Jurnal Informatika Ekonomi Bisnis

http://www.infeb.org

2025 Vol. 7 Iss. 1 Hal: 1-5 e-ISSN: 2714-8491

A Comparative Analysis of Machine Learning Models for Obesity Prediction

Gregorius Airlangga[1]✉

[1]Atma Jaya Catholic University of Indonesia

gregorius.airlangga@atmajaya.ac.id

Abstract

Obesity is a global health challenge with significant implications for public health systems and individual well-being. Predictive modeling using machine learning (ML) offers a powerful approach to identify individuals at risk of obesity and inform early intervention strategies. This study evaluates the performance of ten ML models, including Logistic Regression, Support Vector Machines, Decision Trees, K-Nearest Neighbors, Naive Bayes, Random Forest, Gradient Boosting, AdaBoost, XGBoost, and LightGBM, in predicting obesity using a publicly available dataset. A rigorous preprocessing pipeline, incorporating missing value handling, categorical encoding, normalization, and outlier detection, was applied to ensure data quality and compatibility with ML algorithms. Performance metrics such as accuracy, precision, recall, and F1-score were evaluated using 10-fold stratified cross-validation. Among the models, LightGBM demonstrated the highest test accuracy (99.19%) and F1-score (99.20%), outperforming Gradient Boosting and Random Forest, which also showed competitive results. The study highlights the superior predictive capabilities of ensemble methods while underscoring the trade-offs between model complexity and interpretability. Logistic Regression provided a strong baseline, demonstrating the importance of preprocessing, but was outperformed by advanced ensemble techniques. This research contributes to the growing field of ML-driven healthcare solutions, offering valuable insights into the strengths and limitations of various predictive models. The findings support the integration of advanced ML techniques in public health systems and pave the way for future research on hybrid and explainable models for obesity prediction and management.

Keywords: Obesity Prediction, Machine Learning Models, Ensemble Techniques, LightGBM, Healthcare Analytics.

1. Introduction

Obesity is recognized as a global health crisis with significant socioeconomic and clinical implications [1], [2], [3]. The rapid increase in obesity rates has spurred extensive research into its underlying causes, prevention strategies, and prediction mechanisms [4]. Predictive modeling in healthcare, particularly for conditions like obesity, has gained traction in recent years due to advancements in machine learning (ML) techniques [5]. These models have demonstrated the potential to transform obesity management by identifying at-risk individuals and informing timely interventions, thus contributing to reducing the overall disease burden [6], [7], [8].

The availability of datasets containing comprehensive information on behavioral, physical, and dietary habits has paved the way for data-driven insights into obesity prediction [9]. However, the complexity of obesity-related data, often characterized by missing values, mixed data types, and high dimensionality, presents challenges in developing robust and accurate predictive models [10], [11], [12]. Traditional statistical approaches, while insightful, often fall short of capturing the nuanced patterns and relationships inherent in such datasets [13]. This underscores the need for advanced ML models capable of handling these challenges and delivering reliable predictions [14]. The state of the art in obesity prediction has seen significant contributions from ensemble and hybrid ML techniques [15]. Studies have highlighted the efficacy of models such as Random Forest (RF), Gradient Boosting Machines (GBM), and Support Vector Machines (SVM) in achieving high predictive accuracy [16]. Additionally, boosting techniques, including XGBoost and LightGBM, have further pushed the boundaries of model performance by leveraging iterative improvements and feature importance mechanisms [17]. Despite these advancements, there remains a gap in systematically comparing a wide range of advanced ML models using standardized evaluation metrics, particularly in the context of obesity prediction [18]. Such comparisons are crucial for understanding the relative strengths and limitations of different algorithms, thereby enabling researchers and practitioners to make informed decisions [11], [19].

This study addresses the gap by conducting an extensive evaluation of ten ML models, including Logistic Regression, Decision Tree, K-Nearest Neighbors, Naive Bayes, RF, GBM, AdaBoost, SVM, XGBoost, and LightGBM, using a publicly available obesity dataset. The dataset captures a diverse array of features relevant to obesity prediction, including demographic, behavioral, and physiological variables. To ensure the reliability of the experimental results, a rigorous preprocessing pipeline was implemented. This included

Accepted: 27-01-2025 | Revision: 06-02-2025 | Publication: 31-03-2025 | doi: 10.37034/infeb.v7i1.1089

handling missing values, encoding categorical variables, and normalizing numerical features. The target variable, representing obesity classes, was encoded as an integer for compatibility with ML algorithms.

