



Obesity classification: a comparative study of machine learning models excluding weight and height data

Ahmed Cihad Genc^{1,2*} , Erkut Arıcan³ 

SUMMARY

OBJECTIVE: Obesity is a global health problem. The aim is to analyze the effectiveness of machine learning models in predicting obesity classes and to determine which model performs best in obesity classification.

METHODS: We used a dataset with 2,111 individuals categorized into seven groups based on their body mass index, ranging from average weight to class III obesity. Our classification models were trained and tested using demographic information like age, gender, and eating habits without including height and weight variables.

RESULTS: The study demonstrated that when trained on demographic information, machine learning can classify body mass index. The random forest model provided the highest performance scores among all the classification models tested in this research.

CONCLUSION: Machine learning methods have the potential to be used more extensively in the classification of obesity and in more effective efforts to combat obesity.

KEYWORDS: Artificial intelligence. Classification. Machine learning. Obesity.

INTRODUCTION

Across the world, the two approaches that are mostly used to determine obesity include waist circumference and body mass index (BMI) screenings. BMI is an index directly calculated using the variables of height and weight¹. BMI is ubiquitous in clinical practice due to its ease of use. It correlates with increased body weight as well as levels of body fat. BMI is used as an indirect and easily obtained marker to get body fat and usually gives an even better estimation of fat than weighing². The prevalence of obesity has increased rapidly in the last decade, making it a public health problem that also increases comorbidities such as diabetes, hypertension, and stroke³. BMI effectively categorizes individuals into seven distinct levels, beginning with insufficient weight for BMI below 18.5, followed by normal weight ranging from 18.5 to 24.9, overweight level I between 25 and 29.9, overweight level II from 30 to 34.9, obesity type I from 35 to 39.9, obesity type II ranging from 40 to 44.9, and continuing with obesity type III for BMI values of 45 and above⁴. Comorbidities and severity vary depending on the obesity level⁵.

Artificial intelligence has gained global recognition and includes machine learning (ML), which utilizes sophisticated neural networks. ML can be categorized into two primary types,

with unsupervised learning functioning without labeled data, while supervised learning relies on labeled data for guidance⁶.

Among the ML models that perform classification, one is based on logistic regression⁷. Logistic regression is a statistical method employed to predict the likelihood of a specific outcome by analyzing one or more predictor variables⁸. Classification and regression trees are collective under the tree-based model. They stand out for their readability and can be used on numerical and categorical data⁹. Random forests, which belong to the ensemble learning techniques, are quite useful in supervised learning due to their high accuracy and efficiency in handling large datasets with many features¹⁰. Support vector machines (SVMs) are highly efficient for classification and regression, particularly in handling numerous features and complex nonlinear relationships between features and targets¹¹. Further mathematical development describes SVMs and, finally, the ways in which, guided by inside support vectors, they maximize the entire width margin that separates the classes¹². K-Nearest Neighbors (KNN) assigns labels to unlabeled data points based on their nearest neighbors in a labeled dataset and, despite not requiring knowledge of joint distributions, can achieve an error probability at least as low as the Bayes error rate, with the potential for twice the

¹Bahcesehir University, Graduate School, Department of Artificial Intelligence – İstanbul, Türkiye.

²Sakarya University, Faculty of Medicine, Department of Internal Medicine – Sakarya, Türkiye.

³Bahcesehir University, Department of Computer Engineering, İstanbul, Türkiye.

*Corresponding author: genccihad@gmail.com

Conflicts of interest: the authors declare there is no conflicts of interest. Funding: none.

Received on July 29, 2024. Accepted on October 13, 2024.

Bayes error rate in multi-category situations. This indicates its effectiveness even with small sample sizes¹³. KNN is also one of the most conventional nonparametric regression techniques because of its ease of use and high accuracy¹⁴. The KNN algorithm is a simple technique that depends on k nearest data points of a given dataset to predict new observations¹⁵. Naive Bayes is a straightforward but effective probabilistic classifier that uses Bayes' theorem to classify events and predict them even when the assumption of feature independence is almost false^{16,17}. Some of the most essential AI techniques are neural networks, which mimic the individual's brain and help learn and adapt to various tasks in dealing with data^{18,19}.

