Convergence of Mathematics and Statistics
Volume 01, Issue 1, March 2026, pages 46-53

DOI: https://doi.org/xx.xxxxx/xxxxxxxxx-xx
E-ISSN : xxxx-xxxxx

Open Access - Research Article

Application of Machine Learning Algorithms for Obesity Prediction and Identification of Influencing Factors Using SHAP

Wahyuni[1] *, **Sri Sulastri**[1]

[1]Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya (60111), Indonesia

Article Info

Article history:

Received: November 16th, 2025
Revised: November 25th, 2025
Accepted: December 27th, 2025
Published: January 18th, 2026

Published by:

ARCSTICS PUBLISHING

ABSTRACT

Obesity has become a serious global health problem due to lifestyle changes from traditional lifestyles to sedentary lifestyles characterized by low physical activity and high-calorie, low-fiber diets. This condition is exacerbated by various factors, such as gender, family history, physical activity, and calorie intake, which collectively increase the risk of obesity in adolescents and adults. Therefore, early obesity prediction efforts are needed to support more effective preventive decision-making. This study aims to develop an obesity prediction model using Decision Tree, Random Forest, CatBoost, and Artificial Neural Network (ANN) algorithms to predict obesity probability and find the best method from the four algorithms. As a result of this analysis, obesity classes were predicted with success rates of 77.00%, 85.00%, 86.00%, and 84.00%, respectively. CatBoost was the most successful method for this dataset and classified obesity with a success rate of 86.00%. Analysis using SHAP values revealed that the features Food Intake Between Meals and Consumption of Fast Food had the greatest influence in increasing the probability of overweight/obesity. Conversely, features such as Physical Exercise and Height contributed negatively to the probability of obesity.

Keywords : *Obesity, decision tree, random forest, catboost, artificial neural network, SHAP.*

***Corresponding Author:** Email: wahyuniw2679@gmail.com*

1. Introduction

Modernization has had a negative impact on people's lifestyles, one of which is the shift from a traditional lifestyle to a sedentary lifestyle. This shift is characterized by low physical activity and changes in diet that tend to be high in energy intake (fat, protein, carbohydrates) and low in fiber [1]. This condition contributes to an increase in obesity cases, which can be experienced not only by adults but also by adolescents [2], [3], [4]. Obesity has become a serious concern due to its negative impact on health, including the risk of cardiovascular disease, type 2 diabetes, hypertension, and various other metabolic disorders. Given the magnitude of the negative impact caused by obesity, researchers have been motivated to develop algorithms that can predict obesity levels [5]. Therefore, early prediction of obesity is needed to support more effective preventive decision-making.

In an effort to predict obesity, several studies have utilized machine learning and deep learning algorithms. Machine learning generally refers to the process of learning from data to make predictions or group data [6], [7]. According to Janiesch, in many applications, deep learning models are able to outperform machine learning models in terms of accuracy and prediction performance [8].

Based on these issues, this study aims to utilize Decision Tree, Random Forest, CatBoost, and Artificial Neural Network (ANN) algorithms to predict obesity. The Decision Tree and Random Forest algorithms have the advantage of identifying significant predictor variables and providing transparent interpretations of the main factors that influence obesity. CatBoost is able to handle categorical variables effectively and speed up the model training process,

making it efficient for obesity datasets that contain many categorical variables. MARS is a nonparametric multivariate regression approach developed by Friedman to address high-dimensional data problems [9], [10]. Meanwhile, Artificial Neural Network (ANN) has the ability to learn complex and non-linear data patterns, enabling the recognition of deep relationships between social and physical activity factors and obesity risk. In addition, this study will also identify the features that most influence the prediction model formed using SHAP (SHapley Additive exPlanations).

2. Method Details

This study will evaluate the performance of four methods, namely Random Forest, Decision Tree, CatBoost, and ANN, based on accuracy, precision, and recall values to determine which method is more effective in classifying obesity levels.

