Current Research in Behavioral Sciences 2 (2021) 100053

Contents lists available at ScienceDirect

Current Research in Behavioral Sciences

journal homepage: www.elsevier.com/locate/crbeha

A machine learning approach for obesity risk prediction

Faria Ferdowsy*, Kazi Samsul Alam Rahi, Md. Ismail Jabiullah, Md. Tarek Habib

Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

ARTICLE INFO

Keywords:
Obesity risk
Overweight
Machine learning
Prediction system
Disease
classifier
Logistic regression

ABSTRACT

In modern times, obesity has become a significant threat all over the world. Obesity means an unnatural or excessive amount of fat that is present in our bodies. People are constantly moving towards an unhealthy lifestyle, eating excessive junk food, late-night sleep, spend a long time sitting down. Especially, adolescents are being affected because of their unconscious attitudes. It is a medical problem known as a very complex disease. It promotes the spread of complex illnesses, stroke, heart disease, liver cancer. Consequently, as an aware multitude of Bangladesh, we have to move forward to prevent this risk of obesity. The purpose of this paper is to move towards a machine-learning-based pathway for predicting the risk of obesity using machine-learning algorithms. The great thing about this paper is that people will know the risk of obesity and the reasons behind their obesity. We collect more than 1100 data from many varieties of people of different ages and collect information from both are suffering obesity and non-obesity. For this research, we apply nine prominent machine learning algorithms. We used the algorithm of k-nearest neighbor (*k*-NN), random forest, logistic regression, multilayer perceptron (MLP), support vector machine (SVM), naïve Bayes, adaptive boosting (ADA boosting), decision tree, and gradient boosting classifier, and we have measured the performance of each of these classifications in terms of some prominent performance metrics. From the experimental results, we determine the obesity of high, medium, and low. The Logistic Regression Algorithm achieves the highest accuracy of 97.09% as compared to the other classifiers. In addition, the gradient boosting algorithm gave the poorest accuracy of 64.08% as well as the lowest metric values.

1. Introduction

An excessive amount of body fat is called obesity. Obesity is not only about food genetic, and environmental can be the cause of obesity. In the future, it can be a threat to the world as it is a worldwide health concern. Obesity occurs due to many reasons also, it can be named a disease. Thousands of risks and diseases are associated with obesity. It is one of the most common health problems all over the world. Excessively eating and moving too little is the principal reason for obesity. If people do not burn off their energy through physical activities, such as yoga, exercises, and so on, but take high amounts of energy, particularly fat and sugars, then much of the surplus energy is converted into fat and stored in the body. Most people are not concerned about their obesity as they thought it was one of the general health definitions. Also, they contemplate that it does not affect their health. It is just the outer structure of their body. But the unfortunate reality is most of the diseases are associated with obesity. Sometimes it can cause death, as it has been identified by devastating epidemics for diabetes, cardiovascular disease, malignancy, osteoarthritis, persistent kidney disease, stroke, hypertension, and fatal diseases.

Obesity can grow in both adults and children. The imbalanced energy between calorie intake and expansion is likely to call the fundamental definition of obesity. If we see the obesity rate, then since 1975 worldwide obesity rate is nearly tripled (World Health Organisation). More than 650 million people were obese in 2016, and the overweight rate was 39% from the age of 18 years and older, where 13% were obese (World Health Organisation). In 2016 than 340,000 children were obese, and 34 million children under the age of 5years were obese in 2019 (World Health Organisation). So, from the above information, it would not be wrong to say that obesity will become a leading threat to the whole world in the coming days. Earlier, it considered that well-developed countries have the highest rate of obesity than low and middle-income countries. But along with time, obesity increase in the lower and middle-income country also. Obesity can occur in both boys and girls equally. Bangladesh had faced the 'dual burden' of both nutrition and obesity, which was percept by Imperial College London and the World Health Organization (WHO), established experts. The overweight or obesity rate was 7% for adults and 3% for children in Bangladesh in 1980 (Dugan et al., 2015). Further, the percentage of obesity rate climbed to 4.5% for children and 17% for adults, according to The Institute for Health Metrics and Evaluation (IHME) of the University of Washington (Salahuddin, 2021). The percentage showed that the

* Corresponding author.
E-mail addresses: faria15-9100@diu.edu.bd (F. Ferdowsy), alam25-011@diu.edu.bd (K.S.A. Rahi), drismail.cse@diu.edu.bd (Md.I. Jabiullah).

https://doi.org/10.1016/j.crbeha.2021.100053
Received 26 May 2021; Received in revised form 14 July 2021; Accepted 1 August 2021
2666-5182/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

rate of childhood obesity is increasing slowly for Bangladesh, but the adult obesity rate is increasing fast than child obesity which needs to be controlled. Excess weight and social support can be considered complementary. Women groups are involved in anxiety, stress, and mental pressure for excessive weight. At this point, social support acts as potential support (Abbas et al., 2019). According to WHO, the overweight and obesity rate was minacious among school-aged children in the urban areas in Bangladesh also diabetes country profile of Bangladesh in 2016 physical inactivity was prevailing among 25.1% of the population. Bangladeshi young generation is the current trend of obesity is very high. Because of the impact of the western lifestyle, the internet, children are more likely to consume fast food and other junk foods (Salahuddin, 2021). The principle of this paper is to make people aware of the risk of obesity based on some key factors. The aspects of human life that can bring them obesity and the rate of obesity are presenting here. Body mass index is a calculation that uses a person's weight, age, regular activity, and height to measure body size. Body mass index (BMI) refers to the value of body weight divided by height class. It is extensively referred to as a substitute measure of fat mass due to its availability and ease of use. Obesity can also be measured by a BMI calculator.

The main concern of this paper is to analyze people for obesity and make them aware of the obesity risk factor. This paper aims to predict the obesity risk. The analysis is conducted into two parts where firstly it read the data and then checks the data if it matches the factor with obesity, and then it will show the result. For our analysis, first, we collect raw data sets for our analysis depend on some factors. In addition, we preprocess those data, then we applied nine machine learning supervised algorithms to check the accuracy, sensitivity, specificity, precision, recall, and F_1-score. Then we found which algorithm works more optimal and detect the actual outcome.

