International Journal of Artificial Intelligence in Medical Issues

Volume 3 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v3i1.181

Research Article

Obesity Prediction with Machine Learning Models Comparing Various Algorithm Performances

Yudha Islami Sulistya [1,*]; Maie Istighosah [2]
[1] Telkom University, Purwokerto, Jawa Barat 40257, Indonesia, yudhaislami@telkomuniversity.ac.id
[2] Telkom University, Purwokerto, Jawa Barat 40257, Indonesia, maieistigh@telkomuniversity.ac.id

Correspondence should be addressed to Yudha Islami Sulistya; yudhaislami@telkomuniversity.ac.id

Received 11 January 2025; Revised 20 February 2024; Accepted 25 April 2024; Published 30 May 2025

Copyright © 2025 International Journal of Artificial Intelligence in Medical Issues. This scholarly piece is accessible under the Creative Commons Attribution Non-commercial License, permitting dissemination and modification, conditional upon non-commercial use and due citation.

Abstract:
Obesity poses a significant global health risk due to its links to conditions such as diabetes, cardiovascular disease, and various cancers, underscoring the need for early prediction to enable timely intervention. This study evaluated the performance of seven machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, ExtraTrees, Gradient Boosting, AdaBoost, and XGBoost—in predicting obesity using health and lifestyle data. The models were assessed based on accuracy, precision, recall, and F1-score, with hyperparameter tuning applied for optimization. The results confirmed that the ExtraTrees Classifier was the best performer, achieving an accuracy of 92.6%, precision of 92.7%, recall of 92.8%, and F1-score of 92.7%. Both Random Forest (91.3% accuracy) and XGBoost (89.9% accuracy) also exhibited strong predictive abilities. In contrast, models like Logistic Regression (74.3% accuracy) and AdaBoost (73.0% accuracy) showed lower effectiveness, emphasizing the advantages of ensemble methods such as ExtraTrees in delivering accurate obesity predictions. These findings suggest that ensemble models provide a promising approach for early diagnosis and targeted healthcare interventions.

Keywords: Ensemble Method, Healthcare Planning, Machine Learning, Obesity Prediction.

Dataset link: https://www.kaggle.com/datasets/suleymansulak/obesity-dataset/

1. Introduction

Obesity is increasingly recognized as a critical public health concern worldwide, associated with various chronic diseases, including cardiovascular disease, diabetes, and certain cancers [1]. The prevalence of obesity has been steadily increasing over the past few decades, leading to significant health and economic burdens [2]. Machine learning (ML) techniques offer promising solutions for predicting obesity by identifying complex patterns in health data that traditional statistical methods may overlook [3].

Despite the availability of various ML algorithms, the performance of these models in predicting obesity varies significantly. Challenges arise in determining which algorithm provides the best predictive accuracy for different population groups, especially when handling high-dimensional datasets with multiple influencing factors [4]. Furthermore, there is a need to address the limitations of previous studies, which often focus on specific age groups or lack a comprehensive evaluation of different ML methods [5].

The goal of this study is to evaluate the predictive performance of various ML algorithms, including Logistic Regression, Decision Trees, Random Forest, ExtraTrees, Gradient Boosting, AdaBoost, and XGBoost, in obesity prediction. The models will be compared based on accuracy, precision, recall, and F1-score to determine the most effective approach for predicting obesity risk across different age groups and populations [6].

This study aims to answer the following research questions: 1) Which machine learning algorithm provides the highest accuracy in predicting obesity? 2) Can ensemble methods such as ExtraTrees and Random Forest outperform traditional models like Logistic Regression? The hypothesis is that ensemble methods will outperform simpler models due to their ability to handle complex, non-linear relationships in the data [7].

The research will focus on supervised learning algorithms and datasets containing lifestyle and health-related factors influencing obesity. Although various data pre-processing techniques and feature selection methods will be employed, the generalizability of the findings may be limited to the specific characteristics of the datasets used [8]. Additionally, while the study aims to optimize hyperparameters for each model, factors such as data imbalance or regional variations in obesity determinants may impact the results [9].

This study aims to contribute to the field by providing a comprehensive comparison of multiple machine learning models for obesity prediction, which can help healthcare professionals select appropriate models for early risk assessment [1]. By identifying the most effective algorithms, the study can improve the development of predictive tools for personalized health interventions, potentially leading to better management and prevention of obesity-related health issues [2].

2. Method

This research, as shown in **Figure 1**, focuses on predicting obesity using various machine learning models. It begins with Data Collection, where datasets related to obesity risk factors are gathered. The next stage is Feature Engineering using SMOTE, which balances the dataset through Synthetic Minority Over-sampling Technique and performs feature transformation [10], [11], [12]. The Modelling Split Data step prepares the data by dividing it into training and testing sets.

