Identification of Key Features Contributing to Obesity and Classification of Obesity Classes

|  |  |  |
| --- | --- | --- |
| Rachna Shah  Department of Computer  Information Science  Minnesota State University  Mankato, MN  rachna.shah@mnsu.edu | Aninda Roy  Department of Mathematics and Statistics  Minnesota State University  Mankato, MN  aninda.roy@mnsu.edu | Naseef Mansoor  Department of Computer  Information Science  Minnesota State University  Mankato, MN  naseef.mansoor@mnsu.edu |

|  |  |
| --- | --- |
|  |  |

**Abstract**. Research has found that obesity, a rising epidemic in America, is linked with premature death and disability in adulthood; obesity leads to diabetes, cancer, hypertension, heart ailments and many other chronic diseases. Our research aims to uncover what lifestyle variables most affect obesity by using a dataset from 2019 that contains 2111 records and 17 attributes related to lifestyle and weight. The target attribute in this dataset is the weight class of each patient, which is derived from BMI ranges. Identifying key features from this dataset in relation to the weight class attribute along with applying machine learning classification techniques on this dataset allows us to gather information that could help find risk factors associated with obesity that aren’t normally tracked or studied. We used factor analysis and the chi-squared test to identify key features and then tested various classification methods such as Logistic Regression, Random Forest, SVM, and K-Nearest Neighbor to identify the best-performing model that was capable of accurately classifying the weight class of each available datapoint. We found that age, family history of overweightness, frequency of consumption of vegetables (FCVC), and consumption of food between meals (CAEC) were closely tied to weight class. We found that Random Forest was the best classification model with an accuracy of 96%. All models tested had lower performance when classifying between the Overweight classes (Overweight I and II) and lower Obesity classes (Obesity Type I). This might indicate a decreased level of notable health differentiation between overweight and low-obesity level individuals. The results gathered from this research lead towards the future of healthcare; using this information, we can not only improve personalized healthcare and aid diagnosis, but also create a way for smart devices and the data from smart devices to be used more effectively in that endeavor.

**Keywords:** obesity, machine learning, health informatics, classification

**CCS Concepts:** Computing methodologies → Machine learning → Classification model

1. **Introduction**

Obesity has emerged as one of the most pressing public health concerns worldwide, but particularly in America; obesity has rapidly increased in prevalence over the previous decade in both developed and developing countries. Obesity is described as having an abnormally high level of body fat. While some body fat is good and necessary to maintain health such as essential fat and moderate levels of subcutaneous and visceral fat, having an excess of either of those fat types can be very damaging to the body. Visceral fat in particular is fat that wraps around vital organs in the abdomen and, in moderate levels, is necessary to pad the organs, but in excess is dangerous to your health. The diagnosis of obesity is accompanied by a myriad of related health concerns [1]; high blood pressure, high blood glucose, osteoarthritis, dyslipidemia, diabetes, coronary heart disease, some types of cancer, stroke, mental health issues and sleep-apnea are all linked to being overweight or obese [2]. Seeing as obesity is a health risk with many varying comorbidities, it is a very important field of research in the medical field with implications that stretch far beyond weight management.

The diagnosis of obesity considers a variety of different factors, but the most widely used method for evaluating obesity is the Body Mass Index (BMI). It provides a measure of relative weight that is adjusted for height and can be used to compare within and between populations. BMI is a relatively simple calculation that can be done by hand or with an online tool; the only necessary information is height and weight. While increased caloric intake and decreased energy expenditure are definitely crucial variables, genetic factors also play a significant role in obesity susceptibility. BMI was created as a disease risk indicator; when BMI rises, so does the risk of certain diseases associated with excess weight. However, BMI is a very simple measure, one that doesn’t not factor in muscle mass or any other components of body composition and therefore has a wide margin of error when it comes to disease risk prediction [3]. In 1995, the World Organization of Health (WHO) classified any BMI over 30 as obese. Aside from BMI, another metric for diagnosis is body fat percentage; for women a body fat percentage over 30-35% is considered obese, while for men a body fat percentage over 20-25% is considered obese [4]. Diagnostic tools that aim to measure the body fat metric are direct body fat measurements through the use of calipers to measure subcutaneous fat, using bioelectrical impedance analysis to analyze body composition, or Hydrodensitometry which is a measure of body composition while all body parts are underwater. However, these techniques are all more expensive or time consuming than a BMI measurement. The most accessible and least time consuming of the options would be an at-home scale capable of bioelectrical impedance analysis, but even then the accuracy of the current available devices can be inaccurate by 20-30% [5]. So, while BMI has its flaws, it is ultimately a useful tool for diagnosis. Hopefully with further information about an individual’s lifestyle, in addition to the BMI measurement, there will be an easier way to classify obesity or inclination towards obesity without having to use tools that are time consuming and outside of someone’s regular everyday function. Identifying key lifestyle features that most accurately classify obesity can help healthcare become more well-rounded, holistic, and approachable.