2.1. Random Forest

Random Forest is a predictive data mining technique and also an ensemble machine learning method. The main concept of the ensemble method is that a group of 'weak learners' (trees) come together to form 'strong learners' (random forests). Random Forest is a combination of several decision trees that are combined to obtain accurate predictions [11], [12], [13]

2.2. Simultaneous Test

A decision tree is a predictive model in machine learning that is used to make decisions based on various rules. This model is similar to a decision tree structure, where each node in the tree represents the test results and each leaf represents the predicted value [14].

2.3. CatBoost

CatBoost CatBoost is a machine learning algorithm that is part of the Gradient Boosted Decision Trees (GBDT) family, which falls under the umbrella of ensemble learning. CatBoost is an algorithm that is open to the public for further development in the field of Supervised ML, bringing two innovations: Ordered Target Statistics and Ordered Boosting [15], [16], [17].

$$\frac{\sum_{i=1}^{N} w_i \log\left(\frac{e^{a_{it_i}}}{\Sigma_{j=0}^{M} e^{aij}}\right)}{\sum_{i=1}^{N} w_i}, t \in \{0,\ldots,M-1\} \quad (1)$$

Each boosting iteration in CatBoost forms a tree. At each terminal node in the newly created tree, CatBoost stores M predictions for each M classes. where t_i is the label value in the i^{th} data point of the training data input. ai is the result of applying the model to the i^{th} data point. N is the total number of data points. w is the weight for the i^{th} data point and, by default, is 1. In multi-class classification problems, conversion to probability values is performed using the sigmoid function based on the "RawFormula" value of the CatBoost output. Next, the data will be classified into the class with the highest probability value [18].

2.4. Artificial Neural Network

Artificial Neural Networks (ANN) are popular and effective models used in problem solving and machine learning [19]. ANNs consist of simple processing units called neurons, which are distributed in parallel and have the inherent ability to store and retrieve relevant experimental information. Artificial neurons or processing elements are the basic units that form the foundation of the system [20].

2.5. Research Data

The data used in this study is obesity data sourced from Kaggle and accessible through the Obesity Dataset. The resulting dataset has 14 variables needed to determine obesity. The Obesity Dataset was obtained via the internet using a questionnaire administered to a total of 1610 people living in Türkiye. The distribution of data from the features in the obesity dataset is provided in Appendix Table 1. The data consists of a total of 1,610 individuals. Among them,

898 are women and 712 are men. The youngest participant in this dataset is 18 years old, and the oldest participant is 54 years old.

3. Results and Discussion

3.1. Descriptive Statistics

EDA is performed by displaying visualizations of the dataset in the form of box plots and histograms to find outliers in the dataset and determine the distribution of data in the dataset, especially in the target class.

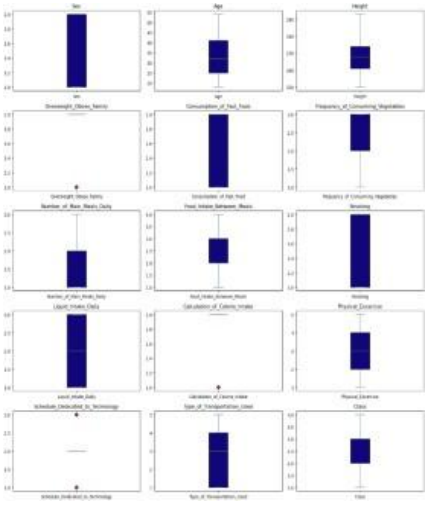


Figure 1. Box Plot of Data Distribution

Figure 1 shows that variables such as "Sex," "Consumption of Fast Food," and 'Smoking' have a dominant categorical distribution pattern. Numerical variables such as "Age," "Height," and "Physical Exercise" have a varied range of values, with several outliers in "Overweight Obese Family" and "Calculation of Calorie Intake." In addition, "Schedule Dedicated to Technology" shows little data variation, dominated by a single category.

