Original Paper

Predicting Overweight and Obesity Status Among Malaysian Working Adults With Machine Learning or Logistic Regression: Retrospective Comparison Study

Jyh Eiin Wong[1], PhD; Miwa Yamaguchi[2], PhD; Nobuo Nishi[2], MD, PhD; Michihiro Araki[2], PhD; Lei Hum Wee[1,3], PhD

[1]Centre for Community Health Studies, Faculty of Health Sciences, Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia

[2]National Institute of Health and Nutrition, National Institutes of Biomedical Innovation, Health and Nutrition, Tokyo, Japan

[3]Faculty of Health and Medical Sciences, School of Medicine, Taylor's University, Selangor, Malaysia

Corresponding Author:
Jyh Eiin Wong, PhD
Centre for Community Health Studies
Faculty of Health Sciences
Universiti Kebangsaan Malaysia
Jalan Raja Muda Abdul Aziz
Kuala Lumpur, 50300
Malaysia
Phone: 60 39289 ext 7683
Email: wjeiin@ukm.edu.my

Abstract

Background: Overweight or obesity is a primary health concern that leads to a significant burden of noncommunicable disease and threatens national productivity and economic growth. Given the complexity of the etiology of overweight or obesity, machine learning (ML) algorithms offer a promising alternative approach in disentangling interdependent factors for predicting overweight or obesity status.

Objective: This study examined the performance of 3 ML algorithms in comparison with logistic regression (LR) to predict overweight or obesity status among working adults in Malaysia.

Methods: Using data from 16,860 participants (mean age 34.2, SD 9.0 years; n=6904, 41% male; n=7048, 41.8% with overweight or obesity) in the Malaysia's Healthiest Workplace by AIA Vitality 2019 survey, predictor variables, including sociodemographic characteristics, job characteristics, health and weight perceptions, and lifestyle-related factors, were modeled using the extreme gradient boosting (XGBoost), random forest (RF), and support vector machine (SVM) algorithms, as well as LR, to predict overweight or obesity status based on a BMI cutoff of 25 kg/m^2.

Results: The area under the receiver operating characteristic curve was 0.81 (95% CI 0.79-0.82), 0.80 (95% CI 0.79-0.81), 0.80 (95% CI 0.78-0.81), and 0.78 (95% CI 0.77-0.80) for the XGBoost, RF, SVM, and LR models, respectively. Weight satisfaction was the top predictor, and ethnicity, age, and gender were also consistent predictor variables of overweight or obesity status in all models.

Conclusions: Based on multi-domain online workplace survey data, this study produced predictive models that identified overweight or obesity status with moderate to high accuracy. The performance of both ML-based and logistic regression models were comparable when predicting obesity among working adults in Malaysia.

(JMIR Form Res 2022;6(12):e40404) doi: 10.2196/40404

KEYWORDS

overweight; obesity; prediction; machine learning; logistic regression; etiology; algorithms; Malaysia; adults; predictive models; accuracy; working adults; surveillance

Introduction

Overweight and obesity are global health issues that are increasingly recognized as major public health concerns in low- and middle-income countries. In Malaysia, 1 in 2 adults, particularly those of working age (ie, aged 30 to 65 years), is either overweight or obese [1]. This is concerning, as obesity prevalence is rising at a very high rate (3.3%) in this country [2]. The increase in overweight and obesity is related to increases in noncommunicable diseases, the mortality rate, and health care costs, as well as decreases in productivity and economic growth [2-5].

Obesity is a chronic, relapsing, multifactorial disease that is attributable to individual or biological, psychological, sociocultural, local, and global environmental factors [6-8]. As obesity is largely preventable, understanding the determinants of and risk factors for obesity is important for the development of population-based strategies to prevent obesity. Identifying individuals at high risk of obesity enables early intervention to modify obesity risk factors. Conventional statistical methods, such as generalized linear or regression models with a low number of predictor variables, have been successful in identifying obesity [9]. However, given the complexity of the etiology of obesity, regression modeling may not be adept at disentangling nonlinear and interdependent relationships among factors for obesity prediction.

Machine learning (ML) is an advanced data analytical method that uses fine-tuned algorithms to characterize and predict outcomes by learning from data without being explicitly programmed to do so. As health data become more available and accessible, ML techniques are increasingly used to perform such complex tasks in obesity research as classifying and predicting obesity at individual and group levels [10-12]. ML techniques have advantages over regression modeling, as they are data driven and do not necessitate a priori assumptions, such as normality, linearity, and multicollinearity. In addition, ML techniques are capable of handling high-dimensional and complex data sources beyond numeric sources, and therefore may be able to provide new insights into unexplored predictor variables [9,13]. Thus, ML techniques are likely to be more accurate than regression models in obesity prediction [14].

A wide range of ML-based algorithms incorporating various predictors and risk factors, training set sizes, and degrees of implementation have been used to predict adult obesity [11,14]. The reported accuracy of ML algorithms to predict adult obesity as a binary outcome ranges broadly, from 0.59 to 0.97 for overall accuracy [15-24] and 0.51 to 0.99 for the area under the curve (AUC) [15,19,20,23,24]. A review suggested that ML-based models predicted childhood and adolescent obesity much better than linear regression [13]. However, studies that have compared the performance of different ML algorithms with regression in adult obesity have reported mixed findings. Some evidence suggests superior performance for ML models compared to regression models [19,21], while some suggests similar or inferior performance [15,17,18,23]. These inconsistencies may partly be due to data quality, variable selection, and the use of different approaches to model fitting, parameter tuning, and validation among studies.

The Malaysia's Healthiest Workplace by AIA Vitality survey is a large, observational online survey of the health and well-being of Malaysian employees [25]. Since 2017 (with the exceptions of 2020 and 2021, because of the COVID-19 pandemic), this annual online workplace survey has collected comprehensive information on Malaysian employees' sociodemographic characteristics, physical and mental health, smoking and alcohol habits, physical activity, diet, musculoskeletal health, and work environment as a database to inform workplace interventions and improve productivity [25]. In this study, we propose an ML-based model to predict overweight and obesity status among employees in Malaysia based on multi-domain variables collected in this large survey. We evaluated the performance of 3 ML algorithms and compared them with logistic regression for the prediction of overweight and obesity status. We hypothesized that ML algorithms would outperform logistic regression models in predicting overweight and obesity status based on BMI.

