


RESEARCH ARTICLE | JUNE 28 2023

# Performance comparison of machine learning algorithms for predicting obesity level

Suwarno ; Nerru Pranuta Murnaka; Puguh Wahyu Prasetyo; Samsul Arifin

 Check for updates

AIP Conference Proceedings 2733, 020002 (2023)

<https://doi.org/10.1063/5.0140856>

  
View  
Online

  
Export  
Citation

CrossMark

500 kHz or 8.5 GHz?  
And all the ranges in between.

Lock-in Amplifiers for your periodic signal measurements



Find out more

 Zurich  
Instruments

# Performance Comparison of Machine Learning Algorithms for Predicting Obesity Level

Suwarno<sup>1,a)</sup>, Nerru Pranuta Murnaka<sup>2,b)</sup>, Puguh Wahyu Prasetyo<sup>3,c)</sup>, Samsul Arifin<sup>4,d)</sup>

<sup>1</sup>Primary Teacher Education Department, Faculty of Humanities, Bina Nusantara University, Jakarta, Indonesia.

<sup>2</sup>Mathematics Education Department, STKIP Surya, Tangerang, Indonesia.

<sup>3</sup>Mathematics Education Department, Faculty of Teacher Training and Education, Universitas Ahmad Dahlan, Indonesia.

<sup>4</sup>Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta, 11480, Indonesia

<sup>a)</sup>Corresponding author: suwarno001@binus.ac.id

<sup>b)</sup> murnaka@gmail.com

<sup>c)</sup> puguh.prasetyo@pmat.uad.ac.id

<sup>d)</sup> samsul.arifin@binus.edu

**Abstract.** Obesity problems have actually come to be a worldwide epidemic that has increased since 1980, with significant repercussions for health and wellness in young adults, adults, and youngsters. Obesity problems are an issue that has actually been expanding steadily which is why daily appear new studies entailing youngsters' excessive weight, specifically those looking for influence elements as well as exactly how to predict the appearance of the condition under these elements; for this reason, early detection is called for. Data mining and also machine learning (ML) algorithms approaches are made use of in obesity problems forecast in our research. We made use of the Obesity Level dataset for our study, accumulated from the UCI Machine Learning Repository. The dataset includes information about 638 patients as well as their matching 17 attributes. We made use of nine ML algorithms on the dataset to predict obesity problems. We found that the model with Logistic Regression algorithm is well on obesity level prediction. The result validated Logistic Regression algorithm has the best performance of accuracy (100%), sensitivity (100%), specificity (100%), as well as AUC (1). The Logistic Regression model was selected because to its best performance, best gain, and fastest total time.

## INTRODUCTION

Obesity is defined as having an excessive quantity of body fat. Weight gain is not just due to food intake; genetics and the environment can all play a role in the development of obesity. Because it is a worldwide health problem, it has the potential to pose a threat to the world in the future. Obesity can be caused by a variety of factors, and it can even be classified as a disease. Obesity is associated with thousands of dangers and illnesses in a variety of sectors. It is one of the most frequent health disorders in the globe, affecting people all over. Obesity is mostly caused by excessive eating combined with insufficient physical activity. If people do not burn off their excess energy by physical activities such as yoga, workouts, and fasting, but instead consume large quantities of energy, particularly fat and sugar, a significant portion of the excess energy is converted to fat and stored in the body as body fat. The majority of individuals are unconcerned with their weight since they believe it is one of the broad definitions of health. Furthermore, they believe that it will have no negative impact on their health. The outside structure of their body is all that is seen of them. However, the unpleasant reality is that obesity is a risk factor for the majority of illnesses. Occasionally, it can result in mortality, as seen by severe epidemics of diabetes, cardiovascular disease, cancer, osteoarthritis, chronic renal disease, stroke, hypertension, and other life-threatening disorders.