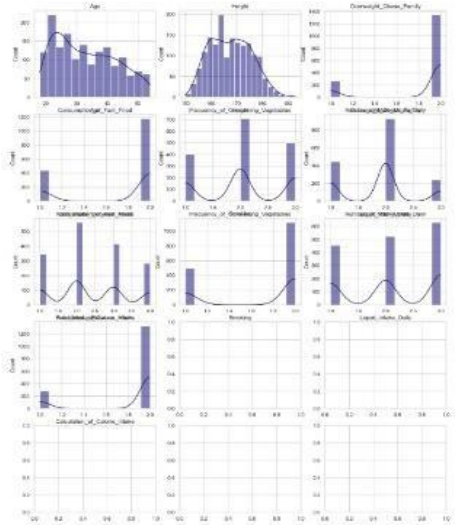


Figure 2. Histogram Data

Figure 2 shows that some variables, such as Calculation of Calorie Intake, have a right-skewed distribution with data concentrated at low values. Variables such as Height have a near-normal distribution. A multimodal distribution pattern is seen in Frequency of Consuming Vegetables and Food Intake Between Meals, indicating several groups in the data. Overweight Obese Family and Smoking show the dominance of certain categories, while Number of Main Meals_Daily has a bimodal pattern, reflecting two different consumption patterns.

3.2. Preprocessing

Data Cleansing

No data cleansing was performed because the data obtained was clean, but there were still outliers.

Data Transform

Attribute names were not changed in order to maintain data authenticity, although changing attribute names is permitted if desired.

Splitting Data

In this process, we divided the data into training data and testing data. In this study, the researcher used a ratio of 80% training data and 20% testing data.

```
X = df_efb.drop(columns=['Class', 'AgeCat'])
Y = df_efb['Class']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
```

Figure 3. Programming

3.3. Model Evaluation

Decision Tree Algorithm

The classification results for the Decision Tree model are presented in this section, with the corresponding confusion matrix given in Table 1. The accuracy rates for the underweight, normal, overweight, and obese classes were calculated using this matrix, and it was found that the model achieved a classification success rate of 77.00%, as shown in Figure 3.

Table 1. Confusion Matrix DT

Decision Tree		Predicted Class			
	(4x4)	Underwe	Normal	Over	Obesity
Actual Class	Underweight	12	1	0	0
	Normal	3	105	15	2
	Overweight	0	18	92	21
	Obesity	0	3	11	39

Table 1 shows a 4x4 confusion matrix that illustrates the model's performance in classifying four categories: Underweight, Normal, Overweight, and Obesity. The main diagonal (12, 105, 92, 39) indicates the number of correct predictions for each category. The Normal class has the highest number of correct predictions (105), while the Underweight class has the lowest number of correct predictions (12), but with a smaller classification error compared to other classes. Several significant errors are seen in the Overweight class, where 18 samples were misclassified as normal.

Random Forest Algorithm

The classification results for the Random Forest model are presented in this section, with the corresponding confusion matrix given in Table 1. The accuracy rates for the underweight, normal, overweight, and obese classes were calculated using this matrix, and it was found that the model achieved a classification success rate of 85.00%, as shown in Figure 3.

Table 2. Confusion Matrix RF

Random Forest		Predicted Class			
	(4x4)	Underwe	Normal	Over	Obesity
Actual Class	Underweight	13	0	0	0
	Normal	0	116	9	0
	Overweight	0	14	102	15
	Obesity	0	1	9	43

Table 2 shows a 4x4 confusion matrix that illustrates the model's performance in classifying four categories: Underweight, Normal, Overweight, and Obesity. The main diagonal (13, 116, 102, 43) indicates the number of correct predictions for each category. The Normal class has the highest number of correct predictions (116), while the Underweight class has the lowest number of correct predictions (13), but without any classification errors. Several significant errors are seen in the Overweight class, where 14 samples are misclassified as Normal and 15 samples are misclassified as Obesity. Overall, the Random Forest model performs better than the Decision Tree model, as it handles overfitting more effectively.

