Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4053

Submitted : 11 June 2024
Accepted : 28 June 2024
Published : 2 July 2024

Prediction of Obesity Categories Based on Physical Activity Using Machine Learning Algorithms

Muhammad Iqbal[1)*], **Lisnawanty**[2)], **Weiskhy Steven Dharmawan**[3)], **Rendi Septian**[4)]
[1)*2)3)]Universitas Bina Sarana Informatika
[4)]Universitas Nusa Mandiri
[1)*]iqbal.mdq@bsi.ac.id, [2)]lisnawanty.lsy@bsi.ac.id, [3)]weiskhy.wvn@bsi.ac.id, [4]14002348@nusamandiri.ac.id

ABSTRACT

Obesity is a global health issue with rising prevalence, marked by excessive fat accumulation that poses health risks. Contributing factors include poor eating habits, lack of physical activity, and genetics, which elevate the risk of chronic diseases like type 2 diabetes, heart disease, stroke, and cancer. This study examines an obesity dataset with seven variables: Age, Gender, Height, Weight, BMI, Physical Activity Level, and Obesity Category. The analysis reveals strong correlations between Body Weight, BMI, and the Obesity Category, while Body Height shows a moderate negative correlation. Various machine learning algorithms were tested, including XGBoost, AdaBoost, Gradient Boosting, and Extra Trees Classification. XGBoost emerged as the top performer, achieving the highest accuracy (0.9961) and an almost perfect AUC (0.9992), making it highly effective for obesity prediction. The study's significance lies in its ability to elucidate the key factors contributing to obesity and their interactions. By recognizing the strong links between Body Weight, BMI, and Obesity Category, healthcare professionals can craft more targeted interventions. Furthermore, the successful application of advanced machine learning algorithms underscores the potential for technology to enhance predictive accuracy and support healthcare decision-making. The findings highlight XGBoost's superior performance, demonstrating its value in predicting obesity and aiding in early diagnosis and prevention strategies. This research emphasizes the critical role of data and technology in tackling obesity and improving public health outcomes.

Keywords: Data Mining; Machine Learning; Obesity; Prediction; XGBoost

1. INTRODUCTION

Obesity is a global health problem that has continued to increase in prevalence in the last few decades (Tandiono & Sanjaya, 2023). According to the World Health Organization (WHO), obesity is defined as excessive and abnormal accumulation of fat that can harm health (Rahmawati et al., 2024) (Lior-sadaka & Greenberg, n.d.). Factors that contribute to obesity include unhealthy diet, lack of physical activity, and genetic factors [4]. The impact of obesity is significant, including a high risk of various chronic diseases such as type 2 diabetes, heart disease, stroke, and some types of cancer (Fitriani & Bahri, 2024) (Wildan et al., 2024).

Physical activity has an important role in managing body weight and preventing obesity. Various studies have shown that increasing physical activity can help with weight loss and improve overall health. However, challenges in measuring and predicting an individual's physical activity levels often become obstacles in efforts to prevent and manage obesity (Wulandari et al., 2024).

In recent years, developments in technology and data science have opened new opportunities to analyze and predict obesity categories based on physical activity (Rahmawati et al., 2024). Machine learning algorithms, which are part of artificial intelligence, have shown great potential in processing large amounts of data and discovering patterns that cannot be identified by conventional methods. By using this algorithm, we can develop a more accurate and effective prediction model to categorize obesity levels based on individual physical activity (Wildan et al., 2024).

This research aims to develop a prediction model for obesity categories based on physical activity data using a machine learning algorithm. By utilizing a public dataset for obesity prediction, this dataset collects various attributes including demographics, living habits, and individual health indicators, with the aim of supporting predictions of obesity prevalence rates (MRSIMPLE, 2020) using 6 input variables (predictors) and 1 output variable (prediction)

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4053

Submitted : 11 June 2024
Accepted : 28 June 2024
Published : 2 July 2024

and categorized into four classifications, namely normal, obese, overweight, and underweight.

