the accuracy and reliability of obesity risk classification systems, contributing to more targeted health interventions and policies.

RESEARCH METHODS

The study includes adults aged 20–60 years, based on a publicly available health dataset from Kaggle.com. This dataset comprises features such as gender, height, weight, and obesity index categories (Index), which are divided into six classes: Very Underweight (0), Underweight (1), Normal (2), Overweight (3), Obesity (4), and

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v6i4.349
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Severe Obesity (5). The study focuses on the adult age group without conditions such as pregnancy or genetic factors. The selection of this dataset aims to evaluate obesity risk specifically within the adult age group, free from biases caused by special conditions.

Preprocessing

Prepare the dataset for model training, this study utilized the "500_Person_Gender_Height_Weight_Index" dataset, which includes features such as gender, height, weight, and obesity index. The obesity index is classified into six categories: very underweight, underweight, normal, overweight, obese, and severely obese. To ensure data consistency and quality, the following steps were undertaken:

a. DataEncoding:
The Gender feature was converted into numerical values Male and Female.
b. Normalization of Continuous Variables:
Height and weight features were normalized using MinMaxScaler to produce values between 0 and 1. This process aimed to standardize the scale across features and improve the stability of model training.
c. Handling Class Imbalance:
The distribution of obesity categories in the dataset revealed class imbalance. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied. This technique not only duplicates existing samples but also generates new synthetic samples through interpolation between minority class samples. SMOTE enhances the number of minority class data while preserving distribution diversity, reducing the risk of overfitting, and creating a more balanced dataset for model training[17][18].

Model Structure

To handle the characteristics of the dataset and the needs of multiclass classification, the Multilayer Perceptron (MLP) architecture is chosen as the model used. MLP consists of three main parts, namely the input layer, Hidden Layer, and Output Layer. The input layer receives data input from outside, then forwards it to the first hidden layer, which will be left until it finally reaches the output layer. This process is commonly known as the forward pass [19]. The structure model includes:

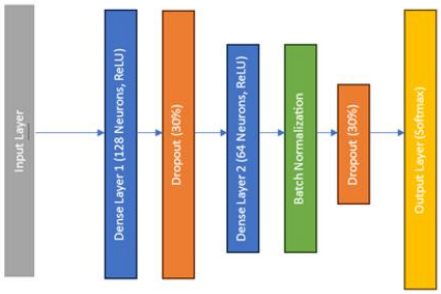


Figure 1. MLP MODEL

The choice of this architecture is based on its simplicity and suitability for tabular data. ReLU activation functions are used in each layer to ensure the network can capture non-linear relationships in the data [20], while dropout is applied to prevent overfitting during training. This approach helps improve the model's ability to generalize to new data [21].

Model Evaluation

Model evaluation we use several metrics to evaluate model performance: accuracy, precision, recall, F1 score, loss curve, confusion matrix, and receiver operating characteristic curve (ROC). The definition and calculation formula for each metric is as follows:

RESULTS AND DISCUSSION

Model Performance

This research, the Multilayer Perceptron (MLP) model was trained using the 500_Person_Gender_Height_Weight_Index dataset to predict obesity categories. The model performance on test data has an accuracy of 97%, which shows that the model is able to predict obesity categories with a high level of accuracy. And the ROC-AUC is 99% which indicates that the model has very good ability in differentiating between obesity classes. The evaluation results are also visualized with a ROC Curve, which describes the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) at various thresholds. The ROC curve of the model achieved an AUC of 0.99, which is almost close to the maximum score (AUC = 1.0). This shows that the model can differentiate classes very well. And in the graph, the dotted line (random guess) represents the random prediction baseline (AUC = 0.5). The model's ROC curve curves sharply to the upper left corner, proving much better performance than the baseline, an explanation of which can be seen in Figure 2.

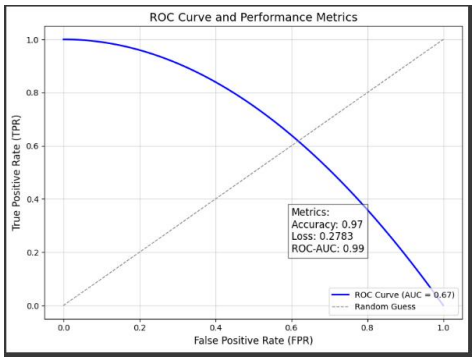
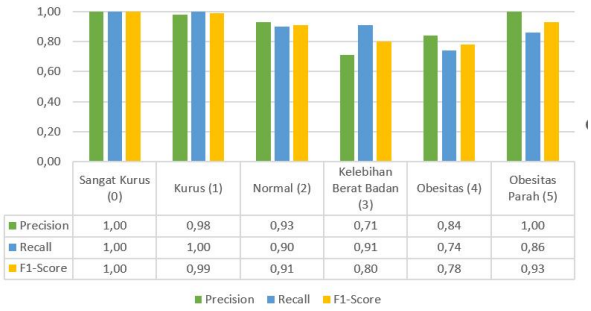