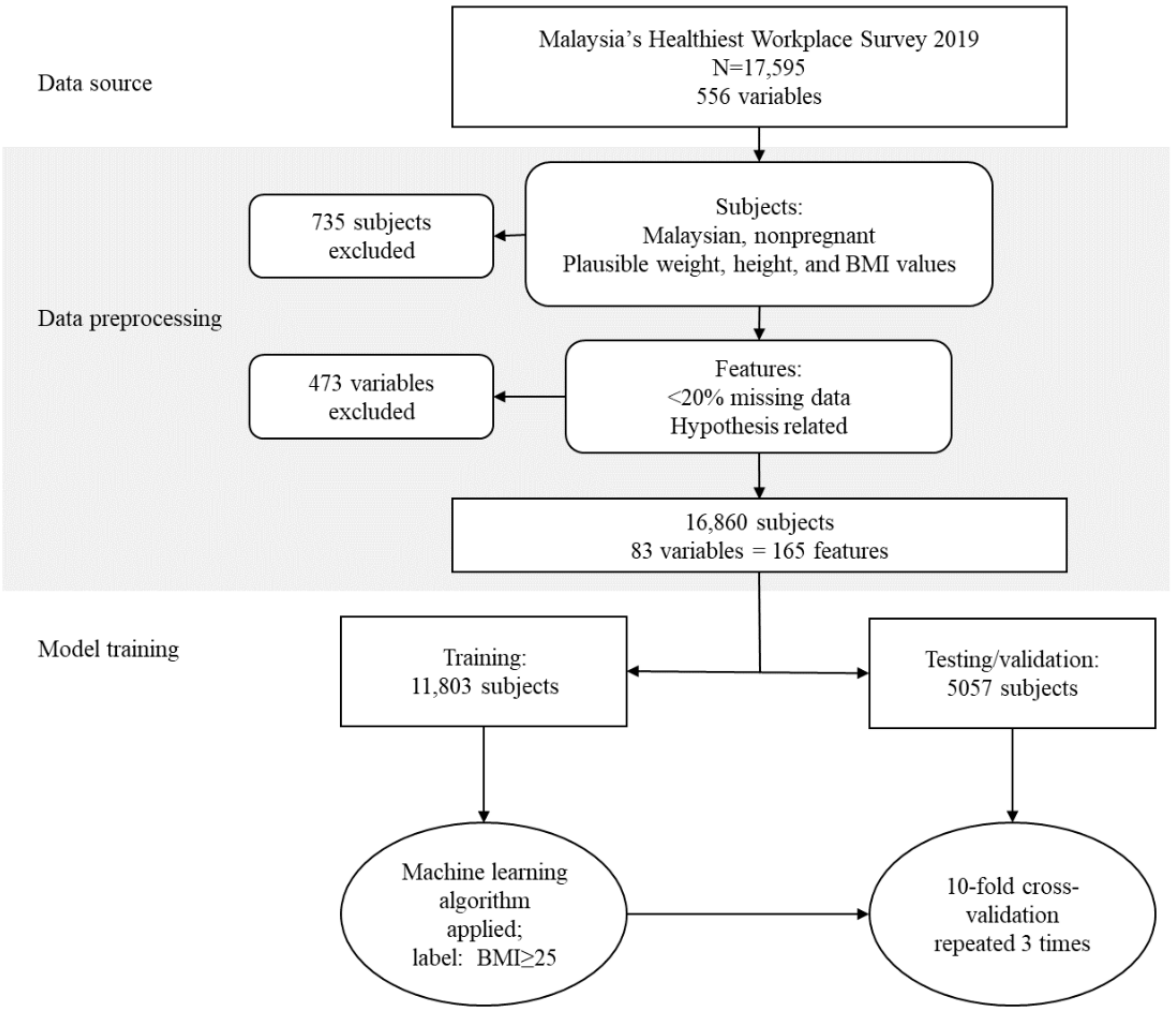
Methods

Study Design and Data

This is a retrospective study of predictive model derivation using data from the Malaysia's Healthiest Workplace by AIA Vitality 2019 survey. This online survey, commissioned by AIA Malaysia and delivered in partnership with RAND Europe, was administered between May and August 2019. The survey, which has taken place annually in Malaysia from 2017 to 2019, aimed to determine workplace productivity and multi-domain factors that influence workplace productivity. Employees from small, medium, and large organizations were invited to answer a 40-minute employee survey questionnaire about their general health, lifestyle behaviors, mental health status, and work environment. The study rationale and methodology have been discussed in detail elsewhere [26-28].

The initial data set comprised data submitted by 17,595 participants from 230 companies. We initially included 16,931 participants resident in Malaysia for whom data were available for body weight and height. If they were women, participants were included if they were not pregnant. Participants with (1) body weight more than 200 kg, (2) height more than 200 cm, or (3) BMI values of more than 60 kg/m^2 or less than 14 kg/m^2 were deemed to have implausible values and were excluded from analysis. After excluding 71 participants who reported implausible weight, height, or BMI, the final data set included 16,860 of 16,931 participants (95.8%) (Figure 1).

Figure 1. Overview of data preprocessing, model development, and model evaluation.



Ethics Approval

The use of the data was approved by the Research Ethics Committee Universiti Kebangsaan Malaysia (JEP-2020-707). As the obtained pooled data were anonymized and deidentified, informed consent from the participants was not required. The study results were presented following the reporting guidelines and recommendations for ML [29,30].

Data Preprocessing

An overview of data preprocessing and model development is illustrated in Figure 1. Data preprocessing involved the selection of participants and variables (features) followed by mean substitution of missing data, one-hot encoding of categorical variables, and min-max scaling for data normalization.

Outcome Variable

The outcome of interest was overweight or obesity status, defined as a BMI of 25 kg/m^2 or more [31]. This was calculated by dividing the self-reported body weight (in kg) by the squared height (in m^2). The cutoff of 25 was chosen as Southeast Asians are reported to have higher body fatness at a lower BMI than Europeans [32,33] and are therefore predisposed to elevated cardiovascular risk factors and other adverse effects of obesity at lower BMI ranges (23 kg/m^2 to 25 kg/m^2), as observed in local studies [34,35]. Further, a recent study suggested that a BMI of 24.8 kg/m^2 is an optimal BMI cutoff to define obesity among Malaysian adults based on percentage of body fat [36].