Since 1980, the number of obese persons has increased. In 2014, about 1900 million persons aged 18 years or older suffered from alteration of their weight. Some of the causes of obesity include an increase in the consumption of high-energy meals and a reduction in physical activity [1]. Obesity is a global public health concern that affects adults, adolescents, and children. Obesity can manifest itself in any age group [2]. According to Hernández [3], the results explain that obesity may be regarded an illness with various contributing elements, with one of the symptoms being an uncontrolled rise in weight as a result of excessive fat and energy intake. In both adults and children, obesity can progress. In most cases, the imbalance in energy between calorie intake and expenditure is considered the core definition of obesity.. The obesity rate has increased by nearly thrice since 1975, according to the World Health Organization. The obesity rate in 2016 reached more over 650 million people, with 39 percent of those aged 18 and older classified as overweight, and 13 percent as obese [4]. As of 2016, more over 340,000 children were overweight, and as of 2019, 34 million children under the age of five were overweight or obesity [4]. Given the evidence shown above, it is reasonable to predict that obesity will become a major concern in the near future.

Several authors have conducted research to evaluate the condition and developed web-based tools to determine a person's obesity level. However, such tools are confined to the computation of the body mass index, neglecting important elements such as family history and time spent exercising. The main objective of this report is to examine individuals for signs of obesity and to educate them about the dangers of being overweight. The purpose of this research is to forecast the risk of obesity. The analysis is divided into two parts: first, it reads the data, and then it verifies the data to see whether it fits the component associated with obesity; second, it displays the results of the analysis. In order to do our research, we must first gather raw data sets, which are dependent on a number of parameters. On top of that, we performed pre-processing on the data and then used nine machine learning supervised algorithms to evaluate the accuracy, sensitivity, specificity, and AUC. Then we discovered which algorithm performs the best and were able to predict the real result.

## LITERATURE REVIEW

This section of the article discusses all of the previous and current research that has been conducted to predict the risk of obesity. We observed their work and attempted to comprehend the way they displayed. Dugan et al. [5] performed a great job predicting childhood obesity. They examined six models in their investigation. Their models include Random Tree, Random Forest, ID3, J48, Nave Bayes, and Bayes Net, all of which were trained on CHICA, a clinical decision support system. They obtained the best result from the model ID3, which was extremely accurate to the tune of 85 percent and highly sensitive to the tune of early 90 percent. Jindal et al. [6] investigate the predictive power of ensemble machine learning algorithms for obesity. Their projected value for obesity was 89.68 percent correct, which enabled them to propose an ensemble machine learning technique for obesity prediction and to apply the ensemble prediction they employed. Additionally, the Python interface makes use of generalized linear models, random forests, and partial least squares to optimize their prediction model.

Machine learning was suggested by Singh and Tawfik [7] to predict the likelihood of being fat or overweight throughout adolescence. Their model was built using seven machine learning methods. A sample of an unchanged, unequally distributed dataset was used to evaluate the performance of all the methods, which included k-NN and J48 pruned tree, Random forest, Bagging, support vector machine, multilayer perception and voting. The MLP algorithm has a precision of 96 percent. There was a 93.96 percent success rate for the  $F_1$  score. Gerl [8] uses a large population cohort to predict several indices of obesity. Additionally, they were able to predict 8 percent of the complete range of BFP with error, while interpreting 73% of its variations based on the age, gender, and the lipidome of each participant.

For children between the ages of 2 and 17, Davila-Payan et al. [9] proposed a Logistic Regression model to assess the likelihood of body mass index in local geographic regions. According to their findings, it is vital to conduct small-scale assessments to develop actionable actions and assist design viable solutions to the issue. Using fuzzy signatures, Manna and Jewkes [10] provided a computational model to grasp and manage complexities on the data of children's obesity and a solution that could handle the risk linked with early obesity and children's motor development risk. A computer paradigm known as fuzzy logic offers a mathematical tool for handling uncertainty and imprecision, which is ubiquitous in human thinking, in the study of fuzzy signatures.

## RESEARCH METHOD

### Data, Feature, and Software Tool

The dataset used in this research was collected from the UCI Machine Learning Repository and represented the prevalence of obesity. In this dataset, you will find information on the estimated prevalence of obesity in people from Mexico, Peru, and Colombia, which is based on their eating habits and physical health. The collection contains information on 638 patients as well as the 17 distinct characteristics that distinguish them from one another. 287 samples of Normal Weight and 351 samples of Obesity Type I make up the collection. Table 1 contains a description of the characteristics of this dataset. A dependent or goal variable is believed to be the property with the word 'outcome,' whilst the remaining sixteen qualities are considered to be independent or feature variables. Results for the obesity level characteristic are represented as a binary value between 0 and 1, with 0 representing Normal Weight and 1 representing Obesity Type I [11]. To improve the accuracy of determining whether or not a patient has obesity, we used data mining and machine learning algorithms in our research. With the help of RapidMiner, a free and open-source machine learning and data mining software tool, we evaluated the performance of the obesity level dataset. Among the features of RapidMiner are tools for preprocessing data as well as clustering, classification, regression, visualization, and feature selection [12]. Dataset description show in Table 1 [13].