One major factor missed by BMI score is the quantity of activity a person gets every day. Physical exercise, or the lack thereof, has been identified as a significant contributor to the obesity pandemic [6]. Any body movement that burns calories counts as physical activity, whether it's for work or play, daily tasks, or the daily commute. Staying active can assist people in maintaining a healthy weight or losing weight. It has also been shown to relieve stress and improve mood, as well as reducing the risk of heart disease, diabetes, stroke, high blood pressure, osteoporosis, and certain malignancies. Sedentary lifestyles have the opposite effect [7]. A slight increase in daily moderate-to-vigorous physical exercise (5–10 minutes) has previously been linked to a lower risk of obesity; people who maintained a healthy weight were more likely to engage in vigorous physical activity on a regular basis [7]. For the purposes of our research, having a measure of a person’s physical activity can key us into the amount of muscle mass they have; this way, we can use this information to cover BMI’s blind spot. For many years, exercise has been an unrecorded factor in people’s lives, however, with the advent of smart watches and other smart devices, it is now possible to analyze the risk of obesity using quantity of exercise as a measure. Along this line, in this modern day and age there are a number of features in people’s everyday lives that are now able to be measured and tracked. This means that with this new data, we can go beyond the simple BMI score, and find out what features best classify a person’s weight class and achieve a level of personalization that is the future of healthcare.

Our research uses a 2019 dataset, published in a peer-reviewed journal, that contains 17 different attributes in order to identify what features and classification methods best identify obesity [8]. Many of the attributes available in the dataset are things that can be tracked or recorded from any simple smart device and accompanying wellness app that are commonplace nowadays. Our research studying the key features that relate to BMI can help develop an understanding of the lifestyle components that make up a person’s weight class. Additionally, our research into finding the optimal classification model for this kind of trackable data aids the field of obesity-healthcare research as it moves towards creating personalized, accessible healthcare for all. This research will further the applications of machine learning in the field of medicine by increasing the ability for healthcare treatment and prevention to be personalized to each individual, along with granting a person further insight into their risk of obesity using easily tracked information on their smartphones.

1. **Literature Review**

Obesity and machine learning are large fields of research in their own respect so our focus was narrowed down to where they intersected in health informatics. One comprehensive, but slightly dated, paper was the 2018 review of the use of machine learning in obesity research [9]. The review covered how machine learning has been used to better understand obesity and help in its prevention and treatment. Some of the machine learning techniques applied in this study were linear and logistic regression, artificial neural networks, deep learning, and decision tree analysis. The data used to test these methods was from the National Health and Nutrition Examination Survey which contained data from 1999 to 2006 with over 25,000 records. The variables collected were race, height, weight, blood pressure, body fat levels, waist measurement, education level, age, gender, and BMI. Although this was a very large dataset, it didn’t take into account a number of lifestyle variables that affect obesity. Part of the significance and novelty of our research is due to its study of previously unstudied attributes such as water intake, vegetable intake, and quantity of exercise. These attributes are all easily trackable using a smart device or wellness app and therefore provide a deeper look into a person’s lifestyle and the factors contributing to their weight class.