Predictor Variables

Initially, the data set consisted of 556 predictor variables. A total of 473 variables that contained redundant information or text information with more than 20% missing or nonapplicable data were removed from the data set. The reduced data set included 83 variables that were grouped into the following 4 main domains: sociodemographic characteristics, job characteristics, status perception, and lifestyle-related behaviors (the list of predictor variables is included in Multimedia Appendix 1).

Categorical variables (n=16) were one-hot encoded into binary variables. For instance, weight satisfaction was assessed by a categorical question that prompted participants to select 1 of 3 statements that best described how they felt about their current body weight. The participants indicated whether they (1) were happy with their weight, (2) were not happy with their weight but had no intention of losing or gaining weight, or (3) wanted to change their weight. This categorical variable was

subsequently encoded into 3 binary variables (ie, "weight_satisfaction_1," "weight_satisfaction_2," and "weight_satisfaction_3"). Finally, prediction models were trained and tested on the final 165 normalized variables. A total of 120 (73%) of these 165 predictor variables were binary (yes/no) variables.

Statistical Analysis Methods

Model Development

The R (version 3.6.1; R Software Foundation) package "caret" (version 6.0-90) was used for model training and validation [37]. Based on a random 70:30 split, a total of 11,803 participants, including 4934 (41.8%) with overweight or obesity, were used to train the model. The remaining 30% of the participants (5057/16,860) ware used to predict the obesity outcome during model validation.

Three supervised, nonlinear ML classifiers were applied, namely extreme gradient boosting (XGBoost), random forest (RF), and a support vector machine (SVM). XGBoost is a tree-based ensemble algorithm that uses a boosting method to create multiple decision trees sequentially. The algorithm combines the predictions of weak decision trees to produce a more robust final model. Improvised on the gradient boosting framework, XGBoost is a popular learning algorithm due to its high predictive power and efficiency in handling continuous and categorical data using relatively low computational power [38]. RF is also an ensemble method but uses a bagging method to train multiple decision trees in parallel using random selection of predictors. The final model merges predictions from each decision tree to predict a class [39]. Finally, SVMs use a kernel-based algorithm to construct a decision boundary or hyperplane that best separates the data into 2 classes in n-dimensional space. SVMs use extreme cases, also known as support vectors, to create an optimal hyperplane that has the maximum margin between the vectors [40].

In this study, logistic regression (LR) was compared with the 3 ML models. Logistic regression is a part of the generalized linear model and is the conventional classifier for categorical outcome responses. The algorithm assumes a linear relationship between the predictor variables and the log odds (probability) of obesity as the outcome in this study. All predictor variables were included in the model, regardless of statistical significance, to maintain comparability across models. The goodness of fit of the logistic regression model was demonstrated by a McFadden R^2 value of 0.3452 and a Nagelkerke R^2 value of 0.3452. The probability produced by the logistic regression was subsequently assigned to a binary outcome (overweight/obese or not), based on the customary probability cutoff point of 0.5.

The details of the package, functions, and parameters used in this study are presented in Multimedia Appendix 2. Using a grid search approach, the best combinations of parameters were employed for each algorithm. All models were tuned using 10-fold cross-validation repeated 3 times. Using the varImp function of the caret library, model-specific metrics were used to identify the best-performing predictors. To present the relative ranking of each predictor, the measures of importance for all models were scaled to have a maximum value of 100.

Model Evaluation

The final trained models were saved and restored for prediction using a separate test data set (n=5057) and for comparison with other models. Classification metrics were obtained from the confusion matrix (confusionMatrix) embedded in the caret package. A prediction of overweight or obesity status was considered a positive prediction. Performance was assessed by 4 main metrics (the first 3 metrics are limited in their discriminating power in selecting the best classifier [41], but they are the most common metrics used in the literature and are therefore presented for comparison with other studies): (1) accuracy, the proportion of correct predictions divided by the total number of instances evaluated; (2) sensitivity (also known as the true positive rate), the proportion of actual positives (ie, overweight or obese status) that were correctly predicted; (3) specificity (also known as the true negative rate), the proportion of actual negatives (ie, no overweight or obese status) that were correctly predicted; and (4) AUC, which represents a tradeoff between sensitivity and specificity and served as the main metric for model evaluation. AUC is extracted from the receiver operating characteristic (ROC) curve, which is the probability plot of the true positive rate (ie, sensitivity) against the false positive rate (ie, 1–specificity). An AUC above 0.5 indicates the model is better capable of distinguishing positives (ie, subjects with overweight or obesity) from negatives. In general, an AUC of 0.7 to <0.8 is considered acceptable, 0.8 to <0.9 excellent, and 0.9 or above outstanding predictive performance [42]. The ROCs and corresponding AUCs were computed and plotted with the pROC package.

The performance metrics of all predictive models are presented as point estimates with 95% CIs. For accuracy, sensitivity, and specificity, 95% CIs were calculated assuming a Gaussian distribution of the proportion. For AUCs, 95% CIs were derived through resampling with the bootstrap percentile method with 2000 repetitions. Model comparisons were made based on the 95% CIs of the 4 performance metrics.

Results

Study Characteristics

The analysis included 16,860 participants, of whom 41% (n=6904) were male and 41.8% (n=7048) had overweight or obese status. The male participants were significantly older, and the distributions for ethnicity, education level, marital status, occupation, individual monthly income, and obesity status were also significantly different by sex (P<.001 for all; Multimedia Appendix 3).

Model Comparisons

Table 1 presents the predictive performance of the ML and logistic regression models. Among the 4 models, the RF and LR models had lower sensitivity but higher specificity. While XGBoost exhibited the best mean accuracy and AUC, overall accuracy was similar across all models based on the 95% CIs. The ROCs of the 4 models are illustrated in Figure 2.