2. Literature review

We wanted to represent our work perfectly, that people get some benefit from our work in a proper way without any difficulty. Therefore, we decided to study this topic, and we read some researcher's work. This section of the paper deals with all of the past and present work that predicted the risk of obesity. We followed their work and tried to understand their demonstrated method, which was applying by them. Dugan et al. (Dugan et al., 2015) did an excellent task to predict obesity in children after age two. They used six models to test for their study. Their models are Random Tree, Random Forest, ID3, J48, Naïve Bayes, and Bayes Net trained on CHICA which are clinical decision support system. They got the best performance from the model ID3, which was highly accurate at 85% and sensitive at nearly 90%.

Jindal et al. (Jindal et al., 2018) works on predicting obesity using ensemble machine learning approaches. Their predicted value of obesity was 89.68% accurate to propose an ensemble machine learning approach to the prediction of obesity and employ the ensemble prediction they used Python interface also used leverages generalized linear model, random forest, and partial least squares for their prediction model.

Hammond et al. (Hammond et al., 2019) and his team used electronic health actual records and used public data which is available for predicting childhood obesity. They trained various machine learning algorithms for binary classification and regression. They showed that they could predict reasonable accuracy by collecting data from the first two years. It detects that children will have obesity by age five. They applied logistic regression, used another application of random forest classifier, and gradient boosting model for predicting their dichotomous measures of low obese/ medium obese/ high obese for prophecy their continuous BMI values they employed LASSO regression. They ran the bootstrap 100 times to get the best performance of the models for the final result.

Dunstan et al. (Dunstan et al., 2019) Research on predicting country level obesity from food sales, and collected data from seventy nine countries, also used three types of machine learning algorithm. Their goal was to identify food sales that could provide us with actual information about the synergic nature of categories. Theirs imitate confirm that they used the five categories, approximately 60% of the countries considered, and 10% (with respect to a complete prevalence range), and below 20% for the 87% of countries can predict an obesity prevalence with an absolute error. They realized that for predicting the obesity the most pertinent food category is baked goods and flours. SVM, RF, and extreme gradient boosting are applied for their model.

Singh and Tawfik (Singh And and Tawfik, 2020) proposed a machine learning Approach to predict the risk of becoming obese or overweight at the adolescence stage. They used seven machine-learning algorithms in their prediction model. Their applied algorithms were *k*-NN, J48 pruned tree, Random forest, and Bagging, support vector machine, multilayer perception, and voting the effectiveness of all the algorithms was tested on a sample of an unaltered, unbalanced dataset. The precision value is 96% for the MLP algorithm. The result of the F_1-score was 93.96%.

Gerl (Gerl et al., 2019) works to predict different measures of obesity based on a large population cohort. For BFP, they identified a perplexing lipidomic signature and also could predict 8% of the full range of BFP with error and interpret 73% of its variants based on age, gender, and lipidome.

Montañezet al. (Montañez et al., 2017) conducted a study of machine learning methods for the prediction of obesity applying genetic profiles which are publicly available. Their applied machine learning algorithm was SVM algorithm, decision tree, decision rule, and *k*-NN algorithm to predict susceptibility to chronic hepatitis using SNPs data. From those algorithms, SVM gave the best result for their prediction model. Their simulation result showed that SVM generated the highest area under the curve value of 90.5%.

Adnan et al. (Muhamad Adnan et al., 2012) work on a hybrid approach for prediction and parameter optimization using Naïve Bayes and genetic algorithms. They got the highest accuracy from genetic algorithm optimization. From their implementation, they identified a weakness of the Naïve Bayes algorithm which is known as "zero value parameters". Their initial test presented that their structure was usable and that it accurately observed 92% of the zero value parameter samples.

Borrell and Samuel (Borrell and Samuel, 2014) research on US Adult Body Mass Index Category and Risk of Mortality: The Impact of Excess Weight and Obesity on the Prevalence of Death, using Cox proportional hazards regression they estimate the rate advancement period for all-cause and the rate of dying and depending on normal-weight counterparts they estimate CVD-specific mortality for adults who are suffering overweight and obese, they state that The CVD mortality rate in obese adults was at least more than 20% which is compared with normal-weight adults.

3. System architecture

The system architecture of the prediction of obesity risk is as depicted in Fig. 1. This architecture is user-friendly, so anyone can use this system architecture. With this application, the user can view the user interface and can give the question and answer through the application. In this application, the user has to deliver some information about their daily activities, food routines, height, weight, etc. After fill-up, the form, the user sends a request to the server for the result, and then from there, the information collected form will go to the expert system. Periodically, the data is pre-processed and follows step the integration, missing value handling, box plot, data normalization, etc. Then prepare the data by splitting it into two sets. One part is the training set, and the other one is the set to test. The result will be resolved, dependent on the user

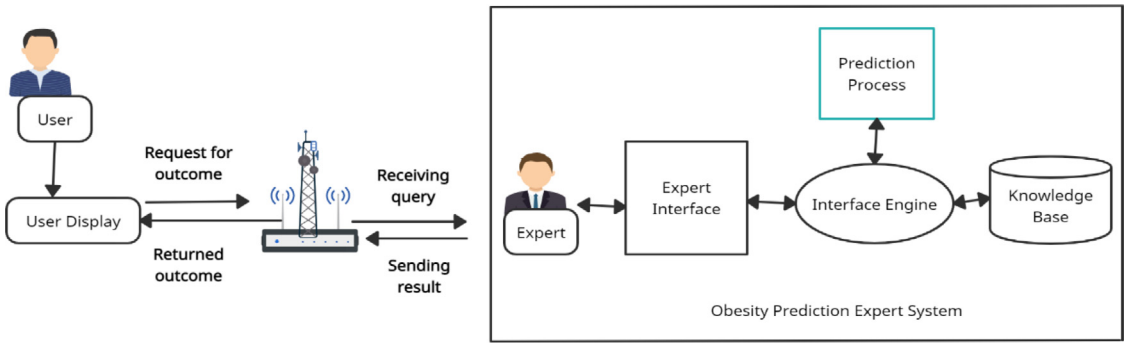

Fig. 1. The system architecture of the prediction of obesity.

input, by deploying a machine learning classifier of logistic regression algorithm to the processed information. Through the web application, the outcome will show in front of the user. The outcome showed the level of obesity which is low, medium, or high.

