Chapter 37
Obesity Prediction Based on Daily Lifestyle Habits and Other Factors Using Different Machine Learning Algorithms

Chalumuru Suresh, B. V. Kiranmayee, Milar Jahnavi, Roshan Pampari, Sai Raghu Ambadipudi, and Sai Srinivasa Preetham Hemadri

1 Introduction

Obesity is regarded as a buildup of fat in the body that can cause serious medical issues. Obesity is not just a cosmetic concern. It is a health condition that increases the risk of developing several illnesses such as high blood pressure, cancer, heart disease, diabetes, and so on.

Body mass index (BMI) is generally considered as the major factor to describe whether a person is suffering with obesity or not. According to the BMI value, there are 4 distinct weighing statuses. An individual is regarded as underweight if his or her BMI is beneath 18.5. A person's BMI is normal if it falls in midway of 18.5 and 24.9 (both included). If a person's BMI is in midway of 25.0 and 29.9 (both included), they are overweight, and if their BMI is 30.0 or higher, they are obese. The body mass index value of an individual is generally calculated by using height in meters and weight in kilograms.

Obesity is not determined just using BMI value, but there are many other factors which are useful to determine obesity in a person. The factors can include unhealthy diet, inactivity, age, gender, lack of sleep, family inheritance, drinking/smoking habits, basal metabolic rate (BMR), resting metabolic rate (RMR), body fat percentage (BFP), protein recommended dietary allowances (RDA). In this study, machine.

C. Suresh · B. V. Kiranmayee · M. Jahnavi (✉) · R. Pampari · S. R. Ambadipudi · S. S. P. Hemadri
Department of CSE, VNR VJIET, Hyderabad, Telangana, India

C. Suresh
e-mail: suresh_ch@vnrvjiet.in

B. V. Kiranmayee
e-mail: kiranmayee_bv@vnrvjiet.in

A. B. Reddy et al. (eds.), *Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems*, Algorithms for Intelligent Systems, https://doi.org/10.1007/978-981-16-7389-4_39

2 Literature Survey

Jindal et al. [1] proposed a model which uses body mass index, basal metabolic rate, resting metabolic rate, body fat percentage, protein recovery dietary allowance, and it predicts whether the person is obese or not. It makes use of machine learning techniques to obtain reliable obesity estimates in a range of scenarios.

Sun et al. [2] included data collection and processing, statistical analysis, and model fitting. It used descriptive statistics, cross-category analysis, model-based analysis, and model-based prediction for developing the obesity prediction model. It developed a predictive model which evaluates children's chances of adhering to one of four BMI categories five years later using currently measured determinants [sex, area of residence, age, body height, and body mass index (BMI)] and specifies the elevated risk group for possible implications using currently measured determinants.

Al Kibria et al. [3] have used dataset consisting of both rural and urban areas. Race, age, gender, relationship status, education level, active contraceptive use, ethnicity, and spirituality were all reported by participants. BMI is also calculated. Based on all these factors, the overweight, underweight, and obesity are predicted.

Singh and Tawfik [4] studied the risk of overweight and obesity in youngsters and found that it can be predicted early. They employed a machine learning algorithm to compute an individual's BMI at three, five, seven, and eleven years of age. The machine learning model will predict if that individual is normal or at risk at the age of 14.

Luhar et al. [5] have forecasted the risk of overweight and obesity levels of an Indian individual of aged 20–69 years using age and sex to 2040. They have also used information about where they live like they live in rural or urban areas.

Cervantes and Palacio [6] applied computational intelligence to estimate the obesity levels in students of age between 18 and 25 years. They have used decision trees, random forest, support vector machine, and k-means algorithm for the estimation.

Thamrin et al. [7] applied logistic regression, classification and regression trees (CART), and Naïve Bayes to estimate the obesity in adults. Logistic regression was utilized as a machine learning model based on the accuracy ratings. It also addressed the problem of data imbalance in predicting obesity status.

Molina et al. [8] employed data mining techniques to create an obesity prediction model. Logistic model tree, random forest, multi-layer perceptron, and support vector machine are among the classification methods employed. The logistic model tree was selected for prediction since it offers the highest precision.