Table 2 compares the performance of XGBoost and LR in predicting obesity by sex. For both algorithms, the models for female participants recorded higher specificity but lower

sensitivity than the models for male participants. Overall accuracy and AUC were similar across all 4 models, with the 2 algorithms showing no sex-specific differences in predictive performance.

The ranking of the most important predictors of the models is summarized in Figure 3. In order of importance, the top 4 predictor variables for the XGBoost ML model were weight satisfaction, ethnicity, age, and gender. For the LR model, the top predictor variables were weight satisfaction, physical health, age, and diet satisfaction.

Table 1. Performance of machine-learning algorithms and logistic regression in obesity prediction.

Metrics	Gradient boosting, mean (95% CI)	Random forest, mean (95% CI)	Support vector machine, mean (95% CI)	Logistic regression, mean (95% CI)
Accuracy[a]	0.73 (0.72-0.75)	0.73 (0.71-0.74)	0.72 (0.71-0.73)	0.71 (0.70-0.72)
Sensitivity[a]	0.67 (0.65-0.69)	0.60 (0.58-0.62)	0.65 (0.62-0.67)	0.56 (0.54-0.58)
Specificity[a]	0.78 (0.76-0.79)	0.82 (0.80-0.83)	0.77 (0.76-0.79)	0.82 (0.81-0.83)
Area under the curve[b]	0.81 (0.79-0.82)	0.80 (0.79-0.81)	0.80 (0.78-0.81)	0.78 (0.77-0.80)

[a]In these rows, 95% CIs were calculated assuming Gaussian distribution of the proportions.

[b]In this row, 95% CIs were derived through resampling with the bootstrap percentile method with 2000 repetitions.

Figure 2. Receiver operating characteristic curves with corresponding AUC values; AUC values for each model are also presented in Table 2. AUC: area under the curve.

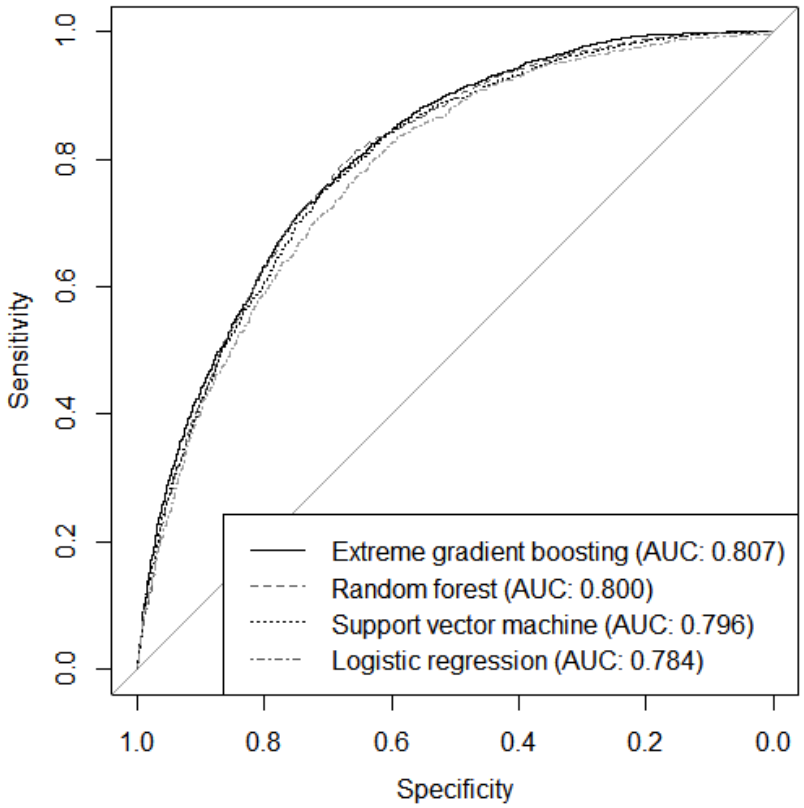


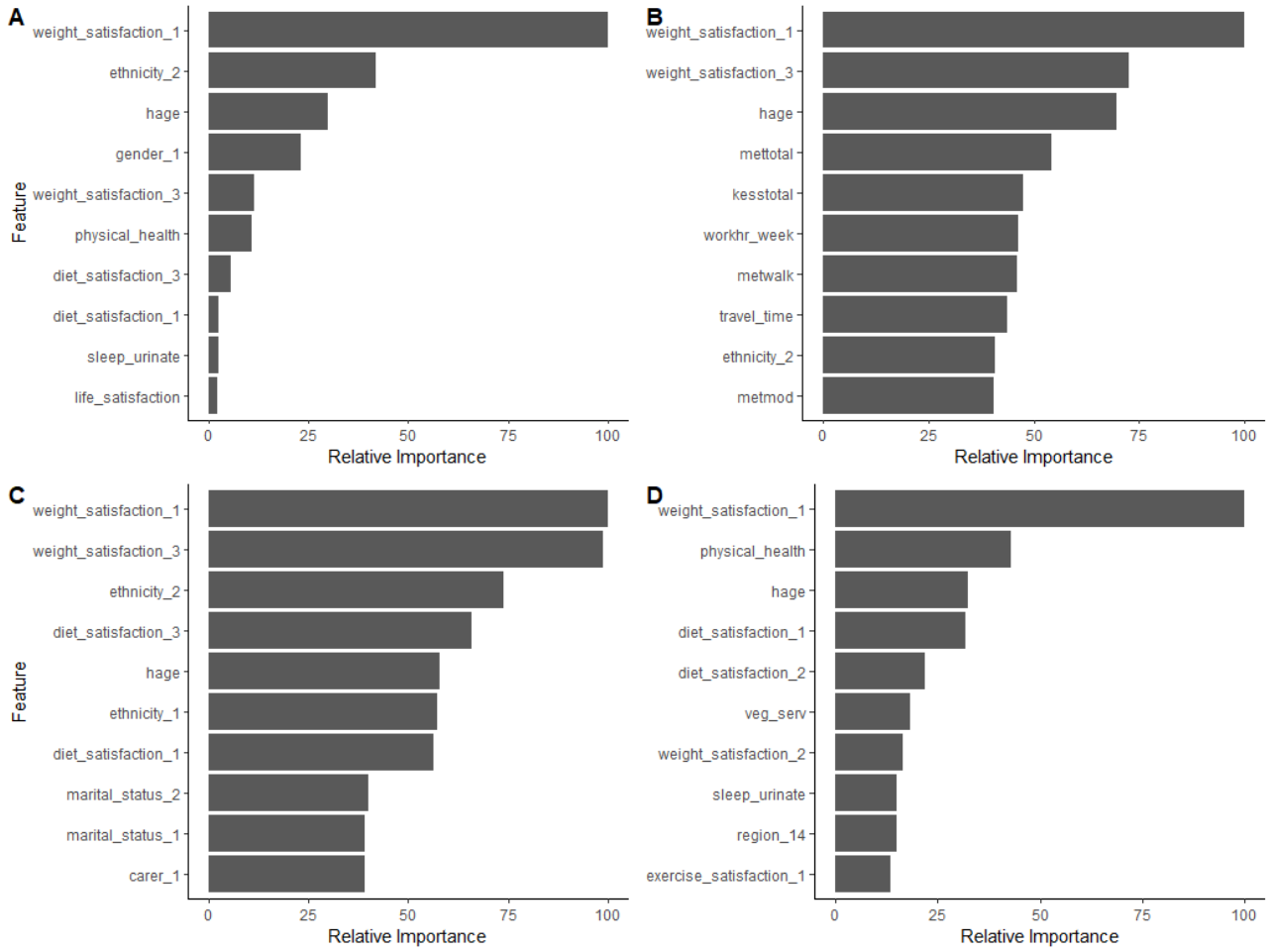
Table 2. Comparison of performance between machine learning and logistic regression in sex-specific obesity prediction.