In the Model Training stage, models like Random Forest and XGBoost are trained with hyperparameter tuning to optimize their performance. Model Evaluation follows, assessing the models using metrics such as accuracy, precision, recall, and F1-score. Then, Model Comparison identifies the best-performing approach based on these metrics. Finally, Conclusion & Recommendations summarize the findings, suggest the optimal model for obesity prediction, and provide directions for future research. This structured approach is designed to enhance predictive accuracy, supporting early intervention efforts in healthcare.

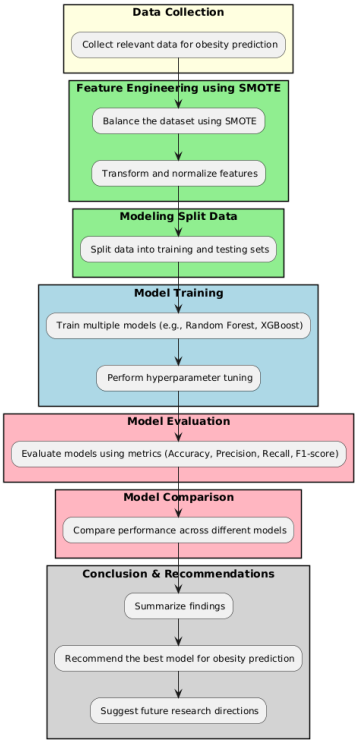

Figure 1: General Research Design

Data Collection:

The dataset used for this research includes various features related to demographic, dietary, and lifestyle factors, which are essential for predicting obesity. In **Table 1**, the features are listed alongside their descriptions, providing a detailed overview of the variables used in the analysis. These features include demographic information such as Sex and Age, dietary habits like Consumption of Fast Food and Frequency of Consuming Vegetables, and lifestyle factors such as Physical Exercise and Schedule Dedicated to Technology. The target variable, Class, represents the obesity classification, which is the outcome the model aims to predict.

The distribution of the target variable, Class, is illustrated in **Figure 2**, showing the number of instances in each obesity category. The dataset is composed of four classes: Underweight (73 samples), Normal (658 samples), Overweight (592 samples), and Obesity (287 samples). The bar chart demonstrates that the majority of the dataset falls within the "Normal" and "Overweight" categories, followed by "Obesity" and a smaller number of "Underweight" cases. This distribution ensures that the model has sufficient data to learn from different obesity levels, although some classes have fewer samples, which could affect classification accuracy for underrepresented categories.

Table 1. Feature Descriptions

Feature	Description
Sex	Gender of the individual (0 for female, 1 for male)
Age	Age of the individual in years
Height	Height of the individual in centimeters

Feature	Description
Overweight_Obese_Family	Presence of overweight or obesity in family (1 if yes, 0 if no)
Consumption_of_Fast_Food	Frequency of consuming fast food (times per week)
Frequency_of_Consuming_Vegetables	Frequency of consuming vegetables (times per week)
Number_of_Main_Meals_Daily	Number of main meals consumed daily
Food_Intake_Between_Meals	Frequency of food intake between main meals (times per day)
Smoking	Smoking status of the individual (1 if smoker, 0 if non-smoker)
Liquid_Intake_Daily	Amount of liquids consumed daily (liters)
Calculation_of_Calorie_Intake	Whether the individual calculates daily calorie intake (1 if yes, 0 if no)
Physical_Excercise	Level of physical exercise performed weekly (hours)
Schedule_Dedicated_to_Technology	Time dedicated to technology use daily (hours)
Type_of_Transportation_Used	Mode of transportation used most frequently (e.g., walking, cycling, driving)
Class	Obesity classification or health status (target variable)

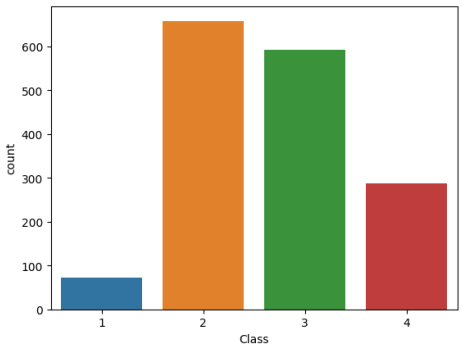

Figure 2. Distribution Class

Figure 3 illustrates the correlation matrix for the features in the dataset, showing the pairwise correlation coefficients between different variables. The correlation coefficient, denoted as r, ranges from -1 to 1, where values closer to 1 indicate a strong positive correlation, values closer to -1 indicate a strong negative correlation, and values near zero suggest no significant linear relationship. The correlation matrix helps identify relationships between features and the target variable, Class, which represents obesity levels [13], [14]. The Equation 1 for calculating the correlation coefficient between two variables X and Y is given by:

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \quad (1)$$

where X_i and Y_i are individual data points, $\bar{X}$ and $\bar{Y}$ are the mean values of the variables X and Y, respectively. The numerator represents the covariance of X and Y, while the denominator is the product of their standard deviations.

