Proceedings of ICBioMed 2025 Symposium: AI for Healthcare: Advanced Medical Data Analytics and Smart Rehabilitation
DOI: 10.54254/2753-8818/2025.AU27039

Machine Learning Analysis of Obesity Levels Based on Lifestyle and Physical Features

Junyuan Zhang

School of Astronomy & Space Science, Nanjing University, Nanjing, China
201840183@smail.nju.edu.cn

Abstract. Obesity has become a major global health challenge, with prevalence rising across all age groups. While Body Mass Index (BMI) is widely used for diagnosis, it has limitations, as it cannot fully capture lifestyle or behavioral risk factors. This study applies machine learning (ML) techniques to classify obesity categories using the UCI dataset "Estimation of Obesity Levels Based on Eating Habits and Physical Condition." The dataset includes 16 lifestyle and demographic features, and experiments were conducted with and without height and weight to evaluate the potential effect of BMI-related variables. Random Forest was adopted as the primary model, with additional classifiers employed as comparative methods. Results show that ensemble tree-based models achieved the highest performance overall. In particular, Random Forest demonstrated strong accuracy even in the absence of height and weight, suggesting that it can capture meaningful lifestyle and behavioral patterns without relying solely on BMI. Feature importance analysis identified age, diet, and physical activity as key predictors. Misclassification analysis further revealed that some severely obese individuals were incorrectly predicted as normal weight, especially those with positive family history but healthier reported habits. These findings highlight the potential of ML to support obesity risk assessment and the importance of addressing information leakage when BMI-related features are included.

Keywords: Obesity, Machine Learning, Random Forest, Feature Importance, Misclassified cases

1. Introduction

Obesity has become a critical global health challenge. According to the World Health Organization (WHO), worldwide obesity has nearly tripled since 1975, and in 2022 more than 1 billion people were living with obesity [1]. The health consequences are severe, including higher risks of diabetes, cardiovascular disease, and certain cancers.

Body Mass Index (BMI), introduced by Adolphe Quetelet in the 19th century and popularized by Ancel Keys in the 1970s, is the most widely used measure for classifying obesity [2,3]. Its simplicity makes it practical, but its limitations are well known: for instance, athletes with high muscle mass may be misclassified as overweight despite low body fat [3]. These shortcomings highlight the need for additional approaches to better capture obesity-related risks.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.AU27039

Machine learning (ML) has recently emerged as a promising tool for healthcare research, including applications in diagnosis, imaging, and personalized prediction. In obesity studies, the use of ML has expanded rapidly, with publications growing exponentially in the past two decades [4]. ML methods, such as decision trees, random forests, and support vector machines, have shown promise in predicting obesity by analyzing multidimensional data, such as lifestyle, diet, physical activity, and genetics.

This study uses Random Forest as the primary method, highlighting its ability to handle both categorical and continuous features while addressing class imbalance. By excluding BMI-derived features (height and weight), this study demonstrates that models can better focus on lifestyle and genetic factors, leading to a more comprehensive classification of obesity. This approach allows for a clearer interpretation of how lifestyle behaviors contribute to obesity risk, rather than relying solely on BMI-related metrics.

This study investigates ML models for obesity classification using a dataset from the UCI repository [5]. The dataset contains 16 lifestyle- and health-related features plus a target variable representing obesity levels [6]. An important consideration is whether to include height and weight, since BMI is directly calculated from them. Retaining these features may inflate apparent accuracy; removing them exposes the real weight of the remaining variables. All experiments were run in both Weka and Python, with multiple algorithms and SMOTE resampling to correct class imbalance [6].

2. Methodology

The dataset used in this study is the Estimation of Obesity Levels Based on Eating Habits and Physical Condition from the UCI Machine Learning Repository [5]. It contains 2,111 samples, each described by 16 input features and one target variable. The target variable, NObeyesdad, categorizes individuals into seven classes: Insufficient Weight, Normal Weight, Overweight Levels I and II, and Obesity Types I, II, and III. Features include demographic information (e.g., gender, age), lifestyle behaviors (e.g., frequency of vegetable consumption [FCVC], physical activity frequency [FAF], water consumption [CH2O]), and family history of overweight [6]. About 77% of the dataset was synthetically generated using SMOTE (Synthetic Minority Over-sampling Technique), which interpolates new synthetic samples between existing minority class neighbors to balance class distributions [7].