	Gradient boosting, mean (95% CI)		Logistic regression, mean (95% CI)	
Metrics	Male participants	Female participants	Male participants	Female participants
Accuracy[a]	0.71 (0.69-0.73)	0.74 (0.72-0.75)	0.70 (0.68-0.72)	0.73 (0.71-0.74)
Sensitivity[a]	0.75 (0.73-0.78)	0.61 (0.58-0.63)	0.72 (0.69-0.75)	0.60 (0.57-0.63)
Specificity[a]	0.66 (0.63-0.69)	0.81 (0.80-0.83)	0.68 (0.65-0.71)	0.80 (0.78-0.81)
Area under the curve[b]	0.78 (0.76-0.80)	0.81 (0.79-0.82)	0.76 (0.74-0.78)	0.79 (0.77-0.80)

[a]In these rows, 95% CIs were calculated assuming Gaussian distribution of the proportions.

[b]In this row, 95% CIs were derived through resampling with the bootstrap percentile method with 2000 repetitions.

Figure 3. Variable importance plots of obesity predictors for extreme gradient boosting (A), random forest (B), support vector machine (C) and logistic regression (D) models. The top 10 predictors are shown for all models.



Discussion

Principal Results

This study applied various ML models and compared their performance to the performance of a conventional logistic regression model in predicting overweight or obesity status among working adults in Malaysia. Our results showed that ML and logistic regression had similarly acceptable or excellent predictive performance, as assessed by the metrics of accuracy (values ranged from 70% to 75%) and AUC (values ranged from 78% to 81%), for both the overall and sex-specific models.

Comparison With Prior Work

Our findings, based on data collected annually as part of a large-scale online survey of employees, compare favorably to those of a recent study by Thamrin et al [23] that also used a large Southeast Asian sample (N=618,898), in Indonesia. That study employed logistic regression, classification and regression trees, and a naive Bayes classifier for obesity prediction based on data for sociodemographic characteristics, diet, physical activity, lifestyle behaviors, and health status from the Indonesian Basic Health Research periodic survey. The study reported accuracy between 70.8% and 72.2% and an AUC between 0.75 and 0.80, which is comparable to the performance

of our models (mean accuracy 71%-73.3% and AUC 0.78-0.81). While there is no definite standard for acceptable accuracy, the models in our study recorded accuracy greater than 70% and AUC greater than 0.7, which is better than the accuracy and AUC of past models that used novel predictors, including genetics [20,24], detailed dietary intake [18,21], and objectively measured physical activity [15].

In this study, the overall performance of the ML models, namely XGBoost, RF, and SVM, was found to be similar to logistic regression, as indicated by the overlapping 95% CIs. This corroborates the findings of a systematic review of 71 studies, which concluded that ML did not offer greater performance benefits than logistic regression for clinical prediction models [43]. Specifically, for obesity prediction, Ferdowsy et al [17] employed 8 algorithms, in addition to logistic regression, in a data set that included 21 well-established risk factors for obesity, such as diet, physical activity, lifestyle behaviors, and disease history. Their study recorded the highest accuracy (97%) with the logistic regression model, which outperformed ML algorithms including k-nearest neighbor, RF, a multilayer perceptron, an SVM, a naive Bayes classifier, adaptive boosting, a decision tree, and a gradient boosting classifier for obesity prediction [17]. Kim et al [18] modeled the effects of 7 dietary factors on overweight or obesity status using data from the Korea National Health and Nutrition Examination Survey. That study showed that the predictive accuracy of logistic regression (0.62486) was higher than that of decision trees (0.54026) and similar to that of a deep neural network model of deep learning (0.62496). Taken together, comparative studies that deal with a small number of strong predictor variables [15,17,18,23] suggest that regression models are likely to perform better than, if not as well as, ML models in obesity prediction.

Another possible reason for this similar performance is that the observed relationships among the significant predictors of obesity in this sample may appear linear on the log-odd scale. Hence, logistic regression was not disadvantaged by assuming linearity in these predictors. In this study, we employed 3 nonlinear ML classifiers due to the fact that many variables, including intrapersonal and socioeconomic factors that affect body weight, such as age, sex, and gender, are nonlinear in nature [44]. However, it could be hypothesized that these nonlinear ML algorithms may have been less proficient at modeling the present data set because the data mostly consisted of binary variables (120/165, 73%).

It is important to acknowledge that different ML algorithms may fit and perform differently when used with different data sets. Guided by previous findings that obesity determinants are different for men and women [19,45], we developed separate, sex-stratified models for overweight or obesity status prediction. However, the predictive accuracy of the sex-specific models was similar to the overall or combined models. This suggests that separate prediction models for each sex are not warranted in this Malaysian adult population.