The ensemble learning model is used to obtain hidden patterns from 1000 available datasets and is expected to provide more accurate predictions, so that it can be used as a tool in efforts to prevent and manage obesity (Nasser & Abu-naser, 2023). It is also hoped that this study will contribute to understanding the relationship between physical activity and obesity and open up opportunities for further research in this area.

This research will begin with the collection and preprocessing of physical activity data and obesity category data. Subsequently, ensemble learning models such as extreme gradient boosting, adaboost, gradient boosting, and extra tree classification will be applied and evaluated to determine the most effective model for predicting obesity categories. Based on initial evaluations, the XGBoost model was selected as the preferred model as it achieved an accuracy of 99% in classifying obesity categories. The results of this research are expected to provide new insights and support efforts to prevent and manage obesity in the community, leveraging the high-performance XGBoost model for obesity prediction.

2. LITERATURE REVIEW

Research on obesity cases by (Admojo & Rismayanti, 2024) concluded that utilizing the predictive capabilities of machine learning to explore the determinants of obesity in populations in Mexico, Peru and Colombia, used the Decision Tree algorithm which was strengthened by 5-fold cross-validation. Comprehensive analysis of lifestyle and physical condition data from 2111 individuals produced accuracy, precision, recall and F1 scores that peaked in the third and fifth folds. The accuracy value of this study reached 95% and these findings confirm the importance of eating habits and physical activity as significant predictors of obesity levels. However, this study recognizes the limitations of self-reported data and the need for a broader dataset that includes a more diverse range of variables.

Studies on obesity using machine learning methods show that of the three algorithms compared (K-Nearest Neighbor, Decision Tree, and Naïve Bayes), the Decision Tree algorithm provides the highest accuracy in predicting the risk of obesity in the adult population, with an accuracy of 93.6%. The biggest risk factors for obesity identified by the Decision Tree model are body weight, gender, age, family history of obesity, and consumption of high-calorie foods. Decision Trees are also recognized for their high interpretability and optimal performance in providing accurate results (Rahmawati et al., 2024). Meanwhile, the application of the Random Forest model to predict the risk of obesity and cardiovascular disease shows that this model is very effective, with an accuracy of 97.23%, in handling complex health data (Nasser & Abu-naser, 2023).

Research using the K-Means clustering algorithm to group individuals based on obesity levels as a prevention effort, with datasets from Kaggle to classify the differences between underweight and overweight individuals. Four main clusters were identified: Cluster 0 consisted of women aged 45–60 years with relatively thin to normal weight; Cluster 1 consists of men aged over 40 years and 55-60 years who are overweight or obese; Cluster 2 consists of the majority of women aged 15-70 years, especially women aged 55-60 years with normal weight; and Cluster 3 consists of many underweight individuals aged 10-45 years, especially those aged 20-25 years. This research shows that men have a higher risk of obesity than women, emphasizing the importance of adopting a healthy lifestyle as a preventive measure (Wildan et al., 2024).

3. METHOD

A technique used in preprocessing categorical data for machine learning algorithms is one-hot encoding. The context of predicting obesity categories based on physical activity in this research, one-hot encoding helps transform categorical variables into a numerical format that can be fed into machine learning models. When inputting physical activity data into machine learning algorithms, categorical data is often converted into a binary matrix using one-hot encoding. This ensures that the algorithm can process the data without assuming any ordinal relationships between the categories.