2.1. Data preprocessing

Categorical variables such as gender and transportation method were mapped to numerical values, while the target variable was encoded from text labels to integers (0–6). Continuous variables such as age, height, and weight were kept as numeric inputs. In Python, StandardScaler was applied to normalize features before training. Two experimental datasets were prepared: one including all 16 features (with height and weight), and another excluding them (without height and weight), since these two directly determine BMI and may cause data leakage.

2.2. Random forest classifier

Random Forest (RF) was chosen as the primary method due to its robustness, nonlinearity, and strong performance in classification tasks [8]. RF constructs an ensemble of decision trees, each trained on a bootstrap sample of the data with randomly selected subsets of features. Final predictions are made through majority voting across all trees.

DOI: 10.54254/2753-8818/2025.AU27039

The workflow of RF is illustrated in Figure 1, which shows the process of bootstrap sampling, tree construction, and aggregation.

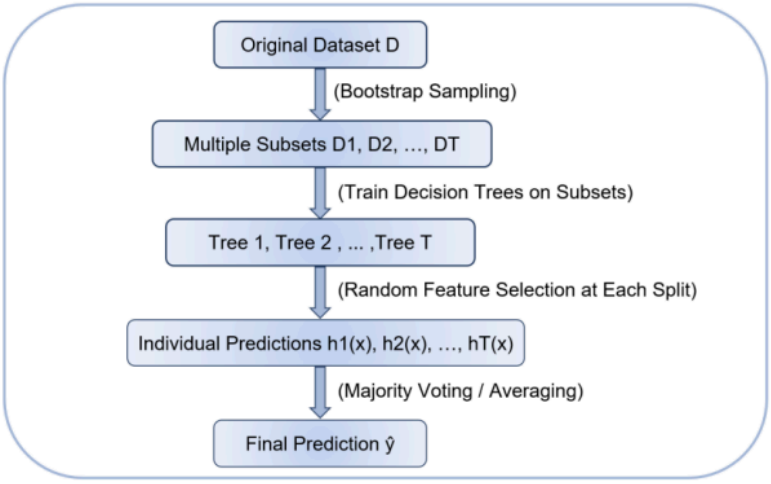

Figure 1. Structure of the random forest algorithm

Other models, including Gradient Boosting, Decision Tree, k-Nearest Neighbors, and Logistic Regression, were implemented as baselines for comparison, but the main analysis focused on the Random Forest.

2.3. Evaluation metrics

Model performance was evaluated using four standard metrics: accuracy, precision, recall, and weighted F1-score. Their definitions are:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3}$$

$$\text{F1}-\text{score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{4}$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. Weighted F1-score averages F1 across all classes with weights proportional to class frequencies, which is particularly suitable for imbalanced datasets.

3. Results

The results from the Weka and Python experiments are summarized in comparative tables and figures. Both sets of experiments evaluated the models under two configurations: one including all 16 features and one excluding height and weight. This design highlighted the influence of anthropometric features compared to lifestyle and demographic variables.

The classification results obtained from multiple models using Weka are summarized in Table 1. Ensemble models, such as Random Forest and Simple Logistic showed superior accuracy compared

DOI: 10.54254/2753-8818/2025.AU27039

to single classifiers like J48 Decision Tree and Naive Bayes. When height and weight were included, the overall weighted accuracy exceeded 92%, with Random Forest reaching the highest level of classification reliability across categories. However, performance dropped notably when these two variables were removed, indicating the degree to which BMI-related information dominates classification. Despite this drop, the models still performed well enough to distinguish between broad categories, such as underweight and obesity, suggesting that lifestyle and behavioral features retain predictive power.