In terms of predictor variables, weight satisfaction appears to be a consistent, novel predictor in all predictive models, together with such well-established risk factors for obesity as ethnicity, age, and gender. Weight satisfaction is an attitudinal component of body image, which reflects individuals' feelings and thoughts about their weight [46]. The variable "weight_satisfaction_1," which represents satisfaction or contentment with current body weight, appears to have had the most influential power in the trained model to predict overweight or obesity status (Figure 3).

This novel finding is consistent with previous studies showing that self-perception of body weight is an important determinant of weight management behaviors and lifestyle practices [47]. However, the relationships between weight satisfaction and weight-related behaviors are complex and multifaceted. Depending on sex, race, ethnicity, accuracy of weight perceptions, and psychological factors, weight satisfaction may promote positive diet and physical activity behaviors or lead to maladaptive or unhealthy weight-control or dieting behaviors [48-50]. As weight satisfaction and dissatisfaction appear to be mostly stable in adulthood [51,52], we posit that this subjective variable may be cognitively easier and more reliable to report than body weight and height among adults. This finding supports the usefulness of including weight satisfaction as a proxy for actual weight status in studies and e-surveys, where anthropometry measurements may not be available or feasible.

Strengths and Limitations

To the best of our knowledge, this is the first study to employ ML to predict overweight or obesity status in an adult working population in Malaysia. This study used rich data from a large annual survey that included a wide, multi-domain set of predictor variables in working adults with a broad range of ages (18 to 88 years) and occupations. Another strength of the study lay in its employment of advanced ML classifiers with careful cross-validation (to avoid model overfitting) and parameter optimization. The variable importance technique afforded novel insights into significant factors that are correlates of overweight or obesity status in a Malaysian working population.

This study was also limited in several ways. First, the study findings do not infer temporality or causality of the observed predictor-obesity relationships due to the use of a cross-sectional design. However, the findings suggest putative variables that could be explored using novel model interpretation techniques such as Shapley additive explanations [53] and could be considered for further testing in longitudinal or trial settings. Second, mislabeling of obesity was likely, due to the reliance on self-reported body weight and height to derive BMI as a surrogate measure of general obesity. Notably, the prevalence of individuals with overweight or obesity in this study (4934/11,803, 41.8%) was lower than the national prevalence of 50.1% [1]. Such errors, or noise, may have reduced the performance of the models. Therefore, the current findings represent conservative estimates of predictive accuracy. Finally, we acknowledge that the generalizability of our models is limited, as validation was based on testing data that came from the same sample. Validating the models with an external data set would more closely approximate the real performance of the prediction models. Future work is needed to confirm the external validity and reproducibility of the models in other data sets, such as the Malaysia's Healthiest Workplace surveys from 2018 or later.

Conclusions

Using a multi-domain set of predictors from a large online employee survey, we constructed models that were able to predict overweight or obesity status in a Malaysian working population with moderate to high accuracy. Weight satisfaction was the most prominent factor, followed by ethnicity, age, and gender, in differentiating individuals with overweight or obese status. Among the 3 ML models (XGBoost, RF, and SVM), XGBoost had the highest accuracy and AUC, but the overall performance of all ML-based models was similar to the logistic regression model for obesity prediction.

This study is complementary to and extends the growing literature showing that ML may be used to predict overweight or obesity status based on online survey data with reasonable accuracy. Besides unveiling distinctive factors that influence weight status in this Asian population, this work also produced potential models or algorithms that can be used to screen for overweight or obesity status in community settings, especially when body weight and height data are not available. A natural progression of this study would be to test the performance of the produced models in an external data set to establish the external validity of the findings.

Acknowledgments

The authors would like to thank AIA Malaysia for granting permission to use the data. The study data source was funded by AIA Malaysia (grants NN-2019-152 and NN-2021-004). Payment of the article processing fee for this publication was supported by the National Institutes of Biomedical Innovation, Health and Nutrition. JEW, MY, NN, and MA jointly conceived the study concept and design. LHW and MA contributed to data acquisition. JEW performed data processing, variable selection, and data analysis under the supervision of MA. JEW drafted the manuscript. All authors contributed to data and result interpretation. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Predictor variables used in the dataset (n=165).
[DOCX File , 37 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Description of software packages, methods and tuning parameters for model development.
[DOCX File , 29 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Study sample characteristics (n=16860).
[DOCX File , 33 KB-Multimedia Appendix 3]

References

1. National Health and Morbidity Survey 2019: Non-communicable diseases, healthcare demand and healthy literacy. Volume I: NCDs – Non-Communicable Diseases: Risk Factors and other Health Problems. Ministry of Health Malaysia. 2020. URL: https://iku.moh.gov.my/images/IKU/Document/REPORT/NHMS2019/Report_NHMS2019-NCD_v2.pdf [accessed 2022-11-24]
2. Lobstein T, Brinsden H, Neveux M. World Obesity Atlas 2022. World Obesity Federation. 2022. URL: https://www.worldobesity.org/resources/resource-library/world-obesity-atlas-2022 [accessed 2022-11-21]
3. Tremmel M, Gerdtham U, Nilsson PM, Saha S. Economic burden of obesity: a systematic literature review. Int J Environ Res Public Health 2017 Apr 19;14(4):435 [FREE Full text] [doi: 10.3390/ijerph14040435] [Medline: 28422077]
4. Goettler A, Grosse A, Sonntag D. Productivity loss due to overweight and obesity: a systematic review of indirect costs. BMJ Open 2017 Oct 05;7(10):e014632 [FREE Full text] [doi: 10.1136/bmjopen-2016-014632] [Medline: 28982806]
5. GBD 2015 Obesity Collaborators, Afshin A, Forouzanfar MH, Reitsma MB, Sur P, Estep K, et al. Health effects of overweight and obesity in 195 countries over 25 years. N Engl J Med 2017 Jul 06;377(1):13-27 [FREE Full text] [doi: 10.1056/NEJMoa1614362] [Medline: 28604169]
6. Bray GA, Kim KK, Wilding JPH, World Obesity Federation. Obesity: a chronic relapsing progressive disease process. A position statement of the World Obesity Federation. Obes Rev 2017 Jul;18(7):715-723. [doi: 10.1111/obr.12551] [Medline: 28489290]
7. Swinburn BA, Sacks G, Hall KD, McPherson K, Finegood DT, Moodie ML, et al. The global obesity pandemic: shaped by global drivers and local environments. Lancet 2011 Aug 27;378(9793):804-814. [doi: 10.1016/S0140-6736(11)60813-1] [Medline: 21872749]