An ensemble learning method of gradient boosting algorithms in this research is using XGBoost (Extreme Gradient Boosting). The advantages of using XGBoost for this research are high performance, handling complex relationship, and scalability. According to (Septiana Rizky et al., 2022), XGBoost is a sophisticated gradient tree boosting method that can efficiently solve large problems with very limited computing resources. XGBoost is capable of being a powerful, efficient, and useful solution to solve various classification problems. So that this research is more focused and structured based on stages from beginning to end, the research methodology is explained in Figure 1.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4053

Submitted : 11 June 2024
Accepted : 28 June 2024
Published : 2 July 2024

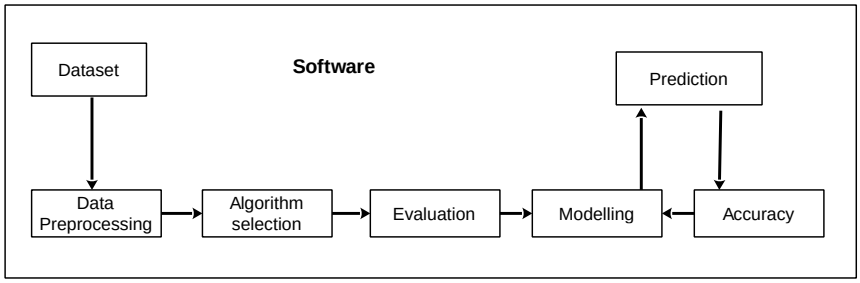

Fig. 1 Research Methodology

This methodology outlines the steps taken in the machine learning process, starting from dataset collection to evaluating model accuracy:

Dataset

The first stage is collecting data about obesity cases. This dataset consists of seven variables, namely age, gender, height, weight, mass load index (BMI), physical activity level and obesity category which consists of four classifications, namely norm, under obesity, obesity and over obesity.

Data Preprocessing

To overcome problems such as missing data, duplication, and inconsistencies. At this stage, the data can also be converted into a more suitable format for further analysis. After preprocessing, the dataset is divided into two parts: training data and testing data. Training data is used to train the model, while testing data is used to evaluate model performance, at this stage the ratio of training data and testing data is 80:20.

Algorithm Selection

In this stage, a machine learning algorithm is selected and applied to the training data. Algorithm testing in this research uses ensemble learning such as extreme gradient boosting, adabost, gradient boosting and extra tree classification.

Evaluation

The trained model is then evaluated using testing data. This evaluation aims to measure model performance based on certain metrics, such as accuracy, precision, recall, and F1-score.

Model

The model that has been evaluated and optimized is then saved for use in predicting new data. This model is the final result of the training and evaluation process.

Prediction

The saved model is used to make predictions on new data. These prediction results are then analyzed to ensure that the model works well in real situations.

Accuracy

The accuracy of the model is measured to assess the reliability and effectiveness of the predictions produced. This stage is important to ensure that the model can provide consistent and reliable results.

By following this structured methodology, research is expected to produce machine learning models that are accurate and reliable, and provide valuable insights for data-based decision making.

4. RESULT

This research analyzes the obesity dataset using seven main variables: Age, Gender, Height, Weight, Body Mass Index (BMI), Physical Activity Level, and Obesity Category based on 1000 datasets consisting of four categories, in

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4053

Submitted : 11 June 2024
Accepted : 28 June 2024
Published : 2 July 2024

detail table 1 contents from the research dataset.

Table 1
Research Dataset

	Age	Gender	Height	Weight	BMI	Physical Activity Level	Obesity Category
0	56	Male	173.575	71.982	23.891	4	Normal Weight
1	69	Male	164.127	89.959	33.395	2	Obese
2	46	Female	168.072	72.930	25.817	4	Overweight
3	32	Male	168.459	84.886	29.912	3	Overweight
4	60	Male	183.568	69.038	20.487	3	Normal Weight
	...	...	...	...	...	...	...
995	18	Male	155.588	64.103	26.480	4	Overweight
996	35	Male	165.076	97.639	35.830	1	Obese
997	49	Female	156.570	78.804	32.146	1	Obese
998	64	Male	164.192	57.978	21.505	4	Normal Weight
999	66	Female	178.537	74.962	23.517	1	Normal Weight

Source : [8]

Here is a brief explanation of each attribute in the table:

Age: Measured in years.
Gender: Categorized as "Male" or "Female".
Height: The height of the individual measured in centimeters.
Weight: The weight of the individual measured in kilograms.
BMI (Body Mass Index): An index calculated from a person's weight and height. Formula: BMI = weight (kg) / (height (m))². It is used to categorize weight status.
Physical Activity Level: The individual's level of physical activity, often measured on a scale from 1 to 5, where higher numbers indicate higher levels of physical activity.
Obesity Category: The category that classifies individuals based on their BMI. Common categories include "Normal Weight", "Overweight", "Obese", and "Underweight".