ANN Algorithm

The classification results for the ANN model are presented in this section, with the corresponding confusion matrix given in Table 3. The accuracy rates for the underweight, normal, overweight, and obese classes were calculated using this matrix, and it was found that the model achieved a classification success rate of 84.00%, as shown in Figure 3.

Table 3. Confusion Matrix ANN

ANN		Predicted Class			
	(4x4)	Underwe	Normal	Over	Obesity
Actual Class	Underweight	1	6	3	5
	Normal	5	52	49	26
	Overweight	6	49	45	18
	Obesity	2	23	24	8

Table 3 shows the 4x4 confusion matrix above, which illustrates the model's performance in classifying four classes: Underweight, Normal, Overweight, and Obesity. The main diagonal (1, 52, 45, 8) shows the number of correct predictions for each class. The Normal class has the highest number of correct predictions (52), while the Underweight class has the lowest number of correct predictions (1), with many classification errors. Significant errors are seen in the Overweight and Obesity classes, where many samples are misclassified as Normal and vice versa. Overall, the ANN performance in this matrix shows that the model has difficulty distinguishing between closely related classes, such as Normal and Overweight.

CatBoost Algorithm

The classification results for the Decision Tree model are presented in this section, with the corresponding confusion matrix given in Table 1. The accuracy rates of the underweight, normal, overweight, and obese classes were calculated using this matrix, and it was found that the model achieved a classification success rate of 77,00%.

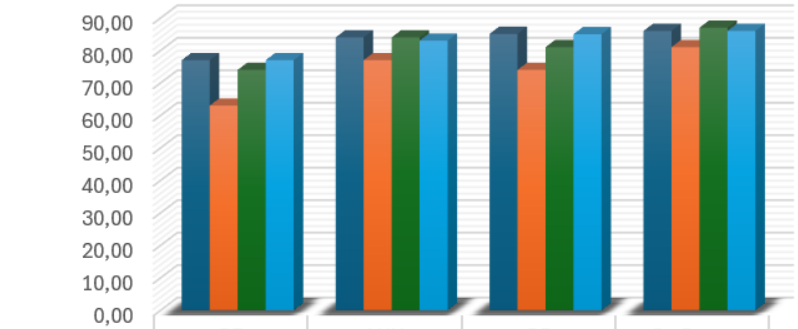
Table 4. Confusion Matrix CatBoost

CatBoost		Predicted Class			
	(4x4)	Underwe	Normal	Over	Obesity
Actual Class	Underweight	13	0	0	0
	Normal	1	113	11	0
	Overweight	0	16	104	11
	Obesity	0	1	6	46

Table 4 shows a 4x4 confusion matrix that illustrates the model's performance in classifying four categories: Underweight, Normal, Overweight, and Obesity. The main diagonal (13, 113, 104, 46) indicates the number of correct predictions for each category. The Normal class has the highest number of correct predictions (113), while the Underweight class has the lowest number of correct predictions (13), but without any classification errors. Several significant errors are seen in the Overweight class, where 16 samples were misclassified as Normal and 11 samples were misclassified as Obesity.

3.4. Predictive Classification Model Performance

The accuracy, precision, recall, and F1 scores of each model were obtained from the analysis results, and the results and graphs for all models are shown in Figure 1.



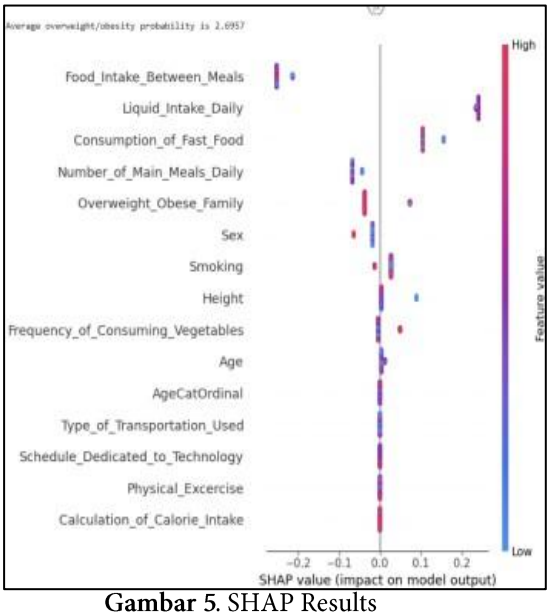
	DT	ANN	RF	CatBoost
Accuracy (%)	77,00	84,00	85,09	86,00
Precision(%)	63,00	77,00	74,00	81,00
Recall (%)	74,00	84,00	81,00	87,00
F1 Score(%)	77,00	83,00	85,00	86,00