Aim

This study investigates the capabilities of ML models to predict obesity and its levels without using height and weight variables, thereby improving obesity management and prevention strategies.

METHODS

A total of 2,111 individuals, categorized into seven groups based on BMI, form the open-access obesity dataset from Peru, Mexico, and Colombia, which estimates obesity classes using physical and dietary indicators²⁰. The dataset contains 17 variables, such as age, gender, height, weight, family history of overweight, frequency of physical activity (FAF), frequent consumption of high-caloric food (FAVC), daily consumption of water (CH_2O), frequent consumption of vegetables (FCVC), number of meals per day (NCP), smoking, consumption of alcohol (CALG), consumption of food between meals (CAEC), caloric consumption monitoring (SCC), time using technology devices (TAU), type of transportation used (MTRANS), and obesity classes (NOBeyesdad). Obesity as a target variable is classified into different categories, including underweight, average weight, overweight (Level I and Level II), and obesity (Type I, Type II, and Type III). The flowchart of the study is visualized in Figure 1.

Data preprocessing

The dataset undergoes thorough preprocessing, including handling missing values, encoding categorical variables, normalizing and scaling numerical data, and selecting key features to improve model performance. Additionally, outlier detection and treatment are performed to ensure data accuracy, with adjustments to mitigate any undue influence on the analysis.

Machine learning model

A 20-fold cross-validation is applied, where the dataset is split into 20 parts, and the model is trained and tested 20 times.

Each subset is used once as validation data to prevent overfitting, and the final performance estimate is obtained by averaging the results from all rounds. The two most influential parameters in the model are identified using One Rule (OneR). Seven different ML models are applied to the classification tasks. Seven different ML models are applied to the classification tasks. ML classification models such as decision trees, SVM, KNN, neural networks, logistic regression, random forest, and Naive Bayes have been applied. Specific settings such as batch size, base classifiers, distance metrics, and kernel functions are adjusted to enhance model performance across different algorithms.

Used software and statistical analysis

Various ML methods are applied to classify a seven-category variable using WEKA, version 3.9.6, and Python, version 3.8.10. Libraries such as sci-kit-learn and Pandas are used for data manipulation and analysis, and tools like Matplotlib are used for result visualization. Descriptive analyses, including visual and analytical techniques, summarize the characteristics of the study population. The numerical data are expressed as mean and standard deviations, and the categorical data are expressed as frequencies and proportions.

RESULTS

The study population consists of 1,068 males (50.6%) and 1,043 females (49.4%), with a mean age of 24.3 ± 6.3 years. The mean height is 1.7 ± 0.09 m, and the mean weight is 86.6 ± 26.2 kg. A family history of overweight is present in 81.7% of participants, and 88.4% report FAVC. The NCP is 2.69 ± 0.78 , and the mean CH_2O is 2 ± 0.6 L. The FAF is low, with a mean of 1.01 ± 0.85 , while the mean TAU is 0.66 ± 0.60 h. Public

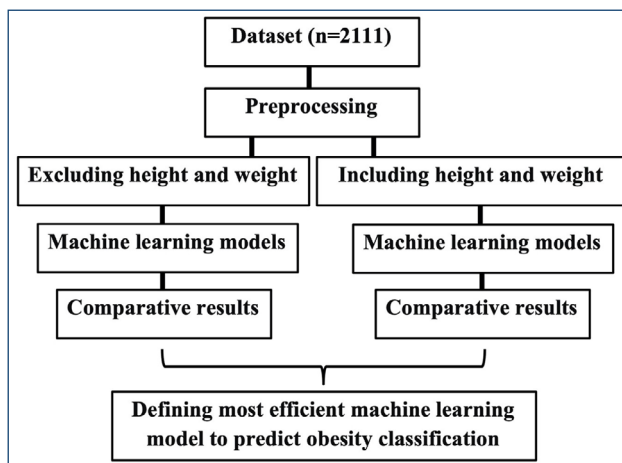


Figure 1. Flowchart of the study.

transportation is the most common mode of transport, used by 74.9% of the study population. Smoking history is reported by 2.1% of individuals. Regarding BMI classifications, 12.9%

Table 1. Descriptive statistics of the study population of 2,111 people: Demographic characteristics, behavioral characteristics, and body mass index classifications.