These attributes provide important information that can be used for further analysis regarding factors related to obesity. Physical activity level represents the amount of daily physical activity and is a method of estimating each person's total energy expenditure. Physical activity level is an important factor in predicting obesity type, as it directly affects a person's energy expenditure and overall health. In the context of machine learning algorithms to predict obesity categories based on physical activity, physical activity level is often measured and classified to provide meaningful input to the model (Xaverius Widiantoro & Sinaga, 2020). There are five level of physical activity level: 1) Sedentary (little to no physical activity); 2) Lightly Active (Low levels of physical activity); 3) Moderately Active (regular exercise included in the daily routine); 4) Active (high levels of physical activity, including frequent and intense exercise); 5) Very Active (Very high levels of physical activity, typically involving professional or semi-professional athletes, or highly active lifestyles).

Attribut Correlations

The next step is to examine the correlation between the seven variables to determine if there is a relationship between each pair of variables and the strength of that correlation. Below are the results of the correlation test shown in the figure 2.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4053

Submitted : 11 June 2024
Accepted : 28 June 2024
Published : 2 July 2024

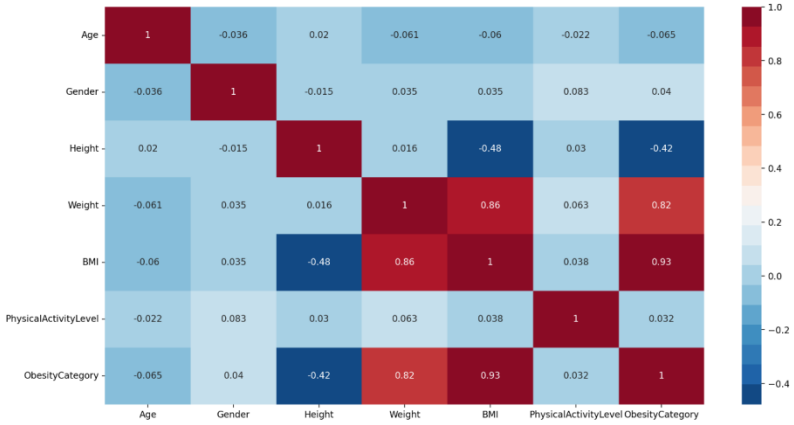

Fig. 2 Attribut Correlations

In Figure 2, the correlation matrix displays the relationships between various variables in the dataset, namely Age, Gender, Height, Weight, BMI, Physical Activity Level, and Obesity Category. Here is an explanation of the relationships between these variables:

Age: Has a weak negative correlation with all other variables, indicating that age does not have a strong relationship with the other variables in this dataset.
Gender: Shows very weak or almost no correlation with other variables, suggesting that gender does not significantly affect other variables such as height, weight, BMI, physical activity level, and obesity category.
Height: Has a moderate negative correlation with BMI (-0.48) and Obesity Category (-0.42), indicating that lower height tends to be associated with higher BMI and obesity category.
Weight: Has a very strong positive correlation with BMI (0.86) and Obesity Category (0.82), indicating that higher weight is correlated with higher BMI and obesity category.
BMI: Has a very strong positive correlation with Obesity Category (0.93), indicating that higher BMI is closely related to a higher obesity category.
Physical Activity Level: Shows a very weak correlation with other variables, indicating that physical activity level does not significantly affect other variables in this dataset.
Obesity Category: Has a strong positive correlation with weight (0.82) and a very strong correlation with BMI (0.93), indicating that higher obesity categories are correlated with higher weight and BMI.