4. Research methodology

We have used 80% of the data for training data and the rest of the data were used for test data from our collected dataset. We collected 1100 data from different places and classes. We have shown our data collection and data preprocessing techniques in the previous section. We have used several machine-learning supervised algorithms such as *k*-NN, logistic regression, SVM, naïve Bayes, classification and regression trees (CART), random forest, multilayer perception (MLP), adaptive boosting (Ada Boost), and gradient boosting machine (GBM). Three times we calculated the accuracy. Before using principal component analysis (PCA), we calculated accuracy on the processed data, and that was the first time we calculated accuracy, after using PCA, we calculated it for the second time, and then finally the accuracies were calculated on the unprocessed data using the algorithm. We have evaluated metrics like sensitivity, specificity, precision, recall and F_1-score and the classifiers based on accuracy. The following flow diagram of these working processes has been described, which is given below in Fig. 2.

From our use of machine learning algorithms, the first one is *k*-NN. Based on the supervised learning technique, *k*-NN is one of the simplest machine learning algorithms. It is mostly used for the classification problems. This algorithm can be used for regression also including classification. *k*-NN doesn't make any assumption on the underlying data because it's a non-parametric algorithm, it works for similarity measures. The equation is given below.

$$\left(\sum_{i=1}^{k}\left(|x_i - y_i|\right)^q\right)^{1/q} \tag{1}$$

SVM is a flexible supervised machine learning algorithm, and it is used for both regression and classifications problems. It has the most unique way of implementation from other machine learning algorithms. A support vector machine edifies a set of hyperplanes in an infinite dimensional space, which can be used for regression, classification and the other outlier detection (12). To solve real-world problems, the SVM algorithm can be used. SVM builds a maximum margin separator, which is used for making decision boundaries with the largest possible distance. We can find out the separator, by using this equation which is given below.

$$W.X + b = 0 \tag{2}$$

The classification algorithm of logistic regression which is used to assign observations to an individual set of classes. It is a predictive analysis algorithm supported the concept of probability. Logistic Regression uses a more complex cost function which is defined by sigmoid function. The hypothesis of logistic regression estimates the limit of the cost function between 0 and 1. Linear functions fail to illustrate it in such a way because its value can be greater than 1 or less than 0, which is not possible according to logistical regression estimates (13).

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

A naïve Bayes is used for classification task, and it is a probabilistic machine learning model. It is mostly used in sentiment analysis, spam filtering, recommendation systems etc. Also, it is fast and easy to implement. This algorithm is based on the Bayes theorem and basic statistics. Class probabilities and conditional probabilities are used in the naïve Bayes model. It extends attributes using Gaussian distribution (Han et al., 2012). The Gaussian distribution with mean and standard deviation is described in below:

$$p\left(x = v|C_k\right) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \tag{4}$$

MLP refers to multilayer perception, it is a class of feed forward artificial neural network. When MLP had a single hidden layer at that time, it was referred to as "vanilla" neural network. It formed of at least three layers of nodes which are input, output, and hidden layer, each node uses nonlinear activation function except the input node. MLP utilizes back propagation for training, which is a supervised learning technique (15).

Another type of supervised machine learning algorithm is Decision trees where according to a certain parameter the data is continuously split. Decision nodes and leaves are the two entities of the decision tree. The decision tree needs a small pre-processing, and it can easily control the categorical features without preprocessing (16).

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{5}$$

Boosting algorithms is a set of low accurate classifiers. It creates a highly accurate classifier. These algorithms track the failed accurate prediction model. Also, it is less affected by the over fitting problem. In 1996, Ada Boost classifiers were proposed by Freund and Schapire, which is one of the ensemble boosting classifiers. To increase the accuracy, it combines multiple classifiers. It trained the data sample in each iteration and set the weights of classifiers. In that way, it ensures the accurate predictions of unusual observations (17).

$$\text{error}\left(M_i\right) = \sum_{j=1}^{d} w_j \times \text{err}\left(X_j\right) \tag{6}$$

For developing any model quickly, random forest is the great choice, to see how it performs training is given early in the model development process. It provides a pretty good indicator, assign to our feature. With some limitation, it is a fast, simple and flexible algorithm (18).

Gradient boosting machines are highly customizable to the special needs of the algorithm. The fundamental idea behind this algorithm is