The correlation matrix in **Figure 3** reveals some notable relationships. For instance, Age has a strong positive correlation with Class ($r = 0.58$), indicating that as age increases, there is a tendency for obesity levels to rise. Similarly, Number of Main Meals Daily shows a moderate positive correlation with Class ($r = 0.51$), suggesting that individuals consuming more main meals daily are more likely to fall into higher obesity categories. Conversely, Frequency of Consuming Vegetables ($r = -0.54$) and Consumption of Fast Food ($r = -0.38$) exhibit negative

correlations with Class, implying that higher vegetable intake and lower fast-food consumption are associated with lower obesity levels.

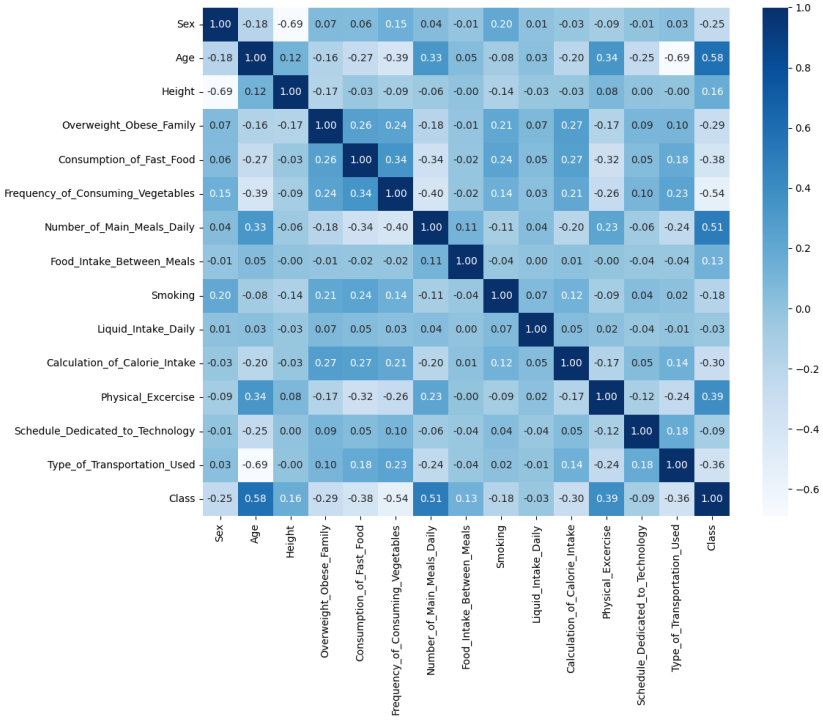

Figure 3. Feature Correlation

Analyzing the correlations further, Physical Exercise has a negative correlation with Class ($r = -0.39$), indicating that more physical activity tends to be associated with lower obesity risk. Additionally, Type of Transportation Used shows a negative correlation ($r = -0.36$), suggesting that active modes of transportation (e.g., walking, cycling) may contribute to lower obesity levels. Other features like Smoking and Food Intake Between Meals show weaker correlations with Class, indicating they may have a less significant impact on obesity prediction compared to other variables in the dataset. Overall, the correlation analysis helps in feature selection by identifying variables that have meaningful associations with the target variable.

Feature Engineering:

The Synthetic Minority Over-sampling Technique (SMOTE) is a commonly used method for addressing class imbalance in datasets, especially in classification tasks [15], [16]. It generates synthetic samples for the minority class by interpolating between existing minority class instances. The Equation 2 for generating a synthetic sample x_{new} is given by:

$$x_{\text{new}} = x_i + \lambda \times (x_j - x_i) \tag{2}$$

where x_i is a minority class sample, x_j is one of its k-nearest neighbours, and λ is a random number between 0 and 1. By generating new samples along the line segments connecting minority class samples and their neighbors, SMOTE

effectively increases the number of instances in the minority class, thus balancing the dataset and improving the model's ability to learn from underrepresented classes.