The experimental methodology incorporated a ten-fold stratified cross-validation approach to mitigate overfitting and ensure robust performance evaluation. Advanced scoring metrics, including accuracy, precision, recall, and F1 scores, were employed to provide a comprehensive assessment of model performance. The results of the study offer valuable insights into the comparative efficacy of different ML models, highlighting the trade-offs between model complexity, computational efficiency, and predictive accuracy. By systematically comparing the performance of state-of-the-art algorithms, this study lays the groundwork for future investigations into model optimization and real-world deployment in clinical and public health settings. The remainder of this article is structured as follows: The next section provides a Research Method section, it outlines the dataset, preprocessing steps, and experimental framework employed in this study. The Results section presents a comprehensive analysis of model performance, while the Discussion contextualizes these findings in relation to existing literature. Finally, the Conclusion highlights the key contributions of this work and proposes directions for future research.

2. Research Method

The study utilized a publicly available dataset on obesity prediction that included diverse attributes capturing demographic, behavioral, and physiological variables. The dataset comprised N samples and M features, with the target variable, Obesity, representing the classification labels for various obesity classes, the dataset can be downloaded from certain source [20]. The dataset included a mix of categorical and numerical features, requiring preprocessing steps to ensure its compatibility with machine learning algorithms. The preprocessing of the dataset was conducted systematically to address challenges such as missing values, class imbalance, and varying scales of numerical features. Missing values were addressed by adopting a complete case analysis approach, where instances with missing values were removed. This approach was chosen due to the small proportion of missing data (< 5%), ensuring minimal loss of information while maintaining the integrity of the dataset. Outliers were detected using the interquartile range (IQR) method, where values lying beyond Q1 - 1.5 x IQR or Q3 + 1.5 x IQR were considered outliers and excluded from the analysis.

Categorical features in the dataset were transformed into numerical representations using label encoding. For a categorical feature x_i, the transformation was defined as $\hat{x_i} = \{0,1,\dots,C-1\}$, where C denotes the number of unique categories. This method preserved ordinal relationships, if present, while facilitating the application of machine learning algorithms. Numerical features x_j were standardized using z-score normalization, defined as $z_j = \frac{x_j - \mu_j}{\sigma_j}$, where μ_j and σ_j represent the mean and standard deviation of feature j, respectively. This normalization ensured that all numerical features had a mean of zero and a standard deviation of one, thereby mitigating the influence of features with larger magnitudes on the model's performance. The target variable, *Obesity*, was encoded as an integer variable $y \in \{0,1,\dots,k-1\}$, where k represents the number of obesity classes. This encoding facilitated multi-class classification tasks and ensured compatibility with the machine learning models employed in the study.

The experimental framework involved splitting the dataset into training and testing subsets, where the training set comprised 80% of the data ($X_{\text{train}}, y_{\text{train}}$) and the testing set comprised 20% ($X_{\text{test}}, y_{\text{test}}$). The models were evaluated using a stratified 10-fold cross-validation approach. For each fold, the training data was further partitioned into 90% for training and 10% for validation, ensuring that the class distribution remained consistent across folds. The cross-validation process aimed to minimize overfitting and provide robust estimates of model performance. The machine learning algorithms evaluated in this study included Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, K-Nearest Neighbors, Naive Bayes, Gradient Boosting Machine, AdaBoost, XGBoost, and LightGBM. Each algorithm was trained to minimize a loss function $\mathcal{L}$, specific to its architecture. For instance, Logistic Regression minimized the cross-entropy loss, defined as Equation 1.

$$\mathcal{L}_{\text{lgsi}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log\left(\widehat{y_{i,k}}\right) \quad (1)$$

Where $y_{i,k}$ is the true label and $\widehat{y_{i,k}}$ is the predicted probability for class k. Tree-based methods such as Random Forest and Gradient Boosting minimized impurity measures like Gini index or entropy at each node, defined as Equation 2.

$$\mathcal{L}_{\text{gn}} = \sum_{k=1}^{K} p_k(1-p_k) \quad (2)$$

Where p_k is the proportion of samples belonging to class k at a given node. Evaluation metrics included accuracy, precision, recall, and F1-score, calculated as follows. Accuracy, denoted as Equation 3 represents the proportion of correctly classified samples.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (3)$$