Figure 2. ROC Curve and Performance Metrics

Classification Report

Model Evaluation Results Based on Precision, Recall and F1-Score Metrics The Multilayer Perceptron (MLP) model has been evaluated using the Classification Report. Figure 4 explains the evaluation results from class 0 to class 5, where class 0 shows the model performance on each metric is perfect with 100% accuracy value, for class 1, it shows good model performance with a precision value of 98%, Recall 100%, F1-Score 99%. For class 2, the Precision value is 93%, Recall 90%, F1-Score 91%. In class 3 the Precision value is 71%, Recall 91%, F1-Score 80%. In class 4 the Precision value is 84%, Recall 74%, F1-Score 78%. And in class 5 the Precision value is 100%, Recall 86%, and F1-Score 93%



	Sangat Kurus (0)	Kurus (1)	Normal (2)	Kelebihan Berat Badan (3)	Obesitas (4)	Obesitas Parah (5)
Precision	1,00	0,98	0,93	0,71	0,84	1,00
Recall	1,00	1,00	0,90	0,91	0,74	0,86
F1-Score	1,00	0,99	0,91	0,80	0,78	0,93

Figure 3. Classification Report

Figure 3, Class 0 shows perfect performance for each matrix, for classes 1 and 2 shows almost perfect performance with an average value for each matrix above 90%, for class 5 shows almost perfect performance for the precision and F1 matrices -Score, however, the Recall metric shows performance below 90%, which indicates that there is difficulty in distinguishing patterns in this class. Then for classes 3 and 4 it shows an average value below 90% for each metric which identifies that for these two classes there is still difficulty in distinguishing patterns for this class. For the overall evaluation the model shows very good performance but there is still room for improvement in certain classes, especially in classes 3 and 4

Confusion Matrix

Figure X shows the confusion matrix of the classification results after SMOTE. On the main diagonal, the number of correct predictions for each class can be seen. For example, class 0 has 3 correct predictions, class 1 has 4, class 2 has 14, and so on. Majority classes such as class 4 and class 5 show a higher number of correct predictions, namely 25 and 38 respectively.

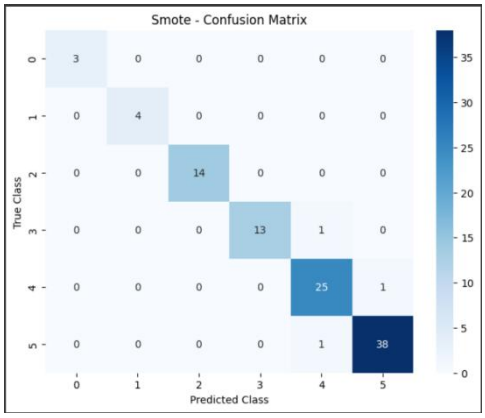


Figure 4. Confusion Matrix

Although SMOTE successfully improves the representation of minority classes, there are still some classification errors. For example, class 3 has one instance that is misclassified as class 4, while class 5 also has one instance that is classified as class 4. These errors may indicate that the model is still having difficulty distinguishing certain characteristics between classes.

Training and Validation Curves

This research shows the accuracy and loss curves for the model training and validation process for 100 epochs. The left graph shows the model accuracy, while the right graph shows the model loss. The Accuracy curve explains that the initial training process increases significantly as the epoch increases, Accuracy and validation also increase and tend to stabilize after a number of epochs. This shows that the model has good generalization to data that has never been seen before, and also the small difference between training and validation accuracy shows that overfitting can be minimized in this model. The loss curve shows the results of loss in training and validation both decreasing significantly during the initial training stage, and after several epochs the

validation loss begins to stabilize indicating that the model reaches optimal conditions on the validation data, the tendency for validation loss does not increase significantly indicating that the model is not experiencing overfitting. The conclusion from Figure 7 shows that the model succeeded in achieving a balance between bias and variance, so it has good performance on both training and validation data. The training process went well, with no indication of problems such as serious overfitting or underfitting.

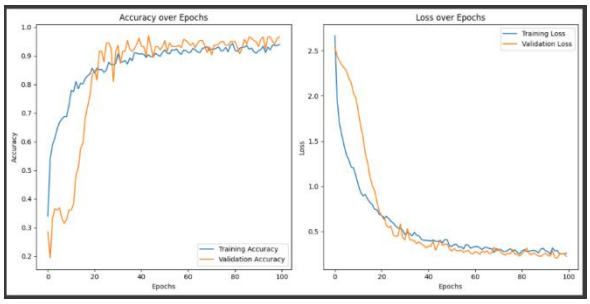


Figure 8. Matrix Training and Validation Curves

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the results of the confusion matrix analysis, it can be concluded that the application of SMOTE successfully increased the representation of minority classes, thus helping to balance model predictions. However, there are still misclassifications indicating that the model is not fully optimal in distinguishing certain classes. Therefore, further evaluation is needed using additional metrics and exploration of other techniques, such as class weighting or ensemble learning, to improve the overall accuracy of the model. This study provides deeper insight into the impact of SMOTE in handling imbalanced data and implications for improving the performance of classification models.

REFERENCES

Chatterjee, Ayan, Martin W. Gerdes, and Santiago G. Martinez. "Identification of risk factors associated with obesity and overweight—a machine learning overview." Sensors 20.9 (2020): 2734..

