Prosiding Seminar Nasional Indonesia
Vol. 4 No. 1 2026, hal., 38-41
https://sociohum.net/index.php/PROSIDINGNASIOANAL
e-ISSN: 3026-5169

OBESITY RISK CLASSIFICATION BASED ON GRADIENT BOOSTING USING MEDICAL DATA

Achmad Lukman Prayogidianto
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Email: achmadlukmanp@gmail.com

Abstract

Obesity is a growing global health issue that significantly contributes to the development of chronic diseases such as diabetes, cardiovascular disorders, and metabolic syndromes, making early detection essential to support preventive healthcare strategies. This study aims to implement the Gradient Boosting algorithm to classify obesity risk based on medical data obtained from the UCI Machine Learning Repository, which includes information on eating habits, physical activity, and individual characteristics. The research methodology involves several stages, including data preprocessing, transformation, normalization, class mapping, and dataset splitting into training and testing sets with a ratio of 70:30. The Gradient Boosting model is constructed using multiple decision trees in an iterative manner to improve classification performance, categorizing individuals into obese and non-obese classes. Model evaluation is conducted using accuracy, precision, recall, and F1-score metrics. The experimental results indicate that the model achieves an accuracy of over 90%, with a relatively small gap between training and testing performance, demonstrating good generalization capability without overfitting. These findings confirm that Gradient Boosting is an effective approach for obesity risk classification and has strong potential to support intelligent healthcare systems in enabling data-driven decision-making for early prevention and treatment.

Keywords: Gradient Boosting, Classification, Obesity Risk, Medical Data, Machine Learning

INTRODUCTION

Obesity is one of the major public health challenges whose prevalence continues to increase globally. This condition not only affects an individual's physical health but also reduces overall quality of life. Obesity is closely associated with various chronic diseases, including diabetes mellitus, cardiovascular diseases, and other metabolic disorders. Therefore, early identification of obesity risk is crucial to minimize long-term health impacts.

The rapid development of information technology has opened opportunities to utilize medical data as a source of valuable information in healthcare systems. Data such as eating habits, physical activity levels, and demographic characteristics can be analyzed to identify patterns related to obesity risk. In this context, machine learning emerges as a powerful tool that can process large amounts of data and generate predictive models based on historical patterns.

Classification techniques in machine learning are widely used to categorize data into specific groups based on their characteristics. One of the effective algorithms is Gradient Boosting, which is an ensemble learning method that combines multiple weak learners to produce a strong predictive model. This algorithm works iteratively by minimizing prediction errors from previous models, resulting in improved accuracy.

Based on these considerations, this study aims to apply the Gradient Boosting algorithm to classify obesity risk using medical data. The results are expected to contribute to the development of intelligent health information systems that can assist healthcare professionals in making more accurate and data-driven decisions.

METHOD

This study adopts a systematic approach consisting of data collection, preprocessing, modeling, and evaluation stages. The dataset used is obtained from the UCI Machine Learning Repository, containing 2,111 records and 17 attributes that describe eating habits, physical activities, and individual characteristics from several countries. The data preprocessing stage plays a crucial role in ensuring data quality before modeling, including data cleaning to handle missing values, transformation of categorical variables into numerical form using one-hot encoding, and normalization using the Min-Max Scaler to ensure all features are within the same range. Additionally, the original dataset, which consists of seven obesity categories, is simplified into two main classes, namely obese and non-obese, to focus the model on binary classification. After preprocessing, the dataset is divided into training data (70%) and testing data (30%), where the training data is used to build the model and the testing data is used for performance evaluation. The classification model is developed using the Gradient Boosting Classifier with parameters such as n_estimators = 100, learning_rate = 0.1, and max_depth = 4, where the algorithm builds multiple decision trees sequentially, with each new tree aiming to correct errors from previous ones. Model evaluation is conducted using several performance metrics, including accuracy, precision, recall, and F1-score, to provide a comprehensive assessment of the classification performance.

RESULTS AND DISCUSSION

The results of this study indicate that the Gradient Boosting model demonstrates strong performance in classifying obesity risk based on medical data. The model achieves an accuracy of 90.38% on testing data, with a relatively small difference compared to the training accuracy, indicating good generalization capability and no overfitting. Further evaluation using classification metrics shows balanced performance across both classes, where the precision reaches approximately 93% for the non-obese class and 88% for the obese class, while recall values are around 89% and 92%, respectively. The F1-score values for both classes are also consistent, approximately 0.91 and 0.90, indicating a good balance between precision and recall. In addition, the

confusion matrix analysis reveals that the model correctly classifies most instances, with 303 true negatives and 270 true positives, while the number of misclassifications remains relatively low, consisting of 37 false positives and 24 false negatives. These results demonstrate that the Gradient Boosting algorithm is effective in capturing complex patterns within the dataset through its iterative learning mechanism, where each subsequent model improves the errors of the previous one. Therefore, the model is considered reliable and suitable for medical data classification tasks, particularly in identifying obesity risk, and shows strong potential for implementation in healthcare systems to support early detection and decision-making processes.