TABLE 1. Dataset Description

Attributes	Values
Sex	H: Male M: Female
Age	Integer Numeric Values
Height	Integer Numeric Values (Mt)
Weight	Integer Numeric Values (Kg)
Family with overweight / Obesity	Yes No
Fast Food Intake	Yes No
Vegetables Consumption Frequency	S: Always A: Sometimes CN: Rarely
Number of main meals daily	1 to 2: UD 3: TR More than 3: MT
Food intake between meals	S: Always CS: Usually A: Sometimes CN: Rarely
Smoking	Yes No
Liquid intake daily	MU: Less than one liter UAD: Between 1 and 2 liters MD: More than 2 liters
Calories Consumption Calculation	Yes No
Physical Activity	UOD: 1 to 2 days TAC: 3 to 4 days COS: 5 to 6 days NO: No physical activity

Attributes	Values
Schedule dedicated to technology	CAD: 0 to 2 hours TAC: 3 to 5 hours MC: More than 5 hours
Alcohol consumption	NO: No consumo de alcohol CF: Rarely S: Weekly D: Daily
Type of Transportation used	TP: Public transportation MTA: Motorbike BTA: Bike CA: Walking AU: Automobile CA: Walking
IMC	WHO Classification
Vulnerable	Based on the WHO Classification

The attributes related with eating habits are: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The attributes related with the physical condition are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS), other variables obtained were: Gender, Age, Height and Weight. Finally, all data was labeled and the class variable NObesity was created with the values of: Normal Weight and Obesity Type I based on Equation [14] and information from WHO and Mexican Normativity.

### Data Preprocessing and Missing Values Identification

Preprocessing aids in the transformation of data in order to build a better machine learning model that is more accurate. In order to enhance the quality of data, preprocessing conducts a number of operations, including outlier rejection and filling missing values. Data normalization and feature selection are also performed. 287 samples were classed as Obesity Type I, while 351 samples were classified as Normal Weight. We were not find the missing values in the datasets using JASP statistical software.

**TABEL 2.** Descriptive Statistics

	Valid	Missing	Mean	Std. Deviation	Minimum	Maximum
Gender	638	0				
Age	638	0	24.020	6.998	14.000	61.000
Height	638	0	1.686	0.097	1.500	1.980
Weight	638	0	79.053	18.578	42.300	125.000
family_history_with_overweight	638	0				
FAVC	638	0				
FCVC	638	0	2.253	0.515	1.000	3.000
NCP	638	0	2.570	0.841	1.000	4.000
CAEC	638	0				
SMOKE	638	0				
CH2O	638	0	1.994	0.644	1.000	3.000
SCC	638	0				
FAF	638	0	1.104	0.959	0.000	3.000
TUE	638	0	0.676	0.687	0.000	2.000
CALC	638	0				
MTRANS	638	0				

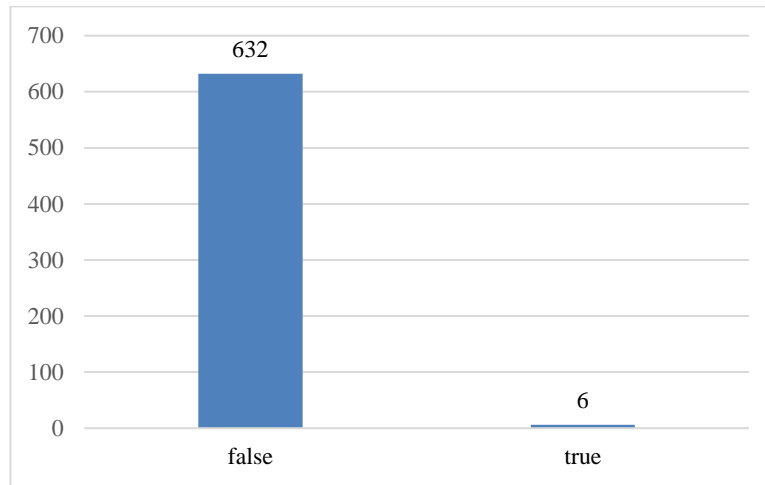
**TABEL 2.** Descriptive Statistics

	Valid	Missing	Mean	Std. Deviation	Minimum	Maximum
NObeyesdad	638	0				

Note. Not all values are available for *Nominal Text* variables

## Outlier Identification and Removal

We used the RapidMiner to filter the dataset in order to discover outliers based on interquartile ranges, which we then analyzed. We can see in Fig. 1 that there are six outliers in the data, which indicates that there are six outliers in the data. Following the removal of these outliers, there were 632 datasets remaining.