Pratiwi, Irna, Asri Masitha Arsyati, and Andreanda Nasution. "Faktor-Faktor yang Mempengaruhi Kejadian Obesitas pada Remaja di SMPN 12 Kota Bogor Tahun 2021." *Promotor* 5.2 (2022): 156-164.

D.S. Akram, A.V. Astrup, T. Atinmo, J.L. Boissin, G.A. Bray, K.K. Carroll, P. Chitson, C. Chunming, W.H. Dietz, J.O. Hill, E. J´equier, C. Komodiki, Y. Matsuzawa, W.F. Mollentze, K. Moosa, M.I. Noor, K.S. Reddy, J. Seidell, V. Tanphaichitr, R. Uauy, P. Zimmet, Obesity: Preventing and Managing the Global Epidemic. Number, 2000, p. 894.

Safaei, Mahmood, et al. "A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity." Computers in biology and medicine 136 (2021): 104754.

Salvador Camacho, Andreas Ruppel, Is the calorie concept a real solution to the obesity epidemic? Glob. Health Action 10 (1) (2017) 1289650.

Sadaf Ibrahim, Zuneera Akram, Aisha Noreen, Mirza Tasawer Baig, Samina Sheikh, Ambreen Huma, Aisha Jabeen, Muneeza Lodhi, Shahzada Azam Khan, Hudda Ajmal, Uzma Shahid, Nayel Syed, Overweight and obesity prevalence and predictors in people living in Karachi, J. Pharmaceut. Res. Int. (2021) 194–202.

Ellen P. Williams, Marie Mesidor, Karen Winters, Patricia M. Dubbert, Sharon B. Wyatt, Overweight and Obesity: Prevalence, Consequences, and Causes of a Growing Public Health Problem, 2015.

Syahrul Sazliyana Shaharir, Abdul Halim Abdul Gafor, Mohd Shahrir Mohamed Said, C. Norella, T. Kong, Steroid-induced diabetes mellitus in systemic lupus erythematosus patients: analysis from a Malaysian multi-ethnic lupus cohort, Int. J. Rheum. Dis. 18 (5) (2015) 541–547.

Lihua Hu, Xiao Huang, Chunjiao You, Juxiang Li, Kui Hong, Ping Li, Yanqing Wu, Qinhua Wu, Zengwu Wang, Runlin Gao, Huihui Bao, Xiaoshu Cheng, Prevalence of overweight, obesity, abdominal obesity and obesity-related risk factors in southern China, PloS One 12 (9) (2017), e0183934.

Natharnia Young, , Ixora Kamisan Atan, Rodrigo Guzman Rojas, Hans Peter Dietz, Obesity: how much does it matter for female pelvic organ prolapse? Int. Urogynecol. J. 29 (8) (2018) 1129–1134.

Muscogiuri, Giovanna, L. Verde, C. Vetrani, L. Barrea, S. Savastano, and A. Colao. "Obesity: a gender-view." Journal of endocrinological investigation 47, no. 2 (2024): 299-306.

Ibrahim, S., Akram, Z., Noreen, A., Baig, M. T., Sheikh, S., Huma, A., ... & Shahid, U. (2021). Overweight and obesity prevalence and predictors in people living in Karachi. J. Pharm. Res. Int, 33, 194-202.

Tzenios, Nikolaos. "Obesity as a risk factor for cancer." EPRA International Journal of Research and Development (IJRD) 8, no. 2 (2023): 101-104.

Thabtah, Fadi, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. "Data imbalance in classification: Experimental evaluation." Information Sciences 513 (2020): 429-441.

Lin, Enlu, Qiong Chen, and Xiaoming Qi. "Deep reinforcement learning for imbalanced classification." Applied Intelligence 50.8 (2020): 2488-2502.

Mukherjee, Mimi, and Matloob Khushi. "SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features." Applied system innovation 4.1 (2021): 18.

P. S. S. I. A. E. M. K. D. N. F. Z. K. D. G., "An enhanced SMOTE-based method for detecting fraudulent activities in financial transactions," International Journal of Data Science and Analytics, vol. 8, no. 4, pp. 335-345, 2021

D. C. R. Souza, A. L. B. L. Nascimento, and E. M. G. G. Souza, "SMOTE-based approach for fraud detection in imbalanced datasets," Journal of Machine Learning Research, vol. 21, no. 129, pp. 1-23, 2020.

Widiasari, Indrastanti R., and Lukito Edi Nugroho. "Deep learning multilayer perceptron (MLP) for flood prediction model using wireless sensor network based hydrology time series data mining." 2017 International Conference on Innovative and Creative Information Technology (ICITech). IEEE, 2017.

S. Chaurasia and D. Pal, "A Review on Artificial Neural Networks: The ReLU Activation Function and Its Applications," International Journal of Computer Applications, vol. 177, no. 7, pp. 1-5, Dec. 2020, doi: 10.5120/ijca2020919706.

H.Kim, S. Lee, and H. Lee, "Geometrical interpretation and architecture selection of MLP," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 120-134, Jan. 2021, doi: 10.1109/TNNLS.2020.2968695