Table 1. Weka model performance with and without height/weight features

Dataset	Model	Accuracy	Precision	Recall	F1-score
with height and weight	J48 Decision Tree	93.75%	0.938	0.937	0.938
	Naive Bayes	67.41%	0.662	0.674	0.665
	Random Forest	95.64%	0.959	0.956	0.957
	SimpleLogistic	96.35%	0.964	0.964	0.964
	SMO (SVM)	84.51%	0.844	0.845	0.842
without height and weight	J48 Decision Tree	78.11%	0.78	0.781	0.781
	Naive Bayes	55.19%	0.548	0.552	0.521
	Random Forest	86.64%	0.871	0.866	0.867
	SimpleLogistic	62.01%	0.614	0.62	0.601
	SMO (SVM)	63.33%	0.628	0.633	0.617

To further evaluate model performance, experiments were also conducted in Python. The outcomes, with and without height and weight features, are reported in Table 2. Python experiments yielded similar trends. Random Forest and Gradient Boosting consistently outperformed the other models, with Gradient Boosting slightly surpassing Random Forest when height and weight were present. Specifically, with all features, Gradient Boosting achieved an accuracy of over 95%, while Random Forest reached about 94%. This can be explained by the boosting mechanism's ability to iteratively correct misclassified instances, providing an edge over bagging-based methods. Decision Tree and KNN achieved moderate accuracy, while Logistic Regression lagged significantly, particularly when height and weight were excluded. Without BMI-related variables, Random Forest still achieved an accuracy of about 84%, while Logistic Regression dropped below 60%. This illustrates the limitations of linear models in capturing the nonlinear relationships between lifestyle factors and obesity categories.

DOI: 10.54254/2753-8818/2025.AU27039

Table 2. Python model performance with and without height/weight features

Dataset	Model	Accuracy	Precision	Recall	F1-score
with height and weight	Random Forest	0.938486	0.9461	0.938486	0.939822
	Gradient Boosting	0.954259	0.956681	0.954259	0.954724
	Decision Tree	0.913249	0.915395	0.913249	0.913711
	KNN	0.880126	0.88251	0.880126	0.877785
	Logistic Regression	0.760252	0.755597	0.760252	0.756549
without height and weight	Random Forest	0.837539	0.845767	0.837539	0.839148
	Gradient Boosting	0.793375	0.80221	0.793375	0.793201
	Decision Tree	0.746057	0.746713	0.746057	0.744655
	KNN	0.719243	0.712709	0.719243	0.695246
	Logistic Regression	0.561514	0.54607	0.561514	0.535467

To further evaluate classification quality, confusion matrices were generated. As illustrated in Figure 2, the confusion matrix demonstrates that while most classes are predicted correctly, there are some misclassifications in the normal weight category. For Random Forest, most misclassifications occurred in borderline categories such as Normal Weight versus Overweight I, or Obesity Types I versus II. In contrast, severe obesity categories such as Obesity Type III were classified with near-perfect accuracy, reflecting the distinctiveness of extreme cases. Figure 3 visualizes the relative importance of input features. Feature importance analysis also indicated that variables such as age, frequency of vegetable consumption (FCVC), physical activity frequency (FAF), number of main meals (NCP), time using technology devices (TUE), and water consumption (CH2O) contributed significantly to the predictions when height and weight were removed.

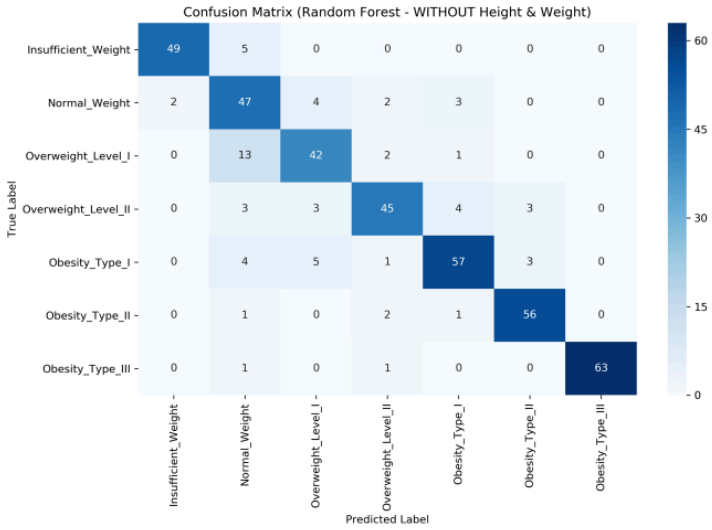

Figure 2. Confusion matrix of random forest classifier (without height and weight)