Precision, defined as Equation 4, measures the proportion of positive predictions that are correct.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (4)$$

Recall, expressed as Equation 5, quantifies the model's ability to identify true positives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (5)$$

The F1-score, which balances precision and recall, is defined as Equation 6.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \quad (6)$$

To ensure a fair comparison across models, hyperparameters were tuned for each algorithm using grid search with cross-validation, optimizing the respective evaluation metric. For ensemble models such as Random Forest, the number of trees n and the maximum tree depth d were tuned to balance model complexity and performance. For boosting algorithms, learning rates η and the number of boosting rounds T were optimized to achieve convergence without overfitting. The results of the experiments were aggregated across folds to compute mean and standard deviation for each metric, providing a comprehensive view of each model's performance. The testing set was used to validate the best-performing models, ensuring that the results were generalizable to unseen data. This robust framework ensured that the study addressed critical challenges in obesity prediction, including handling imbalanced data, optimizing feature representation, and evaluating model performance comprehensively. By employing advanced machine learning techniques and rigorous evaluation metrics, this study contributes to the development of reliable predictive models for obesity classification.

3. Result and Discussion

As presented in Table 1, the results of this study present a detailed comparative evaluation of various machine learning models for obesity prediction. The models were assessed using multiple performance metrics, including training and testing accuracy, precision, recall, and F1-score, to provide a comprehensive understanding of their strengths and weaknesses. This section discusses the performance of each model, highlights key trends, and elaborates on their implications for obesity prediction.

Table 1. Software dan Hardware Supporting Table

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score
LightGBM	1.000000	0.991943	0.992098	0.991943	0.991952
Gradient Boosting	1.000000	0.984362	0.984813	0.984362	0.984332
Logistic Regression	0.981525	0.976319	0.977266	0.976319	0.976350
Random Forest	0.999474	0.970151	0.971050	0.970151	0.970184
Decision Tree	0.995842	0.959262	0.960194	0.959262	0.959284
SVM	0.977683	0.937946	0.940440	0.937946	0.938069
AdaBoost	0.906416	0.891051	0.911806	0.891051	0.894537
KNN	0.915996	0.887257	0.887326	0.887257	0.886268
Naive Bayes	0.713301	0.711495	0.710872	0.711495	0.675717

3.1. Model Performance Analysis

The LightGBM model achieved the highest overall performance, with a perfect training accuracy of 1.000 and a testing accuracy of 0.992. Its precision, recall, and F1-score were all 0.992, demonstrating both high sensitivity and specificity. These results highlight the robustness of LightGBM in generalizing unseen data while maintaining a strong ability to capture complex patterns in the dataset. The Gradient Boosting model followed closely with a testing accuracy of 0.984 and comparable precision (0.985), recall (0.984), and F1-score (0.984). Both boosting models exhibit exceptional capabilities in handling imbalanced datasets and leveraging iterative learning to optimize performance.

Logistic Regression performed surprisingly well for a linear model, achieving a testing accuracy of 0.976 alongside high precision (0.977), recall (0.976), and F1-score (0.976). This competitive performance suggests a degree of linear separability within the dataset, though the model's inability to capture non-linear relationships limited its ability to outperform advanced ensemble methods. Random Forest achieved a testing accuracy of 0.970 with precision, recall, and F1-score values close to 0.970, confirming its effectiveness in handling high-dimensional data. However, its reliance on ensemble averaging resulted in slightly lower performance compared to boosting algorithms.

Decision Tree, with a testing accuracy of 0.959 and F1-score of 0.959, demonstrated reasonable predictive capabilities but suffered from overfitting due to its reliance on a single tree structure. Support Vector Machine (SVM) achieved a testing accuracy of 0.938, with precision (0.940), recall (0.938), and F1-score (0.938). While SVM's kernel-based approach is effective in handling non-linear decision boundaries, its performance lagged ensemble methods, likely due to challenges in hyperparameter optimization.

AdaBoost displayed moderate performance, with a testing accuracy of 0.891, precision of 0.912, recall of 0.891, and F1-score of 0.895. Despite its ability to focus on misclassified samples, AdaBoost's iterative approach was less effective in this dataset, potentially due to noise or feature redundancy. K-Nearest Neighbors (KNN) achieved a testing accuracy of 0.887 and F1-score of 0.886, reflecting its sensitivity to noisy data and high-dimensional feature spaces. Naive Bayes recorded the lowest performance, with a testing accuracy of 0.711 and F1-score of 0.676, primarily due to its strong

independence assumption, which does not hold for interdependent features in the dataset.

3.2. Training and Generalization Performance

The training accuracy of LightGBM, Gradient Boosting, and Random Forest models approached 1.000, indicating their ability to learn intricate patterns in the training data. The high testing accuracy of these models suggests that overfitting was effectively mitigated through their regularization mechanisms. In contrast, Decision Tree exhibited a larger gap between training and testing accuracy, highlighting its susceptibility to overfitting without ensemble averaging. Logistic Regression's strong performance demonstrates its suitability for datasets with linearly separable features, though its simplicity limits its applicability to more complex patterns. The performance gap between SVM and ensemble models suggests that hyperparameter tuning and kernel selection play critical roles in maximizing SVM's potential for this dataset.

3.3. Evaluation Metrics Analysis

The use of multiple evaluation metrics provided a nuanced understanding of the models' performance. Precision and recall were particularly relevant for obesity prediction, as they emphasize the models' ability to correctly classify minority classes. For instance, LightGBM's precision of 0.992 indicates its effectiveness in minimizing false positives, while its recall of 0.992 demonstrates its capacity to identify true positives across all obesity classes. F1-score, as the harmonic mean of precision and recall, offered a balanced measure of model performance. The consistently high F1-scores of LightGBM (0.992), Gradient Boosting (0.984), and Logistic Regression (0.976) validate their reliability and robustness in handling multi-class obesity classification. These metrics are particularly valuable for applications where class imbalance can skew accuracy-based evaluations.

3.4. Implications for Obesity Prediction

The superior performance of boosting algorithms, particularly LightGBM and Gradient Boosting, underscores their suitability for complex, high-dimensional datasets. Their ability to iteratively refine predictions and optimize feature importance makes them ideal for addressing the challenges of obesity prediction. Logistic Regression's competitive performance highlights its potential as a baseline model for initial analysis, especially in scenarios where interpretability is a priority. The underperformance of Naive Bayes and KNN reflects the importance of selecting models that align with the dataset's characteristics. Naive Bayes' independence assumption and KNN's sensitivity to noise and dimensionality limited their effectiveness. These findings emphasize the need for careful model selection and the importance of leveraging ensemble methods for robust predictions.

4. Conclusion

This study examined machine learning techniques for obesity prediction using a public dataset, comparing ten models, including traditional and advanced ensemble methods. LightGBM and Gradient Boosting outperformed others, with LightGBM excelling in handling complex data and imbalances. Effective preprocessing, including handling missing values, encoding, and normalization, was crucial for model performance. While simpler models like Logistic Regression offer interpretability, ensemble methods better capture data complexities. The findings highlight machine learning's potential in healthcare, particularly for personalized interventions and public health monitoring. However, the dataset's limited diversity suggests future research should include broader populations. Additionally, improving ensemble model interpretability through SHAP or LIME could enhance clinical applicability.

References

[1] Anekwe, C. V., Jarrell, A. R., Townsend, M. J., Gaudier, G. I., Hiserodt, J. M., & Stanford, F. C. (2020). Socioeconomics of obesity. *Current Obesity Reports, 9*, 272–279. https://doi.org/10.1007/s13679-020-00398-7

[2] Buoncristiano, M., Williams, J., Simmonds, P., Nurk, E., Ahrens, W., Nardone, P., Rito, A. I., Rutter, H., Bergh, I. H., & Starc, G. (2021). Socioeconomic inequalities in overweight and obesity among 6- to 9-year-old children in 24 countries from the World Health Organization European region. *Obesity Reviews, 22*, e13213. https://doi.org/10.1111/obr.13213

[3] Ryan, D., Barquera, S., Barata Cavalcanti, O., & Ralston, J. (2021). The global pandemic of overweight and obesity: Addressing a twenty-first century multifactorial disease. In *Handbook of Global Health* (pp. 739–773). Springer. https://doi.org/10.1007/978-3-030-45009-0_39

[4] Goel, A., Reddy, S., & Goel, P. (2024). Causes, consequences, and preventive strategies for childhood obesity: A narrative review. *Cureus, 16*(7), e64985. https://doi.org/10.7759/cureus.64985

[5] Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks, 3*, 58–73. https://doi.org/10.1016/j.ijin.2022.05.002

[6] Sarma, S., Sockalingam, S., & Dash, S. (2021). Obesity as a multisystem disease: Trends in obesity rates and obesity-related complications. *Diabetes, Obesity and Metabolism, 23*, 3–16. https://doi.org/10.1111/dom.14290

[7] Kepper, M. M., Walsh-Bailey, C., Brownson, R. C., Kwan, B. M., Morrato, E. H., Garbutt, J., de Las Fuentes, L., Glasgow, R. E., Lopetegui, M. A., & Foraker, R. (2021). Development of a health information technology tool for behavior change to address obesity and prevent chronic disease among adolescents: Designing for dissemination and sustainment using the ORBIT model. *Frontiers in Digital Health, 3*, 648777. https://doi.org/10.3389/fdgth.2021.648777

[8] Hoffman, R. K., Donze, L. F., Agurs-Collins, T., Belay, B., Berrigan, D., Blanck, H. M., Brandau, A., Chue, A., Czajkowski, S., & Dillon, G. (2024). Adult obesity treatment and prevention: A trans-agency commentary on the research landscape, gaps, and future opportunities. *Obesity Reviews*, e13769. https://doi.org/10.1111/obr.13769

[9] Ayub, H., Khan, M.-A., Naqvi, S. S. A., Faseeh, M., Kim, J., Mehmood, A., & Kim, Y.-J. (2024). Unraveling the potential of attentive Bi-LSTM for accurate obesity prognosis: Advancing public health towards sustainable cities. *Bioengineering*, *11*(6), 533. https://doi.org/10.3390/bioengineering11060533

[10] Siddiqui, H., Rattani, A., Woods, N. K., Cure, L., Lewis, R. K., Twomey, J., Smith-Campbell, B., & Hill, T. J. (2021). A survey on machine and deep learning models for childhood and adolescent obesity. *IEEE Access*, *9*, 157337–157360. https://doi.org/10.1109/ACCESS.2021.3131128

[11] Shakti, M. A. S., Vijayalakshmi, M., Kumar, N., & Vaidhehi, M. (2024). Analysis on various machine learning framework for obesity level prediction. In *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)* (pp. 1–6). IEEE. https://doi.org/10.1109/IC-CGU58078.2024.10530812

[12] Gozukara Bag, H. G., Yagin, F. H., Gormez, Y., González, P. P., Colak, C., Gülü, M., Badicu, G., & Ardigò, L. P. (2023). Estimation of obesity levels through the proposed predictive approach based on physical activity and nutritional habits. *Diagnostics*, *13*(18), 2949. https://doi.org/10.3390/diagnostics13182949

[13] Castro, C., Leiva, V., Lourenço-Gomes, M. C., & Amorim, A. P. (2023). Advanced mathematical approaches in psycholinguistic data analysis: A methodological insight. *Fractal and Fractional*, *7*(9), 670. https://doi.org/10.3390/fractalfract7090670

[14] Asif, S., Yi, W., ur-Rehman, S., ul-ain, Q., Amjad, K., Yi, Y., Si, J., & Awais, M. (2024). Advancements and prospects of machine learning in medical diagnostics: Unveiling the future of diagnostic precision. *Archives of Computational Methods in Engineering*, 1–31. https://doi.org/ 10.1007/s11831-024-10148-w

[15] Ganie, S. M., Reddy, B. B., Hemachandran, K., & Rege, M. (2024). An investigation of ensemble learning techniques for obesity risk prediction using lifestyle data. *Decision Analytics Journal*, *5*, 100539. https://doi.org/10.1016/j.dajour.2024.100539

[16] Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., & Habib, M. T. (2021). A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, *2*, 100053. https://doi.org/10.1016/j.crbeha.2021.100053

[17] Verma, O. P., Verma, S., & Perumal, T. (2024). *Advancement of intelligent computational methods and technologies*. CRC Press.

[18] An, R., Shen, J., & Xiao, Y. (2022). Applications of artificial intelligence to obesity research: Scoping review of methodologies. *Journal of Medical Internet Research*, *24*(12), e40589. https://doi.org/10.2196/40589

[19] Wu, Y., Li, D., & Vermund, S. H. (2024). Advantages and limitations of the body mass index (BMI) to assess adult obesity. *International Journal of Environmental Research and Public Health*, *21*(6), 757. https://doi.org/10.3390/ijerph21060757

[20] Palechor, F. M., & De la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief*, *25*, 104344. https://doi.org/10.1016/j.dib.2019.104344