In **Figure 4**, the results after applying SMOTE to balance the dataset are shown. The count plot displays a uniform distribution across all four classes (1: Underweight, 2: Normal, 3: Overweight, and 4: Obesity), indicating that each class now has approximately the same number of samples. This balanced distribution ensures that the model will not be biased towards any particular class during training, which helps improve the classification performance across all categories. The balanced dataset provides a stronger foundation for training machine learning models, allowing them to better generalize to new, unseen data.

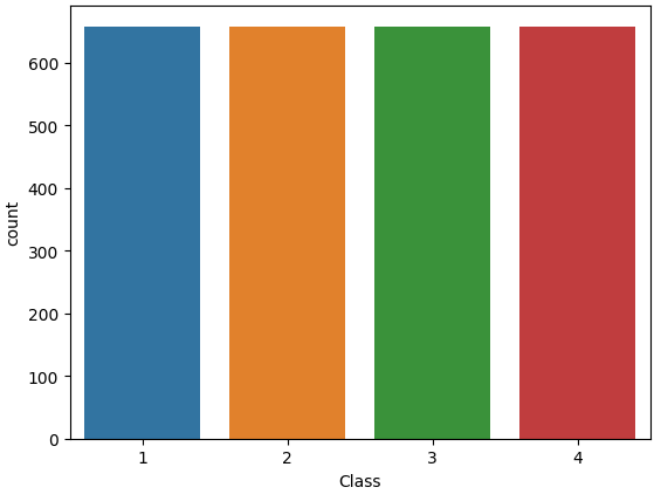

Figure 4. After Balance Dataset

Algorithm and Modelling

The first algorithm, Logistic Regression, is a linear model that estimates the probability of an instance belonging to a particular class using the logistic function. It is typically used for binary classification but can be extended to multiclass problems [17], [18]. The prediction is based on the log-odds of the outcome being linearly related to the input features. The logistic regression model uses the sigmoid function, represented is Equation 3:

$$P(y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \tag{3}$$

where β_0 is the intercept, β_i are the coefficients for the features x_i, and $P(y=1)$ is the probability of the positive class. Decision Tree Classifier uses a tree-like structure for making decisions. It splits the dataset into subsets based on the feature that provides the maximum information gain or reduces impurity the most [19]. The impurity can be measured using metrics such as Gini index or entropy. The Equation 4 for entropy at a node is.

$$H(X) = -\sum_{i=1}^{n} P_i \log_2 P_i \tag{4}$$

where P_i is the probability of class i at a particular node. Decision Trees are intuitive and easy to visualize but can overfit if the tree is too deep. The Random Forest Classifier is an ensemble learning method that combines multiple Decision Trees to create a stronger model. Each tree is trained on a random subset of the data, and the final prediction is the majority vote of all the trees. This method introduces randomness by selecting a random subset of features at each split, which reduces variance [20], [21]. The prediction of a Random Forest can be expressed as Equation 5

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} f_m(x) \tag{5}$$

where M is the number of trees and $f_m(x)$ is the prediction from the m-th tree. ExtraTreesClassifier, or Extremely Randomized Trees, is another ensemble method similar to Random Forest but with more randomization. Instead of finding the best split for each node, ExtraTrees randomly selects the split points. This approach aims to reduce variance and prevent overfitting by increasing randomness [22], [23]. The prediction process is the same as that for Random Forest, but with extra randomness during the tree-building process.

The Gradient Boosting Classifier is a boosting method that builds trees sequentially, with each new tree focusing on the residual errors of the previous trees. The goal is to minimize a loss function by combining the predictions of many weak learners. The Equation 6 for updating the model at each iteration is.

$$F_m(x) = F_{m-1}(x) + \gamma \cdot h_m(x) \tag{6}$$

where $F_m(x)$ is the model at iteration m, $h_m(x)$ is the new tree, and γ is the learning rate. Gradient Boosting is powerful but requires careful tuning to avoid overfitting. AdaBoostClassifier, or Adaptive Boosting, works by adjusting the weights of instances based on whether they are correctly classified [24]. The model gives higher weights to misclassified instances, forcing the new classifier to focus on harder cases. The weight update Equation 7 for a sample is given by.

$$w_i = w_i \times e^{\alpha \cdot I(\widehat{y_i} \neq y_i)} \tag{7}$$

where α is a measure of the classifier's error, I is an indicator function, and w_i is the weight of the i-th instance. Finally, XGB (Extreme Gradient Boosting) is an optimized version of Gradient Boosting, designed for speed and performance [25]. It includes regularization terms to avoid overfitting, making it more robust. The objective function for XGBoost is Equation 8.