DOI: 10.54254/2753-8818/2025.AU27039

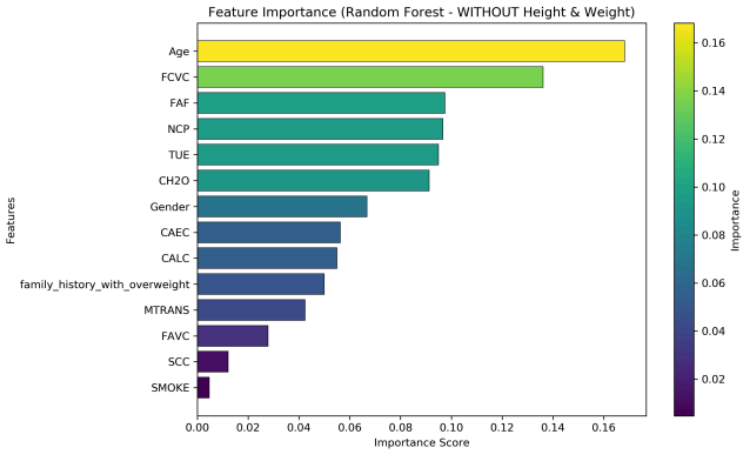

Figure 3. Feature importance ranking in random forest (without height and weight)

Finally, when examining the confusion matrix (Figure 2), it was observed that some samples were misclassified. In particular, several individuals were predicted as Normal Weight despite actually belonging to Obesity Type I, II or III. These representative misclassified cases are listed in Table 3. One individual had a BMI exceeding 40, yet the model underestimated their condition. These errors often corresponded to younger individuals who reported moderately healthy lifestyle habits, suggesting that reliance on self-reported behavior may sometimes obscure clinical risk. Such misclassifications highlight the importance of cautious interpretation when applying ML models in health contexts.

DOI: 10.54254/2753-8818/2025.AU27039

Table 3. Misclassified cases: predicted normal weight but actual obesity

No.	Gender	Age	Height	Weight	Family history with overweight	FAVC
1	Female	32	1.67	90	yes	yes
2	Male	20	1.8	114	yes	yes
3	Female	18.267696	1.706343	93.348843	yes	yes
4	Male	21	1.79	105	yes	yes
5	Male	26	1.85	105	yes	yes
6	Male	18	1.87	173	yes	yes
No.	FCVC	NCP	CAEC	SMOKE	CH2O	SCC
1	3	1	Sometimes	no	2	no
2	2	3	Frequently	no	2	no
3	1.70825	3	Sometimes	no	1	no
4	2	3	Always	no	1	no
5	3	3	Frequently	no	3	no
6	3	3	Frequently	no	2	no
No.	FAF	TUE	CALC	MTRANS	True label	Pred label
1	2	0	Sometimes	Automobile	Obesity Type I	Normal Weight
2	0	1	no	Public Transportation	Obesity Type II	Normal Weight
3	0.899864	1	Sometimes	Public Transportation	Obesity Type I	Normal Weight
4	0	0	Sometimes	Public Transportation	Obesity Type I	Normal Weight
5	2	2	Sometimes	Public Transportation	Obesity Type I	Normal Weight
6	2	1	Sometimes	Public Transportation	Obesity Type III	Normal Weight

4. Discussion

Random Forest emerged as the most robust model across experiments, consistently outperforming Gradient Boosting, Decision Trees, and Logistic Regression. This advantage became clearer when height and weight were excluded: while all models showed reduced accuracy, Random Forest retained relatively strong performance, reflecting its capacity to capture nonlinear relationships and interactions in heterogeneous health data.

The decision to train models both with and without height and weight highlighted the risk of data leakage. Since BMI categories are derived directly from these measures, including them led to inflated performance metrics, whereas removing them forced the model to rely on behavioral and demographic features. This approach provided more realistic insight into which lifestyle and environmental variables drive obesity risk.

Feature importance analysis indicated that age, vegetable consumption (FCVC), physical activity (FAF), meal frequency (NCP), screen time (TUE), and water consumption (CH2O) were among the most significant predictors. These results align with established nutritional and clinical evidence linking sedentary behavior, poor diet, and hydration patterns to obesity.

Family history of overweight, though a well-known genetic determinant, was not ranked highly in the global importance scores [9]. This may be due to its binary encoding, which provides less granularity compared to continuous lifestyle features. Nonetheless, case-level misclassification

DOI: 10.54254/2753-8818/2025.AU27039

analysis showed that nearly all severely obese individuals wrongly predicted as normal weight reported a positive family history, suggesting that the variable still holds critical clinical relevance.