8. Ford ND, Patel SA, Narayan KV. Obesity in low- and middle-income countries: burden, drivers, and emerging challenges. Annu Rev Public Health 2017 Mar 20;38:145-164 [FREE Full text] [doi: 10.1146/annurev-publhealth-031816-044604] [Medline: 28068485]
9. DeGregory KW, Kuiper P, DeSilvio T, Pleuss JD, Miller R, Roginski JW, et al. A review of machine learning in obesity. Obes Rev 2018 May;19(5):668-685 [FREE Full text] [doi: 10.1111/obr.12667] [Medline: 29426065]
10. Birkin M, Wilkins E, Morris MA. Creating a long-term future for big data in obesity research. Int J Obes (Lond) 2019 Dec;43(12):2587-2592. [doi: 10.1038/s41366-019-0477-y] [Medline: 31641212]
11. Chatterjee A, Gerdes MW, Martinez SG. Identification of risk factors associated with obesity and overweight-a machine learning overview. Sensors (Basel) 2020 May 11;20(9):2734 [FREE Full text] [doi: 10.3390/s20092734] [Medline: 32403349]
12. Scheinker D, Valencia A, Rodriguez F. Identification of factors associated with variation in US county-level obesity prevalence rates using epidemiologic vs machine learning models. JAMA Netw Open 2019 Apr 05;2(4):e192884 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.2884] [Medline: 31026030]
13. Colmenarejo G. Machine learning models to predict childhood and adolescent obesity: a review. Nutrients 2020 Aug 16;12(8):2466 [FREE Full text] [doi: 10.3390/nu12082466] [Medline: 32824342]
14. Safaei M, Sundararajan EA, Driss M, Boulila W, Shapi'i A. Comput Biol Med 2021 Sep;136:104754 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104754] [Medline: 34426171]
15. Cheng X, Lin S, Liu J, Liu S, Zhang J, Nie P, et al. Does physical activity predict obesity-a machine learning and statistical method-based analysis. Int J Environ Res Public Health 2021 Apr 09;18(8):3966 [FREE Full text] [doi: 10.3390/ijerph18083966] [Medline: 33918760]
16. Delnevo G, Mancini G, Roccetti M, Salomoni P, Trombini E, Andrei F. The prediction of body mass index from negative affectivity through machine learning: a confirmatory study. Sensors (Basel) 2021 Mar 29;21(7):2361 [FREE Full text] [doi: 10.3390/s21072361] [Medline: 33805257]
17. Ferdowsy F, Rahi KSA, Jabiullah MI, Habib MT. A machine learning approach for obesity risk prediction. CRBS 2021 Nov;2:100053. [doi: 10.1016/j.crbeha.2021.100053]
18. Kim H, Lim DH, Kim Y. Classification and prediction on the effects of nutritional intake on overweight/obesity, dyslipidemia, hypertension and type 2 diabetes mellitus using deep learning model: 4-7th Korea National Health and Nutrition Examination Survey. Int J Environ Res Public Health 2021 May 24;18(11):5597 [FREE Full text] [doi: 10.3390/ijerph18115597] [Medline: 34073854]
19. Lee BJ, Kim KH, Ku B, Jang J, Kim JY. Prediction of body mass index status from voice signals based on machine learning for automated medical applications. Artif Intell Med 2013 May;58(1):51-61. [doi: 10.1016/j.artmed.2013.02.001] [Medline: 23453267]
20. Lee Y, Christensen JJ, Parnell LD, Smith CE, Shao J, McKeown NM, et al. Using machine learning to predict obesity based on genome-wide and epigenome-wide gene-gene and gene-diet interactions. Front Genet 2021;12:783845 [FREE Full text] [doi: 10.3389/fgene.2021.783845] [Medline: 35047011]
21. Selya AS, Anshutz D. Machine learning for the classification of obesity from dietary and physical activity patterns. In: Giabbanelli P, Mago V, Papageorgiou E, editors. Advanced Data Analytics in Health. Smart Innovation, Systems and Technologies, vol 93. Cham, Switzerland: Springer; 2018:77-97.
22. Taghiyev A, Altun A, Caglar S. A hybrid approach based on machine learning to identify the causes of obesity. Journal of Control Engineering and Applied Informatics 2020;22(2):56-66 [FREE Full text]
23. Thamrin SA, Arsyad DS, Kuswanto H, Lawi A, Nasir S. Predicting obesity in adults using machine learning techniques: an analysis of Indonesian basic health research 2018. Front Nutr 2021;8:669155 [FREE Full text] [doi: 10.3389/fnut.2021.669155] [Medline: 34235168]
24. Wang H, Chang S, Lin W, Chen C, Chiang S, Huang K, et al. Machine learning-based method for obesity risk evaluation using single-nucleotide polymorphisms derived from next-generation sequencing. J Comput Biol 2018 Dec;25(12):1347-1360. [doi: 10.1089/cmb.2018.0002] [Medline: 30204480]
25. AIA Vitality: The Healthiest Workplace. AIA Group. URL: https://www.aia.com/language-masters/en/en/health-wellness/vitality/healthiest-workplace [accessed 2022-11-24]
26. Toh B, Akmal A, Syed JS. Malaysia's healthiest workplace: AIA Vitality special report. The EDGE Malaysia 2017 Nov 20:1-16.
27. Malaysian Workforce: Sleepless and Overworked? AIA Group. URL: https://www.aia.com.my/en/about-aia/media-centre/press-releases/2019/malaysian-workforce-sleepless-and-overworked.html [accessed 2022-11-24]
28. Toh B, Akmal A, Syed JS. Malaysia's healthiest workplace AIA vitality special report: methodology and demographics. The EDGE Malaysia 2018 Dec 03:1-16.
29. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
30. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes 2020 Oct;13(10):e006556 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.120.006556] [Medline: 33079589]