Overall, the correlation matrix in Figure 2 shows that the most related variables are BMI, weight, and obesity category. Height has a moderate negative relationship with BMI and obesity category, while age, gender, and physical activity level have very weak correlations with other variables.

Preprocessing

In the preprocessing stage, data collected from the obesity dataset is processed to ensure quality and readiness for further analysis. Preprocessing steps include data cleaning by addressing missing values and removing duplications, as well as normalizing and standardizing numerical data such as Height, Weight, and BMI. Additionally, categorical variables such as Gender and Obesity Category are converted to numeric format using coding techniques such as one-hot encoding or label encoding (Mailo & Lazuardi, 2019). In the preprocessing stage we will see the distribution of data from the research dataset in Figure 3

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4053

Submitted : 11 June 2024
Accepted : 28 June 2024
Published : 2 July 2024

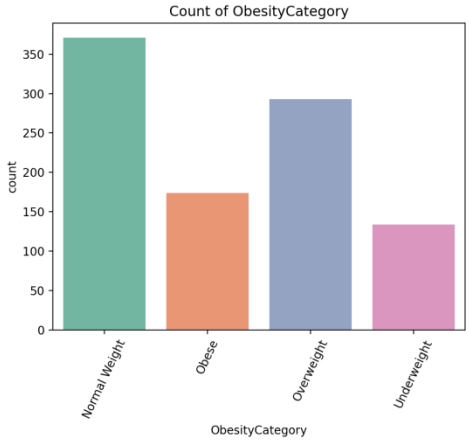

Fig. 3 Dataset Category Distribution

In Figure 3, distribution of obesity categories in the dataset. The “Normal Weight” category had the highest number with over 375 entries, indicating that the majority of individuals in the dataset were of normal weight. The "Overweight" category followed with about 300 entries, followed by "Obese" with about 180 entries. The "Underweight" category had the lowest number with 145 entries.

Splitting Dataset

The dataset is then divided into training (training data) and testing (testing data) subsets to ensure that the training data is used to train the model, while the testing data is used to evaluate the model's performance. In this study, the comparison between training data and testing data was 80:20. In detail in figure 4.

Fig.4 Splitting Dataset

From Figure 4 it can be seen that the total dataset used in this research is 1000 data and divided into two, where 800 data are training data and 200 data are testing data.

Algorithmic Modeling

In this study, we used several machine learning algorithms to model data and predict obesity categories. The algorithms used include Extreme Gradient Boosting (XGBoost), AdaBoost, Gradient Boosting, and Extra Tree Classification.

Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient boosting algorithm optimized for speed and performance. It is often used in data competitions due to its ability to handle large and complex datasets with high efficiency (Elina et al., 2022). XGBoost equation:

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4053

Submitted : 11 June 2024
Accepted : 28 June 2024
Published : 2 July 2024

$$L(\emptyset) = \sum_{i=1}^{n} L(yi, yi) + \sum_{k=1}^{K} \Omega(fk) \qquad (1)$$

Where L is the training loss function, and Ω is the regularization function, and θ are the corresponding model parameters.

AdaBoost

AdaBoost, or Adaptive Boosting, is a boosting technique that combines several weak models to form a strong model. It works by iteratively adjusting data weights to correct errors from previous models (Andreyestha & Subekti, 2020). Adaboost Equation Formula:

$$Ct = \sum_{i=1}^{n} wi.I(yi \neq ht(xi)) \qquad (2)$$

Gradient Boosting

Gradient Boosting is a boosting technique that builds a model incrementally from a weak model optimized for prediction error. It is very effective for regression and classification problems (Cendani & Wibowo, 2022). Gradient Boosting Equation:

$$F^{(m)}(x) = F^{m-1}(x) + v.h^{m}(x) \qquad (3)$$

where $h^{(m)}(x)$ is the new model fitted to the residuals and v is the learning rate.