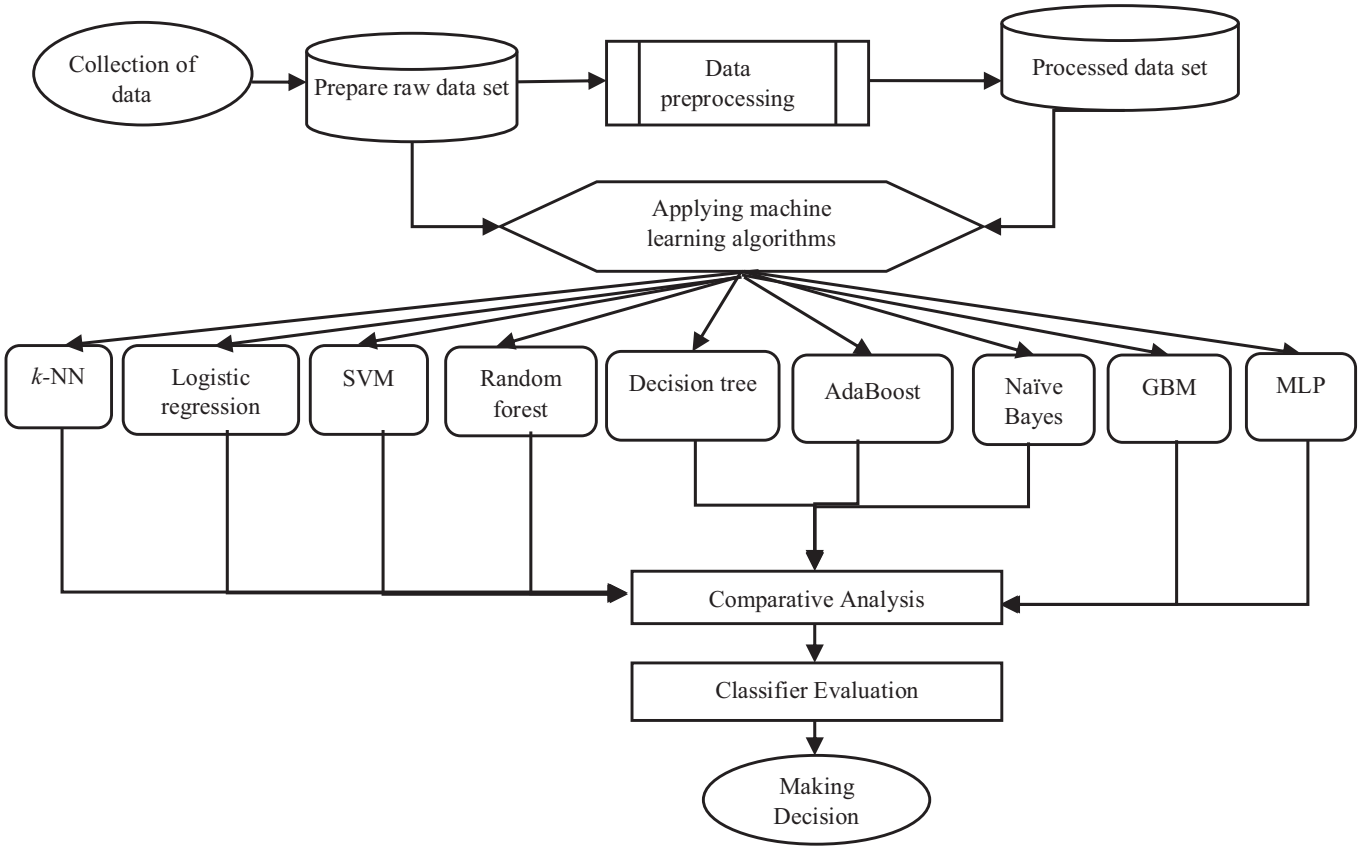

Fig. 2. Methodology of predicting the risk of obesity.

to build the new base-learners to be utmost correlated with the negative gradient of the loss function, associated with the complete ensemble. Gradient boosting machines can successfully capture complex nonlinear function dependencies and it is a powerful method. This family of models has shown considerable success in various practical applications. Moreover, the GBMs are highly flexible and can easily be customized to various real needs (Natekin, 2013).

We have inquired into the performance of the classifiers from the confusion matrix of dimension 3 × 3 since the number of class labels equals three. We have calculated class-wise accuracy, precision, recall, and F_1-score by using Eq. (9) to Eq. (12), respectively. We calculated sensitivity, specificity, precision, recall, F_1-score, ROC curve 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 positives. Recall means as the percentage of the total positive samples that were correctly predicted as positives by the classifier. F_1-score is the measurement of the harmonic mean of recall and precision. It considers both false positive and false negative values for calculation. Sensitivity and specificity are such factors who describe that how valid a test is.

Sensitivity can be measure using the equation of-

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \tag{7}$$

And specificity can be measure by this equation –

$$Specificity = \frac{TN}{FP+TN} \times 100\% \tag{8}$$

Precision can be defined as follows–

$$Accuracy = \frac{No.\ of\ correctly\ classified\ samples}{No.\ of\ tested\ samples} \times 100\% \tag{9}$$

One must examine both precision and recall to fully evaluate the effectiveness of any model.

Precision can be defined as follows –

$$Precision = \frac{TP}{TP+FP} \times 100\% \tag{10}$$

Recall can be defined as follows –

$$Recall = \frac{TP}{TP+FN} \times 100\% \tag{11}$$

F_1-score is the measurement of the harmonic mean of recall and precision. It considers both false positive and false negative values for calculation–

$$F_1\ \text{score} = \frac{2 \times precision \times recall}{precision + recall} \times 100\% \tag{12}$$

Class-wise TP, TN, FP, and FN are computed using Eq. (9) to Eq. (12) as shown by Habib et. al. (Habib et al., 2020).

The nine algorithms which have been used for our research, all of them have certain parameters. These parameters have different values which vary from each other. These parameter values are used for training the model and they are discussed in Table 1.

5. Description of features and data

In machine learning applications, data preprocessing plays a significant role in getting better performance and accurate results. In a similar way, the relevant feature selection process simplifies subsequent tasks and provides better performance in machine learning and data mining.

5.1. Selection of features

In this paper, features of obesity are very significant for detecting the problem. A set of features is blooming by analyzing the major causes of obesity, through which it is feasible to recognize the person who is

Table 1
Detailed specifications of the algorithms used.

Algorithm	Specifications of algorithm
k-NN algorithm	Number of neighbors = 1 Weight function used for prediction, weights = *c*, where *c* is a constant Power parameter, $p = 2$ Distance metric: Minkowski distance = $(\sum_{i=1}^{k}(\lvert x_i - y_i\rvert)^q)^{1/q}$
SVM algorithm	$C = 1.0$ Kernel: radial basis function = $exp(-\gamma \lVert x - -x_n \rVert^2)$ Gamma: $scale = \frac{1}{number\ of\ features \times X.var()}$
Logistic regression algorithm	Penalty = *l*2 $C = 1.0$ Number of random states = 0 Maximum number of iterations = 100
Naïve bayes algorithm	Distribution: Gaussian distribution = $f(x,\ \mu,\ \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ Mean,$\mu_y = \frac{1}{N}\sum_i x^{(i)}$ Variance,$\sigma_y = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$
Decision tree algorithm	Distribution measure: Gini index, $Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$ Maximum depth = 0 Minimum samples split = 2
AdaBoost algorithm	Number of estimators = 100 Learning rate = 1.0 Number of random states = 73
Random forest algorithm	Number of estimators = 100 Maximum depth = 2 Number of random states = 73
Multi-layer perception algorithm	Alpha = 0.0001 Hidden layer configuration = (5,2), Network architecture: 23-5-2-1 Number of random states = 0
Gradient boosting algorithm	Number of estimators = 2 Learning rate = 0.25 Maximum depth = 5 Number of random states = 0

Table 2
Features of obesity prediction.

Feature Name	Evidence Based-on	Feature Name	Evidence Based-on
Height	(23)	Spend time in front of screen	(24)
Weight	(23)	Spend time behind sitting	(24)
Gender	(23)	Eat back-to-back whole day	(24)
Age	(21)	Eating fast food everyday	(23)
Diet	(21)	Eating sugary foods everyday	(23)
Exercise regularly	(22)	Eating healthy food regularly	(23)
Gym regularly	(24)	Diabetes	(20)
Preferred walk a short distance	(25)	Side effects of medicine	(23)
Physical activity against will	(24)	Asthma problems	(21)
Diet before, not now	(24)	Obesity genetically	(23)
Prefer to spend lazy time	(23)	Heart disease	(21)
Depression/ stress	(21)	Social isolation	(21)
Insomnia	(21)	Smoking	(21)
Spend time on social media	(24)		

suffering from obesity. The feature lists of obesity are demonstrated in Table 2.

Class label: *High/Medium/Low*

For considering the risk of obesity, we identify each of these factors of obesity from (20, 21, 22, 23, 24, 25) these articles. Also, we were talking with some specialists on obesity that was emphasis these factors.

Table 3 describes the distribution of data, where we collect three class of data. Here we determine the obesity of high, medium, and low. We found that more than half of the data was a high class. It was 530. Another half of data distributed in medium and low, the numbers are respectively 330 and 240.

Table 3
Class-wise distribution of Data.

Obesity	Number of data
High	530
Medium	330
Low	240

5.2. Data collection and preprocessing method

This section was very difficult for us as it is the main stage for our research, so it was our main concern to collect the accurate data. We collect factors from some articles, journals, also we communicate with dietitians and nutrition specialists for gathering the main factors of obesity. For completing our research, we need to collect data. We collected the data both online and offline. We have collected our necessary data from various sources. For example, club members, students of schools, universities in Dhaka city are called offline data collection. On the other hand, collecting data from social media and made a google form to collect data from different people by distributing a link is called online data collection. We asked them our following questionnaires, and we took the answers from them in the survey paper. According to this, we collected all the necessary data. We were able to collect 1100 data based on 28 factors. Then we labeled the class of each record of the data set by consulting with some nutritionists and student counselors in educational institutions. Thus, our data set is collected.

Data processing is the ability to transform data into a suitable format after collecting data. From the collected data, we got some missing data, categorical data, numerical and text data. Then we decided we would make this data suitable for algorithms through our data processing. As shown in Fig. 3, we first started the work of data cleaning. We checked if there is a null value in the data set, then we encoded the level that converts the text data to numerical data. We solved the missing value problem. Then we checked if there is a noisy value in the data set using a box plot. Here we can see that there was some noisy data in the numerical data. Then we analyzed the correlation matrix as a data integration process. This matrix shows us the ratio of each data connected to each data. We removed noisy values by using outlier quantile detection. Then we dropped our outcome feature that was the obesity column. A separate histogram of each feature helped us with data reduction and data visualization in feature engineering. Through normalization, we completed the data transformation.

6. Experimental evaluation

By gathering 1100 data, our satisfactory data set was prepared. The feature's connectivity to other features is described by a correlation matrix. People can predict the level of obesity through the given factors and could analyze their obesity level.

For evaluation, we used the holdout method to elucidate in (Tan et al., 2006) and (Han et al., 2012) to evaluate our system. According to this method, we divide the entire dataset into two parts, used the training part to train the model (80%), and used the testing part to test the model (20. After empirical experimentation, we found that 80-20 gave use the best ratio. That means 80% for training data and 20% for testing data. Fig. 4 shows the variation of the values, how the accuracy changes.

Table 4 illustrates the correlation between the results attributes and other attributes. Correlation compares which features are correlated with the outcome. Additionally, when we were pre-processing the data at that time, we found some noisy values in our dataset, but they were not suitable for the dataset. We tried to solve the problem of the noisy values. For this solution, we used a box plot and manifested the appropriate consequence. Fig. 5. shows through the box plot, which features give the noisy values. In this figure, the 'diet' feature had the noisy value,

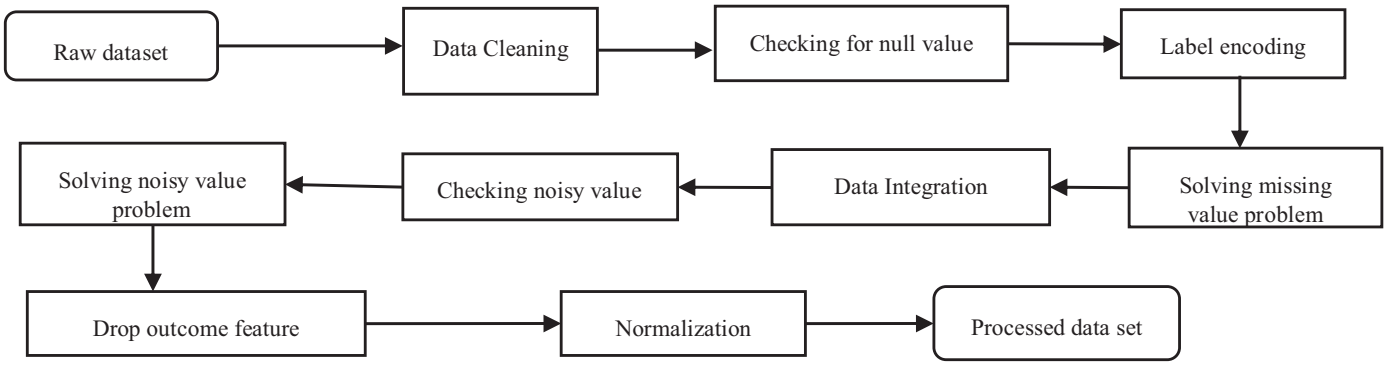

Fig. 3. Steps of data preprocessing of gathered data.

Table 4
Correlation between other features with outcome feature.

Features	Correlation Values	Features	Correlation Values
Prefer to walk a short distance	0.216545	Introvert	-0.052836
Height	0.152514	Physical illness	-0.054771
Gender	0.100857	Spend time in front of screen TV or Mobile	-0.057680
Buy healthy food according to hygiene	0.080971	Like sugary foods	-0.068258
Name	0.038369	Used to diet before but now you don't	-0.075094
Suffering depression	-0.003953	Asthma problems	-0.080893
Smoking	-0.005520	Keep fast food daily routine	-0.083271
Exercise regularly	-0.006557	Spend lazy time	-0.095184
Diet	-0.008761	Eat back-to-back whole day	-0.096575
Heart disease	-0.026166	Side effects of medicine	-0.115766
Time spend on social media per day	-0.034847	Spend time behind sitting	-0.120897
Regular gym	-0.039426	Age	-0.122173
Suffer insomnia	-0.043013	Obesity genetically	-0.234448
Doing physical activity against your will	-0.044772	Weight	-0.571634

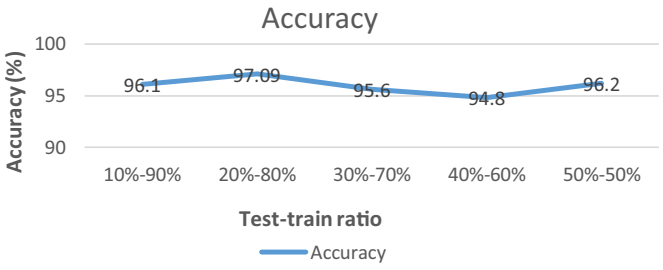

Fig. 4. Variation of the values.

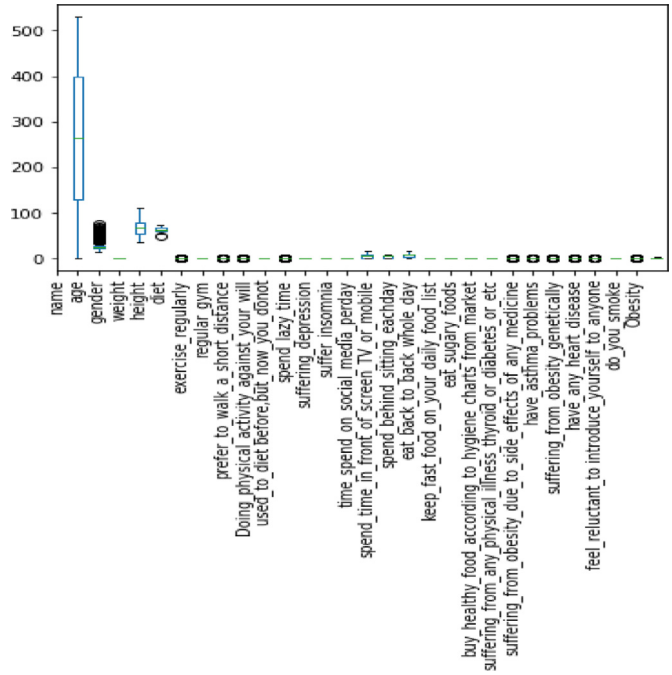

Fig. 5. Box plot with noisy value in 'diet'.

and after applying the outlier quantile we cleaned the noisy value furthermore solved this problem. Fig. 6. shows from the diet feature, the noisy value has been removed.

For our process data set, we ran nine machine learning algorithms where our features number was 28. Then we have used the PCA. PCA is one kind of dimension reduction technique that reduces redundancy from a data set. For uncorrelated variables, PCA allows extraction which is a linear combination of two or more of the original variables. PCA is a statistical procedure, a set of values of linearly uncorrelated variables is converted from a set of observations of possibly correlated variables with the use of an orthogonal transformation which is called PCA. The dimensionality of any model is the independent variable of a model. Only the important variables were selected for the next task after reducing the number of variables using PCA. Scree plot showed in Fig. 7 where the y-axis was explained by variance, and the x-axis showed the number of features. We use the scree plot to determine the number of factors, and the scree plot displays one curve, it checks whether PCA is working well or not. Here, we take 90% variance explained as a threshold, then calculate the principal component number, and we found 21 factors in 90% variance.

It appears that before using PCA, *k*-NN has achieved 83.0% accuracy, SVM has achieved 53.0% accuracy, logistic regression has achieved 50% accuracy, naïve Bayes has achieved 36% accuracy, the random forest has

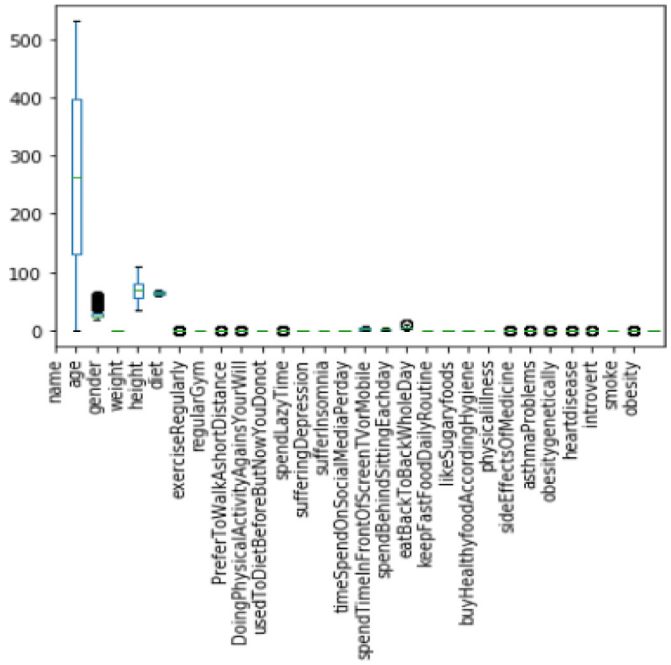

Fig. 6. Box plot without noisy value in 'diet'.

achieved 45% accuracy, decision tree has also achieved 50% accuracy, Ada Boost has achieved 49.0% accuracy, MLP has also acquired 49.0% accuracy, GBM has achieved 50.0% accuracy. After using PCA, we can see that the accuracy of some algorithm has increased and some has decreased and some algorithm has remained unchanged, *k*-NN has accomplished 77.5% accuracy, SVM has achieved 66.02% accuracy, logistic regression has achieved 97.09% accuracy, naïve Bayes has achieved 86.04% accuracy, the random forest has achieved 72.3% accuracy, decision tree has achieved 70.3% accuracy, Ada Boost has reached 70.30% accuracy, MLP has achieved 66.02% accuracy and GBM has achieved 64.08% accuracy. The difference in the accuracy of the algorithm obtained before and after the use of PCA is shown in Fig. 8. We have also calculated the accuracy with an unprocessed data set. *k*-NN has achieved 80.37% accuracy, SVM has achieved 50.4% accuracy, logistic regression has achieved 46.74% accuracy, naïve Bayes has achieved 71.57% accuracy, the random forest has achieved 21.38% accuracy, decision tree has reached 92 % accuracy, Ada Boost has achieved 21.38% accuracy, MLP has achieved 49.49% accuracy and GBM has achieved 49.38% accuracy with the unprocessed data set.

Table 5 narrates the performance of each of the nine algorithms. The performance is determined according to its accuracy, sensitivity, specificity, precision, recall, and F_1-score. The algorithm, which will fit best for our problem domain would be determined according to its performance. The highest performance giver algorithm would be chosen as the most suitable algorithm. It is shown in the table that logistic regression has given the highest accuracy, specificity, precision, and F_1-score. Again, based on sensitivity, and recall MLP performs better. However, the other performances of gradient boosting were not good. So, considering everything, the best performance of the model was found using a logistic regression algorithm.

Table 5
Classifier performance evaluation.

Algorithms	Accuracy	Sensitivity	Specificity	Precision	Recall	F_1-score
k-NN	77.5%	100%	100%	79%	77%	77%
SVM	66.02%	100%	nan	53%	66%	56%
Logistic Regression	97.09%	100%	100%	97%	97%	97%
Naïve Bayes	86.04%	100%	100%	86%	86%	86%
Random forest	72.3%	94.11%	100%	57%	72%	63%
Decision tree	70.3%	90.19%	100%	57%	70%	61%
ADA boosting	70.3%	90.69%	100%	57%	70%	61%
MLP	66.02%	100%	65.38%	49%	66%	56%
Gradient boosting	64.08%	78.43%	100%	55%	65%	57%

7. Comparative analysis of results

We need to compare our work with some other relatives' work, after that we can evaluate the goodness of our proposed obesity prediction system. After following a good amount of research paper, we saw that there was a lot of work done for predicting, but it's difficult to find the work which is relative to ours. A little amount of work, which has been done, on childhood obesity risk prediction, 2 types of diabetics risk prediction from obesity, disease prediction, etc. Nevertheless, we have combated to compare our work with others based on certain parameters. A comparative overview of other works and ours is given below-

Table 6 determines the comparison of our work and other works. Dugan (Dugan et al., 2015) work to predict obesity, they used six models to check the accuracy. Jindal (Jindal et al., 2018) works for predicting obesity using ensemble machine learning approaches. Hammond (Hammond et al., 2019) predicts childhood obesity using electronic health records and used data which is publicly available. They used logistic regression with L1 loss, a model of random forest, and the gradient boosting classifier to predict their dichotomous measures of low obese/medium obese/ high obese to predict their continuous BMI values. They employed LASSO regression, random forest regression, and gradient boosting regression. Dunstan (Dunstan et al., 2019) predicts nationwide obesity from food sales.

From the above table, we can say that our accuracy is better than others and also our work is done in the context of Bangladesh. We col-

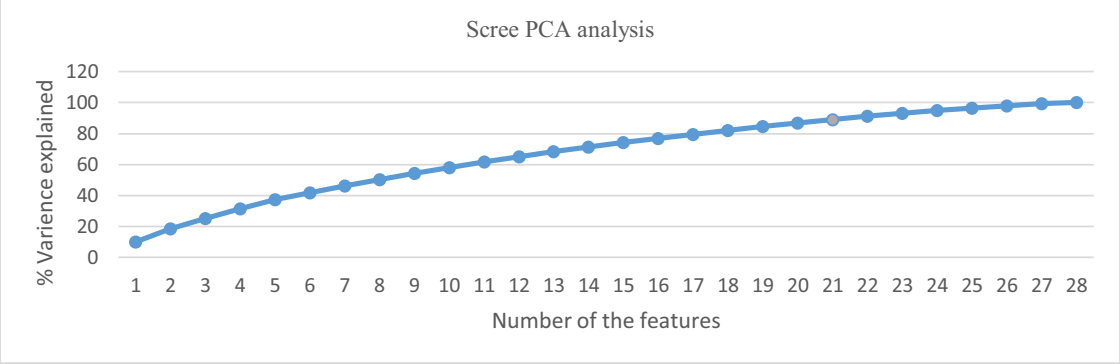

Fig. 7. Scree plot where the number of principal components is shown in gray color.

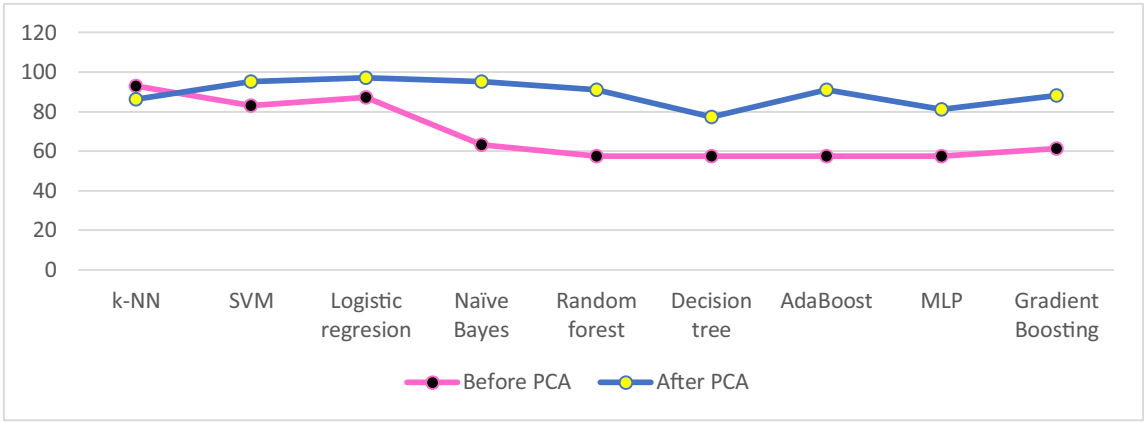

Fig. 8. Comparison of accuracy between before PCA and after PCA.

Table 6
Results of the comparison of our work and other works.

Method/ Work Done	Prediction Dealt with	Problem Domain	Sample Size	Size of Feature Set	Algorithm	Accuracy
This Work	Obesity risk prediction using machine Learning	Prediction	1100	28	Logistic Regression	97.09%
Dugan (Dugan et al., 2015 Aug 12)	Early childhood obesity	Prediction	7519	3	Naïve Bayes	85%
Jindal (Jindal et al., January 2018)	obesity using ensemble machine learning	prediction	5000	4	Random Forest	89.68%
Hammond (Hammond et al., October 7, 2019)	Childhood obesity using electronic health record	prediction	52945	23	LASSO regression	81.8% for girls and 76.1% for boys
Dunstan (Dunstan et al., May 19, 2019)	Nationwide obesity from food sales	prediction	3792	20	Random Forest	10 % for about 60% countries considered and below 20% for 87% of countries

lected data from the people of different classes and ages. But it's true that our data set is not a bigger data set like others but we tried our best to finds the best accuracy applying various machine learning algorithms, and we came out with a satisfactory result.

8. Conclusion

In this paper, our aim was to predict obesity in the context of Bangladesh and we collected data on this basis. We have conducted in-depth research using various machine learning techniques to predict the risk of obesityr. The risk forecast for obesity has been completed by nine explicit classifications. The merits of those classifiers have been measured in terms of six conspicuous performance metrics. The relative merits of the results achieved have been assessed by analyzing the results of similar works thereafter. The accuracy came out from logistic regression with a value of 97.09%. Our future plan is to make this work more rigorous with a bigger data set to cover as much a wider range of low-obese and medium-obese and high-obese people as required for Bangladesh.

Data and Code availability statement

Data originate from a source material that depends largely on the focus of the experiment. We collect this data from general people, doctors, and sufferers of obesity. Data cane be accessed at https://drive.google.com/drive/folders/1iJoEDtLE9XVNceXVIdt-HEjeAYZNbISV?usp=sharing Some data are restricted for access- Data originate from a source material that depends largely on the focus of the experiment. We collect this data from general people, doctors, and sufferers of obesity. Data cane be accessed at https://drive.google.com/drive/folders/1iJoEDtLE9XVNceXVIdt-HEjeAYZNbISV?usp=sharing Some data are restricted for access

Declaration of Competing Interest

The authors declare having no conflict of interest.

References

Abbas, J., Aqeel, M., Abbas, J., Shaher, B., A, J., Sundas, J., Zhang, W, 2019. The moderating role of social support for marital adjustment, depression, anxiety, and stress: Evidence from Pakistani working and nonworking women. J. Affect. Disord. 244, 231–238. doi:10.1016/j.jad.2018.07.071.
Available at- https://en.wikipedia.org/wiki/Support-vector_machine.
Available at- https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148.
Available at- https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html.
Available at- https://www.datacamp.com/community/tutorials/adaboost-classifier-python.
Available at- https://builtin.com/data-science/random-forest-algorithm.
Available at- https://en.wikipedia.org/wiki/Multilayer_perceptron.
Borrell, L.N., Samuel, L., 2014. Body mass index categories and mortality risk in us adults: the effect of overweight and obesity on advancing death. Am. J. Public Health 104 (3). available at- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3953803/ .
Diabetes and obesity, Available at https://www.diabetes.co.uk/diabetes-and-obesity.html.
Dugan, T.M., Mukhopadhyay, S., Carroll, A., Downs, S., 2015. Machine learning techniques for prediction of early childhood obesity. Appl. Clin. Inform. 6 (3). available at- https://www.karger.com/Article/Fulltext/496563 .
Dunstan, J., Aguirre, M., Bastías, M., Glass, T.A, Tobar, F., 2019. Predicting nationwide obesity from food sales using machine learning. Health Inf. J. 26 (issue:1), 652–663. pagesavailable at- https://journals.sagepub.com/doi/full/10.1177/1460458219845959 .
Gerl, M.J., Klose, C., Surma, M.A., Fernandez, C., Melander, O., Männistö, S., Borodulin, K., Havulinna, A.S., Salomaa, V., Ikonen, E., Cannistraci, C.V., Simons, K., 2019. Machine learning of human plasma lipidomes for obesity estimation in a large population cohort. Journal. available at- https://pubmed.ncbi.nlm.nih.gov/31626640/ .

Habib, M.T., Mia, M.J., Uddin, M.S., Ahmed, F., 2020. An in-depth exploration of automated jackfruit disease recognition. J. King Saud University – Comput. Inf. Sci. doi:10.1016/j.jksuci.2020.04.018.

Hammond, R., Athanasiadou, R., Curado, S., Aphinyanaphongs, Y., Abrams, C., Messito, M.J., Gross, R., Katzow, M., Jay, M., Razavian, N., Elbel, B., 2019. Correction: Predicting childhood obesity using electronic health records and publicly available data. J., PLS ONE 14 (4). available at- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0223796 .

Han, J., Kamber, M., Pei, J., 2012. Data Mining Concept and Technique. Morgan Kaufmann, pp. 332–398.

Han, J., Kamber, M., Pei, J., 2012. Data Mining Concepts and Techniques. Elsevier.

Height Available at, https://www.niddk.nih.gov/health-information/weight-management/adult-overweight-obesity/factors-affecting-weight-health.

Height Available at, https://www.niddk.nih.gov/health-information/weight-management/adult-overweight-obesity/factors-affecting-weight-health.

https://onlinelibrary.wiley.com/doi/full/10.1038/oby.2005.103.

Jindal, K., Baliyan, N., Rana, P.S., 2018. obesity prediction using ensemble machine learning approaches. In: Proceedings of the 5th ICACNI 2017, 2, pp. 355–362 available at-.

Montañez, C.A.C., et al., 2017. Machine learning approaches for the prediction of obesity using publicly available genetic profiles. In: 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, pp. 2743–2750. doi:10.1109/IJCNN.2017.7966194.

Muhamad Adnan, M.H.B., Husain, W., Rashid, N.Abdul, 2012. A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction. In: 2012 International Conference on Computer & Information Science (ICCIS), Kuala Lumpur, Malaysia, pp. 281–285. doi:10.1109/ICCISci.2012.6297254s.

Natekin, A., Knoll, A., 2013, available at- https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021/full.

Salahuddin, T., 2021. "The daily star", September 23, 2018 /LAST MODIFIED: 02:04 AM, December 25, 2018, available at- https://www.thedailystar.net/health/obesity-increasing-in-bangladesh-younger-generation-1637107.

Singh And, B., Tawfik, H., 2020. Machine learning approach for the early prediction of the risk of overweight and obesity in young people. Int. Conf. Comput. Sci. 12140, 523–535. available at- https://link.springer.com/chapter/10.1007%2F978-3-030-50423-6_39 .

Social isolation Available at https://www.mayoclinic.org/diseases-conditions/obesity/symptoms-causes/syc-20375742.

Tan, P.-N., Steinbach, M., Kumar, V., 2006. Introduction to Data Mining. AddisonWesley.

What is obesity is Available at, https://www.medicinenet.com/obesity_weight_loss/article.htm.

World Health Organization, available at- https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.