3 Proposed System

Generally, most of the people across the world are facing issues with obesity. Even though it is the major concern for health, people were not interested to know their

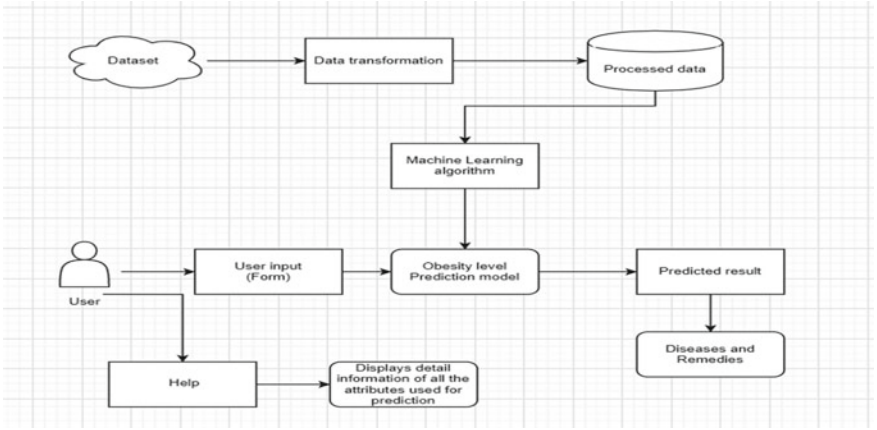


Fig. 1 System architecture for obesity prediction

obesity levels. In recent studies, there were some solutions for this obesity problem which were calculated based on BMI that depends only on weight and height of an individual to know their obesity levels. But there are many diseases that occurred due to obesity which was not predicted earlier. Based on the previous solutions, the implemented solution predicts the obesity levels with diseases along with the remedies they need to take.

In this paper, there are 22 factors that determine the obese levels which include unhealthy diet, inactivity, age, gender, lack of sleep, family inheritance, drinking/smoking habits, basal metabolic rate (BMR), resting metabolic rate (RMR), body fat percentage (BFP), protein recommended dietary allowances (RDA) that gives better results.

A solution is implemented in order to overcome from this problem which includes two phases.

1) Designing a ml model
2) Filling the form (Fig. 1).

3.1 Phase 1 (Designing a ML Model)

Step 1: Importing Dataset. The data contains the values of columns including age, gender, height, weight, obesity of parents, smoking habits, consumption of vegetables, water intake, physical activity, consumption of alcohol that are used to predict obesity, and is stored in a csv file. And the data must be imported into a data frame for further computations. Pandas, which is a Python library, provides a function read_csv; using this function, the data in the csv can be imported into the data frame. The dataset is derived from an article, and it consists of 2111 records with 22 fields.

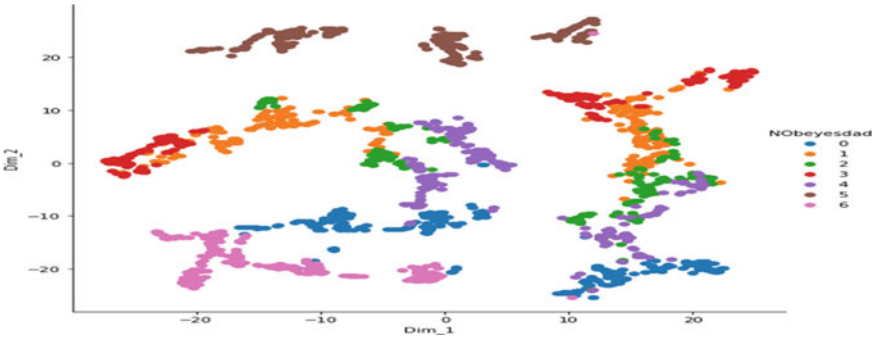


Fig. 2 T-SNE plot, 2-D representation of the dataset using T-SNE

Visualization of data using T-SNE plot. Figure 2 represents a t-distributed stochastic neighbor embedding (T-SNE) which is used to convert multi-dimensional data to two-dimensional (2-D) data. Here, "NObeyesdad" is the target variable.

Step 2: Data preprocessing. The data that is previously imported may contain null values, strings, outliers. So, the data must be properly preprocessed before applying any machine learning model.

Label encoding. Generally, most of the data will be in strings and numbers. But machine can only understand numerical values, so the strings in the data must be converted into numerical values instead of totally removing them. To make this possible, label encoding is used. For example, if there is a gender column, then usually data is stored as M, F. This function converts all the M's to 0 and all the F's to 1 or vice versa and similarly for all other columns having character values.

Handling null values (missing values). After converting all the column values to the numerical, check for any missing values in the data. If there are very few missing values, then delete the rows having null values and proceed with further process. In the other case, they must be filled with either mean, median or mode of their respective columns, and median is most preferred.

Outlier detection and removal. A observation in a random sampling that departs substantially from other values is called an outlier. In certain ways, this concept defers to the analyst in determining what constitutes aberrant behavior. Here, boxplots are used to detect outliers.

Figure 3 shows the outlier ends for each and every column.

Boxplot. Boxplot is used to display how the data is distributed graphically. It looks like a box attached with two strings at both the ends. At the center of the box, the median of the data is represented (Q2), the left apex of the box is the 25th percentile (Q1), and the right apex of the box is the 75th percentile (Q3). (Q3 − Q1) is called the interquartile range or IQR. Then, the left extreme is calculated as Q1 − 1.5 * IQR. The right extreme is calculated as Q3 + 1.5 * IQR. Any point in the data to which boxplot is drawn staying in the range of left extreme and right extreme is picked and used for further processing. And the other points that are less than left extreme and

```
outlier ends for age are  10.867980000000003 35.079212
outlier ends for height are  142.23039999999997 197.61600000000004
outlier ends for Weight are  2.537334499999993 170.3666905
outlier ends for BMI are  6.789753367500005 53.552549307499994
outlier ends for BMR are  643.9741947500006 2991.9270367499994
outlier ends for RMR are  754.51838625 2683.34549625
outlier ends for ProteinRDA are  2.029867600000024 136.29335239999997
outlier ends for Fatpercentage are  -2.144737642500008 60.063408417500014
outlier ends for FCVC are  0.5 4.5
outlier ends for NCP are  2.146845 3.5118929999999997
outlier ends for SMOKE are  0.0 0.0
outlier ends for CH2O are  0.2459012499999993 3.816331250000001
outlier ends for FAF are  -2.18875375 3.97993625
outlier ends for TUE are  -1.5 2.5
```

Fig. 3 Outlier values of dataset

greater than right extreme are considered as outliers and are removed from the rest of the data (Fig. 4).

Boxplots from the data which shows outliers are present in the dataset.

Figure 5 represents a boxplot which visualizes how the data is distributed for the column "BMI." Here, BMI column has no outliers.

Figure 6 represents a boxplot which visualizes how the data is distributed for the column "NCP." Here, NCP column has many outliers.

Target variable separation. The target variable (prediction value column) column must be separated from the rest of the columns to fit the data into a machine learning model.

Figure 7 represents a correlation matrix which shows correlation coefficients between each column. Each box in the table shows how two columns are correlated. The above correlation matrix summarizes data.

Fig. 4 Boxplot

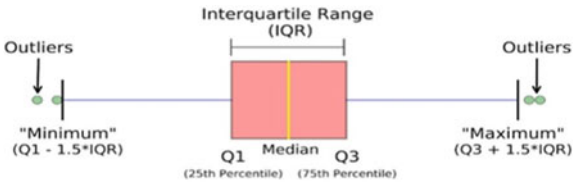
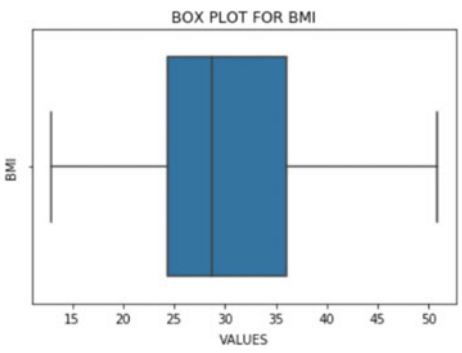


Fig. 5 Outlier boxplot for BMI



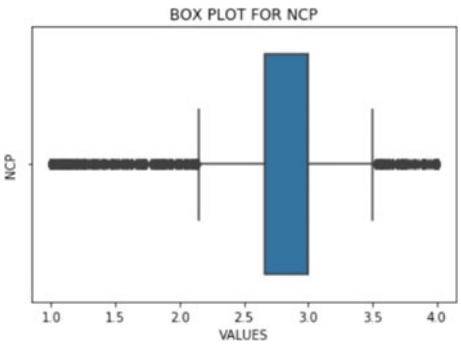


Fig. 6 Outlier boxplot for NCP

Fig. 7 Correlation matrix

Step 3: Splitting data and Applying Machine Learning model.

Splitting data into training and testing data. The data must be partitioned into train data (70% of the total data) and test data (30% of whole data) to test the accuracy of the applied machine learning model. Then applying the machine learning model on the train data and test the accuracy of the ML mode using test data.

Applying ML model. Random forest classifier is one of the best supervised machine learning algorithms. It splits the data into multiple subsets and applies decision tree algorithm individually to each subset. All the results obtained from those decision tree algorithms are collected, and mean (or) mode is calculated depending on the type of target variable. Now, this ML model is put into the back end of the Web site using Django.

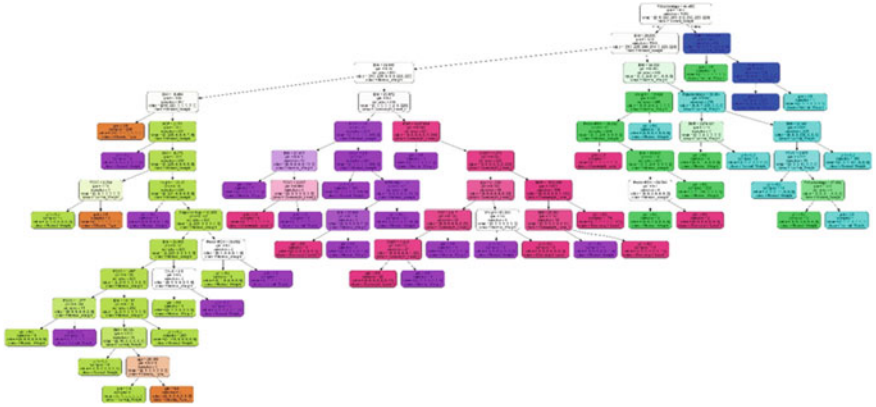


Fig. 8 Decision tree plot based on dataset

3.2 Phase 2 (Filling the Form)

This step is about filling the form which consists of factors that lead to obesity of a person. The factors are age, gender, height, weight, obesity of parents, smoking habits, consumption of vegetables, water intake, physical activity, consumption of alcohol.

After filling the form, click on the submit button where it displays the type of obesity level of that person. The result includes obesity level and risks that may occur due to obesity, and if the cursor is hovered over each risk, it will display remedies for that risk.

Figure 8 represents a decision tree which is a tree like structure, in which any particular internal node denotes a test on a column, a branch between two internal nodes represents a result of the test, and every leaf node represents a class value.

3.3 Comparative Study

On the basis of their accuracy scores, a comparative analysis of numerous machine learning algorithms such as random forest, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN) has been conducted.

Table 1 shows a detailed comparison of several machine learning algorithms focusing on their accuracy scores. From the table, the conclusion is that random forest is more efficient than other algorithms based on the accuracy ratings.

Figure 9 shows a bar graph which plots various machine learning algorithms on *X*-axis and their accuracy values on *Y*-axis.

Random Forest. Random forest uses decision tree algorithm. It is a parallel process unlike decision tree which is a step process. Here, it splits the whole dataset

Table 1 Comparative study of various machine learning algorithms

ML algorithms	Accuracy score (%)
K-nearest neighbor	78.97
Support vector machine (SVM)	96.21
Decision tree	96.96
Random forest	98.48

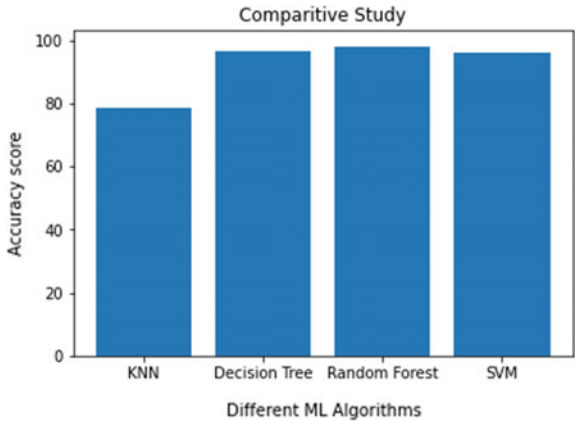


Fig. 9 Comparative study of various machine learning algorithms based on their accuracy scores

into multiple data subsets in which the number of rows of each data subset must be same but not necessarily the number of columns is same.