31. Physical status: the use and interpretation of anthropometry. Report of a WHO Expert Committee. World Health Organization. 1995. URL: https://apps.who.int/iris/bitstream/handle/10665/37003/WHO_TRS_854.pdf [accessed 2022-11-21]
32. Chen KK, Wee S, Pang BWJ, Lau LK, Jabbar KA, Seah WT, et al. Relationship between BMI with percentage body fat and obesity in Singaporean adults - The Yishun Study. BMC Public Health 2021 Jun 01;21(1):1030 [FREE Full text] [doi: 10.1186/s12889-021-11070-7] [Medline: 34074272]
33. Deurenberg-Yap M, Schmidt G, van Staveren WA, Deurenberg P. The paradox of low body mass index and high body fat percentage among Chinese, Malays and Indians in Singapore. Int J Obes Relat Metab Disord 2000 Aug;24(8):1011-1017. [doi: 10.1038/sj.ijo.0801353] [Medline: 10951540]
34. Cheong KC, Yusoff AF, Ghazali SM, Lim KH, Selvarajah S, Haniff J, et al. Optimal BMI cut-off values for predicting diabetes, hypertension and hypercholesterolaemia in a multi-ethnic population. Public Health Nutr 2013 Mar;16(3):453-459. [doi: 10.1017/S1368980012002911] [Medline: 22647482]
35. Zaher ZMM, Zambari R, Pheng CS, Muruga V, Ng B, Appannah G, et al. Optimal cut-off levels to define obesity: body mass index and waist circumference, and their relationship to cardiovascular disease, dyslipidaemia, hypertension and diabetes in Malaysia. Asia Pac J Clin Nutr 2009;18(2):209-216 [FREE Full text] [Medline: 19713180]
36. Aizuddin AN, Chan CM, Anwar AR, Ong YX, Chin K. Performance of body mass index in identifying obesity defined by body fat percentage and hypertension among Malaysian population: a retrospective study. Int J Gen Med 2021;14:3251-3257 [FREE Full text] [doi: 10.2147/IJGM.S316360] [Medline: 34267543]
37. Kuhn M. Building predictive models in R using the caret package. J Stat Soft 2008;28(5):1-26 [FREE Full text] [doi: 10.18637/jss.v028.i05]
38. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
39. Breiman L. Random forests. Mach Learn 2001;45(1):5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]
40. Vapnik VN. The Nature of Statistical Learning Theory. Berlin, Germany: Springer; 1995.
41. Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process 2015 Mar 31;5(2):01-11 [FREE Full text] [doi: 10.5121/ijdkp.2015.5201]
42. Hosmer JD, Lemeshow S, Sturdivant R. Area under the receiver operating characteristic curve. In: Applied Logistic Regression, Third Ed. Hoboken, NJ: John Wiley & Sons; 2013:173-182.
43. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
44. Dogbe W, Salazar-Ordóñez M, Gil JM. Disentangling the drivers of obesity: an analytical framework based on socioeconomic and intrapersonal factors. Front Nutr 2021 Mar 3;8:585318 [FREE Full text] [doi: 10.3389/fnut.2021.585318] [Medline: 33791330]
45. Hammond R, Athanasiadou R, Curado S, Aphinyanaphongs Y, Abrams C, Messito MJ, et al. Predicting childhood obesity using electronic health records and publicly available data. PLoS One 2019;14(4):e0215571 [FREE Full text] [doi: 10.1371/journal.pone.0215571] [Medline: 31009509]
46. Cash T, Smolak L. Body Image: A Handbook of Science, Practice, and Prevention. New York, NY: Guilford Press; 2012.
47. Haynes A, Kersbergen I, Sutin A, Daly M, Robinson E. A systematic review of the relationship between weight status perceptions and weight loss attempts, strategies, behaviours and outcomes. Obes Rev 2018 Mar;19(3):347-363 [FREE Full text] [doi: 10.1111/obr.12634] [Medline: 29266851]
48. Blake CE, Hébert JR, Lee D, Adams SA, Steck SE, Sui X, et al. Adults with greater weight satisfaction report more positive health behaviors and have better health status regardless of BMI. J Obes 2013;2013:291371 [FREE Full text] [doi: 10.1155/2013/291371] [Medline: 23862054]
49. Millstein RA, Carlson SA, Fulton JE, Galuska DA, Zhang J, Blanck HM, et al. Relationships between body size satisfaction and weight control practices among US adults. Medscape J Med 2008 May 19;10(5):119. [Medline: 18596944]
50. Kuk JL, Ardern CI, Church TS, Hebert JR, Sui X, Blair SN. Ideal weight and weight satisfaction: association with health practices. Am J Epidemiol 2009 Aug 15;170(4):456-463 [FREE Full text] [doi: 10.1093/aje/kwp135] [Medline: 19546153]
51. Tiggemann M. Body image across the adult life span: stability and change. Body Image 2004 Jan;1(1):29-41. [doi: 10.1016/S1740-1445(03)00002-0] [Medline: 18089139]
52. Quittkat HL, Hartmann AS, Düsing R, Buhlmann U, Vocks S. Body dissatisfaction, importance of appearance, and body appreciation in men and women over the lifespan. Front Psychiatry 2019;10:864 [FREE Full text] [doi: 10.3389/fpsyt.2019.00864] [Medline: 31920737]
53. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: The 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach, CA.

Abbreviations

AUC: area under the curve
LR: logistic regression
ML: machine learning
RF: random forest
ROC: receiver operating characteristic curve
SVM: support vector machine
XGBoost: extreme gradient boosting

Edited by A Mavragani; submitted 20.06.22; peer-reviewed by V Kumar, M Roccetti; comments to author 30.08.22; revised version received 09.10.22; accepted 11.10.22; published 07.12.22

Please cite as:
Wong JE, Yamaguchi M, Nishi N, Araki M, Wee LH
Predicting Overweight and Obesity Status Among Malaysian Working Adults With Machine Learning or Logistic Regression: Retrospective Comparison Study
JMIR Form Res 2022;6(12):e40404
URL: https://formative.jmir.org/2022/12/e40404
doi: 10.2196/40404
PMID: