

Behavioral and Lifestyle Determinants of Nutritional Status: A Comparative Machine Learning Analysis of BMI and WHR Among University Students

Manh Toan Nguyen

Hanoi University of Physical Education and Sport

Minh Nguyen Xuan

VNU International School

Minh Hai Bui

University of Science and Technology

Duy Manh Le

VNU University of Engineering and Technology

Hoang Dung Bui

University of Science and Technology

Xuan Hai Le

VNU International School

Viet Nga Nguyen Thi

Bac Ninh Sport University of Vietnam

Thu Nguyen Thi

`nguyenthithugdtc@vnu.edu.vn`

Physical Education and Sports Center

Research Article

Keywords: BMI classification, WHR prediction, Machine learning, Student health, Lifestyle factor

Posted Date: May 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9318316/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Behavioral and Lifestyle Determinants of Nutritional Status: A Comparative Machine Learning Analysis of BMI and WHR Among University Students

Nguyen Manh Toan^c, Nguyen Xuan Minh^a, Bui Minh Hai^b, Le Duy Manh^d, Bui Hoang Dung^b,
Le Xuan Hai^a, Nguyen Thi Viet Nga^c, Nguyen Thi Thu^{f,*}

^aVNU International School, Hanoi, Vietnam

^bUniversity of Science and Technology, Hanoi, Vietnam

^cBac Ninh Sport University of Vietnam, Bac Ninh, Vietnam

^dVNU University of Engineering and Technology, Hanoi, Vietnam

^eHanoi University of Physical Education and Sport, Hanoi, Vietnam

^fPhysical Education and Sports Center, Vietnam National University, Hanoi, Vietnam

Abstract

Background: Nutritional status among university students is influenced by multiple behavioral and lifestyle factors, yet the effectiveness of different anthropometric indicators in predictive modeling remains unclear. This study aims to evaluate machine learning approaches for predicting nutritional status using Body Mass Index (BMI) and Waist–Hip Ratio (WHR).

Methods: A cross-sectional dataset of 819 university students was collected through structured surveys, including demographic, anthropometric, health, dietary, and lifestyle variables. Multiple machine learning classifiers, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), were evaluated using Accuracy, weighted F1-score, and cross-validation techniques.

Results: BMI was more predictable than WHR across all models. Random Forest achieved the highest Accuracy for BMI prediction (0.7805) with strong generalization performance, while KNN obtained the highest weighted F1-score (0.6942). In contrast, WHR prediction showed lower performance, with a maximum Accuracy of 0.6829 across models. Feature importance analysis indicated that BMI was associated with a broader range of behavioral and lifestyle factors, whereas WHR depended on fewer and less consistent predictors.

Conclusions: The findings reveal structural differences between BMI and WHR in terms of predictability and associated factors. Machine learning models are more effective in capturing patterns related to BMI than WHR, suggesting that BMI may be a more suitable target for survey-based nutritional assessment. These results provide practical implications for model selection and feature design in public health research.

Keywords: BMI classification, WHR prediction, Machine learning, Student health, Lifestyle factor

*Corresponding author: nguyenthithugdtc@vnu.edu.vn

1. Introduction

The nutritional status of university students has become an increasingly important concern in public health, particularly in developing countries undergoing rapid urbanization and lifestyle transitions. University students represent a vulnerable population group due to significant changes in living environments, academic pressure, and daily habits. These transitions often lead to irregular dietary patterns, reduced physical activity, and unhealthy lifestyle behaviors, thereby increasing the prevalence of underweight, overweight, and obesity [1, 2]. Beyond immediate physical consequences, poor nutritional status is a well-established risk factor for chronic conditions such as cardiovascular diseases, type 2 diabetes, and metabolic disorders [3, 4].

Anthropometric indices such as Body Mass Index (BMI) and Waist–Hip Ratio (WHR) are widely used to assess nutritional status and associated health risks. BMI provides a general measure of body composition, while WHR reflects fat distribution, particularly central obesity, which is more closely linked to metabolic and cardiovascular complications [5, 6]. The joint use of BMI and WHR enables a more comprehensive assessment of both overall nutritional status and health risks related to fat distribution, making them suitable targets for predictive modeling.

A growing body of research has shown that nutritional status among students is influenced by a combination of behavioral and lifestyle factors. Dietary habits, including breakfast consumption, fast food intake, and sugar-sweetened beverage consumption, have been identified as key determinants [7, 8]. Physical activity plays an essential role in maintaining energy balance [9], while sleep patterns are associated with metabolic regulation and body weight outcomes [10, 11]. In addition, psychological factors such as stress and emotional eating behaviors further contribute to variations in dietary intake and nutritional status [12]. These findings indicate that nutritional status is a multifactorial problem involving complex interactions among diverse variables.

In Vietnam, a dual burden of undernutrition and overnutrition has been documented among university students and young adults, with rising rates of overweight and obesity occurring alongside micronutrient inadequacy and low fruit and vegetable intake [13, 14]. Variations in BMI distribution and associated lifestyle factors have been observed across different institutions [15, 16, 17, 18, 19]. However, most existing studies rely on conventional statistical methods, which are often limited in capturing nonlinear relationships and complex interactions among multiple influencing factors. Furthermore, WHR has been less frequently incorporated as a predictive target, despite its importance in assessing health risks related to fat distribution.