Figure 4. Model Accuracy Results

From Figure 4, it can be concluded that the CatBoost model has the highest classification success rate (86.00%), while the Decision Tree model has the lowest (77.00%). Similarly, the CatBoost model shows the highest metric value outside of classification success, while the Decision Tree model shows the lowest value. These results illustrate that choosing the right machine learning model greatly affects prediction performance. The success of the CatBoost model can be explained by its ability to handle unbalanced data, its sensitivity to hyperparameter settings, and its efficient boosting mechanism.

3.5. Feature Contribution Analysis Using SHAP

To understand the contribution of features to the prediction of overweight/obesity probability, SHAP (SHapley Additive exPlanations) analysis was used. The figure below shows the SHAP summary plot, where each feature is sorted based on its impact on the model.



Gambar 5. SHAP Results

From Figure 5, it can be seen that features such as Food Intake Between Meals and Consumption of Fast Food have the greatest influence on overweight/obesity predictions. High values for these features (shown in red) tend to significantly increase the probability of overweight/obesity, as seen from the dominance of positive SHAP values.

Conversely, features such as Physical Exercise and Height show a negative contribution, where high values on these features reduce the probability of overweight/obesity. In addition, the Liquid Intake Daily feature shows an interesting pattern, where high values have a negative impact on the probability of overweight/obesity, while low values make a positive contribution, albeit on a smaller scale.

4. Conclusion

In this study, classification was implemented for obesity using the Decision Tree, Random Forest, CatBoost, and ANN algorithms. Based on the results of the analysis obtained from data processing, the CatBoost algorithm had the highest classification success rate (86.00%), while the Decision Tree model had the lowest (77.00%). The features of Food Intake Between Meals and Consumption of Fast Food had the greatest influence in increasing the probability of overweight/obesity, as indicated by positive SHAP values. Although CatBoost's accuracy was quite high, there was still potential for improving the model's performance, one of which was by selecting more selective features to reduce noise in the data. In addition, testing other classification algorithms could be carried out to explore the possibility of higher accuracy. This research is expected to contribute to the development of a more accurate and interpretative machine learning-based obesity prediction system.

Limitations

There are no specific limitations to report from this study.

Author Contributions Statement

All authors have made significant contributions to this research and fulfill the authorship criteria based on the CRediT (Contributor Roles Taxonomy) guidelines. The specific contributions of each author are as follows :

- Wahyuni
  Conceptualization, Methodology Development, Formal Analysis, Writing – Original Draft
- Sri Sulastri
  Data Curation, Investigation, Visualization, Writing – Review & Editing

Conflict of Interest Statement

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

References

[1] U. L. Wijaya, B. Widjanarko, and R. Indraswari, "Faktor-Faktor yang Berhubungan dengan Perilaku Makan Berisiko Gizi Lebih pada Remaja SMA di Kota Semarang," *Jurnal Kesehatan Masyarakat*, vol. 8, no. 3, 2020.

[2] A. Mutia, J. Jumiyati, and K. Kusdalinah, "POLA MAKAN DAN AKTIVITAS FISIK TERHADAP KEJADIAN OBESITAS REMAJA PADA MASA PANDEMI COVID-19," *Journal of Nutrition College*, vol. 11, no. 1, 2022, doi: 10.14710/jnc.v11i1.32070.