Variable	Mean±SD/n (%) n=2,111
Age, years	24.3±6.3
Gender	
Male	1,068 (50.6%)
Female	1,043 (49.4%)
Height, m	1.7±0.09
Weight, kg	86.6±26.2
Family history of overweight, yes	1,726 (81.7%)
Frequent consumption of high-caloric food (FAVC), yes	1,866 (88.4%)
Frequent consumption of vegetables (FCVC), frequent	2.42±0.534
Number of meals per day (NCP), number	2.69±0.78
Consumption of food between meals (CAEC)	
Sometimes	1,765 (83.6%)
Frequently	242 (11.5%)
Always	53 (2.5%)
No	51 (2.4%)
Smoking history, yes	44 (2.1%)
Daily consumption of water (CH ₂ O), L	2±0.6
Caloric consumption monitoring (SCC), yes	96 (4.5%)
Frequency of physical activity (FAF), frequency	1.01±0.85
Time using technology devices (TAU), h	0.66±0.60
Consumption of alcohol (CALG)	
No	639 (30.3%)
Sometimes	1,401 (66.4%)
Frequently	70 (3.3%)
Always	1 (<0.1%)
Type of transportation used (MTRANS)	
Public transportation	1,580 (74.9%)
Walk	56 (2.7%)
Automobile	457 (21.6%)
Motorbike	11 (0.5%)
Bike	7 (0.3%)
Body mass index (BMI) classes	
Insufficient weight	272 (12.9%)
Normal weight	287 (13.6%)
Overweight level I	290 (13.7%)
Overweight level II	290 (13.7%)
Obesity type I	351 (16.6%)
Obesity type II	297 (14.1%)
Obesity type III	324 (15.4%)

SD: standard deviation; FAVC: frequent consumption of high-caloric food; FCVC: frequent consumption of vegetables; NCP: number of meals per day; CAEC: consumption of food between meals; SCC: caloric consumption monitoring; FAF: frequency of physical activity; TAU: time using technology devices; CALG: consumption of alcohol; MTRANS: type of transportation used; BMI: body mass index.

of the study population falls into the insufficient weight category, 13.6% falls into the normal weight category, and the remainder are classified as overweight or obese, with Obesity Type I being the most prevalent at 16.6% (Table 1).

The OneR model with all variables has an Area Under the Receiver Operating Characteristic Curve (ROC area) of 0.81, a sensitivity of 0.67, a specificity of 0.94, a precision of 0.67, and a recall of 0.67. Without height and weight variables, the ROC area is 0.67, the sensitivity is 0.44, the specificity is 0.90, the precision is 0.42, and the recall is 0.44. The logistic regression classification model with all variables achieves an ROC area of 0.98, a sensitivity of 0.84, a specificity of 0.97, a precision of 0.84, and a recall of 0.84. Without height and weight variables, the ROC area is 0.91, the sensitivity is 0.66, the specificity is 0.94, the precision is 0.65, and the recall is 0.66. The decision tree model with all variables shows an ROC area of 0.98, a sensitivity of 0.94, a specificity of 0.99, a precision of 0.94, and a recall of 0.94. Without height and weight variables, the ROC area is 0.90, the sensitivity is 0.79, the specificity is 0.96, the precision is 0.79, and the recall is 0.79. The SVM model with all variables reports an ROC area of 0.96, a sensitivity of 0.85, a specificity of 0.97, a precision of 0.85, and a recall of 0.85. Without height and weight variables, the ROC area is 0.86, the sensitivity is 0.63, the specificity is 0.94, the precision is 0.63, and the recall is 0.63. The KNN model with all variables shows an ROC area of 0.90, a sensitivity of 0.82, a specificity of 0.97, a precision of 0.82, and a recall of 0.82. Without height and weight variables, the ROC area is 0.88, the sensitivity is 0.78, the specificity is 0.96, the precision is 0.78, and the recall is 0.78. The Naive Bayes model with all variables achieves an ROC area of 0.93, a sensitivity of 0.67, a specificity of 0.94, a precision of 0.66, and a recall of 0.67. Without height and weight variables, the ROC area is 0.86, the sensitivity is 0.55, the specificity is 0.92, the precision is 0.55, and the recall is 0.55. The neural network classification model with all variables presents an ROC area of 0.99, a sensitivity of 0.94, a specificity of 0.99, a precision of 0.94, and a recall of 0.94. Without height and weight variables, the ROC area is 0.86, the sensitivity is 0.61, the specificity is 0.94, the precision is 0.60, and the recall is 0.61. The random forest model with height and weight variables included presents an ROC area of 1.00, a sensitivity of 0.96, a specificity of 0.99, a precision of 0.96, a recall of 0.96, an F-score of 0.96, a Matthews correlation coefficient (MCC) of 0.95, and a precision–recall curve (PRC) of 0.99. Without height and weight variables, the ROC area is 0.98, the sensitivity is 0.87, the specificity is 0.98, the precision is 0.88, the recall is 0.87, the F-score is 0.87, the MCC is 0.85, and the PRC is 0.92 (Table 2).

Table 2. Comparison of performance measurements of various machine learning models with and without height and weight variables.

Model, N=2,111	ROC area	Sensitivity	Specificity	Precision	Recall	F-Score	MCC	PRC
OneR (age)								
With height and weight	0.81	0.67	0.94	0.67	0.67	0.67	0.61	0.51
Without height and weight	0.67	0.44	0.90	0.42	0.44	0.40	0.33	0.28
Logistic regression								
With height and weight	0.98	0.84	0.97	0.84	0.84	0.84	0.81	0.88
Without height and weight	0.91	0.66	0.94	0.65	0.66	0.64	0.59	0.66
Decision tree								
With height and weight	0.98	0.94	0.99	0.94	0.94	0.94	0.93	0.91
Without height and weight	0.90	0.79	0.96	0.79	0.79	0.79	0.75	0.73
Random forest								
With height and weight	1.00	0.96	0.99	0.96	0.96	0.96	0.95	0.99
Without height and weight	0.98	0.87	0.98	0.88	0.87	0.87	0.85	0.92
SVM								
With height and weight	0.96	0.85	0.97	0.85	0.85	0.85	0.82	0.79
Without height and weight	0.86	0.63	0.94	0.63	0.63	0.62	0.56	0.53
KNN								
With height and weight	0.90	0.82	0.97	0.82	0.82	0.82	0.79	0.72
Without height and weight	0.88	0.78	0.96	0.78	0.78	0.78	0.75	0.67
Naive Bayes								
With height and weight	0.93	0.67	0.94	0.66	0.67	0.67	0.61	0.72
Without height and weight	0.86	0.55	0.92	0.55	0.55	0.52	0.47	0.57
Neural network								
With height and weight	0.99	0.94	0.99	0.94	0.94	0.94	0.93	0.97
Without height and weight	0.86	0.61	0.94	0.60	0.61	0.60	0.54	0.58

ROC area: area under the receiver operating characteristic curve; MCC: Matthews correlation coefficient; PRC: precision-recall curve; OneR: one rule; SMV: support vector machine; KNN: K-nearest neighbors.

DISCUSSION

The fact that 50.6% of 2,111 people are male and 49.4% are female and the number of people in the seven groups in the BMI classification, which is the outcome variable, is almost equal increases the confidence in the results of the study. By determining which ML model achieves the highest performance based on daily lifestyle data without knowing height and weight values, we are offering a novel approach to combating obesity.

Some gynecological diseases may increase the risk of obesity²¹. Medical and surgical interventions and lifestyle alterations are also effective in obesity management. Bariatric surgery improves health outcomes and reduces long-term healthcare costs, despite initial high costs^{22,23}. Bariatric surgery may be considered for infertility in women with morbid

obesity²⁴. Liraglutide, in conjunction with intragastric balloon placement, results in better weight loss than intragastric balloon placement alone, and there is no considerable difference in the results between males and females²⁵. ML studies on obesity, which have various treatment alternatives, have become more critical.

A review of childhood obesity discusses the potential of ML to improve traditional prevention and treatment methods. The analysis shows that ML has significant potential to enhance strategies for managing pediatric obesity²⁶. The study used ML techniques to predict metabolic syndrome, focusing on SVM and decision tree models. The findings demonstrated that SVM achieved better performance with a sensitivity of 0.774 and a specificity of 0.74, compared to the decision tree, which recorded a sensitivity of 0.76 and a specificity of 0.72²⁷. In our study, although the SVM was

not the most successful model, it still achieved a sensitivity of 0.63 and a specificity of 0.94 for obesity classification without using height and weight variables.

The analysis with a similar dataset also shows volumes about the need to perform data preprocessing in an ML analysis, specifically data integration and cleaning. Out of the models tried out, which include SVM, random forests, and decision trees, the random forest model was the most accurate, with a 96% accuracy rate²⁸. The above study used the same dataset as the current study but with different methods and results. BMI is one of the main targets for prediction, and it can be obtained based on height and weight; however, we did not include these variables in the dataset to make the objective for ML models more attractive. This exclusion resulted in the expected reduction of prediction accuracy. The OneR method identified weight as the most significant variable; nevertheless, age was the most critical factor when height and weight were excluded. As for the differences between the two studies, our work demonstrates methodological and result-oriented variation even when working with the same dataset.

REFERENCES

1. Chen K, Shen Z, Gu W, Lyu Z, Qi X, Mu Y, et al. Prevalence of obesity and associated complications in China: a cross-sectional, real-world study in 15.8 million adults. *Diabetes Obes Metab*. 2023;25(11):3390-9. <https://doi.org/10.1111/dom.15238>
2. Mei Z, Grummer-Strawn LM, Pietrobello A, Goulding A, Goran MI, Dietz WH. Validity of body mass index compared with other body-composition screening indexes for the assessment of body fatness in children and adolescents. *Am J Clin Nutr*. 2002;75(6):978-85. <https://doi.org/10.1093/ajcn/75.6.978>
3. Elmaleh-Sachs A, Schwartz JL, Bramante CT, Nicklas JM, Gudzone KA, Jay M. Obesity management in adults: a review. *JAMA*. 2023;330(20):2000-15. <https://doi.org/10.1001/jama.2023.19897>
4. Nam GE, Kim YH, Han K, Jung JH, Rhee EJ, Lee WY, et al. Obesity fact sheet in Korea, 2020: prevalence of obesity by obesity class from 2009 to 2018. *J Obes Metab Syndr*. 2021;30(2):141-8. <https://doi.org/10.7570/jomes21056>
5. Kim KK, Haam JH, Kim BT, Kim EM, Park JH, Rhee SY, et al. Evaluation and treatment of obesity and its comorbidities: 2022 update of clinical practice guidelines for obesity by the Korean society for the study of obesity. *J Obes Metab Syndr*. 2023;32(1):1-24. <https://doi.org/10.7570/jomes23016>
6. Kufel J, Bargieł-Łączek K, Kocot S, Koźlik M, Bartnikowska W, Janik M, et al. What is machine learning, artificial neural networks and deep learning?-Examples of practical applications in medicine. *Diagnostics (Basel)*. 2023;13(15):2582. <https://doi.org/10.3390/diagnostics13152582>
7. Kleinbaum DG. Logistic regression: a self-learning text. New York: Springer Science & Business Media; 2013. p. 291.
8. Cox DR. The regression analysis of binary sequences. *J R Stat Soc Series B Stat Methodol*. 1958;20:215-32. <https://doi.org/10.1111/j.2517-6161.1958.tb00292.x>
9. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1:81-106. <https://doi.org/10.1007/BF00116251>
10. Sun Z, Wang G, Li P, Wang H, Zhang M, Liang X. An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Syst Appl*. 2024;237:121549. <https://doi.org/10.1016/j.eswa.2023.121549>
11. Joachims T. Text categorization with support vector machines: learning with many relevant features. In: *Machine Learning: ECML-98*. Berlin: Springer Berlin Heidelberg; 1998. p. 137-42.
12. Zhou Y, Cui S, Wang Y. Machine learning based embedded code multi-label classification. *IEEE Access*. 2021;9:150187-200.
13. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory*. 1967;13:21-7.
14. Tang Y, Chang Y, Li K. Applications of K-nearest neighbor algorithm in intelligent diagnosis of wind turbine blades damage. *Renewable Energy*. 2023;212:855-64. <https://doi.org/10.1016/j.renene.2023.05.087>
15. Sumayli A. Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil: Gaussian process regression (GPR), multilayer perceptron (MLP), and K-nearest neighbor (KNN) regression models. *Arab J Chem*. 2023;16:104833. <https://doi.org/10.1016/j.arabjc.2023.104833>
16. Bayes T. Naive bayes classifier. *Art Sourc Contribut*. 1968;1-9.
17. Cao Y, Ikenoya Y, Kawaguchi T, Hashimoto S, Morino T. A real-time application for the analysis of multi-purpose vending machines with machine learning. *Sensors (Basel)*. 2023;23(4):1935. <https://doi.org/10.3390/s23041935>
18. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biol*. 1943;5:115-33.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-44. <https://doi.org/10.1038/nature14539>
20. Palechor FM, Manotas AH. Obesity or CVD risk (classify/regressor/cluster) Kaggle. 2023. [cited on 2024 Aug 24]. Available from: <https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster>

CONCLUSION

In the clinical approach to obesity, which is now treatable, ML algorithms can be incorporated alongside medical algorithms. Our model successfully predicted BMI with 87% sensitivity, 99% specificity, and 88% precision, even without height and weight information. Further randomized controlled clinical trials are needed for more definitive results.

ACKNOWLEDGMENTS

We would like to thank all individuals who suffer from obesity and struggle with it and the scientists who play an active role in the treatment of obesity. This study is part of master's thesis at Bahcesehir University.

AUTHORS' CONTRIBUTIONS

ACG: Formal Analysis, Conceptualization, Investigation, Methodology, Validation, Writing - original draft. EA: Project administration, Supervision, Writing - review & editing.

21. Medeiros SF, Junior JMS, Medeiros MAS, Yamamoto AKLW, Medeiros CLW, Silva Carvalho AB, et al. Combined oral contraceptive use and obesity in women with polycystic ovary syndrome. A meta-analysis of randomized clinical trials. *Arch Gynecol Obstet*. 2024;310(4):2223-33. <https://doi.org/10.1007/s00404-024-07637-5>
22. Turri JAO, Anokye NK, Dos Santos LL, Júnior JMS, Baracat EC, Santo MA, et al. Impacts of bariatric surgery in health outcomes and health care costs in Brazil: interrupted time series analysis of multi-panel data. *BMC Health Serv Res*. 2022;22(1):41. <https://doi.org/10.1186/s12913-021-07432-x>
23. Turri JAO, Anokye NK, Dos Santos LL, Júnior JMS, Baracat EC, Santo MA, et al. Correction to: impacts of bariatric surgery in health outcomes and health care costs in Brazil: interrupted time series analysis of multi-panel data. *BMC Health Serv Res*. 2022;22(1):83. <https://doi.org/10.1186/s12913-022-07486-5>
24. Soares Júnior JM, Lobel A, Ejzenberg D, Serafini PC, Baracat EC. Bariatric surgery in infertile women with morbid obesity: definitive solution? *Rev Assoc Med Bras* (1992). 2018;64(7):565-67. <https://doi.org/10.1590/1806-9282.64.07.565>
25. Yilmaz A, Hanlioğlu S, Demiral G, Karatepe O. The efficacy of liraglutide combined with intragastric balloon on weight loss. *Rev Assoc Med Bras* (1992). 2023;69(12):e20230571. <https://doi.org/10.1590/1806-9282.20230571>
26. Alghalyini B. Applications of artificial intelligence in the management of childhood obesity. *J Family Med Prim Care*. 2023;12(11):2558-64. https://doi.org/10.4103/jfmpc.jfmpc_469_23
27. Karimi-Alavijeh F, Jalili S, Sadeghi M. Predicting metabolic syndrome using decision tree and support vector machine methods. *ARYA Atheroscler*. 2016;12(3):146-52. PMID: 27752272
28. Dutta RR, Mukherjee I, Chakraborty C. Obesity disease risk prediction using machine learning. *Int J Data Sci Anal*. 2024;1-10. <https://doi.org/10.1007/s41060-023-00491-9>