**FIGURE 1.** Count of Outlier

## Feature Selection

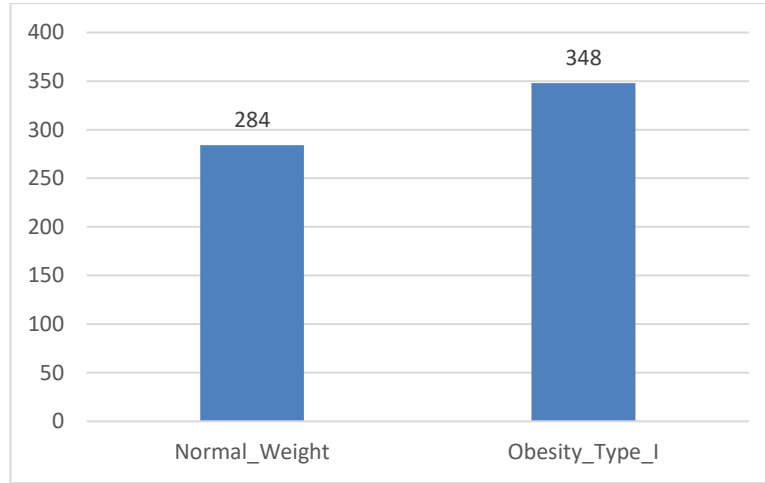
In order to determine the most significant attributes/features, Pearson's correlation approach is widely utilized in various applications. This approach calculates the correlation coefficient, which is a measure of how well the output and input properties correlate with one another. The value of the coefficient remains within the range of -1 and 1. A correlation coefficient greater than or equal to 0.5 indicates a significant correlation, whereas a correlation coefficient of zero shows no correlations. Table 3 shows the results of finding the correlation coefficient in Weka using the correlation filter, which is implemented in the software package. For relevant qualities, we deleted the features that contain NaN correlation. Hence CAEC, CALC, and MTRANS are removed.

**TABEL 3.** Correlation Between Other Features With Outcome Feature.

Attribute	Correlation	Attribute	Correlation
Age	0.291	Gender	0.051
CAEC	NaN	Height	0.095
CALC	NaN	MTRANS	NaN
CH2O	0.188	NCP	-0.195
FAF	-0.143	SCC	-0.222
family_history_with_overweight	-0.528	SMOKE	-0.094
FAVC	0.359	TUE	0.01
FCVC	-0.142	Weight	0.823

## Normalization

The algorithm's computation time was increased by performing feature scaling by normalizing the data from 0 to 1 range [15]. As a consequence of normalization, the mean and standard deviation for each attribute are displayed in Table 4. In Fig. 2, we can see that, after finishing preprocessing, we get 632 samples, of which 284 samples were classified as Normal Weight and 348 samples were classified as Obesity Type I. After the preprocessing stage the correlation coefficient between 'Weight' and the outcome is 0.823. As a result, these are highly correlated.



**FIGURE 2.** Count of Obesity Level After Preprocessing

**TABEL 4.** Mean and Standard Deviation After Normalization

	Valid	Missing	Mean	Std. Deviation	Minimum	Maximum
Age	632	0	24.003	6.973	14.000	61.000
CH2O	632	0	2.002	0.684	1.000	3.000
FAF	632	0	1.100	0.976	0.000	3.000
family_history_with_overweight	632	0				
FAVC	632	0				
FCVC	632	0	2.245	0.544	1.000	3.000
Gender	632	0				
Height	632	0	1.693	0.101	1.500	2.000
NCP	632	0	2.563	0.864	1.000	4.000
NObesyedad	632	0				
SCC	632	0				
SMOKE	632	0				
TUE	632	0	0.691	0.722	0.000	2.000
Weight	632	0	79.100	18.599	42.300	125.000