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}i) + \sum k = 1^T \Omega(f_k) \tag{8}$$

where L is the loss function, Ω is a regularization term for the tree f_k, and T is the total number of trees. XGBoost's additional features like parallel computation and tree pruning make it one of the most efficient algorithms. The dataset is split into training and testing sets using the train_test_split function from the sklearn.model_selection module. The data splitting is performed to ensure that the model can be trained on a subset of the data (training set) and then evaluated on a different subset (testing set) to measure its generalization performance. The dataset is divided in an 80-

20 ratio, where 80% of the data is used for training and 20% is used for testing. The random_state parameter is set to 42 to ensure reproducibility of the results.

Performance evaluation is essential for determining the effectiveness of the machine learning model in predicting obesity. Several metrics are used, including accuracy, precision, recall, and F1-score, which help assess the model's ability to correctly classify individuals into different obesity categories (Underweight, Normal, Overweight, and Obesity). Where accuracy measures the proportion of correctly classified instances out of the total number of instances, and is given by the Equation 9.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

where TP (True Positive) represents correctly classified individuals in a specific obesity category, TN (True Negative) indicates correctly classified individuals in non-target categories, FP (False Positive) denotes individuals incorrectly classified as belonging to the target category, and FN (False Negative) is the number of individuals misclassified as not belonging to the target category. Precision, calculated as $\text{Precision} = \frac{TP}{TP+FP}$, shows how many of the individuals predicted to belong to a specific obesity category actually do, reflecting the model's ability to avoid false positives. Recall, given by $\text{Recall} = \frac{TP}{TP+FN}$, measures the proportion of actual individuals in a specific obesity category correctly identified by the model, indicating the model's sensitivity to true positives. The F1-score, calculated as the harmonic mean of precision and recall using the Equation 10.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

provides a balance between precision and recall, making it particularly valuable when dealing with imbalanced datasets. These metrics collectively offer a comprehensive evaluation of the model's performance, highlighting its strengths and areas for improvement in accurately classifying individuals into the different obesity categories.

3. Result and Discussion

Results

The performance evaluation of different machine learning models reveals significant variations in their ability to classify obesity categories. According to **Table 2**, the ExtraTreesClassifier emerges as the top-performing model with the highest accuracy (0.926), precision (0.927), recall (0.928), and F1-score (0.927), indicating its superior predictive power. Both the Random Forest Classifier and XGB follow closely, showcasing strong performance with accuracies of 0.913 and 0.899, respectively. These results suggest that ensemble methods, particularly those based on decision trees, are highly effective for this classification task. On the other hand, simpler models like Logistic Regression and AdaBoostClassifier demonstrate lower performance, achieving accuracy scores of 0.743 and 0.730, respectively, indicating their limitations in capturing complex patterns within the dataset.

The bar charts in **Figure 5** visually compare the metrics across models, reinforcing the findings. The charts clearly illustrate the superior performance of the ExtraTreesClassifier, Random Forest, and XGB across all four metrics, while

models such as GradientBoostingClassifier show moderate results. In contrast, Logistic Regression and AdaBoostClassifier consistently exhibit the lowest scores, reflecting their reduced effectiveness in this context. The visual representation highlights the advantage of using ensemble approaches, especially tree-based methods, which excel in managing diverse feature interactions and contribute to higher classification accuracy for obesity prediction.

Table 2. Performance each Model Machine Learning

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.74383	0.74488	0.74803	0.74558
Decision Tree Classifier	0.87476	0.87536	0.87666	0.87485
Random Forest Classifier	0.91271	0.91321	0.91485	0.91315
ExtraTreesClassifier	0.92600	0.92734	0.92839	0.92658
GradientBoostingClassifier	0.86338	0.86324	0.86700	0.86327
AdaBoostClassifier	0.73055	0.72511	0.73312	0.72631
XGB	0.89943	0.89973	0.90189	0.90032