Extra Tree Classification

Extra Tree Classification is a variant of the Random Forest algorithm that uses additional decision trees. It builds multiple decision trees from different subsets of data and combines the results to improve prediction accuracy (Ishaq et al., 2021). Extra Tree Classification equation:

$$\hat{y} = mode(\{hb(x)\}^{B} b = 1 \qquad (4)$$

where $hb(x)$ is the prediction of the bth tree and B is the total number of trees.

By using a combination of these algorithms, we can compare the performance and accuracy of each model in predicting obesity categories. This analysis helps in determining the best model that can be used for more accurate and reliable predictions.

DISCUSSIONS

At this stage we can find findings obtained from the calculation results of the proposed model to conclude that the model is the best in analyzing obesity cases based on the dataset. It is hoped that these findings can become recommendations for the best model in predicting obesity cases.

Evaluation

In the evaluation stage, we assess the performance of machine learning models that have been built using various algorithms such as Extreme Gradient Boosting (XGBoost), AdaBoost, Gradient Boosting, and Extra Tree Classification. Evaluation is carried out to measure how well each model predicts obesity categories based on relevant evaluation metrics, such as accuracy, precision, recall, and F1-score with the formula:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (5)$$

$$Sensitivity = Recal = TPRate = \frac{TP}{TP+FN} \qquad (6)$$

$$Precission = \frac{TP}{TP+FP} \qquad (7)$$

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4053

Submitted : 11 June 2024
Accepted : 28 June 2024
Published : 2 July 2024

$$F1 = \frac{2\ x\ Recall\ x\ Preccission}{Recall + Precission} \quad (8)$$

$$AUC = \frac{TPRate + FPRate}{2} \quad (9)$$

Where:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

The results of the calculations using the above formulation are in table 2.

Table 2
Model Evaluations Result

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	TT (Sec)
Extreme Gradient Boosting	0.9961	0.9992	0.9961	0.9663	0.9962	0.9946	10.600
Ada Boost Classifier	0.9949	0.0000	0.9949	0.9951	0.9948	0.9928	0.2240
Gradient Boosting Classifier	0.9949	0.0000	0.9949	0.9951	0.9948	0.9928	0.8790
Extra Trees Classifier	0.9743	0.9990	0.9743	0.9752	0.9742	0.9639	0.2300

Based on the evaluation results of the ensemble learning model in table 2, it can be concluded that the Extreme Gradient Boosting (XGBoost) model shows the best performance with high accuracy (0.9961) and almost perfect AUC (0.9992), indicating very precise prediction capabilities. This model also has very good recall, precision, and F1 scores (0.9961, 0.9963, and 0.9962). AdaBoost Classifier and Gradient Boosting Classifier also showed good performance with accuracy of 0.9949 each, but with unmeasured AUC (0.0000). These two models have very high precision and recall values, even though the training time is faster than XGBoost. Extra Trees Classifier shows slightly lower performance with an accuracy of 0.9743, but is still reliable with an AUC of 0.9990. Overall, XGBoost has the best performance followed by AdaBoost and Gradient Boosting and Extra Trees Classifier has the smallest performance.

Feature Importance

Feature importance is a technique in machine learning that is used to identify and measure the contribution of each feature to model predictions. In the case of obesity classification, feature importance helps us understand which features are most influential in determining whether someone is included in the obese category or not (Nawawi et al., 2024).

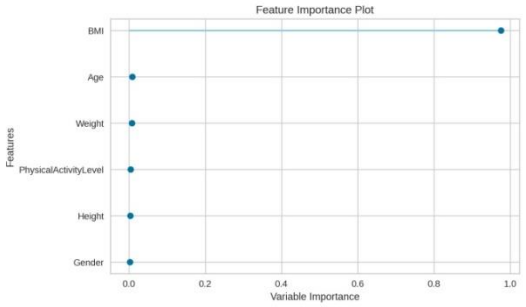

Fig. 5 Feature Importance Plot

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4053

Submitted : 11 June 2024
Accepted : 28 June 2024
Published : 2 July 2024

In figure 4 The main conclusion from these results is that BMI is the main determining factor in obesity classification. Other features such as age, weight, physical activity level, height, and gender did not contribute significantly to this model. Therefore, in the context of obesity classification, the main focus should be on measuring and monitoring BMI.

Confussion Matrix

Confusion matrix is a very useful evaluation tool in machine learning, especially for classification, which provides a detailed description of the performance of the prediction model in this research. The selected extreme gradient boosting (XGBoost) model. In the context of obesity classification, the confusion matrix allows us to clearly see how many correct predictions (true positives and true negatives) and errors the model made (false positives and false negatives) (Hozairi et al., 2021). By using a confusion matrix, we can evaluate the extent to which our model is successful in identifying individuals who are truly obese and those who are not, as well as understand areas where the model may be having difficulty, thereby allowing further refinement and increased prediction accuracy. The following is the confusion matrix produced by the model in Figure 6.

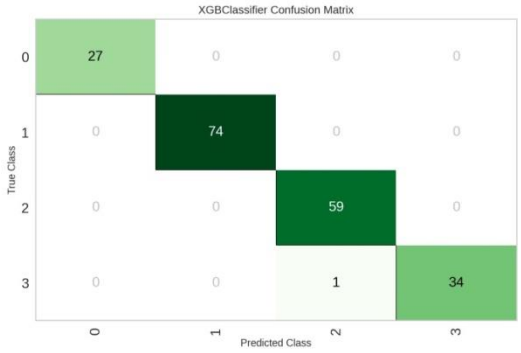

Fig. 6 Confussion Matrx Result

Based on the confusion matrix in Figure 7 generated by the XGBoostClassifier model in the classification of obesity, the model demonstrates excellent performance. The predictions are almost entirely accurate for each class, with 27 correct predictions for class 0 (normal), 74 correct predictions for class 1 (obesity), 59 correct predictions for class 2 (overweight), and 34 correct predictions for class 3 (underweight). There is only one misclassification where one instance of class 3 was predicted as class 2. This indicates that the model has a high classification capability, with a very low error rate, signifying outstanding performance in accurately identifying the correct obesity classes.

5. CONCLUSION

This research analyzes the obesity dataset which consists of seven main variables, namely Age, Gender, Height, Body Weight, Body Mass Index (BMI), Physical Activity Level, and Obesity Category, with the Obesity Category variable as the target variable. From the results of the analysis and evaluation, it was found that the variables Body Weight and BMI have a very strong correlation with the Obesity Category, indicating that individuals with a higher body weight and BMI tend to fall into a higher obesity category. Height had a moderate negative correlation with BMI and Obesity Category, indicating that individuals with lower height tended to have a higher BMI and fall into a higher obesity category. To model the data and predict obesity categories, several machine learning algorithms have been used, including Extreme Gradient Boosting (XGBoost), AdaBoost, Gradient Boosting, and Extra Trees Classification. Based on performance evaluation, the algorithm proposed and selected in this research is Extreme Gradient Boosting (XGBoost), because it shows the highest accuracy (0.9961) and almost perfect AUC (0.9992). XGBoost also shows excellent recall, precision, and F1 score values, making it the best choice for predicting obesity categories in this dataset. By using XGBoost, obesity category predictions can be made very accurately and efficiently, providing a strong basis for further interventions in the treatment and prevention of obesity.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4053

Submitted : 11 June 2024
Accepted : 28 June 2024
Published : 2 July 2024

6. REFERENCES

Admojo, F. T., & Rismayanti, N. (2024). Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation : A Study on Eating Habits and Physical Conditions. *Indonesian Journal of Data and Science*, *5*(1), 37–44.