The misclassified cases also highlight a risk of underestimation. Younger individuals with partial healthy habits were sometimes predicted as normal despite being obese, which could delay necessary interventions. This reinforces the need for richer genetic and behavioral data to reduce such errors.

Overall, the findings confirm that while anthropometric features dominate prediction, Random Forest effectively identifies other meaningful risk factors, underscoring its value for obesity classification in both clinical and public health contexts.

5. Conclusion

This study explored the application of machine learning methods to classify obesity categories using basic personal characteristics (e.g., age, gender), lifestyle, and hereditary features. Ensemble models such as Random Forest and Gradient Boosting achieved the highest accuracy, confirming their suitability for handling the nonlinear and multifactorial nature of obesity. While including height and weight improved performance due to their direct link to BMI, it also risked data leakage. Removing these variables reduced accuracy but revealed meaningful predictors, such as age, vegetable consumption, physical activity, and hydration. This suggests that although anthropometric factors dominate, behavioral and lifestyle variables still provide valuable insights into obesity risk.

Misclassification analysis showed that some individuals with severe obesity were predicted as normal weight, particularly when self-reported behaviors appeared healthier. This highlights the limitations of binary encoding for hereditary factors and of self-reported data more broadly. Improving feature engineering and incorporating richer genetic information could mitigate such errors.

From a clinical perspective, these findings echo recent efforts to refine obesity assessment. The World Health Organization recognizes the limitations of BMI, while the Lancet Commission recommends integrating waist circumference and waist-to-height ratio into diagnostic frameworks [10]. Similarly, the 2025 guidelines from China's National Health Commission propose adjusted thresholds for older adults, reflecting changes in body composition with age [11]. The results align with these perspectives by showing that models can identify meaningful predictors even without direct BMI-related features.

Future work should expand the dataset with larger, real-world samples to improve generalizability and reduce biases introduced by synthetic oversampling methods such as SMOTE. Integrating longitudinal data or wearable device outputs could enable more dynamic and personalized predictions. Moreover, deep learning approaches may capture complex temporal patterns, while explainable AI (XAI) could enhance clinical interpretability [12]. Finally, cross-national datasets should be examined to account for cultural, dietary, and socioeconomic diversity, ensuring models remain robust across populations.

In summary, machine learning provides a promising approach for obesity classification and risk stratification, but its effectiveness depends on careful feature design, avoidance of data leakage, and alignment with evolving clinical standards.

References

[1] World Health Organization. (2022) Obesity and overweight. [Online]. Available: https: //www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[2] A. Keys. (1972) Indices of relative weight and obesity, Journal of Chronic Diseases, vol. 25, no. 6, pp. 329–343.

DOI: 10.54254/2753-8818/2025.AU27039

[3] Nuttall, Frank Q. (2015) Body Mass Index: Obesity, BMI, and Health: A Critical Review, Nutrition Today, vol. 50, no. 3, pp. 117–128.
[4] M. Safaei, E. A. Sundararajan, M. Driss, W. Boulila, and A. Shapi'i.(2021) A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity Computers in Biology and Medicine, vol. 136, p. 104754. [Online]. Available: https: //doi.org/10.1016/j.compbiomed.2021.104754
[5] UCI Machine Learning Repository. Estimation of obesity levels based on eating habits and physical condition dataset [Online]. Available: https: //archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition
[6] E. Alejo, E. García-Peñalvo, and M. J. Rodríguez-Conde. (2019) Estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico, Data in Brief, vol. 27, pp. 104344.
[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. (2002) SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, vol. 16, pp. 321–357.
[8] L. Breiman. (2001) Random Forests. Machine Learning, 45, 5–32.
[9] R. J. Ward et al. (2021) Genetic contributions to obesity and related traits, Obesity, vol. 29, no. 5, pp. 783–795.
[10] The Lancet Diabetes & Endocrinology Commission (2025) Definition and diagnostic criteria of clinical obesity, Lancet Diabetes Endocrinol, vol. 13, no. 2, pp. 89–101.
[11] National Health Commission of China. (2025) Guidelines on BMI and nutrition management in elderly adults WS/T 868-2025. [Online]. Available: https: //www.nhc.gov.cn/wjw/yingyang/202507/
[12] LeCun, Y., Bengio, Y. & Hinton. (2015) G. Deep learning. Nature, 521, 436–444.