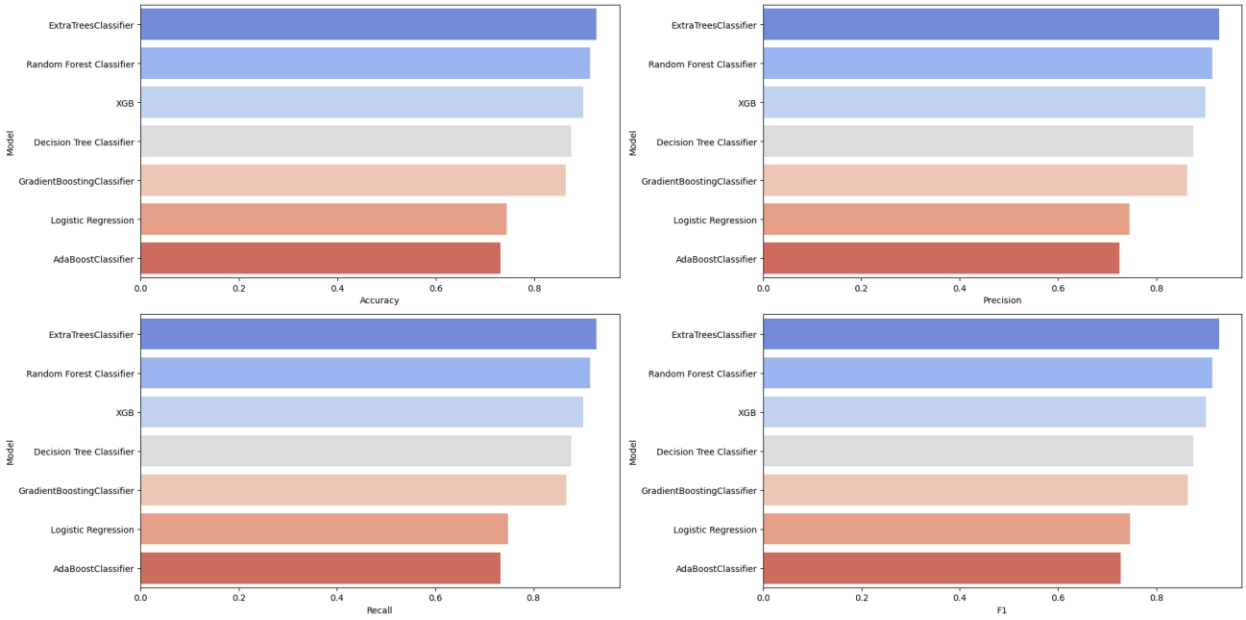

Figure 5. Performance each Model Machine Learning (Bar Chart)

The confusion matrix provides an overview of the classification performance of the best model, with rows representing actual classes and columns indicating predicted classes. The diagonal values (124, 120, 111, and 133) show the correctly classified instances for each obesity category (0: Underweight, 1: Normal, 2: Overweight, 3: Obesity), demonstrating the model's effectiveness in accurately identifying most cases. **Figure 6** indicates particularly strong performance in the "Underweight" and "Obesity" categories, where almost all instances were classified correctly.

However, there are notable misclassifications, especially between adjacent categories. For example, the model misclassified 16 instances of "Overweight" as "Normal" and 11 as "Obesity," suggesting difficulties in distinguishing between these classes. These off-diagonal values in **Figure 6** highlight the challenge of separating nearby obesity levels, which may arise due to similar feature distributions or less distinct boundaries. Overall, the matrix reflects the model's strengths while pointing out specific areas for improvement in handling neighboring categories.

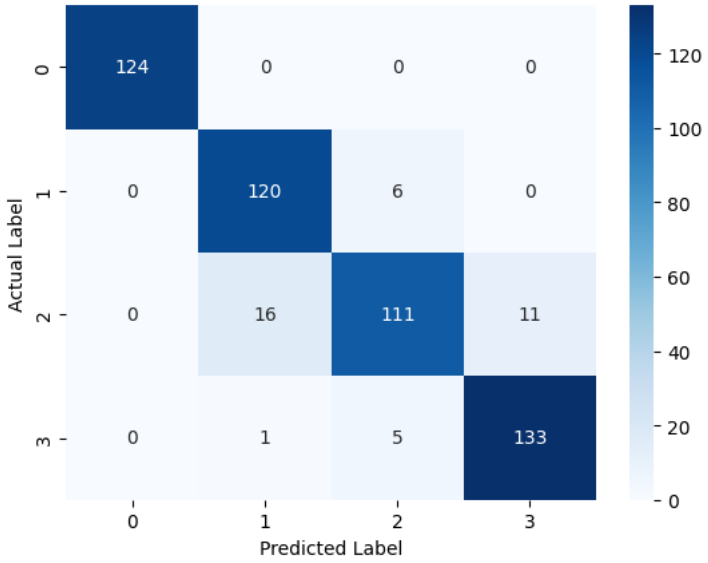

Figure 6. Confusion Matrix with the Best Model

Discussion

The results indicate that ensemble-based models, particularly the ExtraTreesClassifier, achieved the highest performance across all metrics, with notable accuracy, precision, recall, and F1-score values. This aligns with existing research suggesting that ensemble techniques, which combine multiple decision trees, are more effective in handling complex datasets with numerous features and interactions. The strong performance of models like Random Forest and XGB further supports the view that ensemble learning methods can better capture non-linear relationships compared to simpler algorithms such as Logistic Regression or AdaBoost, which showed lower predictive accuracy.