*Note.* Not all values are available for *Nominal Text* variables

## Dataset Training, Testing, and Model Evaluation

Once the data has been cleaned and preprocessed, it is ready for training and testing. We utilize Auto Model to choose the best model among nine machine learning techniques in this research. The auto model partitions the original

dataset 60% and 40% (Training and Testing). The auto model's validation is a multiple hold out set validation. The model will be trained on 60% of the data, while the other 40% will be separated into seven subgroups. Once trained, the model will be used to generate predictions separately on each of the seven subgroups, and the performance of these seven subsets will be averaged.

Nine machine learning models were evaluated in this study utilizing the Confusion Matrix and the Receiver Operating Characteristic Curve (ROC). The Confusion Matrix contains information on misclassification. In this study, the author used a confusion matrix to assess the performance of the classifier. Known as the confusion matrix, this approach is widely used to measure accuracy in the context of data mining or decision support systems concepts. Table 5 details the Confusion Matrix [16].

**TABEL 5.** Confusion Matrix

Classification		Predicted Class	
		Class=Yes	Class=No
Observed Class	Class=Yes	True Positive-TP	False Negative-FN
	Class=No	False Positive-FP	True Negative-TN

To determine the accuracy of nine machine learning algorithms in detecting patterns, it is necessary to perform accuracy tests on their predictions. The accuracy of model predictions is measured by the accuracy formula [17], which is as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \quad (1)$$

In addition, measurements of sensitivity and specificity, which are statistical measures of the performance of binary classifications, were obtained in order to assess the accuracy of the model's predictions. In contrast, specificity measures the proportion of 'true positives' that are correctly recognized, whereas sensitivity measures the proportion of 'true negatives' that are correctly detected. The following are the sensitivity formula and specificity values [16].

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \quad (3)$$

The Receiver Operating Characteristic Curve (ROC) is used to graphically examine the outcomes of nine machine learning algorithms in this study (ROC Curve). The receiver operating characteristic (ROC) graph depicts the connection between the observed class and the anticipated class. Calculating the area under the ROC curve is a method of determining the accuracy of the ROC classification. Table 6 summarizes the accuracy criteria for diagnostic tests utilizing AUC [18].

**TABEL 6.** AUC Criteria

AUC	Interpretation
0.90 – 1.00	Excellent Classification
0.80 – 0.90	Good Classification
0.70 – 0.80	Fair Classification
0.60 – 0.70	Poor Classification
0.50 – 0.60	Failure

In our research, nine algorithms were implemented, each of which included a set of parameters that needed to be satisfied. Each of these parameters has a different value from the others. These parameter values are utilized in the model's training process.



## RESULTS

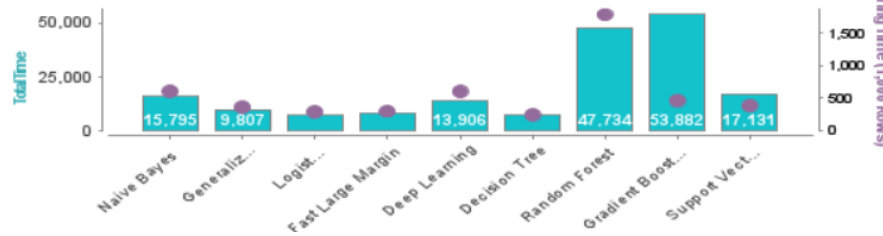
### Finding The Best Model

We performed nine machine learning algorithms on our process data set, which included a total of 14 features, for our process data set. The performance and of each of the nine algorithms is described in detail in Table 7. The length of time required to conduct pattern recognition describe in Figure 3. The accuracy, sensitivity, and specificity scores are used to assess the overall performance of the system. The algorithm that will perform the best for our issue domain will be chosen based on its overall performance and suitability. The algorithm with the greatest performance giver would be selected as the best appropriate algorithm. Performance analysis revealed that Logistic Regression had higher accuracy, sensitivity, and specificity scores than any other model evaluated. The other characteristics of the generalized linear model, on the other hand, were not favorable. As a result, taking everything into consideration, the Logistic Regression technique was used to choose the model with the greatest performance.