[3] A. R. Kansra, S. Lakkunarajah, and M. S. Jay, "Childhood and Adolescent Obesity: A Review," 2021. doi: 10.3389/fped.2020.581461.

[4] A. Nadjamuddin and H. H. Dukalang, "Hubungan Body Image Dan Pola Makan Mahasiswi Ma'had Al-Jami'ah IAIN Sultan Amai Gorontalo," *Tamaddun Journal of Islamic Studies*, vol. 1, no. 1, pp. 80–88, 2022.

[5] L. Setiyani, A. N. Indahsari, and R. Roestam, "Analisis Prediksi Level Obesitas Menggunakan Perbandingan Algoritma Machine Learning dan Deep Learning," *JTERA (Jurnal Teknologi Rekayasa)*, vol. 8, no. 1, 2023, doi: 10.31544/jtera.v8.i1.2022.139-146.

[6] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, "A guide to machine learning for biologists," 2022. doi: 10.1038/s41580-021-00407-0.

[7] M. Isnaini and S. Sulastri, "Analisis Dampak Covid-19 Terhadap Tingkat Inflasi di Indonesia," *Jurnal MSA (Matematika dan Statistika serta Aplikasinya)*, vol. 11, no. 2, pp. 57–63, 2023.

[8] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, 2021, doi: 10.1007/s12525-021-00475-2.

[9] H. H. Dukalang, "Modeling Infant Mortality Rate with Multivariate Adaptive Regression Spline Approach," *Journal of Applied Informatics and Computing*, vol. 1, no. 2, pp. 19–28, 2017.

[10] S. Sulastri, B. W. Otok, and A. Choiruddin, "Multivariate adaptive bivariate regression splines (MABRS) binary response for modeling stroke and hypertension in RSKD Dadi City Makassar," *Commun. Math. Biol. Neurosci.*, vol. 2025, p. Article-ID, 2025.
[11] S. Haykin, *Neural Networks and Learning Machines*, vol. 3. 2008. doi: 978-0131471399.
[12] A. Sekulić, M. Kilibarda, G. B. M. Heuvelink, M. Nikolić, and B. Bajat, "Random forest spatial interpolation," *Remote Sens (Basel)*, vol. 12, no. 10, 2020, doi: 10.3390/rs12101687.
[13] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, no. 1, 2020, doi: 10.1177/1536867X20909688.
[14] A. B. Alpiansah and Y. Ramdhani, "Optimasi Fitur dengan Forward Selection pada Estimasi Tingkat Obesitas menggunakan Random Forest," *SISTEMASI*, vol. 12, no. 3, 2023, doi: 10.32520/stmsi.v12i3.3125.
[15] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00369-8.
[16] W. Chang, X. Wang, J. Yang, and T. Qin, "An Improved CatBoost-Based Classification Model for Ecological Suitability of Blueberries," *Sensors*, vol. 23, no. 4, 2023, doi: 10.3390/s23041811.
[17] H. Kim, S. Park, H. J. Park, H. G. Son, and S. Kim, "Solar Radiation Forecasting Based on the Hybrid CNN-CatBoost Model," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3243252.
[18] A. N. A. Aldania, A. M. Soleh, and K. A. Notodiputro, "A Comparative Study of CatBoost and Double Random Forest for Multi-class Classification," *Jurnal RESTI*, vol. 7, no. 1, 2023, doi: 10.29207/resti.v7i1.4766.
[19] G. N. Silva *et al.*, "Artificial neural networks compared with bayesian generalized linear regression for leaf rust resistance prediction in arabica coffee," *Pesqui Agropecu Bras*, vol. 52, no. 3, 2017, doi: 10.1590/s0100-204x2017000300009.
[20] I. A. Ozkan, M. Koklu, and I. U. Sert, "Diagnosis of urinary tract infection based on artificial intelligence methods," *Comput Methods Programs Biomed*, vol. 166, 2018, doi: 10.1016/j.cmpb.2018.10.007.