Another paper covering different machine learning algorithms tested on obesity data tested out a larger array of models [10]. In this paper, 1100 data points were collected and 9 different machine learning algorithms were applied such as k-nearest neighbor (k-NN), random forest, logistic regression, Support Vector Machine(SVM) and more. The performance of these algorithms was then compared, and they found that logistic regression has the highest accuracy of 97.09% while gradient boosting gave the lowest accuracy. This paper is useful for its comparison of machine learning algorithms, however it was conducted on a dataset with only a few features and with only three classes to separate weight— low, medium and high. Our dataset is larger, with more features, and with seven classes that are true to medical weight classes— this way our study more clearly bridges the gap between lifestyle variables to medical diagnoses.

An older study from 2016 used machine learning to study the relationship between BMI and psychological measures [11]. They used machine learning algorithms such as gradient boosting, KNN, SVM classification and regression tree, random forest and many other algorithms in order to predict BMI levels as well as BMI status based on psychological characteristics. When the various machine learning algorithms were trained on positive and negative psychological variables separately, the results revealed that positive psychological variables have no effect on BMI, whereas negative psychological variables do. This study shows us that machine learning approaches can help us enhance our understanding of obesity and aid in our ability to forecast it more accurately. The downside of this study is that the psychological measures are fairly clinical and removed from real life. Our study uses variables that compose day-to-day lifestyle choices in order to enhance our understanding of obesity. This way, the results from our research can be used to make healthcare more accessible and personalized.

Another study looked into the relationship between obesity and physical activity [12]; this was investigated using classification methods such as naive Bayes, radial basis function, k-nearest neighbors, classification via regression, random subspace, decision table, and others. These machine learning techniques offer a novel way to examine multi-factorial data, which may then be used to develop predictions about the complex interrelationships that are likely to influence obesity risk. The findings show that lack of physical activity particularly moderate to vigorous intensity is a key risk factor for being overweight or obese. Much like our research this study aims to better understand the complex factors that influence a person’s risk for obesity. Unfortunately, this paper is narrow in its focus on exercise. Our dataset includes physical activity, but it also covers things like quantity of food intake, type of food intake, family history of obesity, and smoking status. Using this expansive list of variables we aim to better understand which features have the greatest connection to obesity and how all of the features together aid in the classification of obesity.

Our research approaches obesity from a lifestyle standpoint. Rather than classifying obesity based on only weight and height, we want to fill in the blind spots of BMI. Using data that encapsulates a person’s lifestyle will help round out the BMI metric, so that the diagnosis of obesity can be more accurate, and so that the lifestyle patterns behind obesity can be understood and used to create individualized prevention and treatment recommendations. Our research aids this larger goal by exploring the relationship between lifestyle variables, such as water intake, activity level and smoking status, and the target variable, weight class. Finding key features along with finding the best classification model for our lifestyle-based dataset will aid in one day creating a system that will allow people to be more aware and conscious of their lifestyle choices as well as provide them with more efficient treatments and prevention methods, all while remaining accessible for anyone with a smart device.

1. **Methods and Materials**

*3.1.**Data Set Description*

The data used in this research project was collected from individuals from the countries of Mexico, Peru and Colombia [8]. The dataset consists of 2111 observations with 17 variables where Age, Height and Weight are numerical variables and the remaining are categorical. The data was collected from individuals ages between 14-61 by using a web platform with a survey where anonymous users answered each question. Those variables were divided in terms of their eating habits and physical condition. The attributes related with eating habits are: Frequent of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of Main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of Alcohol (CALC). The attributes related to physical condition are Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS). The records are labeled with the class variable NObesity (Obesity level), that allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. These weight classes align with the weight classes WHO uses to assign labels to different BMI ranges (Table 1).

*3.2. Data Preprocessing*

The initial data collection consisted of 485 records with an overrepresentation of the “Normal” weight class. In order to normalize the dataset and balance the representation of weight classes, 77% of the data was generated synthetically using the Weka tool and the filter SMOTE [8]. The filter is required to indicate the class for generation of synthetic data, the number of nearest neighbors used, the percentage that you need to increase the selected class and the random seed used for random sampling. Other aspects analyzed were the identification of atypical and missing data. Finally, after the filter was applied to each category, the final result was 2111 records.