Andreyestha, & Subekti, A. (2020). Analisis Sentiment pada Ulasan Film Dengan Optimasi Ensemble Learning. *JURNAL INFORMATIKA*, *7*(1), 5–8.

Cendani, L. M., & Wibowo, A. (2022). Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes. *Jurnal Masyarakat Informatika*, *13*(1), 33–44.

Elina, S., Yulianti, H., Soesanto, O., & Sukmawaty, Y. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *Journal of Mathematics: Theory and Applications*, *4*(1), 21–26.

Fitriani, D. N., & Bahri, S. (2024). Prediksi tingkat obesitas menggunakan neural network : pendekatan klasifikasi biner. *PARAMETER JURNAL MATEMATIKA, STATISTIKA DAN TERAPANNYA*, *03*(01), 85–92.

Hozairi, H., Anwari, A., & Alim, S. (2021). Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes. *Network Engineering Research Operation*, *6*(2), 133. https://doi.org/10.21107/nero.v6i2.237

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques. *IEEE Access*, *9*, 39707–39716. https://doi.org/10.1109/ACCESS.2021.3064084

Lior-sadaka, I., & Greenberg, D. (n.d.). *Assessing Screening Methods and Machine Learning for Predicting Childhood Overweight and Obesity : A Population-Based Study Background : Methods : Results : Conclusions :* 1–18.

Mailo, F. F., & Lazuardi, L. (2019). Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia. *Journal of Information Systems for Public Health*, *4*(1).

MRSIMPLE. (2020). *Obesity Prediction*. https://doi.org/10.34740/kaggle/dsv/7479144

Nasser, M. S. A., & Abu-naser, S. S. (2023). Predictive Modeling of Obesity and Cardiovascular Disease Risk : A Random Forest Approach. *International Journal of Academic Information Systems Research (IJAISR)*, *7*(12), 26–38.

Nawawi, H. M., Hikmah, A. B., Mustopa, A., & Wijaya, G. (2024). Model Klasifikasi Machine Learning untuk Prediksi Ketepatan Penempatan Karir. *Jurnal Saintekom (Sains, Teknologi, Komputer Dan Manajemen)*, *14*(1), 13–25.

Rahmawati, M., Lestari, A. F., & Hardani, S. (2024). Phyton-Based Machine Learning Algorithm to Predict Obesity Risk Factors in Adult Populations. *Paradigma*, *26*(1), 51–57.

Septiana Rizky, P., Haiban Hirzi, R., Hidayaturrohman, U., Hamzanwadi Selong Jl TGKH Muhammad Zainuddin Abdul Madjid Pancor, U., & Timur, L. (2022). Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang. In *J Statistika* (Vol. 15, Issue 2). www.unipasby.ac.id

Tandiono, S. M., & Sanjaya, S. A. (2023). Machine Learning Approach of Obesity Level Classification: A Systematic Literature Review of Methods and Factors. *G-Tech : Jurnal Teknologi Terapan*, *8*(1), 196–208.

Wildan, A., Burhansyah, H. A., & Ferdiansyah, C. (2024). Prediction of Obesity Classification Using K-Means Clustering. *Journal of Dinda Data Science, Information Technology, and Data Analytics*, *4*(1), 14–22.

Wulandari, A., Mulya, A., & Dermawan, T. (2024). Application of Artificial Neural Network , K-Nearest Neighbor and Naive Bayes Algorithms for Classification of Obesity Risk Cardiovascular Disease. *IJATIS: Indonesian Journal of Applied Technology and Innovation Science*, *1*(February), 9–15.

Xaverius Widiantoro, F., & Sinaga, F. (2020). A Concept Analysis: Physical Activity Level. In *Malahayati International Journal of Nursing and Health Science* (Vol. 03, Issue 1). https://ejurnalmalahayati.ac.id/index.php/nursing/article/view/2413/pdf

* Corresponding author