**TABEL 7.** Classifier Performance Evaluation

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Classification Error	Gains	Total Time	Training Time (1,000 Rows)
Naive Bayes	87.9	79.2	98.8	12.1%	118.0	16 s	255 ms
Generalized Linear Model	86.2	75.3	100	13.8%	112.0	10 s	854 ms
Logistic Regression	100	100	100	0.0%	164.0	7 s	462 ms
Fast Large Margin	100	100	100	0.0%	164.0	8 s	176 ms
Deep Learning	91.2	84.3	100	8.8%	128.0	14 s	896 ms
Decision Tree	87.3	80.3	96.2	12.7%	116.0	7 s	190 ms
Random Forest	87.9	78.4	100	12.1%	118.0	48 s	315 ms
Gradient Boosted Trees	89.0	80.3	100	11.0%	122.0	54 s	1 s
Support Vector Machine	100	100	100	0.0%	164.0	17 s	171 ms

**Runtimes (ms)**

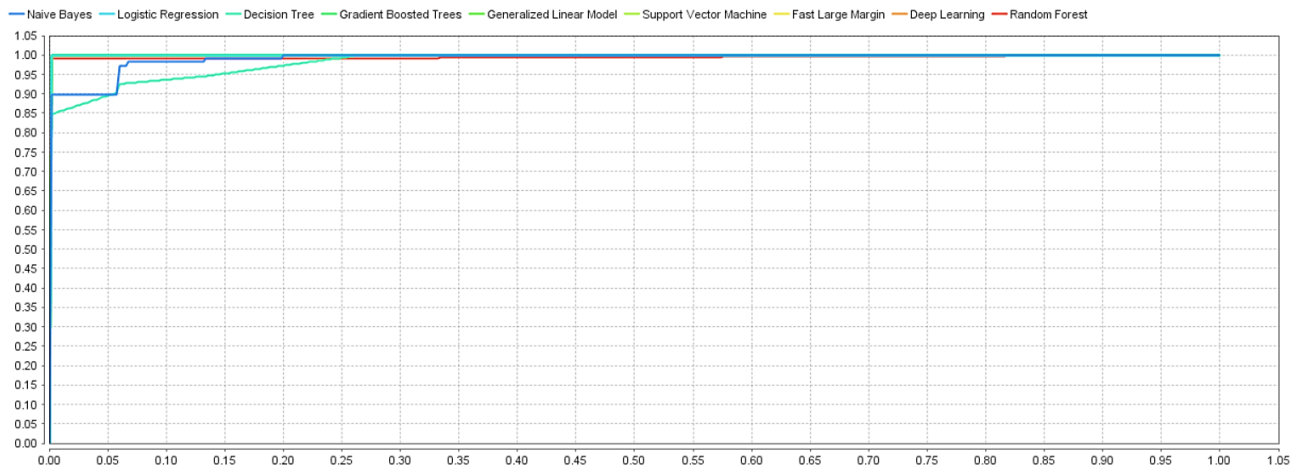


**FIGURE 3.** Runtimes Comparison

At every given classification threshold, the model's performance is shown in the form of a ROC curve (receiver operating characteristic curve). The term "Area under the ROC Curve" refers to this area. Thus, AUC is a two-dimensional measurement of the complete two-dimensional area under the entire ROC curve from (0,0) to (1,1). The AUC scores and ROC curve comparison for all model of the experiment are shown in Table 8 and Fig. 4. As a consequence of these findings, it is known that the Logistic Regression model is effective in recognizing patterns. The AUC value indicates that this model given excellent classification. The Logistic Regression model was selected because to its best performance, best gain, and fastest total time.

**TABEL 8.** AUC Scores Comparison

Model	AUC Score
Naive Bayes	0.992
Generalized Linear Model	1
Logistic Regression	1
Fast Large Margin	1
Deep Learning	1
Decision Tree	0.984
Random Forest	0.995
Gradient Boosted Trees	1
Support Vector Machine	1



**FIGURE 4.** ROC Comparison

## Experimental Evaluation

After selecting the best model, the researcher randomly splits the data into training and testing groups 10 times, ensuring that the same experiment produces distinct data groups. The training and testing data were divided into three categories: 70% training data (30% testing data), 80% training data (20% testing data), and 90% training data (10% testing data). The Logistic Regression model's validation is 10-fold cross-validation. Figure 5 illustrates the rapidminer process design. Table 9 contains the results of the performance evaluation for the optimum parameter values for both training and testing data. Overall, all experimental results showed 100% average accuracy.

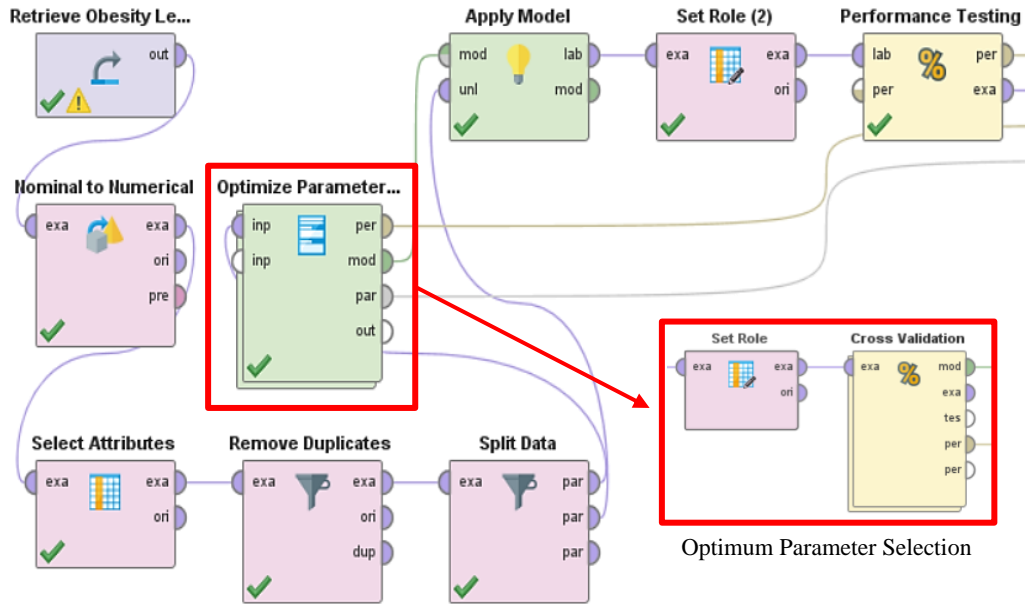


FIGURE 5. RapidMiner Process Design

TABEL 9. Accuracy for All Experiments

Ex p.	Number of Lambdas	70% Training	30% Testing	Number of Lambdas	80% Training	20% Testing	Number of Lambdas	90% Training	10% Testing
1	50	100	100	20	100	100	20	100	100
2	80	100	100	50	100	100	50	100	100
3	0	100	100	0	100	100	0	100	100
4	60	100	100	30	100	100	30	100	100
5	90	100	100	60	100	100	60	100	100
6	10	100	100	10	100	100	10	100	100
7	70	100	100	40	100	100	40	100	100
8	100	100	100	70	100	100	70	100	100
9	20	100	100	90	100	100	90	100	100
10	40	100	100	80	100	100	100	100	100

### Research Limitation

It is important to note that this research has various limitations, one of which is that the data utilized for this investigation must meet the characteristics listed in Table 1. Besides from that, this discovery requires scientific confirmation from professionals in several fields, particularly the medical area. Furthermore, as of right now, this research has only analyzed 14 characteristics for the classifications algorithm. This study focuses exclusively on two types of obesity, Normal Weight and Obesity Type I. For further research, additional analysis may be conducted to predict other levels of obesity (Insufficient Weight, Overweight Level I, Overweight Level II, Obesity Type II and Obesity Type III).

## CONCLUSION

Obesity level early diagnosis is a big challenge for the health care industry. In our research, we developed a technique that is highly accurate in predicting obesity level. RapidMiner was used to preprocess the data. We eliminated three features using the feature reduction technique. In the Obesity Type I dataset, we utilized 14 input variables (Age, CH2O, FAF, Family History, FAVC, FCVC, Gender, Height, NCP, SCC, SMOKE, TUE, and Weight) and one output variable (target). We evaluated nine different machine learning algorithms to predict the type of obesity using the Naive Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and SVM. We analyzed performance using a variety of metrics.. To get the best accuracy rates possible for diagnosing obesity, we used data mining techniques to achieve this aim. When it comes to uncovering information, data mining is a highly effective tool. Based on the experimental findings and comparisons to past approaches, it can be concluded that Logistic Regression delivers more accurate results than the previous experiment. This conclusion is based on the Accuracy (100%), Sensitivity (100%), Specificity (100%), AUC (1) values, as well as the overall time necessary for data training. According to the research findings, the Logistics Regression model is a feasible classification system to consider. This study demonstrates that it is feasible to predict whether or not a person is classified as Obesity Type I by using the Logistics Regression model.

## ACKNOWLEDGMENTS

This research is supported by research funding from Research and Technology Transfer Office (RTTO) Bina Nusantara University.

## REFERENCES

1. OMS. Organización Mundial de la Salud. *Obesidad y sobrepeso* (2016).
2. H. M. Gutiérrez, *Diez problemas de la población de jalisco: Una perspectiva sociodemográfica (Primera Edición ed.* (Dirección de Publicaciones del Gobierno de Jalisco, Guadalajara, México, 2010).
3. G. M. Hernández, *Prevalencia de sobrepeso y obesidad, y factores de riesgo, en niños de 7-12 años, en una escuela pública de Cartagena septiembre - octubre de 2010* (Universidad Nacional de Colombia, Bogotá – Colombia, 2011).
4. World Health Organization, see <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>.
5. T. M. Dugan, S. Mukhopadhyay, A. Carroll and S. Downs, S. “Machine learning techniques for prediction of early childhood obesity,” *Appl. Clin. Inform.* **6**(3), 503-520 (2015).
6. K. Jindal, N. Baliyan, P. S. Rana, “Obesity prediction using ensemble machine learning approaches,” in *Proceedings of the 5<sup>th</sup> ICACNI 2017*, **2**, pp. 355–362 (2018).
7. B. Singh and H. Tawfik, “Machine learning approach for the early prediction of the risk of overweight and obesity in young people,” *Int. Conf. Comput. Sci.* **12140**, pp. 523–535 (2020).
8. M. J. Gerl, C. Klose, M. A. Surma, C. Fernandez, O. Melander, S. Männistö, K. Borodulin, A. S. Havulinna, V. Salomaa, E. Ikonen, C. V. Cannistraci and K. Simons, K. Machine learning of human plasma lipidomes for obesity estimation in a large population cohort,” *PLoS Biol.* **17**(10), (2019).
9. C. Davila-Payan, M. DeGuzman, K. Johnson, N. Serban and J. Swann, “Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data,” *Prevent. Chronic Dis.* **12**, E32-E32 (2015).
10. S. Manna and A.M. Jewkes, “Understanding early childhood obesity risks: An empirical study using fuzzy signatures,” in *Proceedings of the IEEE International Conference on Fuzzy Systems* (IEEE Xplore Press, Beijing, China, 2014), pp.1333-1339.
11. F. M. Palechor, and A. de la Hoz Manotas, *UCI Machine Learning Repository: Estimation of obesity levels based on eating habits and physical condition Data Set*, UCI Machine Learning Repository 2021, see <https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+> (2019, August 27).
12. M. North, *Data Mining for the Masses, Third Edition: With Implementations in RapidMiner and R 3<sup>rd</sup>* (CreateSpace Independent Publishing Platform, South California, 2018).

13. E. De-La-Hoz-Correa, F. Mendoza Palechor, A. De-La-Hoz-Manotas, R. Morales Ortega, and A. B. S. Hernández, "Obesity level estimation software based on decision trees," *J. Comp. Science*, **15**(1), 67-77 (2019).
14. F. M. Palechor and A. de la Hoz Manotas, A, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico" *Data in Brief*, **104344** (2019).
15. H. Benhar, A. Idri, J. Fernández-Alemán, "Data preprocessing for decision making in medical informatics: potential and analysis," in *the Proceeding of the World Conference on Information Systems and Technologies*, pp. 1208–1218 (2018).
16. A. A. Abdillah and Suwarno, "Diagnosis of Diabetes Using Support Vector Machines with Radial Basis Function Kernels," *Int. J. of Tech.*, **5**, pp. 849-858 (2016).
17. F. Ferdowsy, K. S. A. Rahi, M. I. Jabiullah, and M. T. Habib, "A machine learning approach for obesity risk prediction," *Curr. Res. Behav. Sci.*, **2**, p. 100053 (2021).
18. Gorunescu, F. *Data mining: concepts and techniques* (Springer, Germany, 2011).