With the increasing availability of multidimensional survey data, Machine Learning (ML) has emerged as a promising approach for modeling complex health-related problems. Recent studies have demonstrated that classification algorithms such as Random Forest, Support Vector Machine, and K-Nearest Neighbors can effectively predict BMI and obesity categories using behavioral and demographic data among university students, often outperforming traditional statistical models [20, 21]. Unlike traditional methods, ML algorithms can effectively capture nonlinear relationships, handle high-dimensional feature spaces, and improve predictive performance without strong parametric assumptions [22]. Models such as Decision Tree [23], Random Forest [24], Support Vector Machine [25], and K-Nearest Neighbors [26] have been widely applied in classification tasks within healthcare and epidemiology.

Despite these advances, two key gaps remain. First, there is a lack of integrated studies that simultaneously consider multiple behavioral, lifestyle, and psychological factors in predicting both BMI and WHR. Second, systematic comparisons of different Machine Learning models on the same real-world dataset are still limited, making it difficult to identify the most effective approach for this problem.

To address these gaps, this study formulates nutritional status prediction as a supervised classification problem using real-world survey data collected from university students. The study aims to predict both BMI categories and WHR risk levels based on demographic, behavioral, and lifestyle features. Multiple Machine Learning models, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), are implemented and systematically evaluated using consistent performance metrics and cross-validation.

The main contributions of this study are as follows. First, it provides an integrated analysis of multiple factors influencing nutritional status, incorporating dietary behavior, physical activity, sleep, and psychological variables within a unified framework. Second, it presents a systematic comparison of diverse Machine Learning models for predicting both BMI and WHR, offering insights into their relative performance and suitability for different prediction tasks. These findings contribute to a better understanding of nutritional health among university students and support the development of data-driven intervention strategies.

2. Related Work

Studies on university students' nutritional status have examined a range of factors, from physical measurements to behavioral patterns, and increasingly, computational methods for prediction. Anthropometric indices, particularly Body Mass Index (BMI) and Waist–Hip Ratio (WHR), remain central to assessing nutritional status and health risks. While BMI is widely used to classify underweight, normal weight, overweight, and obesity, WHR provides additional insight into fat distribution, especially central adiposity, which is closely linked to cardiovascular and metabolic conditions [5, 6]. Research suggests that combining BMI and WHR offers a more comprehensive evaluation of health risk [3], yet many studies continue to rely solely on BMI, limiting the understanding of body fat-related risk. Recent work integrating waist circumference and hip circumference into predictive models for blood pressure among young adults further highlights the potential of WHR-related measures in capturing cardiometabolic risk [27].

Behavioral and lifestyle factors are also recognized as key determinants of nutritional outcomes. Frequent consumption of fast food, sugary beverages, and irregular meal patterns, such as skipping breakfast, have been consistently associated with overweight and obesity [7, 8]. Physical inactivity and insufficient sleep further exacerbate metabolic imbalance. Psychological factors, including stress and emotional eating, influence dietary decisions and caloric intake [12]. Most prior research, however, tends to analyze these factors in isolation, without considering their combined or interactive effects.

Analytical approaches in this area have evolved from traditional statistical methods to more sophisticated computational techniques. Classical methods like Chi-square tests and ANOVA are widely used for their interpretability and theoretical grounding but are limited in capturing nonlinear interactions and complex, high-dimensional relationships. Machine Learning models—including Logistic Regression, Decision Tree [23], Random Forest [24], Support Vector Machine, and K-Nearest Neighbors [26]—offer greater flexibility, allowing for the modeling of nonlinear patterns and interactions across multiple variables [22]. Recent studies have demonstrated that such algorithms can effectively classify BMI and obesity categories using behavioral and demographic data among university students, often outperforming traditional statistical models in predictive accuracy [20, 21]. More advanced frameworks, including deep hybrid learning with attention-enhanced feature fusion, have begun to explore BMI assessment from multiple anthropometric and lifestyle variables in student populations [28].

Despite these methodological advances, several gaps remain. Studies integrating multiple behavioral, lifestyle, and psychological factors are scarce, WHR is often underutilized as a target, and systematic comparisons of Machine Learning models under consistent conditions are limited. Addressing these gaps requires a comprehensive framework that incorporates diverse determinants and evaluates multiple predictive models on the same dataset, enabling a deeper understanding of factors affecting nutritional status and more robust predictive performance.

3. Method

This study formulates the prediction of nutritional status as a supervised multi-class classification problem. The primary objective is to predict two distinct but related target variables: Body Mass Index (BMI) categories—Thinness, Normal, Overweight, and Obese—and Waist-Hip Ratio (WHR) risk levels—Low, Moderate, and High. These targets are derived from anthropometric measurements commonly used in public health to assess overall nutritional status and central obesity, which is strongly associated with metabolic and cardiovascular risks. The predictive framework is built on a rich set of features, including demographic characteristics (e.g., gender, age, academic major, family background), dietary patterns (e.g., meal frequency, breakfast consumption, intake of fast food and sugar-sweetened beverages), lifestyle behaviors (e.g., physical activity level, screen time, sleep patterns), and psychological factors (e.g., stress, emotional eating behaviors under different emotional states).