CONCLUSION

This study concludes that the K-Means clustering method is effective in grouping regencies/cities in Indonesia based on waste management performance using variables such as waste generation, waste reduction, and waste handling. The results demonstrate that each cluster represents distinct characteristics, reflecting differences in the effectiveness of waste management across regions. The evaluation using Silhouette Score and Davies-Bouldin Index indicates that the clustering model produces a sufficiently good structure, confirming its reliability in identifying patterns within the data. In addition, the implementation of results through a Streamlit-based user interface enhances accessibility and facilitates user understanding of the clustering outcomes. Therefore, this study provides a data-driven approach that can support policymakers in identifying priority regions and formulating more targeted waste management strategies. Future research is recommended to incorporate additional variables, explore alternative clustering methods, and conduct further analysis to improve the accuracy and depth of insights related to waste management in Indonesia.

REFERENCES

[1] N. R. Aditya, Data Mining. Bandung, Indonesia: UNIKOM, 2015. [Online]. Available: https://repository.unikom.ac.id/47451/1/Pertemuan%203%20-%20Materi%20%5BDM%20-%202015%5D.pdf

[2] A. R. Arrahimi, M. K. Ihsan, D. Kartini, M. R. Faisal, and F. Indriani, "Teknik Bagging dan Boosting Pada Algoritma CART Untuk Klasifikasi Masa Studi Mahasiswa," Jurnal Sains dan Informatika, vol. 5, Jun. 2019. [Online]. Available: https://jsi.politala.ac.id/index.php/JSI/article/view/171/94

[3] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasfikasi pada Algoritma C4.5, Random Forest, SVM, dan Naive Bayes," Jurnal Media Informatika Budidarma, vol. 5, pp. 640–651, 2021. [Online]. Available: https://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/2937

[4] EKRUT, "Dataset Adalah: Pengertian, Tipe, Perbedaan dengan Database, dan 10 Web Penyedia," Sep. 28, 2022. [Online]. Available: https://www.ekrut.com/media/dataset-adalah

[5] Direktorat Pengendalian Penyakit Tidak Menular, Pedoman Umum Pengendalian Obesitas. Jakarta, Indonesia: Kementerian Kesehatan RI, 2015. [Online]. Available: https://extranet.who.int/ncdccs/Data/IDN_B11_Buku%20Obesitas-1.pdf

[6] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA, USA: Elsevier, 2011.

[7] F. P. Mendoza and A. de la 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, p. 104344, 2019, doi: 10.1016/j.dib.2019.104344.

[8] R. Arthana, "Mengenal Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan," Apr. 5, 2019. [Online]. Available: https://rey1024.medium.com/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8

[9] P. Pujiastuti, "Obesitas dan Penyakit Periodontal," Stomatognatic (J.K.G Unej), vol. 9, pp. 82–85,2012.[Online].Available:https://jurnal.unej.ac.id/index.php/STOMA/article/download/2112/1715

[10] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review Paper," IJCIT, vol. 5, pp. 75–82, Apr. 2020. [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/ijcit/article/view/7951/pdf

[11] R. Tibshirani, J. Friedman, and T. Hastie, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[12] I. F. Yulianti and P. R. Sihombing, "Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia," Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer, vol. 20, pp. 417–426, May 2021. [Online]. Available: https://journal.universitasbumigora.ac.id/index.php/matrik/article/view/1174/703

[13] Gregorius Airlangga, "Machine Learning-Based Obesity Classification: A Comparative Study Using Self-Reported Survey Data and Ensemble Learning Models," Jurnal Teknologi Informatika dan Komputer, vol. 11 no. 1, 2025.

[14] Shalini K., A.V.K. Shanthi, C. Shakila, N. Chamudeeswari, "Machine Learning Approaches For Obesity Level Classification," International Journal of Environmental Sciences, vol. 11, 2025.

[15] Achmad Rivai Syahputra, Rian Hidayat, Fathur Rismansyah, dkk., "Komparasi Algoritma Machine Learning (SVM, Random Forest, dan Regresi Logistik) untuk Prediksi Tingkat Obesitas," Jurnal Ilmiah Teknik Informatika dan Komunikasi, vol. 5 no. 3, 2025