These findings are consistent with previous studies that highlight the advantage of tree-based ensemble models in classification tasks, particularly when dealing with imbalanced data or complex feature spaces. The use of techniques like SMOTE for data balancing has also proven effective in improving model performance, as it reduces bias toward majority classes. Practical implications of these results include the potential use of these models in healthcare settings to predict obesity risk more accurately, allowing for early intervention and personalized health recommendations. However, the model's occasional misclassifications between adjacent obesity categories (e.g., "Normal" and "Overweight") suggest that further refinement is needed to enhance precision.

The research has some limitations, including the specific demographic characteristics of the dataset, which may affect the generalizability of the results. Additionally, the analysis mainly focused on supervised learning algorithms,

without exploring other techniques like deep learning. For future research, it is recommended to incorporate more diverse datasets from different populations and investigate hybrid approaches combining ensemble learning with deep learning methods to improve classification accuracy. Further exploration of advanced data balancing techniques could also enhance the model's performance across all obesity categories.

4. Conclusion

In conclusion, the research demonstrates that ensemble-based models, particularly the ExtraTreesClassifier, outperformed other algorithms in predicting obesity, with the highest accuracy (92.6%), precision (92.7%), recall (92.8%), and F1-score (92.7%). These results confirm the hypothesis that tree-based ensemble methods are more effective for handling complex data and non-linear relationships compared to simpler models like Logistic Regression (accuracy 74.3%) or boosting methods such as AdaBoostClassifier (accuracy 73.0%). The use of data balancing techniques, such as SMOTE, played a crucial role in improving the model's performance by addressing class imbalance. Misclassifications were still present, particularly between neighboring categories like "Normal" and "Overweight," indicating areas where the model's sensitivity could be improved.

The findings address the research questions by showing that the ExtraTreesClassifier, followed by Random Forest and XGB, provides the best classification results for obesity prediction. These models effectively distinguish between different obesity categories, confirming that ensemble techniques are well-suited for this task. The study contributes to the field by providing a comparative analysis of various machine learning algorithms and demonstrating the practical value of ensemble methods in healthcare applications for predicting obesity. The research offers valuable insights for healthcare professionals seeking to implement machine learning-based tools for early intervention and personalized patient management.

To further enhance the accuracy and generalizability of obesity prediction models, future research should consider using more diverse datasets from different populations, which would help in validating the model's applicability across various demographic groups. Additionally, integrating hybrid approaches, such as combining deep learning techniques with ensemble models, may lead to improved classification performance. Exploring advanced data augmentation or balancing techniques beyond SMOTE could also further mitigate the issues associated with class imbalance, enhancing the model's effectiveness in real-world settings.

References:

[1] M. Safaei, E. A. Sundararajan, M. Driss, W. Boulila, and A. Shapi'i, "A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity," Sep. 01, 2021, Elsevier Ltd. doi: 10.1016/j.compbiomed.2021.104754.

[2] B. Bonnechère, A. Cuevas-Sierra, J. Jeon, S. Lee, and C. Oh, "Age-specific risk factors for the prediction of obesity using a machine learning approach." [Online]. Available: https://knhanes.kdca.go.kr/knhanes

[3] W. Lin, S. Shi, H. Huang, J. Wen, and G. Chen, "Predicting risk of obesity in overweight adults using interpretable machine learning algorithms," Front Endocrinol (Lausanne), vol. 14, 2023, doi:

10.3389/fendo.2023.1292167.

[4] K. Fujihara et al., “Machine learning approach to predict body weight in adults,” Front Public Health, vol. 11, 2023, doi: 10.3389/fpubh.2023.1090146.

[5] X. Pang, C. B. Forrest, F. Lê-Scherban, and A. J. Masino, “Prediction of early childhood obesity with machine learning and electronic health record data,” Int J Med Inform, vol. 150, Jun. 2021, doi: 10.1016/j.ijmedinf.2021.104454.

[6] A. Fernandes, S. Dahikar, K. Chopra, and K. Saxena, “Comparison of Machine Learning Algorithms for Obesity Prediction,” in 2023 3rd Asian Conference on Innovation in Technology, ASIANCON 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ASIANCON58793.2023.10270246.

[7] E. R. Cheng, R. Steinhardt, and Z. Ben Miled, “Predicting Childhood Obesity Using Machine Learning: Practical Considerations,” BioMedInformatics, vol. 2, no. 1, pp. 184–203, Mar. 2022, doi: 10.3390/biomedinformatics2010012.

[8] J. Dunstan, M. Aguirre, M. Bastías, C. Nau, T. A. Glass, and F. Tobar, “Predicting nationwide obesity from food sales using machine learning,” Health Informatics J, vol. 26, no. 1, pp. 652–663, Mar. 2020, doi: 10.1177/1460458219845959.

[9] F. Musa, F. Basaky, and O. E.O, “Obesity prediction using machine learning techniques,” Journal of Applied Artificial Intelligence, vol. 3, no. 1, pp. 24–33, Jun. 2022, doi: 10.48185/jaai.v3i1.470.

[10] Y. Bao and S. Yang, “Two Novel SMOTE Methods for Solving Imbalanced Classification Problems,” IEEE Access, vol. 11, pp. 5816–5823, 2023, doi: 10.1109/ACCESS.2023.3236794.

[11] K. Kim, “Noise Avoidance SMOTE in Ensemble Learning for Imbalanced Data,” IEEE Access, vol. 9, pp. 143250–143265, 2021, doi: 10.1109/ACCESS.2021.3120738.

[12] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, “Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data,” IEEE Access, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316.

[13] C. Wang et al., “Prediction of the Consolidation Coefficient of Soft Soil Based on Machine Learning Models,” Soil Mechanics and Foundation Engineering, vol. 61, no. 3, pp. 223–229, Jul. 2024, doi: 10.1007/s11204-024-09966-8.

[14] T. Nie, M. Zhao, Z. Zhu, K. Zhao, and Z. Wang, “Estimating feature importance in circuit network using machine learning,” Multimed Tools Appl, vol. 83, no. 11, pp. 31233–31249, Mar. 2024, doi: 10.1007/s11042-023-16814-8.

[15] D. Elreedy, A. F. Atiya, and F. Kamalov, “A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning,” Mach Learn, vol. 113, no. 7, pp. 4903–4923,

Jul. 2024, doi: 10.1007/s10994-022-06296-4.

[16] Y. He, X. Lu, P. Fournier-Viger, and J. Z. Huang, “A novel overlapping minimization SMOTE algorithm for imbalanced classification,” Frontiers of Information Technology and Electronic Engineering, Sep. 2024, doi: 10.1631/FITEE.2300278.

[17] H. Liu, X. Li, F. Chen, W. Härdle, and H. Liang, “A comprehensive comparison of goodness-of-fit tests for logistic regression models,” Stat Comput, vol. 34, no. 5, Oct. 2024, doi: 10.1007/s11222-024-10487-5.

[18] K. Nishiura, E. H. Choi, E. Choi, and O. Mizuno, “Two improving approaches for faulty interaction localization using logistic regression analysis,” Software Quality Journal, vol. 32, no. 3, pp. 1039–1073, Sep. 2024, doi: 10.1007/s11219-024-09677-1.

[19] A. J. Albert, R. Murugan, and T. Sripriya, “Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology,” Research on Biomedical Engineering, vol. 39, no. 1, pp. 99–113, Mar. 2023, doi: 10.1007/s42600-022-00253-9.

[20] M. Li et al., “Protein-Protein Interaction Sites Prediction Based on an Under-Sampling Strategy and Random Forest Algorithm,” IEEE/ACM Trans Comput Biol Bioinform, vol. 19, no. 6, pp. 3646–3654, Nov. 2022, doi: 10.1109/TCBB.2021.3123269.

[21] N. S. Thomas and S. Kaliraj, “An Improved and Optimized Random Forest Based Approach to Predict the Software Faults,” SN Comput Sci, vol. 5, no. 5, Jun. 2024, doi: 10.1007/s42979-024-02764-x.

[22] M. G. Lanjewar, J. S. Parab, A. Y. Shaikh, and M. Sequeira, “CNN with machine learning approaches using ExtraTreesClassifier and MRMR feature selection techniques to detect liver diseases on cloud,” Cluster Comput, vol. 26, no. 6, pp. 3657–3672, Dec. 2023, doi: 10.1007/s10586-022-03752-7.

[23] S. Kaushik, M. Balachandra, D. Olivia, and Z. Khan, “Unveiling the epilepsy enigma: an agile and optimal machine learning approach for detecting inter-ictal state from electroencephalogram signals,” International Journal of Information Technology (Singapore), 2024, doi: 10.1007/s41870-024-02078-4.

[24] I. Colakovic and S. Karakatič, “Adaptive Boosting Method for Mitigating Ethnicity and Age Group Unfairness,” SN Comput Sci, vol. 5, no. 1, Jan. 2024, doi: 10.1007/s42979-023-02342-7.

[25] H. T. Wu, Z. L. Zhang, and D. Dias, “Prediction on compression indicators of clay soils using XGBoost with Bayesian optimization,” J Cent South Univ, 2024, doi: 10.1007/s11771-024-5681-9.