A diverse set of five widely used Machine Learning models is employed to ensure a comprehensive evaluation of predictive performance. The models were selected to capture different learning paradigms, from linear and interpretable methods to nonlinear, instance-based, and ensemble approaches. This diversity allows for the assessment of each model’s ability to learn complex relationships, handle interactions among features, deal with high-dimensional and potentially noisy data, and generalize effectively to unseen samples. The chosen models include Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM).

3.1. Logistic Regression

Logistic Regression (LR) is a parametric, generalized linear model designed for probabilistic classification. It models the conditional probability of class membership by applying the sigmoid function to a linear combination of input features:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$$

where $\mathbf{w}$ is the weight vector representing the influence of each feature, and b is the bias term. For multi-class problems, the one-vs-rest (OvR) strategy is employed, wherein a separate binary classifier is trained for each class against all others. Model parameters are estimated by minimizing the cross-entropy loss function with L2 regularization:

$$\mathcal{L} = - \sum_{i=1}^n \sum_{k=1}^K y_{ik} \log(\hat{y}_{ik}) + \lambda \|\mathbf{w}\|_2^2$$

Logistic Regression is chosen for its simplicity, interpretability, and ability to quantify the contribution of individual features. Each coefficient directly reflects the effect of a demographic, behavioral, or psychological factor on the probability of belonging to a specific BMI or WHR

category. Moreover, its probabilistic outputs facilitate threshold-based decision-making, which is valuable in clinical and public health contexts where risk stratification is important. Despite its linear assumptions, LR provides a baseline for comparison with more complex nonlinear methods.

3.2. Decision Tree

Decision Tree (DT) is a non-parametric, hierarchical model that recursively partitions the feature space into regions that are increasingly homogeneous with respect to the target variable. At each internal node, a feature and a corresponding threshold are selected to maximize a splitting criterion, such as the Gini impurity:

$$Gini = 1 - \sum_{k=1}^K p_k^2$$

where p_k is the proportion of samples belonging to class k at the node. The splitting process continues until stopping criteria are met, such as a maximum tree depth, minimum number of samples per leaf, or minimal improvement in impurity reduction. Decision Trees are capable of capturing complex nonlinear relationships and interactions between features without requiring prior assumptions about data distribution or feature scaling. They are also highly interpretable, as the learned rules can be visualized as if-then statements, making it easier to communicate which factors contribute to BMI and WHR outcomes. However, single trees are prone to overfitting, particularly when the dataset is small or noisy, which necessitates careful tuning of hyperparameters such as tree depth and minimum leaf size.

3.3. Random Forest

Random Forest (RF) is an ensemble learning method that constructs a collection of Decision Trees trained on bootstrapped samples of the dataset. At each node split, only a random subset of features is considered, which introduces diversity among trees. Final predictions are obtained via majority voting across all trees:

$$\hat{y} = \text{mode}(y_1, y_2, \dots, y_T)$$

Random Forest mitigates overfitting relative to a single Decision Tree by averaging across multiple models, improving generalization performance. It is robust to noisy and high-dimensional data, capable of capturing nonlinear patterns and interactions among features. RF also provides feature importance metrics, either based on the mean decrease in impurity or permutation importance, which can identify the most influential factors in predicting BMI and WHR. This model is particularly suitable for complex public health datasets where interactions among diet, lifestyle, and psychological variables play critical roles.

3.4. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning method. For a query sample, the algorithm identifies the k closest training samples using a distance metric, typically Euclidean distance:

$$d(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^d (x_j - x_{ij})^2}$$

The predicted class is determined by majority vote among the k neighbors. KNN does not require an explicit training phase but relies on careful feature scaling to prevent bias from variables with different units. It is effective at capturing local, nonlinear relationships and subtle interactions that may be missed by global linear models. KNN is particularly useful for datasets with complex or irregular patterns, although its computational cost increases with dataset size. The choice of k influences model bias and variance, with smaller k capturing finer patterns and larger k producing smoother decision boundaries.

3.5. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a margin-based classifier that constructs an optimal hyperplane to separate classes while maximizing the margin between them. For soft-margin classification, the optimization problem is formulated as:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$$

subject to

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

where ξ_i are slack variables that allow some misclassification, and C controls the trade-off between margin width and training error. Only support vectors, the samples closest to the decision boundary, determine the hyperplane, which makes SVM robust to irrelevant features. A linear kernel is used to ensure computational efficiency and reduce the risk of overfitting, given the moderate number of features. SVM is selected for its strong generalization capability, performance in high-dimensional spaces, and ability to handle classes that are not completely linearly separable.

3.6. Rationale for Model Selection

The combined use of linear (Logistic Regression), nonlinear (Decision Tree, Random Forest), instance-based (KNN), and margin-based (SVM) models allows a thorough evaluation of predictive performance across different assumptions and learning paradigms. This approach enables comparison of how well each model handles feature interactions, nonlinearity, data heterogeneity, and class separability. It also provides insights into which factors—demographic, dietary, lifestyle, or psychological—are most predictive of BMI and WHR categories. Evaluating multiple models under consistent preprocessing and validation procedures ensures that the selected approach balances predictive accuracy, interpretability, and practical applicability for nutritional assessment and intervention planning.

4. Experiments

4.1. Dataset

The dataset was collected through a structured survey administered to 819 university students, capturing a broad spectrum of demographic, health-related, behavioral, and lifestyle characteristics. The variables included in the study are systematically organized into five main groups, as presented in Table 1. The sample exhibits a noticeable gender imbalance, with males accounting for 71.06% and females 28.94%. Demographic variables include gender, academic major,

birth order, and family size, providing contextual information about participants' socio-family background, which may influence dietary behaviors and lifestyle choices.

Anthropometric indicators are used to assess nutritional status through Body Mass Index (BMI) and Waist–Hip Ratio (WHR). The BMI distribution shows that most students fall within the normal category (73.26%), while thinness, overweight, and obesity account for 19.29%, 6.35%, and 1.1%, respectively. In contrast, WHR classification indicates that a considerable proportion of students are at moderate risk (67.16%), followed by high risk (17.34%) and low risk (15.51%). This suggests that although overall body weight appears largely normal, fat distribution may still present potential health concerns.

Table 1: Summary of variables included in the study

Variable Group	Variables
Demographics	Gender, Academic major, Birth order, Family size
Anthropometric	BMI category (Thinness, Normal, Overweight, Obese), WHR risk level (Low, Moderate, High)
Health	Chronic conditions (Respiratory, Digestive, Other), Self-perception of weight, Satisfaction with weight, Desire to change weight
Eating behavior	Emotional eating under different emotional states, Food preferences (Sweets, Fatty foods, Lean meat, Eggs, Seafood, Fruits & vegetables, Packaged foods, Fast food), Weekly eating frequency for various foods, Eating habits (following appetite, eating before sleep)
Lifestyle	Physical activity (extracurricular exercise), Duration of exercise, Exercise timing, Exercise location, Transportation to campus, Daily screen time

Health-related variables further describe participants' conditions, including self-reported chronic diseases, self-perception of body weight, satisfaction with current weight, and the desire to change weight. These variables provide additional behavioral and psychological context associated with nutritional status.

Eating-related factors capture multiple dimensions of dietary behavior. Emotional eating patterns are assessed under different psychological states such as stress, boredom, and happiness. Food preferences are recorded across various categories, including sweets, fatty foods, protein-rich foods, fruits and vegetables, packaged foods, and fast food. The dataset also includes weekly consumption frequency for key food groups such as grains, protein sources, dairy products, snacks, sugary beverages, and fried foods. General eating habits, including eating according to personal preference and consuming food before sleep, are also incorporated.

Lifestyle-related variables describe physical activity and daily routines, including participation in extracurricular exercise, duration and timing of exercise, preferred exercise locations, transportation modes to campus, and daily screen time. These variables reflect both physical activity levels and sedentary behavior patterns.

Based on these features, two target variables are defined: BMI categories (Thinness, Normal, Overweight, Obese) and WHR risk levels (Low, Moderate, High). This dual-target framework enables the analysis of both overall body composition and fat distribution, providing a comprehensive assessment of nutritional status and associated health risks.

4.2. Data Preprocessing and Evaluation

To ensure methodological rigor, variables directly associated with the prediction targets were removed to prevent data leakage. Specifically, for both BMI and WHR classification tasks, the following variables were excluded: weight, height, BMI values and categories, waist and hip measurements, WHR values and categories, and self-perception-related weight assessment questions (e.g., perceived weight status, satisfaction, and desire to change weight). These variables are either direct determinants or proxies of the target labels and would artificially inflate model performance if retained.

After feature selection, categorical variables such as gender, academic major, transportation mode, and exercise-related categories were encoded into numerical representations. Ordinal variables derived from Likert-scale responses (e.g., emotional eating behaviors) were normalized into the range $[0,1]$, ensuring consistency in scale and interpretability. Frequency-based variables (e.g., weekly food consumption) were also scaled proportionally to reflect relative behavioral intensity.

Continuous numerical features, including normalized anthropometric proxies, time-encoded variables, and aggregated behavioral indices, were standardized using the StandardScaler method. This transformation ensures that each feature has zero mean and unit variance, which is essential for distance-based and margin-based algorithms such as K-Nearest Neighbors (KNN) and Support Vector Machine (SVM). Without standardization, features with larger magnitudes would dominate the distance computation, leading to biased model behavior.

The dataset was split into 80% training and 20% testing subsets. Model performance was evaluated using multiple classification metrics, including Accuracy, weighted Precision, weighted Recall, and F1-score. The use of weighted metrics is particularly important due to potential class imbalance in BMI and WHR categories, where certain classes (e.g., overweight or high-risk groups) may be underrepresented. Weighted averaging ensures that each class contributes proportionally to its frequency, providing a more realistic assessment of model performance.

To further assess model robustness and generalization capability, five-fold cross-validation was applied on the training data. In this procedure, the training set is partitioned into five equal subsets; in each iteration, four subsets are used for training while the remaining subset is used for validation. This process is repeated five times each subset serves as the validation set once. The final performance is computed as the average across all folds, reducing variance caused by a single train-test split and providing a more stable estimate of model effectiveness.

4.3. Results

Table 2: Performance comparison for BMI classification

Model	Accuracy	F1-score	CV Mean
Logistic Regression	0.7317	0.6779	0.6550
Decision Tree	0.5732	0.6041	0.5450
Random Forest	0.7805	0.6896	0.7191
KNN	0.7561	0.6942	0.6885
SVM	0.7744	0.6913	0.7023

The comparative performance of machine learning models on BMI classification (Table 2) reveals clear differences in their ability to capture the underlying structure of the data. Ensemble and margin-based methods, particularly Random Forest and SVM, consistently outperform linear

and single-tree approaches. This pattern indicates that the relationship between input variables and BMI categories is inherently nonlinear and involves complex interaction effects that simpler models fail to capture effectively.

Random Forest achieves the highest Accuracy (0.7805) and cross-validation mean (0.7191), demonstrating both strong predictive capability and robustness to sampling variability. The relatively small gap between test accuracy and cross-validation performance suggests limited overfitting and good generalization. SVM also performs competitively (Accuracy 0.7744), confirming that BMI classes are separable in a higher-dimensional feature space. KNN, while slightly lower in accuracy (0.7561), achieves the highest F1-score (0.6942), indicating better balance between precision and recall, particularly for minority classes. This suggests that local neighborhood structure plays a significant role in BMI classification.

In contrast, Logistic Regression exhibits moderate performance (Accuracy 0.7317), reflecting its limitation in modeling nonlinear relationships. Decision Tree performs the worst (Accuracy 0.5732), with a low cross-validation mean (0.5450), indicating high variance and instability.

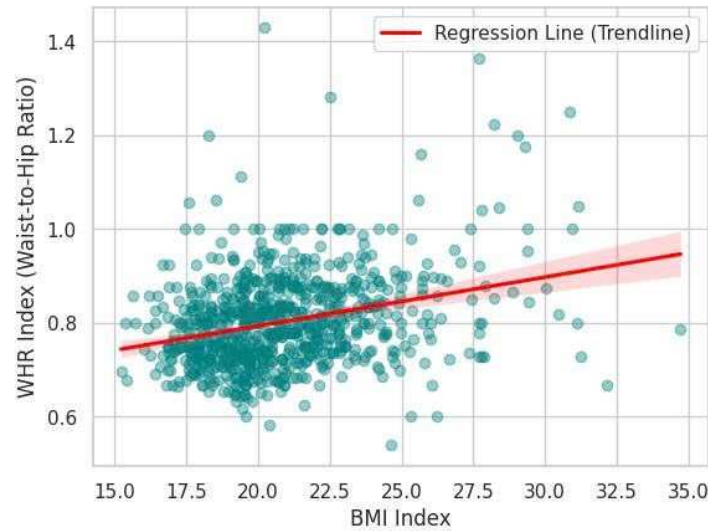


Figure 1: Scatter plot illustrating the relationship between BMI and WHR. Each point represents an individual observation, while the regression line indicates the overall positive trend.

Feature importance analysis for BMI classification (Figure 2) provides deeper insight into model behavior. Tree-based models emphasize demographic and lifestyle-related variables. Birth Year and Favorite Sport appear consistently important, suggesting the role of age and physical activity preferences. Sleep-related variables (e.g., Nap Wakeup, Bedtime, Wakeup Time) contribute moderately, indicating indirect influence through daily routines.

KNN emphasizes behavioral and psychological eating patterns. Food Neophilia is the most influential variable, followed by All-day Eating Capacity, Daily Screen Time, Snack Impact on Main Meals, and Positive Emotion Eating. These results indicate that BMI is closely associated with complex behavioral profiles rather than isolated dietary components.

Linear models distribute importance more evenly across variables. Logistic Regression highlights dietary-related variables such as sausage intake, packaged food, lean meat, and fast food,

while SVM assigns similar weights across eating behavior and lifestyle variables. This uniform distribution limits their ability to capture dominant nonlinear effects.

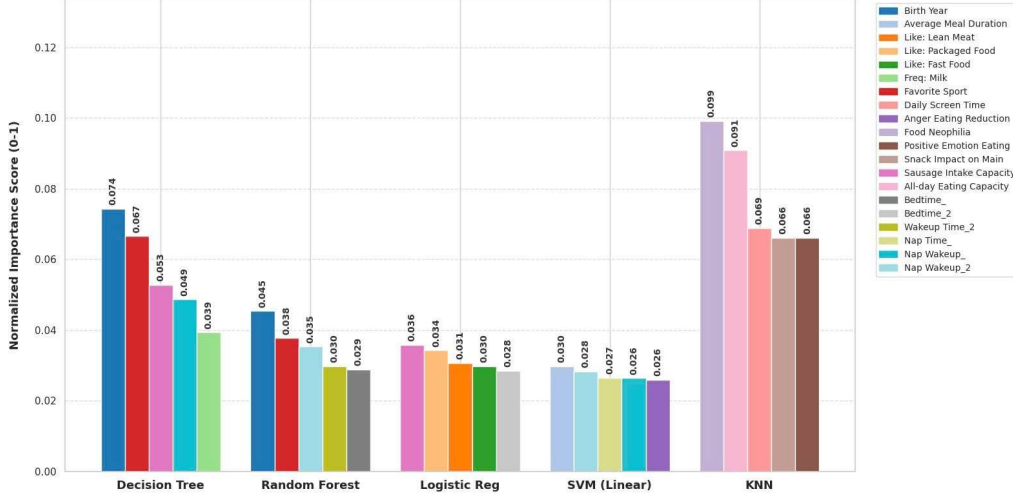


Figure 2: Feature importance comparison for BMI classification across models.

Table 3: Performance comparison for WHR classification

Model	Accuracy	F1-score	CV Mean
Logistic Regression	0.5793	0.5247	0.6153
Decision Tree	0.5427	0.5570	0.4901
Random Forest	0.6829	0.5543	0.6656
KNN	0.6402	0.5652	0.5893
SVM	0.6829	0.5543	0.6595

For WHR classification (Table 3), all models exhibit lower performance compared to BMI, confirming that WHR is more difficult to predict. Random Forest and SVM achieve the highest Accuracy (0.6829), but their F1-scores remain moderate, indicating difficulty in handling class imbalance.

KNN achieves the highest F1-score (0.5652), suggesting its effectiveness in capturing local behavioral patterns. Logistic Regression and Decision Tree perform poorly, reflecting weak linear relationships and instability.

Feature importance analysis for WHR (Figure 3) reveals a more concentrated and model-dependent structure. KNN shows two dominant variables: Eating When Full and Wakeup Time, followed by preference for fatty food, night snacking, and protein intake frequency. This indicates reliance on a small number of strong behavioral signals.

SVM captures broader behavioral and social factors such as eating capacity, sugary drink craving, and family-related variables, but with moderate importance values. Random Forest distributes importance across many weak predictors, indicating the absence of dominant features. Logistic Regression shows the lowest importance values overall, confirming weak linear relationships.

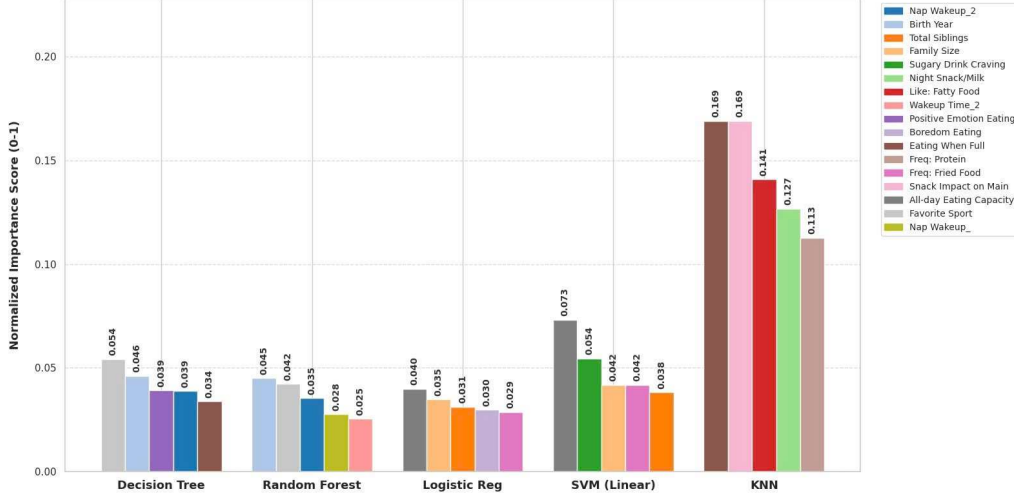


Figure 3: Feature importance for WHR classification across models.

Overall, BMI is influenced by a broad set of interacting variables, particularly complex eating behaviors and lifestyle patterns, resulting in strong nonlinear relationships. In contrast, WHR depends on fewer dominant but less stable factors, leading to lower predictive performance and greater variability across models.

5. Discussion

The experimental results are consistent with the patterns observed in the Results section and provide meaningful insights into the predictive mechanisms of BMI and WHR. A key finding is the clear performance gap between the two targets, indicating that BMI is structurally easier to model than WHR due to stronger and more stable relationships with observed features.

For BMI classification, Random Forest achieves the best overall performance (Accuracy 0.7805; CV Mean 0.7191), confirming its effectiveness in capturing nonlinear interactions among demographic, behavioral, and lifestyle variables. This aligns with the theoretical advantages of ensemble learning, where multiple trees reduce variance and improve generalization [24]. The relatively small difference between test accuracy and cross-validation results suggests that the model maintains stability and avoids overfitting. SVM also demonstrates strong performance (Accuracy 0.7744), supporting its capability to construct optimal separating hyperplanes in complex feature spaces [25].

KNN achieves the highest F1-score (0.6942), indicating better balance between precision and recall. As shown in the Results, KNN heavily relies on behavioral variables such as Food Neophilia, All-day Eating Capacity, and emotional eating patterns. This suggests that BMI is strongly associated with behavioral similarity, where individuals with comparable eating habits and lifestyles form local clusters in the feature space. This observation is consistent with the fundamental principle of instance-based learning [26], where classification depends on proximity in feature space rather than global structure.

Logistic Regression shows moderate performance, reflecting its limitation in modeling non-linear relationships. Decision Tree performs the worst due to high variance and instability, which is consistent with its sensitivity to noise and tendency to overfit when used as a single model [23]. The absence of dominant features in linear and single-tree models further explains their lower predictive performance.

Importantly, the feature importance analysis indicates that BMI is influenced by a wide range of interacting variables. Both demographic factors (e.g., Birth Year) and lifestyle variables (e.g., sleep patterns, physical activity preferences) contribute meaningfully, while behavioral eating patterns dominate in KNN. These findings are consistent with previous studies on university students, which emphasize that nutritional status is shaped by multiple factors including diet, lifestyle, and daily habits [1, 8]. Studies in Vietnam also report similar trends, where BMI and nutritional status are associated with lifestyle and behavioral factors among students [15, 19, 17].

In contrast, WHR classification is significantly more challenging, as reflected by lower performance across all models (maximum Accuracy 0.6829). This supports the observation in the Results that WHR exhibits weaker and less consistent relationships with the available features. While Random Forest and SVM achieve the highest accuracy, their F1-scores remain moderate, indicating difficulty in correctly classifying minority classes.

A key distinction revealed in the Results is the difference in feature importance structure. For WHR, KNN shows a highly concentrated importance distribution, with dominant variables such as Eating When Full and Wakeup Time. This indicates that WHR prediction depends on a small number of strong behavioral signals in localized regions of the data. However, other models (Random Forest, SVM) display more distributed and low-magnitude importance values, suggesting that WHR lacks stable global predictors. This inconsistency explains both the lower predictive performance and reduced generalization.

The weaker predictability of WHR is consistent with prior research, which shows that fat distribution is influenced by more complex physiological and metabolic factors compared to overall body mass [3, 9]. Unlike BMI, which is strongly associated with dietary intake and lifestyle behaviors, WHR reflects central fat accumulation and may depend on factors not captured in survey-based data, such as genetics or hormonal regulation. This also explains the large dispersion observed in the BMI–WHR relationship, where individuals with similar BMI can have significantly different WHR values.

The role of lifestyle factors such as sleep and daily routines, highlighted in the Results through moderate feature importance, is also supported by previous studies. Research has shown that sleep quality and duration are associated with nutritional status and metabolic outcomes in students [10, 11]. Similarly, dietary behaviors such as snacking, meal timing, and food preferences have been linked to obesity and nutritional imbalance [7, 29]. These findings reinforce the interpretation that behavioral variables contribute significantly to BMI prediction but are insufficient to fully explain WHR variation.

Overall, the results demonstrate that BMI is influenced by a broad and interacting set of behavioral and lifestyle factors, leading to strong nonlinear relationships that can be effectively captured by ensemble and margin-based models. In contrast, WHR depends on fewer but less stable predictors, resulting in weaker patterns and lower model performance.

Several limitations should be acknowledged. First, the dataset exhibits demographic imbalance, which may affect generalizability. Second, the use of self-reported survey data introduces potential bias and measurement error. Third, the absence of biological and clinical variables limits the ability to model complex targets such as WHR. These limitations are consistent with challenges reported in nutritional studies among university populations [30, 18].

Future research should focus on improving both data and modeling approaches. Incorporating more comprehensive variables, including physiological and longitudinal data, may enhance WHR prediction. Techniques for handling class imbalance could improve model performance, particularly for minority groups. Additionally, advanced models such as gradient boosting may further capture subtle nonlinear interactions.

In summary, the findings confirm that model effectiveness depends strongly on the structural properties of the target variable. BMI can be predicted with relatively high accuracy using behavioral and lifestyle features, while WHR remains a more complex and less predictable indicator due to its dependence on latent and weakly observable factors.

Declarations

Ethics approval and consent to participate

This study was reviewed and approved by the Institutional Review Board (IRB) of the University of Medicine and Pharmacy, Vietnam National University, Hanoi. The IRB operates under the code IRB-VN 01016 (granted by the Ministry of Health of Vietnam) and IRB00013221 (granted by the U.S. Department of Health and Human Services – OHRP). Ethical approval was granted on October 31, 2022.

All procedures involving human participants were conducted in accordance with the ethical standards of the institutional research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Written informed consent was obtained from all participants prior to their inclusion in the study.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by the authors. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

- [1] B. Alipour, et al., Nutritional status and its determinants among university students, PLoS One 10 (5) (2015) e0127198.

- [2] A. S. Al-Rethaiaa, et al., Obesity and overweight prevalence among university students: a cross-sectional study, *Nutrition in Clinical Practice* 36 (3) (2021) 712–720.
- [3] H. Al-Domi, et al., Metabolic syndrome and nutritional status, *Best Practice & Research Clinical Endocrinology & Metabolism* 27 (2) (2013) 251–261.
- [4] A. S. Al-Rethaiaa, et al., Health care utilization and nutritional status among university students, *INQUIRY* 60 (2023).
- [5] B. Ganesh, et al., Prevalence and risk factors of overweight/obesity among school children in United Arab Emirates: a cross-sectional study, *Journal of Family Medicine and Primary Care* 10 (10) (2021) 3725–3731.
- [6] M. Al-Thani, et al., Prevalence of overweight/obesity, anaemia and their associations among female university students in Dubai, United Arab Emirates, *Journal of Nutritional Science* 8 (2019) e38.
- [7] H. Al-Domi, et al., Breakfast habits and nutritional status in university students, *International Journal of Food Sciences and Nutrition* 54 (5) (2003) 403–413.
- [8] H. M. Al-Hazaa, et al., Dietary habits and obesity among university students, *Preventive Medicine* 51 (2) (2010) 123–128.
- [9] B. Alipour, et al., Physical activity and nutritional status in students, *Journal of Research in Health Sciences* 14 (3) (2014).
- [10] H. M. Al-Hazaa, et al., Sleep quality and its association with nutrition status among university students, *Sleep Health* 3 (6) (2017) e301–e302.
- [11] A. S. Al-Rethaiaa, et al., Sleep duration and its association with nutritional status among adolescents, *Metabolism* 85 (2018) 94–101.
- [12] B. Alipour, et al., The relationship between nutritional status and quality of life in university students, *PLoS One* 15 (11) (2020) e0242751.
- [13] T. T. Thuy, et al., Nutritional status and associated factors among university students in Vietnam: A cross sectional study, *Global Health Action* 14 (1) (2021) 1–10.
- [14] K. V. Nguyen, et al., Differential nutrient inadequacy among Vietnamese youth: A national cross sectional analysis, *Public Health Nutrition* (2025).
- [15] N. H. Long, H. M. Tuan, N. T. Trung, N. T. Son, D. C. Nhu, Quality of life and nutrition status among first-year students of Vietnam National University, *Tạp chí Y học dự phòng* 24 (6) (2014) 96–102.
- [16] H. L. Nguyễn, et al., Tình trạng dinh dưỡng và chất lượng cuộc sống của sinh viên năm thứ nhất Đại học Quốc gia Hà Nội, *Tạp chí Y học dự phòng* (2014).
- [17] H. N. Q. Trương, et al., Nghiên cứu các chỉ điểm nhân trắc học và tình hình dinh dưỡng của sinh viên khóa 26 Đại học Duy Tân, *Vietnam Journal of Diabetes and Endocrinology* (49) (2022) 104–109.

- [18] N. Thị Pháp, T. Thị Vân Khanh, P. Văn Phú, Tình trạng dinh dưỡng và một số yếu tố liên quan của sinh viên y khoa năm thứ 4 và thứ 5 tại Đại học Tây Nguyên, Tạp chí Y học Việt Nam 512 (2) (2022).
- [19] H. T. L. Ngọc, N. T. T. Hoa, L. T. Hương, Tình trạng dinh dưỡng và một số yếu tố liên quan của sinh viên Y1 Trường Đại học Y Hà Nội, Tạp chí Nghiên cứu Y học 146 (10) (2021) 192–197.
- [20] R. Baharudin, et al., Body mass index classification using machine learning algorithms among university students, International Journal of Advanced Computer Science and Applications 13 (6) (2022) 123–130.
- [21] A. G. Silva, et al., Obesity classification: a comparative study of machine learning algorithms using lifestyle and demographic data, Scientific Reports 11 (2021) 12345.
- [22] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, 2009.
- [23] J. R. Quinlan, Induction of Decision Trees, Machine Learning 1 (1) (1986) 81–106.
- [24] L. Breiman, Random Forests, Machine Learning 45 (1) (2001) 5–32.
- [25] C. Cortes, V. Vapnik, Support-Vector Networks, Machine Learning 20 (3) (1995) 273–297.
- [26] T. Cover, P. Hart, Nearest Neighbor Pattern Classification, IEEE Transactions on Information Theory 13 (1) (1967) 21–27.
- [27] E. Lima, et al., Predicting increased blood pressure using machine learning, Computational and Mathematical Methods in Medicine 2014 (2014) 1–8.
- [28] Y. Li, et al., A deep hybrid learning framework with attention enhanced feature fusion for body mass index assessment in university students, Computers in Biology and Medicine 169 (2025) 108321.
- [29] A. S. Al-Rethaiaa, et al., Food insecurity and nutritional status among students, International Journal of Adolescent Medicine and Health 34 (4) (2022).
- [30] V. Thị Nhung, Tình trạng dinh dưỡng và một số chỉ số sinh học của sinh viên đại học chính quy năm thứ 2 Trường Đại học Điều dưỡng Nam Định, Tạp chí Y học Việt Nam 516 (1) (2022).