It is important to notice that data was preprocessed (deletion of missing data, handling atypical data, data normalization, etc.) before using SMOTE, since the neighbor selected to generate the synthetic data could contain noise or disturbances, and the data produced would have low quality. Nevertheless, using the filter SMOTE has a positive impact when data is unbalanced and the balancing process decreases the probability of skewed learning in favor of a majority class.

*3.3. Feature Selection*

Feature Selection is the method of reducing the input variables going into your model by using only relevant data and getting rid of the possible noise in the dataset. In order to identify significant variables we used the Chi-square association test to find out which factors in the dataset were significantly related to weight class (Table 2). The null hypothesis in this scenario is that the dependent variable is not related to weight class. The Chi-square association test works by comparing the distribution that we observe between two variables to the distribution that we expect if there is no relationship between the categorical variables. Relationships that have a p-value less than the significance level of 0.05, means that the variables are significantly related. We also used the rankFeatures function in R with Random Forest as the method to rank the impact of each dependent variable on weight class (Figure 1).

*3.4. Classification Methods*

One of the machine learning algorithms used on this dataset was K-Nearest Neighbors (k-NN). K-NN doesn’t make any assumption on the underlying data because it is a non-parametric algorithm; it works by looking for similarity measures. The number of neighbors to consider can be a user-defined constant K as in the case of K-nearest neighbors, or it can be based on the density of points in a certain radius specified [13]. Accuracy was used to select the optimal model using the largest value. The final k value used for selecting the optimal model was k=7.

A Support Vector Machine (SVM) is a flexible supervised machine learning algorithm, and it is used for both regression and classifications problems. A SVM edifies a set of hyperplanes in an infinite dimensional space, which can be used for regression, classification, and outlier detection. In our research we used SVM with a linear kernel which means a set of linear hyperplanes were created to classify the obesity classes and the tuning parameter was held constant at a value of 1.

The Logistic Regression classification algorithm is a predictive ML technique that relies on analyzing different possible probabilities to model the relationship between the dependent and independent variables. Logistic Regression uses a more complex cost function which is defined by a sigmoid function. The hypothesis of logistic regression estimates the limit of the cost function between 0 and 1. In our study, a Penalized (regularized) multinomial logistic model is used, which is a parametric model that is commonly estimated using maximum likelihood estimation. We set the maximum iterations to use for estimating the model to 100. The final, optimized value used for the model was decay = 1e-04.

The Random Forest algorithm is a ML technique that works by generating multiple decision trees on different sub-samples of the data and then predicts the accuracy or loss score by taking the mean of these values. This process helps minimize any possible over-fitting of the model [14]. In this algorithm, the split for each node is determined from a subset of predictor variables that are randomly chosen at the given node [15]. This is done using the Out-Of-Bag (OOB) approach specified by Breiman [16]. Here, accuracy also used to select the optimal model and the final value used for the model was mtry=16.

*3.5. Model validation and performance measures*

The analysis for each method was run with 10 cross-folds validation, meaning the dataset was randomly divided into 10 parts and 10 runs of the model were executed with each run consisting of one of the 10 parts being the testing data and the remaining 90% being the training data. Performance statistics were then averaged across all 10 runs. We have inquired into the performance of the classifiers from the confusion matrix of dimension 7 ×7 since the number of class labels equals seven. We calculated sensitivity, specificity, precision, *F*1-score of each algorithm along with accuracy.

Accuracy is defined as the percentage of the total samples that were correctly recognized by the classifier. Precision is defined as the percentage of total predicted positive samples by the classifier that was actually positive. *F*1-score is the measurement of the harmonic mean of sensitivity and precision. It considers both false positive and false negative values for calculation. Sensitivity, the true positive rate, and Specificity, the true negative rate describe how valid a model’s results are (Table 3).

1. **Figures and Tables**

Table 1: BMI classification according to WHO

|  |  |
| --- | --- |
| Weight Class | BMI Range |
| Underweight | Less than 18.5 |
| Normal | 18.5 to 24.9 |
| Overweight | 25.0 to 29.9 |
| Obesity I | 30.0 to 34.9 |
| Obesity II | 35.0 to 39.9 |
| Obesity III | Higher than 40 |

Table 2: Chi-square association with obesity and different attributes.

|  |  |  |
| --- | --- | --- |
| Attributes | Chi-squared | P-value |
| Family history with overweight | 621.98 | 2.328e-16 |
| FAVC | 233.34 | 2.568e-16 |
| CAEC | 802.98 | 2.326e-16 |
| CALC | 338.58 | 2.218e-16 |
| SMOKE | 32.138 | 1.535e-05 |
| CH20 | 163.32 | 2.26e-16 |
| SCC | 123.02 | 2.238e-16 |
| FAF | 273.71 | 1.567e-05 |
| TUE | 177.64 | 2.2e-16 |
| FCVC | 566.93 | 2.236e-16 |
| MTRANS | 292.59 | 2.683e-16 |

\*Significance level 0.05

Table 3: Performance of the classifiers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Evaluation Criteria | RF | LR | SVM | KNN |
| Correctly classified instances | 506 | 483 | 416 | 378 |
| Incorrectly Classified Instances | 19 | 42 | 109 | 147 |
| Accuracy (%) | 96.38 | 92 | 79.24 | 72 |

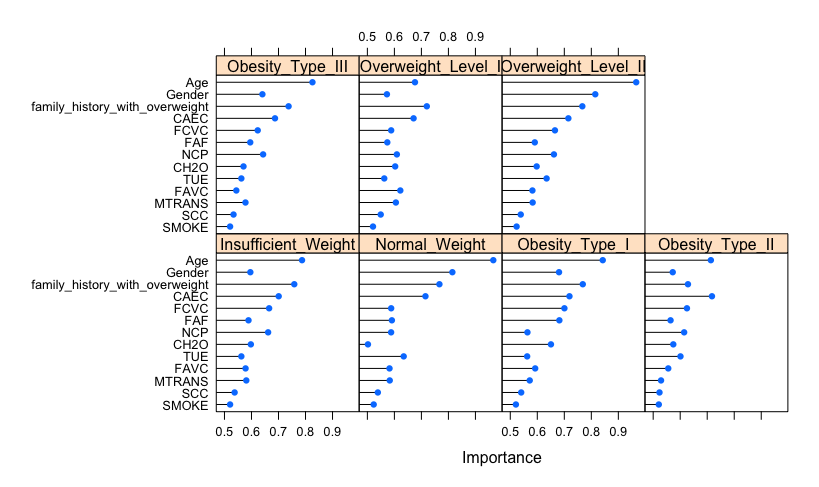


Figure 1: Feature selection from classification model

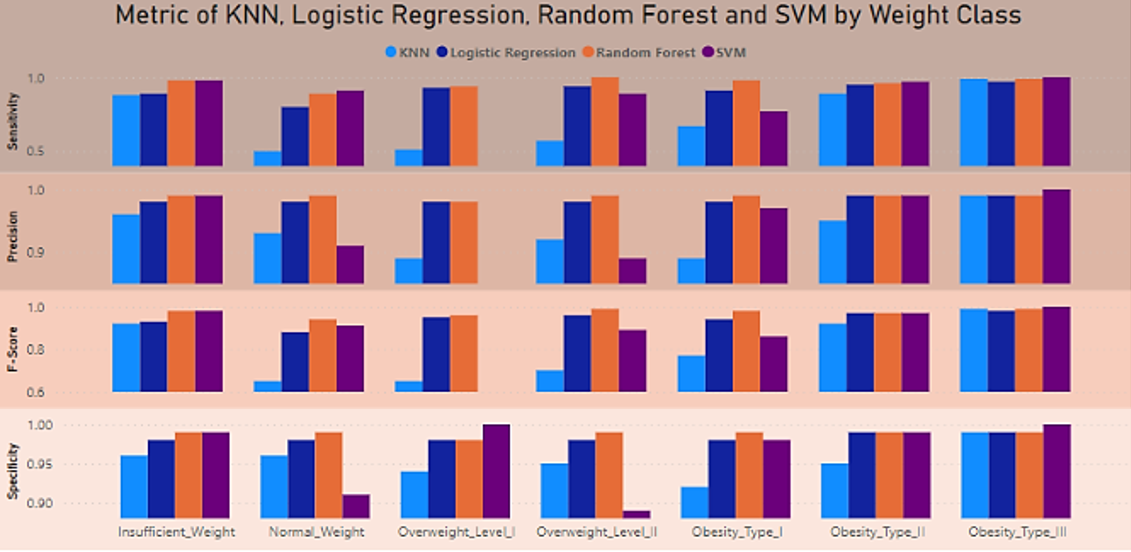


Figure 2: Comparison of Accuracy Measures for KNN, LR, RF, and SVM.

1. **Results and Discussion**

The feature selection showed that of the 14 variables we tested against weight class, 11 of them had a significant (p<0.05) relationship with weight class (Table 2). Further exploring that data, when we ranked the features we found that the top three features with the closest relationship to weight class were: Age, Gender, Family history of overweightness (Figure 1). The top five lifestyle variables associated with weight class were: Consumption of food between meals (CAEC), Frequency of consumption of vegetables (FCVC), Amount of Physical activity (FAF), Number of main meals (NCP), and Consumption of water (CH20). Surprisingly variables like frequent consumption of high caloric food (FAVC) and calorie consumption (SCC), while significant, were not part of the top variables that related to weight class. It is interesting to note that inherent, biological characteristics were more connected to a person’s weight class than simply a combination of calorie intake and amount of physical activity. In fact, calorie intake wasn’t even in the top five most important lifestyle variables. The top lifestyle variable was consumption of food between meals— snacking. This makes sense because there has been a lot of research over the years studying the relationship between snacking and obesity, and while the results are still a bit contentious, the research seems to show a relationship between snacking and higher levels of obesity [17]. All the top lifestyle variables associated closely with weight class are all attributes that are frequently recorded and tracked on wellness apps these days; there are apps for recording water intake, there are apps for meal tracking and calorie counting, and there are a myriad of apps that record a person’s physical activity throughout the day. Seeing as these attributes are a good indicator of weight class, they would be useful to add more support to someone’s BMI number, along with filling in the blind spots of BMI. If patients were to have a record of the variables that make up their lifestyle, they would have greater insight into their health, and greater ability to advocate for their needs in healthcare rather than simply being recommended dieting as a response to their high BMI. The ability to personalize prevention and treatment for obesity based on lifestyle variables and how they relate to weight class would greatly enhance a patient’s medical power, both through the personalization of their healthcare but also through the awareness of how their personal actions affect their health.

Our study also aimed to find the best classification model for our chosen dataset of lifestyle variables and we found that Random Forest was the best performing classification model (Table 3). To measure the performance and effectiveness of the classifiers, we evaluated the accuracy of all classifiers. The RF model had the highest value of correctly classified instances (506) and a lowest value of incorrectly classified instances (19) in comparison to all the other tested classification models resulting in the highest accuracy (96.38%). In contrast, the KNN model had the lowest number of correctly classified instances (378) and the highest incorrectly classified samples (147), resulting in an accuracy of 72%. The second and third best classifiers are LR, with an accuracy of 92%, and SVM, with an accuracy of 79.24%. These results agree with the current research in the field, although some obesity studies found LR [10, 11] to be the best classifier. However those studies had fewer variables and used large demographic data, unlike the dataset we used for our study.

After creating the classification models, we further analyzed the efficiency of our models using sensitivity, specificity, precision, *F*1-score as our metrics. RF maintained averaged as the model with the best metrics across all weight classes. It is also notable that most models had lower performance when classifying weight classes between Normal, Overweight I & II, and Obesity I. As shown in Figure 2, SVM and KNN especially had lower performance when classifying between those intermediate weight classes. This might indicate a decreased level of lifestyle differentiation between those who have a BMI between 25 and 34. However, since we can’t assume causation from our data, we must rely on further research to explore that relationship.

1. **Conclusions**

Our research has found that the top attributes most closely associated with weight class are more so biological than lifestyle based. The top lifestyle attributes most closely associated with weight class include consumption of food between meals, frequency of consumption of vegetables, amount of physical activity, number of main meals, and consumption of water. These variables align with common health advice: eat your vegetables, don’t snack between meals, move your body every day, drink lots of water. However, it’s also interesting to note that these results don’t confirm calorie counting and exercise as the primary attributes related to weight. Our results also show that the best classification model for this kind of lifestyle data is Random Forest, which had the highest overall metrics between all the classification models we tested. The metrics also show that models such as SVM and KNN struggled to classify weight classes that did not lie on either extreme of the BMI range. This research is important as we try to build accurate and computationally efficient classifiers for medical applications. In the future, longitudinal studies observing lifestyle variables in relation to weight and health would be very useful to expand on the results found in our research.

**References**

1. Thamrin, Sri & Arsyad, Dian & Kuswanto, Hedi & Lawi, Armin & Nasir, Sudirman. (2021). Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018. Frontiers in Nutrition. 8. 669155. 10.3389/fnut.2021.669155.
2. Singh, B., & Tawfik, H. (2020). Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People. *Computational Science – ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV*, *12140*, 523–535. <https://doi.org/10.1007/978-3-030-50423-6_39>
3. Barzinji, A. O., Ma, C., Du, W., & Ma, J. (2021, September). A Machine Learning Approach to Predict the Trend of Obesity Prevalence at a Global Level. In *2021 IEEE/ACIS 6th International Conference on Big Data, Cloud Computing, and Data Science (BCD)* (pp. 25-30). IEEE.
4. Lopez-Jimenez, F., & Miranda, W. R. (2010). Diagnosing obesity: Beyond BMI. Virtual Mentor, 12(4), 292–298. https://doi.org/10.1001/VIRTUALMENTOR.2010.12.4.CPRL1-1004
5. Franco-Villoria, M., Wright, C. M., McColl, J. H., Sherriff, A., Pearce, M. S., Adamson, A., Dale, A., Drewett, R., Le Couteur, A., McArdle, P., Parkinson, K., & Reilly, J. J. (2016). Assessment of adult body composition using bioelectrical impedance: comparison of researcher calculated to machine outputted values. BMJ Open, 6(1). <https://doi.org/10.1136/BMJOPEN-2015-008922>
6. Barzinji, A. O., Ma, C., Du, W., & Ma, J. (2021, September). A Machine Learning Approach to Predict the Trend of Obesity Prevalence at a Global Level. In *2021 IEEE/ACIS 6th International Conference on Big Data, Cloud Computing, and Data Science (BCD)* (pp. 25-30). IEEE.
7. Dubnov-Raz, G., Maor, S., & Ziv-Baran, T. (2021). Pediatric obesity and body weight following the COVID-19 pandemic. *Child: Health, Care, Development*. https://doi.org/10.1111/cch.12939
8. Palechor, F. M., & Manotas, A. de la H. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief*, *25*, 104344. https://doi.org/10.1016/J.DIB.2019.104344
9. Rajput, K., Chetty, G., & Davey, R. (2018, December). Obesity and co-morbidity detection in clinical text using deep learning and machine learning techniques. In *2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)* (pp. 51-56). IEEE.
10. DeGregory, K. W., Kuiper, P., DeSilvio, T., Pleuss, J. D., Miller, R., Roginski, J. W., Fisher, C. B., Harness, D., Viswanath, S., Heymsfield, S. B., Dungan, I., & Thomas, D. M. (2018). A review of machine learning in obesity. *Obesity Reviews : An Official Journal of the International Association for the Study of Obesity*, *19*(5), 668. <https://doi.org/10.1111/OBR.12667>
11. Mitchell, T. B., & Steele, R. G. (2016). The Effect of Body Mass Index, Negative Affect, and Disordered Eating on Health-Related Quality of Life in Preadolescent Youth. *Journal of Pediatric Psychology*, *41*(7), 768–776. <https://doi.org/10.1093/JPEPSY/JSV163>
12. Cheng, X., Lin, S. Y., Liu, J., Liu, S., Zhang, J., Nie, P., Fuemmeler, B. F., Wang, Y., & Xue, H. (2021). Does Physical Activity Predict Obesity-A Machine Learning and Statistical Method-Based Analysis. *International journal of environmental research and public health*, *18*(8), 3966. <https://doi.org/10.3390/ijerph18083966>
13. Nadkarni, S., & Shenoy, P. P. (2001). A Bayesian Network Approach to Making Inferences in Cognitive Maps. *European J. Operational Research*, *128*(3).
14. Livingston, F. (2005). Implementation of Breiman’s random forest machine learning algorithm. *ECE591Q Machine Learning Journal Paper*, 1-13.
15. Nadkarni, S., & Shenoy, P. P. (2004). A causal mapping approach to constructing Bayesian networks. *Decision support systems*, *38*(2), 259-281.
16. Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Brooks. *Wadsworth and Brooks, Monterey, CA*.
17. Mattes, R. D. (2018). Snacking: A cause for concern. *Physiology & Behavior*, *193*, 279–283. https://doi.org/10.1016/J.PHYSBEH.2018.02.010
18. Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., & Habib, M. T. (2021). A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, *2*, 100053.
19. Celik, Y., Guney, S., & Dengiz, B. (2021, July). Obesity Level Estimation based on Machine Learning Methods and Artificial Neural Networks. In *2021 44th International Conference on Telecommunications and Signal Processing (TSP)* (pp. 329-332). IEEE.
20. Delnevo, G., Mancini, G., Roccetti, M., Salomoni, P., Trombini, E., & Andrei, F. (2021). The Prediction of Body Mass Index from Negative Affectivity through Machine Learning: A Confirmatory Study. *Sensors*, *21*(7), 2361. https://doi.org/10.3390/s21072361
21. Edwards, W. (1998). Hailfinder: Tools for and experiences with Bayesian normative modeling. *American Psychologist*, *53*(4), 416.
22. Edwards, W., & Fasolo, B. (2001). Decision technology. *Annual review of psychology*, *52*(1), 581-606.
23. I. H. Witten and E. Frank. 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
24. Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.
25. Kunal Rajput, G. C. (2018, December). Obesity and Co-Morbidity Detection in Clinical Text Using Deep Learning and Machine Learning Techniques. *Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE).*
26. la Othman Barzinji, C. M. (2021, September). A Machine Learning Approach to Predict the Trend of Obesity Prevalence at a Global Level. *Big Data, Cloud Computing, Data Science & Engineering (BCD)*.
27. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.
28. Paul Wiechmann, K. L. (2017, November). Identifying Discriminative Attributes to Gain Insights Regarding Child Obesity in Hispanic Preschoolers Using Machine Learning Techniques. *International Conference on Tools for Artificial Intelligence (ICTAI)*.
29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & others (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825--2830.
30. Pang, X., Forrest, C. B., Lê-Scherban, F., & Masino, A. J. (2019, December). Understanding early childhood obesity via interpretation of machine learning model predictions. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)* (pp. 1438-1443). IEEE.
31. Quinlan, R. C. (1993). 4.5: Programs for machine learning morgan kaufmann publishers inc. *San Francisco, USA*.
32. Xueqin Pang, C. B.-S. (2019, December). Understanding Early Childhood Obesity via Interpretation of Machine Learning Model Predictions. *International Conference on Machine Learning and Applications (ICMLA)*.
33. Wiechmann, P., Lora, K., Branscum, P., & Fu, J. (2017, November). Identifying discriminative attributes to gain insights regarding child obesity in hispanic preschoolers using machine learning Techniques. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 11-15). IEEE.
34. Yaren Celik, S. G. (2021, Jun 10th). Obesity Level Estimation based on Machine Learning Methods and Artificial Neural Networks. *International Conference on Telecommunications and Signal Processing (TSP)*.
35. Casimiro Aday Curbelo. Montañez *et al*. Machine learning approaches for the prediction of obesity using publicly available genetic profiles, *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017.
36. De-La-Hoz-Correa, E., Mendoza-Palechor, F. E., De-La-Hoz-Manotas, A., Morales-Ortega, R. C., & Adriana, S. H. B. (2019). Obesity Level Estimation Software based on Decision Trees. Journal of Computer Science, 15(1), 67–77. https://doi.org/10.3844/JCSSP.2019.67.77