CAUCHY: Jurnal Matematika Murni dan Aplikasi
Volume 10(1) (2025), Pages 263-277
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Obesity Prediction Using Synthetic Minority Oversampling Technique for Numeric and Continous and XGBoost Approaches

Tiara Azahra Wika Putri*, Umu Sa'adah, Ummu Habibah

Department of Mathematics, Faculty of Science, University of Brawijaya

Email: tiaraazahra@student.ub.ac.id

Corresponding Author: u.saadah@ub.ac.id

ABSTRACT

Obesity is a global health problem that requires serious attention, because it can cause various chronic diseases. This study aims to evaluate the effect of oversampling techniques, namely SMOTE-NC (Synthetic Minority Oversampling Technique for Nominal and Continuous), on the performance of the XGBoost algorithm in predicting obesity. The dataset used was taken from the University of California, Irvine Machine Learning Repository, and includes variables relevant to obesity. Before applying SMOTE-NC, the data underwent pre-processing to address class imbalance and missing values. The XGBoost algorithm was then applied to build a predictive model, with performance evaluation using accuracy, precision, recall, and F1-score metrics. The results showed that the application of SMOTE-NC significantly improved the accuracy of the model, reaching 98.30%, followed by precision of 98.32%, recall of 98.02%, and f1-score of 97.30%. In addition, the application obtained the best stability with a standard deviation of the accuracy metric of 0.004%. Feature analysis also identified that weight, height, and age are key factors in obesity prediction. These findings highlight the potential of machine learning techniques in improving public health prediction and the importance of early intervention to prevent obesity. This study provides a basis for further studies on the application of this method in a broader context.

Keywords: Obesity; Ensemble Learning; SMOTE-NC; XGBoost.

INTRODUCTION

Obesity is identified as a significant public health challenge that continues to improve. Obesity sufferers are predicted to increase to reach 1 billion sufferers by 2030 (world obesity atlas) [1]. Facts from the World Health Organization (WHO) state that around 67% of the population of the American continent is obese [2]. The number of obese people worldwide has increased since 1980 [3]. Obesity sufferers in Indonesia in 2017 reached 1 in 4 adults [4]. The impact of obesity is not only for sufferers but also for the wider community. Obesity sufferers can cause several losses in the health sector to the economy [5]. Research on the relationship between obesity and non-communicable diseases (NCDs) was conducted in the Ethiopian community with 8.5% of women and 2.4% of men living with obesity and the results showed that obesity and NCDs have a

Submitted: December 17, 2024 Reviewed: December 19, 2024 Accepted: January 27, 2025
DOI: https://doi.org/10.18860/cauchy.v10i1.30818

significant relationship, as a cause of health disorders such as hypertension and high cholesterol levels [6]. Several complications in obesity sufferers can reduce the quality of life of sufferers and can also shorten the life expectancy of sufferers [5]. Obesity is a contributor to 4 million deaths in 2015[7]. The disease with the highest death rate is cardiovascular disease followed by diabetes experienced by obesity sufferers[7]. In addition, obesity sufferers are also susceptible to mental health disorders, due to the many negative stigmas from society, discrimination, and ending in decreased self-confidence [5]. Obesity also causes a significant economic burden [5]. Sufferers will experience several complications of the disease that require quite high medical costs [8]. The ability to predict and classify the level of obesity in each individual needs to be done early on. This step can allow for improvements in several factors that trigger obesity.

Although various methods have been developed in predicting and managing obesity. However, many approaches have limitations. Innovations in technology and data analysis such as Machine Learning (ML) methods are becoming relevant. The ability to analyze large amounts of data and complex patterns can decide on more precise and targeted prevention [9]. Several previous studies have been conducted using the ML approach with basic and Ensemble algorithms. The Ensemble algorithm is a combination of several basic algorithms with the aim of increasing accuracy, commonly referred to as Ensemble Learning (EL) [10]. Research conducted in 2021 used several algorithms from the ML approach, namely K-NN (K-Nearest Neighbors), SVM (Support Vector Machine), LR (Logistic Regression), NB (Naïve Bayes), DT (Decision Tree), AdaBoost, RF (Random Forest), MLP (Multi Layer Preceptron), and Gradient Boosting [11]. The use of obesity sample data in the study was 1100 data which were then preprocessed without including the data balancing process. The data is divided into 3 obesity classes, namely low, medium and high. Based on the results of the study, the best accuracy results were obtained using the LR algorithm. In another study, a comparison of ML algorithms in the case of classification using obesity datasets, with the addition of the XGBoost algorithm in another study conducted in the same year [12]. The study did not involve data balancing techniques in the preprocessing stage. The results showed that the best algorithm was performed by XGBoost with a difference of 1.33% better than LR. The study used obesity data in children and XGBoost in the case of feature selection. In the following 2 years, the XGBoost algorithm was applied to predict obesity using the Chilean adolescent population dataset in the classification case [13]. Data preparation at the preprocessing stage did not go through the data balancing process and the data splitting process used a proportion of 60:40. The results of the study obtained a ROC (Receiver Operating Characteristic) curve value of 87.5%. Based on these results, there are several shortcomings that can be fixed to improve the quality of predictions. Efforts to improve the prediction process are carried out using data balancing techniques at the preprocessing stage. In this technique, data with a minority class is synthesized until it reaches the amount of data from the majority class [14]. In 2021, the SMOTE-NC algorithm was applied in the preprocessing stage to predict customer churn with the results showing that the best prediction accuracy was 76% [15]. The study did not focus on the application of data balancing techniques to increase accuracy. The data balancing technique was then carried out in 2022, by combining it with the basic technique of ML, namely SVM [16]. The results of its application increased by 1%. The application of the SMOTE-NC technique was then carried out in 2023 on one of the EL algorithms, namely CatBoost, with an increase in prediction accuracy of 7% after applying SMOTE-NC at the stage before prediction [17]. The accuracy reached 93% and was the highest accuracy value of other data balancing techniques. In the two previous studies, the data used was not obesity data. Based on

these results, it can be used as material to consider this study regarding the application of SMOTE-NC to obesity prediction using one of the ML approaches, namely the EL technique. Then the algorithm is combined with the aim of obtaining increased accuracy in predicting obesity, such as XGBoost is expected to increase the efficiency of the computational process and prediction performance [18]. XGBoost's performance is considered better in producing accurate obesity predictions in a shorter time compared to other algorithms [19]. The aim of the study is that obesity predictions can be carried out as optimally as possible. The results of increasing prediction accuracy can be used as a means of considering the next steps to prevent an increase in obesity sufferers.Methods

METHODS

There are several steps applied in this research. Figure 1 shows the initial research process, namely the input obesity dataset. Then the dataset is processed through the preprocessing stage by going through 3 main processes. Once the data is ready and balanced, predictions are made using the XGBoost algorithm. The next stage is a prediction based on two approaches, namely using several X variables after carrying out the feature selection stage and using all X variables. The two treatments are then compared to obtain optimal performance results from obesity prediction.

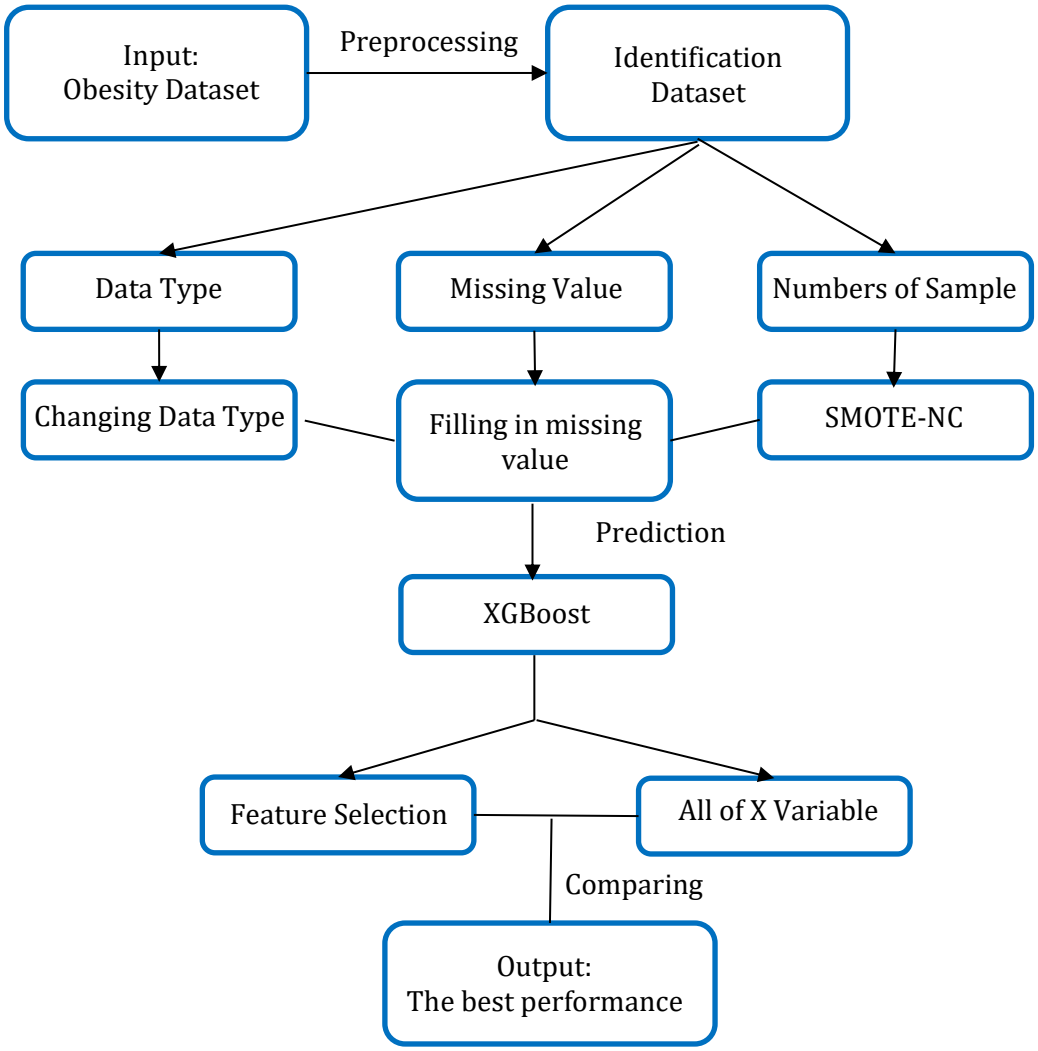


Figure 1. Research Method

Preprocessing Data

Data preprocessing is the initial step before performing data prediction analysis. This aims to help prepare data in the prediction process [20]. The initial step in data preprocessing is to import data from the repository [21]. This study uses an obesity-related dataset that contains some information about the variables used. The obesity dataset has a Y variable in the form of each individual's obesity level, such as NW (Normal Weight), OT_I (Obesity Type I), OT_II (Obesity Type II), OT_III (Obesity Type III), OW (Overweight), UW (Underweight) [22]. This stage is carried out in several steps, namely data identification, changing data types, and balancing data on the Y variable. Data identification includes checking data types, missing values, and the number of data for each class on the Y variable [23].

These stages are taken based on the needs of data processing. If there is non-uniformity in all data variables, it is necessary to change the data type, this also applies to the results of identifying missing values and the amount of data in the Y variable. Categorical data is changed to numeric data to match the data type. An example of this change is in the category in the Y variable, namely NW, by using the label encoding method, the data becomes numeric, namely 0 (null) [23]. The way label encoding works is by starting to identify columns with categorical data types. Then the column will be mapped from each category to numbers. For example, in the Y variable, the NW class as categorical changes to 0 as numeric. The label encoding process can be easily used and applied. Data with all its variables of numeric type is then checked for missing values and the amount of data. The check aims to determine whether or not there are missing values and the balance of data in the dataset. For example, in a variable there is a missing value, then to fill the gap, it can be done using Equation (1), namely by using the method of providing its average value. $\bar{X}$ is the average value obtained from the total number of data x_i with $i \in 1, \dots, n$. The number of data samples is n[24].

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

The amount of unbalanced data in the Y variable is balanced using the data balancing technique. This stage aims to improve data quality in helping the prediction process to be more accurate and improve model performance. The balanced obesity data is then divided into two parts, namely training data and testing data. The proportion of data separation is 80% training data and 20% testing data. The data separation process is carried out manually with a certain proportion for each class. After the separation process, the number of each class in the training data will be the same, namely 80% of the total data. This also applies to testing data.

SMOTE-NC

Data balance in obesity datasets can affect model performance, especially in minority classes [14]. The SMOTE-NC technique is a resampling technique using the oversampling method and can work on mixed data types, namely nominal and continuous [14]. The oversampling method works by synthesizing data samples in the minority data class until the amount of data in the minority class is the same as the amount of data in the majority class [14]. The initial step in implementing the SMOTE-NC technique is to define data samples that are included in the minority class. Suppose there are m data samples. The data samples have $X = [x_1, x_2, \dots, x_m]$ and $Y = [y_1, y_2, \dots, y_m]$. The process of forming data synthesis x_i' requires searching for the nearest neighbor z_{ij} which runs as far as r as shown in Equation (2) and Equation (3). This process can be run on continuous

data types, in a different way when using nominal or categorical data [25]. The formation of data synthesis is based on searching for the majority value and the intensity of the appearance of the data [25]. Then the data is used as an observation in adding minority class data. After the data on the Y variable is balanced, the prediction process can be continued using the ensemble learning method, namely XGBoost.

$$z_{ij} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)}, \quad (2)$$

$$x_i' = x_i + r \times (z_{ij} - x_i). \quad (3)$$

Technically, the implementation of SMOTE-NC according to Chawla can be done by identifying the minority class, which is indicated by the number of data samples in that class being less than in the majority class. Then the minority class will search for the Nearest Neighbor (NN) to find samples in other minority classes. The NN search is done using the K-NN algorithm by calculating using the Euclidean distance for continuous data. The number of NNs in the data synthesis process is determined by r as a consideration. The data synthesis process is carried out by selecting samples x_i and y_i in the minority class and then searching for the nearest neighbor z_{ij}. This is processed repeatedly until the number of data samples in the minority class is the same as the majority class. Handling data synthesis on categorical or nominal data types is done by selecting the majority value from the selected NN, then the data in the minority class is synthesized according to that value. After the entire data synthesis process is completed, the results are combined with the original data from the minority class. This combination produces the same amount of data in the minority class as the amount of data in the majority class. The existence of a balanced amount of data can reduce overfitting in the prediction process that can occur if the data is trained on a limited minority class.

Ensemble Learning

EL technique is part of ML by combining several models to improve the accuracy and stability of predictions [26]. Prediction using EL technique by fixing the weaknesses of individual models and obtaining better results [26]. There are several types of EL, namely Bootstrap Aggregating (Bagging), Boosting, and Stacking [26]. This study uses the XGBoost Boosting technique. The Boosting model works by fixing errors made by the previous model [26].

XGBoost

XGBoost is an ML algorithm designed to be an EL technique by boosting [27]. This algorithm prioritizes speed and better performance in the prediction model, and is an implementation of the gradient boosting algorithm [27]. The prediction process uses a block system which is used to store data to make the prediction process more efficient [27]. This is because the tree model in XGBoost is designed to improve the shortcomings of the previous tree [27]. Mathematically, the XGBoost algorithm can be described by assuming the dataset $D = \{x_i, y_i\}$ with $i \in 1, \ldots, K$ [19]. The value of K is the number of trees. The predicted results $\hat{y}$ can be calculated using Equation (1). In this equation, it is calculated based on the function f for each k^{th} tree [19].

$$\hat{y}_i = \varphi(x_i) = \sum_{k=1}^{K} f_k(x_i). \quad (1)$$

The initial step in making predictions on obesity datasets is done by determining the initial prediction value using the log-odds value for the classification case [28]. The log-odds function is shown in Equation (2) [28]. The log-odds value is a comparison between the occurrence of an event p and the non-occurrence of the event [28]. This is certainly different from the usual probability which is a comparison value of the occurrence of an event to all events or the total number [28]. The next step is to find the difference between the actual value and the predicted value. This will later be used in calculating the Gradient and Hessian of the loss function. In XGBoost the loss function is part of the objective function L. The log-odds value is transformed to p to obtain Equation (3). This value can be used to calculate the loss function with $l(y_i, p)$.

$$\log(odds) = \frac{p}{1-p}, \tag{2}$$

$$p = \frac{e^{\log(odds)}}{1 + e^{\log(odds)}}. \tag{3}$$

In the XGBoost algorithm, the objective function L is applied to improve the prediction results [19]. The loss function l and the penalty function Ω shown in Equation (4) are two important parts in calculating the objective function [19]. Measuring how good a model is by calculating the error between the actual value and the predicted value is what is used for the loss function [19]. The prediction is better if the value of the loss function is smaller. In Equation (4) is the loss function in predicting the multiclass case, where y_i is the actual value and $\hat{y}_i$ is the predicted value that can be calculated using p [29]. The penalty function Ω is used to prevent the model from becoming too complicated. This is shown in Equation (5). The parameter γT is the penalty on the leaves, where T is the number of leaves and the parameter γ regulates the amount of penalty on each leaf [19]. The smaller the value of the parameter γ, the simpler and more flexible the tree is. This is used to limit the size of the tree so that it does not form too many leaves. ω_j is the weight value on leaf j, the parameter λ is used to set the L2 regularization penalty on ω_j [19]. Too large a weight value can cause instability in the model and the use of the penalty function is to encourage the weight to remain small and help distribute contributions between leaves. The larger the value of λ and the smaller the value of ω_j, the better and more conservative the model will be [15].

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{3}$$

$$l(y_i, \hat{y}_i) = -\frac{1}{n} \sum_{i=1}^{n} y_i \log(\hat{y}), \tag{4}$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2. \tag{5}$$

Gradient and Hessian are used to calculate the weight of the leaf. Calculations can be obtained sequentially with radians being the first derivative and Hessian being the second derivative of the loss function [19]. The Gradient equation g_i and Hessian h_i are obtained using Equation (6) and Equation (7) [19]. The g_i function is used to determine the model update by minimizing the loss function [19]. In contrast to the h_i function as the second derivative of the loss function is used to measure the rate of change of Gradient and provide information about the change in the speed of the loss function to the prediction [15]. Hessian is used to help determine the optimal value in the model update. This can help the model maintain stability in the training process and can build an optimal

tree in each iteration. In Equation (8) is used to calculate the weight of the leaf using g_i and h_i [19]. Then the function is substituted into the loss function and obtained in Equation (9) [19].

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right), \quad (6)$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right), \quad (7)$$

$$\omega_j = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}, \quad (8)$$

$$L = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T. \quad (9)$$

with T being the number of iterations where $T \in 1, \dots, t$.

Evaluation

The evaluation process uses the confusion matrix which then obtains the percentage of model accuracy. The confusion matrix has 4 important components, namely TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) [30]. The TP and TN components produce correct predictions in the positive and negative classes [30]. The opposite applies to FP and FN predicting incorrectly in both classes, namely positive and negative [30]. In the multiclass case, the metrics evaluation process uses the index k which defines the class [29]. The confusion matrix can be used to calculate accuracy, prediction quality (precision), prediction accuracy (recall), and harmony between prediction quality and accuracy (f1-score). Accuracy is used to calculate the overall correct percentage using Equation (10) [30]. Precision is used to predict the proportion of the quality of the correct positive class and recall is used to measure the proportion of positive class data detected by the model and shows the sensitivity of the model in recognizing positive class data. This is shown in Equation (11) and Equation (12) [30]. The harmony of precision and recall can be calculated using the f1-score shown in Equation (13) [30]. The f1-score calculation in the multiclass case is very important, because in the calculation there is a trade-off between precision and recall. This can be used as a consideration for performance in all classes. This overall performance matrix is important in considering the quality of ML algorithm performance in prediction.

$$\text{Accuracy} = \frac{TP_k + TN_k}{TP_k + FP_k + TN_k + FN_k} \quad (10)$$

$$\text{Precision} = \frac{TP_k}{TP_k + FP_k} \quad (11)$$

$$\text{Recall} = \frac{TP_k}{TP_k + FN_k} \quad (12)$$

$$\text{F1} - \text{score} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (13)$$

RESULTS AND DISCUSSION

Explanation and presentation of the results are carried out in this section in the form of analysis. The aim of this research is to obtain analysis results regarding the performance of the XGBoost and SMOTE-NC models in predicting obesity data sets. The presentation of the results of this research is shown in tables and graphs with the aim of facilitating understanding. The evaluation results include several classification metrics such as accuracy, precision, recall, and f1-score. These metrics describe the performance

of the model and the efficiency of the training process that has been carried out. This research process was carried out as many as 25 iterations and 2 treatments on each algorithm in order to avoid bias in making decisions on model performance. The treatment was carried out by comparing the use of SMOTE-NC and not using SMOTE-NC in the preprocessing stage in obesity prediction. This study uses a dataset taken from the UCI Machine Learning platform. This dataset was taken because of its relevance to the research topic, consisting of 17 columns and 2111 rows in the form of nominal and continuous data. The data consists of 16 attribute columns (variable X) and 1 response column (variable Y). Before the classification analysis process is carried out, the obesity dataset goes through a preprocessing stage. This stage includes data identification, equalizing data types by changing data with categorical characters to nominal, balancing data using the SMOTE-NC technique. After this stage is finished, the data is confirmed to be ready for further processing in the model classification analysis.

Preprocessing Data

The preprocessing stage is an important part of every research. This is because at this stage the data undergoes processing and maturation so that it is ready to be used for the analysis process using a certain algorithm. In the first stage, the dataset is identified to obtain information about the characteristics and structure of the data. The identification process includes examining the data type, the number of missing values, and also the number of samples in each class of variable Y data. The identification results for the data type are 9 data with the object data type and the rest are of the int64 type. All columns in the obesity dataset have no missing values. Then, the data with the object type is changed to int64 to match the data type. The results of the data identification process are shown in **Table 1**. Changes were made to all variables with the previous object data type. Examples of changes to the Y variable or the Nobeyesdad variable are presented in **Table 2.**

Table 1. Data Types and Missing Value

Name of Variable	Data Type	Missing Value
Gender	Object	0
Age	int64	0
Height	int64	0
Weight	int64	0
family_history_with_overweight	Object	0
FAVC	Object	0
FCVC	int64	0
NCP	int64	0
CAEC	Object	0
SMOKE	Object	0
CH2O	int64	0
SCC	Object	0
FAF	int64	0
TUE	int64	0
CALC	Object	0
MTRANS	Object	0
NObeyesdad	Object	0

Table 2. The Result of Changing for Data Types on Y Variable

Variable	Types of Data	
	Before	After
NObeyesdad	NW	0
	OT_I	1
	OT_II	2
	OT_III	3
	OW	4
	UW	5

Following are the results of the data identification process to determine the number of data samples for each class in variable Y, then a resampling technique is carried out to balance the number of data samples. This is summarized in Table 3. The results of the study show a significant change in the number of minority class samples compared to the number of samples in the majority class. The class with the largest samples in variable Y is the overweight class with 580 sample data. In the other classes, it is the minority class which is then resampled to 580 sample data in each class. The application of SMOTE-NC can perform data synthesis on data with mixed types such as in the obesity dataset, namely nominal and continuous. Data on the X variable will be synthesized based on the Y variable data. For example, in the Y variable the fifth data is included in the minority class and data synthesis will be carried out, then in the X variable the fifth data will also occur data synthesis. This process will occur until the minority class has reached the target number to balance with largest number of classes.

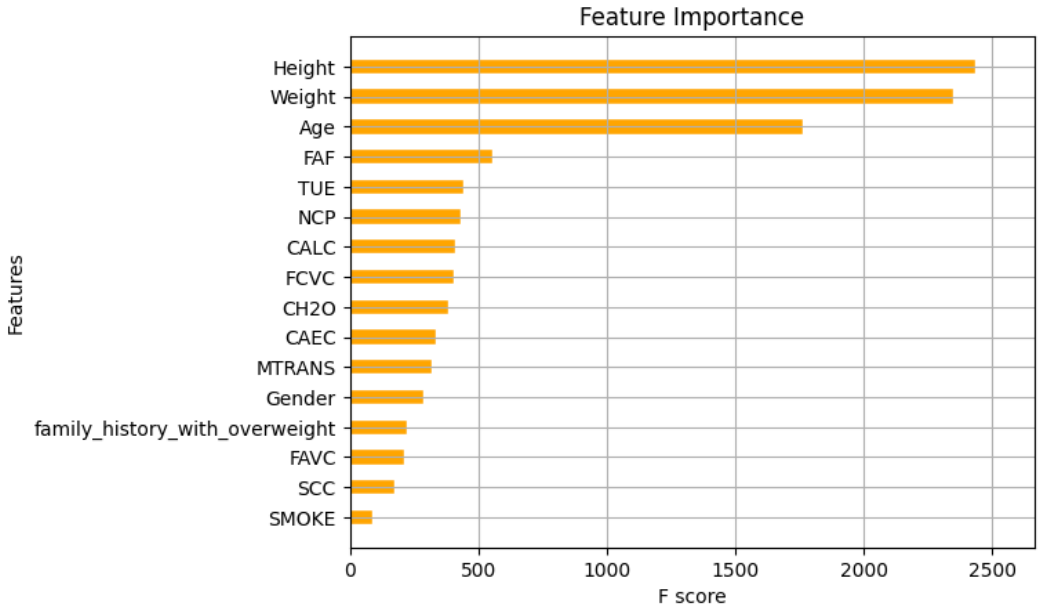
Table 3. Number of Data Samples in Y Variable

No	Class	Number of Samples	
		Before SMOTE-NC	After SMOTE-NC
1.	NW	287	580
2.	OT_I	351	580
3.	OT_II	324	580
4.	OT_III	297	580
5.	OW	580	580
6.	UW	272	580

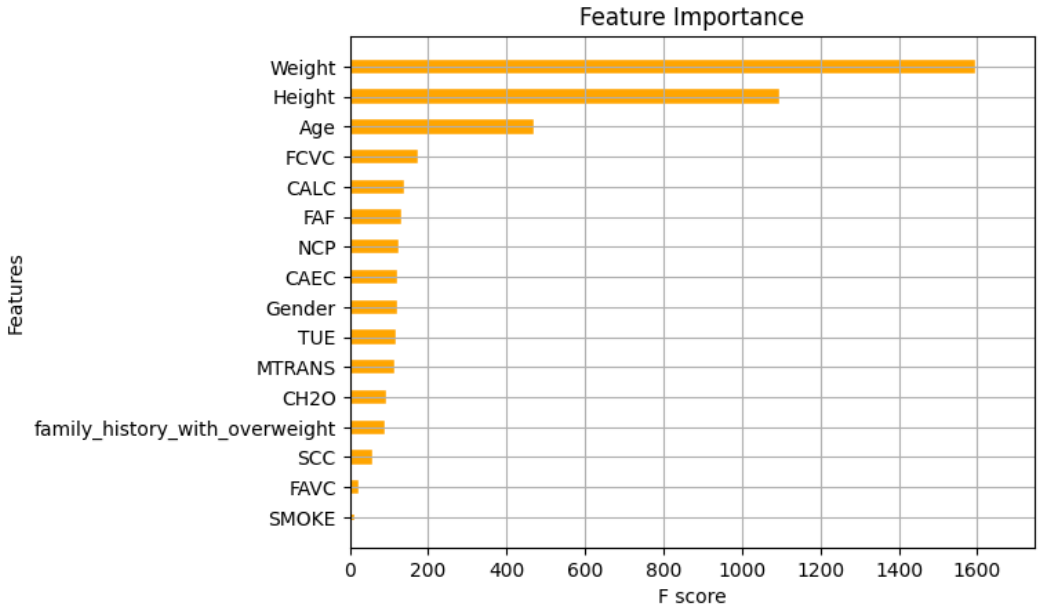
Ensemble Learning Method and Evaluation

After the obesity dataset has undergone data preprocessing and the data has been balanced, the prediction process can be carried out using the ML algorithm, namely XGBoost. Evaluation using several model performance metrics was carried out on both treatments, namely applying SMOTE-NC and not. The calculation of the evaluation of the obesity prediction performance results is shown in Table 4. The high accuracy results in both treatments indicate their ability to classify data correctly. The prediction results for XGBoost without using SMOTE-NC show that its performance is much lower than using

SMOTE-NC. The average accuracy of XGBoost performance without using SMOTE-NC has a lower accuracy difference of 19.78% in the use of all features in the X variable.



(a),



(b).

Figure 2. Feature importance for XGBoost model (a) without SMOTE-NC (b) with SMOTE-NC.

If observed more deeply, the fastest processing time is carried out by the XGBoost algorithm without using SMOTE-NC. This does not have a significant impact if processing time is not the biggest consideration in the study. In addition, the highest percentage in other evaluation metrics in the form of a precision value of 98.32%, recall of 98.02% and f1-score of 97.30% was also obtained in the XGBoost treatment using SMOTE-NC. The obtained values indicate that the performance of XGBoost and SMOTE-NC can predict classes correctly at high precision values, can correctly detect a class based on a recall value of 98.02%, and show solid performance in the process of predicting and detecting a

class correctly. These values are obtained based on the mean of the prediction process which is repeated 25 times to obtain accuracy in the evaluation metrics. This study also considers the use of feature selection, namely feature importance. This is used to determine the variables or features that have the most influence on the obesity prediction process. Based on Figure 2, it is known that the order of the three highest features is the most influential in the obesity prediction process. Figure 2 (a) shows a picture of the results of the feature importance level for treatment without SMOTE-NC. The figure shows that the feature with the highest influence is obtained by Height, followed by Weight and Age. In addition, in Figure 2 (b) it is known that the Weight variable has the greatest influence in the prediction process, followed by Height and Weight. There are 3 types of combination features with the same high level of influence in the obesity prediction process.

Further analysis is carried out based on the acquisition of information regarding the magnitude of the influence of variables on the obesity prediction process, it can be seen from the results of the prediction process in all treatments in Table 4 that the stability of performance can be measured using the value of the standard deviation. The smaller the standard deviation value, the more stable the prediction performance. Based on this, if an observation is made on the results of the standard deviation shown in Table 4, the treatment that obtained the best level of performance stability was in the prediction using SMOTE-NC on XGBoost by including all X variables. The results of the standard deviation on all metrics are around 0.004%.

Table 4. Comparation of Mean and STD Result for XGBoost

Variable	Method	Algorithm	Acc (%)	Pre (%)	Rcl (%)	F1S (%)	T (s)
All of X Variabel	Without SMOTE-NC	Mean	0,7854	0,7840	0,7823	0,7809	3,169
		STD	0,0234	0,0249	0,0212	0,0236	0,3394
	With SMOTE-NC	Mean	**0,9830**	**0,9832**	**0,9802**	**0,9730**	3,468
		STD	0,004899	0,004787	0,004892	0,004898	0,3374
Weight, Height, Age	Without SMOTE-NC	Mean	0,7081	0,6982	0,6982	0,6976	2,990
		STD	0,0234	0,0249	0,0212	0,0236	0,8034
	With SMOTE-NC	Mean	**0,9803**	**0,9804**	**0,9790**	**0,9803**	3,256
		STD	0,005472	0,005423	0,005041	0,004898	0,4630

The performance results of XGBoost in the case of multiclass classification and balanced data using SMOTE-NC using only 3 X variables in the prediction process are shown in Table 4. The results after reducing the feature variables other than Height, Weight and Age showed that the performance in the treatment without SMOTE-NC decreased, with an average accuracy of around 70.81% and followed by other metrics. The table shows that when using SMOTE-NC, the results increased from not using SMOTE-NC to reach an average accuracy of around 98.03%. There are advantages after feature selection, namely shorter computing time than before. Although after the feature selection process there was a decrease in accuracy, especially in the treatment without the SMOTE-NC technique, this did not have a significant impact on the treatment using the SMOTE-NC technique. The decrease in accuracy can be caused because not all variables are used in the prediction process, but the decrease is still in good accuracy for the treatment with the SMOTE-NC method, which is 98.03%.

Table 5. Calculation of Confidence Interval, Lower Bound, and Upper Bound

Variable	Method	Confidence Interval 90%	Lower Bound	Upper Bound
All of X Variable	Without SMOTE-NC	0,006877	0,77852	0,79227
	With SMOTE-NC	0,001438	**0,98156**	**0,98443**
Weight, Height, Age	Without SMOTE-NC	0,006870	0,70122	0,71497
	With SMOTE-NC	0,001606	0,97869	0,98140

In addition, a search for the confidence interval of these results can be carried out. This is explained in Table 5 that the analysis of the use of variables in two different approaches produces a confidence interval of accuracy metrics with certain limits. The results of the approach using all variables obtained that the confidence interval in the treatment without using SMOTE-NC ranged from 77.85% -79.22%. The results of this range were not better than the treatment when using SMOTE-NC. This treatment obtained a confidence interval in the range of 98.15% -98.44%. The approach using only 3 variables showed results that were not higher than the approach using all variables. The decrease in results ranged from 70.12% - 71.49% in the treatment without SMOTE-NC and 97.86% - 98.14% in the treatment using SMOTE-NC. However, in all approaches, whether using SMOTE-NC or not, the best results were obtained when the SMOTE-NC technique was applied. Figure 3 shows the best range of accuracy results in the XGBoost treatment using SMOTE-NC with the approach using all X variables. These results indicate that the combination of XGBoost and SMOTE-NC results in increased prediction performance in terms of obesity in particular. In general, based on the important features obtained in obesity prediction, it is known that the 3 features can be a reference for the general public that it is important to maintain body weight so that it is always within normal weight limits, especially at an increasingly older age. Based on this, it is an initial step in preventing and reducing obesity rates in the world.

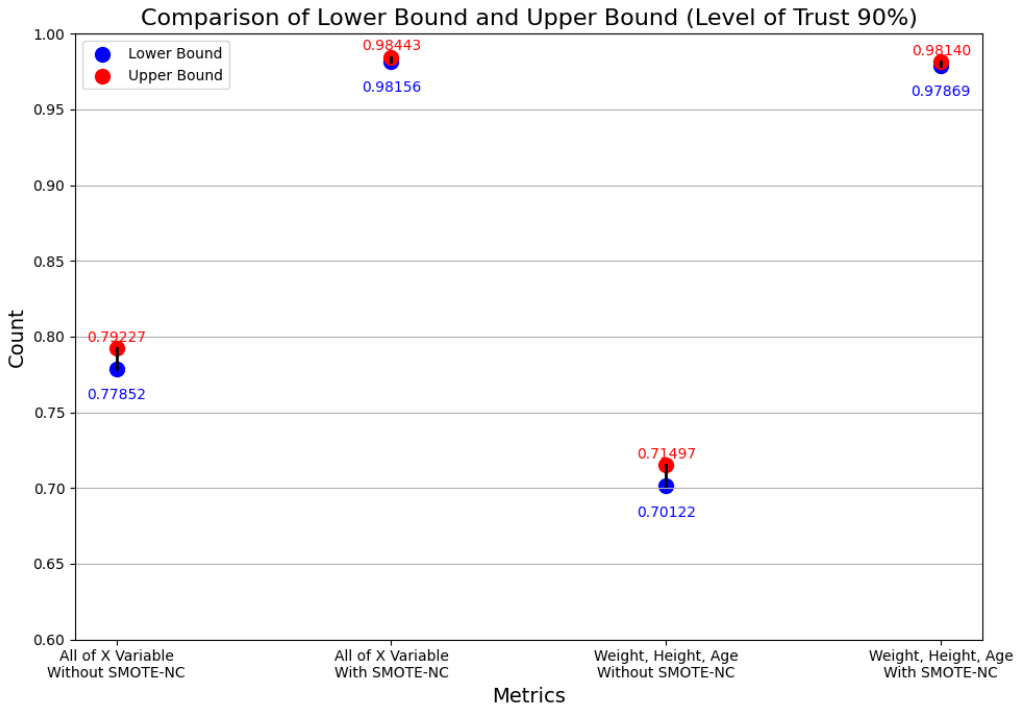


Figure 3. Comparing The Accuracy Results of Confidence Interval

CONCLUSIONS

Based on the results and discussion chapters, it is obtained that both treatments and approaches can work well in the obesity prediction classification process, especially on datasets that have passed the data balancing stage using the SMOTE-NC method. The results of the evaluation metrics on the XGBoost algorithm using SMOTE-NC by including all X variables are the best results. The average results of the evaluation metrics on the XGBoost treatment with SMOTE-NC are around 98.22% accuracy, 98.32% precision, 98.02% recall, and 97.30% f1-score. The prediction process using all X variables in the XGBoost technique with SMOTE-NC is 0.2 seconds slower than after applying feature selection and using only 3 X variables. At the level of model stability in predicting obesity, a lower standard deviation is obtained in the XGBoost model with SMOTE-NC, especially in the use of all X variables. The overall results show that XGBoost supported by SMOTE-NC and using all X variables is the best choice in the case of obesity dataset classification. This can be seen from various aspects such as accuracy, speed, and stability. In addition, the use of the SMOTE-NC technique also has a significant effect on the performance of the model classification. The application of SMOTE-NC can improve model performance in cases of obesity classification. The results of the application of feature selection obtained that the level of importance of maintaining body weight is very important to do and this will also affect age. Prevention of obesity is done by starting the body with a normal and healthy weight.

REFERENSI

[1] WOF, World Obesity Atlas 2022, London: World Obesity Federation 2022, 2022.

[2] WHO, "Obesity and Overweight," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[3] E. Laurence, "Obesity Statistics And Facts In 2024," 2024. [Online]. Available: https://www.forbes.com/health/weight-loss/obesity-statistics/.

[4] Kemenkes, "1 Dari 4 Penduduk Indonesia Mengalami Obesitas," Kementrian Kesehatan Indonesia, 2018. http://surl.li/dajvxd

[5] J. Cawley, "An economy of scales: A selective review of obesity's economic causes, consequences, and solutions," *Journal of Health Economics,* no. Elsevier B.V, pp. 244-268, 2015. doi:http://dx.doi.org/10.1016/j.jhealeco.2015.03.001

[6] B. A. Ejigu dan F. N. Tiruneh, "The Link between Overweight/Obesity and Noncommunicable Diseases in Ethiopia: Evidences from Nationwide WHO STEPS Survey 2015," *International Journal of Hypertension,* pp. 1-11, 2023. doi:https://doi.org/10.1155/2023/2199853

[7] G. 2. O. Collaborators, "Health Effects of Overweight and Obesity in 195 Countries over 25 Years," *Bill & Melinda Gates Foundation,* vol. 1, no. 377, pp. 13-27, 2017. doi:10.1056/NEJMoa1614362

[8] R. Peters, N. Ee, J. Peters, N. Beckett, A. Booth, K. Rockwood dan K. J. Anstey, "Common risk factors for major noncommunicable disease, a systematic overview of reviews and commentary: the implied potential for targeted risk reduction," *Therapeutic Advances in Chronic Disease,* vol. 10, pp. 1-14, 2019. doi:https://doi.org/10.1177/2040622319880392

[9] N. L. Rane, M. Paramesha, S. P. Choudhary dan J. Rane, "Machine Learning and Deep Learning for Big Data Analytics: A Review of Methods and Applications," *Partners*

Universal International Innovation Journal (PUIIJ), vol. 03, no. 02, p. 172, 2024. doi:DOI: 10.5281/zenodo.12271006

[10] P. Mahajan, S. Uddin, F. Hajati dan M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," *healthcare,* no. 11, pp. 1-21, 2023. doi:https://doi.org/10.3390/healthcare11121808

[11] F. Ferdowsy, K. S. A. Rahi, M. I. Jabiullah dan M. T. Habib, "A Machine Learning Approach for Obesity Risk Prediction," *Current Research in Behavioral Sciences,* pp. 1-9, 2021. doi:https://doi.org/10.1016/j.crbeha.2021.100053

[12] X. Pang, C. B. Forrest, F. Le-Scherban dan A. J. Masino, "Prediction of early childhood obesity with machine learning and electronic health record data," *International Journal of Medical Informatics,* pp. 1-8, 2021. doi:https://doi.org/10.1016/j.ijmedinf.2021.104454.

[13] M. Calderon-Diaz, L. J. Seret-Castillo, E. A. Vallejos-Cuevas, A. Espinoza, R. Salas dan M. A. Macias-Jimenez, "Detection of Variables for The Diagnosis of Overweight and Obesitty in Young Chileans Using Machine Learning Techniques," dalam *The 1th International Workshop on Human-Centric Innovation and Computational Intelligence (IWHICI 2023)*, 2023. doi:10.1016/j.procs.2023.03.135

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall dan W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002. doi:DOI: https://doi.org/10.1613/jair.953.

[15] I. A. Rahmayanti, Sediono, T. Saifudin dan E. Ana, "Applying Smote-Nc On Cart Algorithm To Handle Imbalanced Data In Customer Churn Prediction: A Case Study Of Telecommunications Industry," *Jurnal Ilmiah Indonesia,* no. 6, pp. 1321-1338, 2021. doi:https://doi.org/10.47709/cnahpc.v6i2.3790.

[16] Q. H. Doan, S.-H. Mai, Q. T. Do dan D.-K. Thai, "A cluster-based data splitting method for small sample and class imbalance problems in impact damage classification," *Applied Soft Computing,* no. 120, pp. 1-17, 2022. doi:https://doi.org/10.24433/CO.3297630.v1.

[17] D. L. Zou, X. Fang, L. Xu dan L. L. Wu, "Study on the relationship of design parameters and damage modes for RC slabs subjected to large-scale hard missile impacts base on task-driven approach," *Structures,* vol. 58, pp. 1-17, 2023. doi:https://doi.org/10.1016/j.istruc.2023.105635

[18] S. Chalichalamala, N. Govindan dan R. Kasarapu, "An extreme gradient boost based classification and regression tree for network intrusion detection in IoT," *Bulletin of Electrical Engineering and Informatics,* vol. 3, no. 13, pp. 1741-1751, 2024. doi: 10.11591/eei.v13i3.6843

[19] T. Chen dan C. Guestrin, "XGBoost: A Scalable Tree Boosting System," dalam *KDD '16*, San Francisco, 2016. doi: http://dx.doi.org/10.1145/2939672.2939785

[20] S. B. Kotsiantis, D. Kanellopoulos dan P. E. Pintelas, "Data Preprocessing for Supervised Leaning," *International Journal Of Computer Science,* vol. 1, no. 1, pp. 111-117, 2006.

[21] Suraya, M. Sholeh dan U. Lestari, "Evaluation of Data Clustering Accuracy usingK-Means Algorithm," *International Journal of Multidisciplinary Approach Research and Science,* vol. 01, no. 02, pp. 385-396, 2024 doi:https://doi.org/10.59653/ijmars.v2i01.504

[22] F. M. Palechor dan A. d. l. H. Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico," *Data in brief,* vol. 25, pp. 1-5, 2019. doi:https://doi.org/10.1016/j.dib.2019.104344

[23] S. A. Alasadi and W. S. Bhaya, "Review of Data Preprocessing Techniques in Data Mining," *Journal of Engineering and Applied Sciences,* vol. 16, no. 12, pp. 4102-4107, 2017. https://www.researchgate.net/publication/320161439.
[24] I. Miller dan M. Miller, John E. Freund's Mathematical Statistics Eighth Edition, Edinburgh Gate: Pearson Education Limited, 2014
[25] D. T. Utari, "Integration Of Svm And Smote-Nc For Classification Of Heart Failure Patients," *Barekeng: Journal of Mathematics and Its Applications,* vol. 4, no. 17, pp. 2263-2272, 2023. doi: https://doi.org/10.30598/barekengvol17iss4pp2263-2272
[26] Z. H. Zhou, Ensemble Method Foundations and Algorithm, Boca Raton: Taylor and Francis Group, CRC press, 2012.
[27] A. F. L. Ptr, M. M. Siregar dan I. Daniel, "Analysis of Gradient Boosting, XGBoost, and CatBooston Mobile Phone Classification," *Journal of Computer Networks, Architecture and High Performance Computing,* vol. 2, no. 6, pp. 661-670, 2024. doi:https://doi.org/10.47709/cnahpc.v6i2.3790
[28] D. Hu, C. Wang dan A. M. O'Connor, "A method of back-calculating the log odds ratio and standard error of the log odds ratio from the reported group-level risk of disease," *Plos One,* pp. 1-8, 2020.
[29] M. Grandini, E. Bagli dan G. Visani, "Metrics for Multi-Class Classification: an Overview," *Cornell University,* 2020. doi:https://doi.org/10.48550/arXiv.2008.05756
[30] Ž. Đ. Vujović, "Classification Model Evaluation Metrics," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 6, no. 12, pp. 1-8, 2021. doi:10.14569/IJACSA.2021.0120670