For each data subset, random forest algorithm applies decision tree algorithm individually and extracts the results from each decision tree. It calculates the final output by mode of all the outputs from the decision tree. It is called bootstrapping because the data is split and then the result is calculated by aggregation.

Here, the root node selection in the random forest and decision tree algorithms is done with the parameter Gini index. Greater the Gini index value, the higher the level of the attribute in the random forest and decision tree.

4 Experimental Analysis

The home page of the Web site is depicted in Fig. 10, which describes information about obesity and various risks associated with being underweight, overweight, or obese.

Figure 11 shows the survey page which has to be filled by the user. This form contains 21 fields which are used to predict various obesity levels.

Figure 12 shows the result which is predicted by the machine learning model based on the input given by the user. It also displays possible risks (diseases) associated with respective to the result.

Fig. 10 Home page

Fig. 11 Survey form

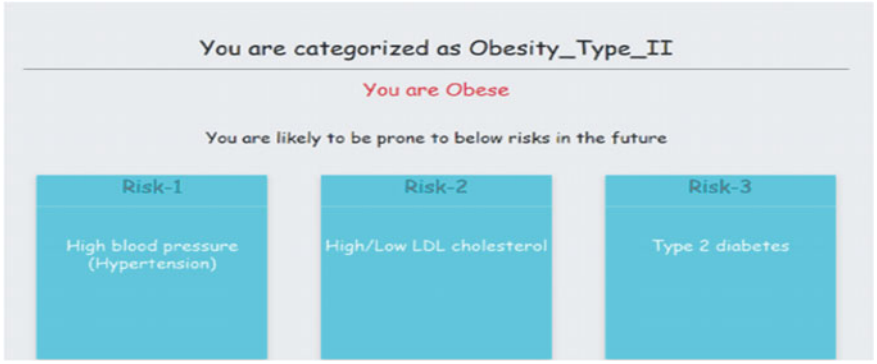


Fig. 12 Result page

Fig. 13 Result page with remedies

Figure 13 shows the remedies if the cursor is hovered above the risk. Detailed description about the remedies can be accessed by clicking on the hyperlink.

5 Conclusion

Obesity is considered as the most common disease in the modern world. As such, it is important for the individuals to know which category of weight class they belong to so that they can be healthy. So, in this study, individual's weight is classified into seven categories based on 21 factors. Based on the results of the classification, the risks and diseases that individuals may have are predicted. Based on the diseases, the remedies are also provided. Early recognition of the risks associated with obesity and overweight will be helpful to prevent many diseases.

References

1. K. Jindal, N. Baliyan, P.S. Rana, Obesity prediction using ensemble machine learning approaches, in *Recent Findings in Intelligent Computing Techniques*, ed. by P. Sa, S. Bakshi, I. Hatzilygeroudis, M. Sahoo. Advances in Intelligent Systems and Computing, vol. 708 (Springer, Singapore, 2018), pp. 355–362
2. Y. Sun, Y. Xing, J. Liu, Five-year change in body mass index category of childhood and the establishment of an obesity prediction model. Sci. Rep. **10**, 10309 (2020)

3. G. Al Kibria, K. Swasey, M.Z. Hasan, Prevalence and factors associated with underweight, overweight and obesity among women of reproductive age in India. Glob. Health Res. Policy **4**, 24 (2019)
4. B. Singh, H. Tawfik, Machine learning approach for the early prediction of the risk of overweight and obesity in young people, in *Computational Science—ICCS 2020*, ed. by V. Krzhizhanovskaya. Lecture Notes in Computer Science, vol. 12140 (Springer, Cham, 2020), pp. 523–535
5. S. Luhar, I.M. Timæus, R. Jones, S. Cunningham, S.A. Patel, S. Kinra, L. Clarke, R. Houben, Forecasting the prevalence of overweight and obesity in India to 2040. PLoS ONE (2020)
6. R.C. Cervantes, U.M. Palacio, Estimation of obesity levels based on computational intelligence. Inform. Med. Unlocked **21**, 100472 (2020)
7. S.A. Thamrin, D.S. Arsyad, H. Kuswanto, A. Lawi, S. Nasir, Predicting obesity in adults using machine learning techniques: an analysis of Indonesian basic health research 2018. Front. Nutrition **8**, 669155 (2021)
8. E.D. Molina, D.K. Kevin, M.P. Fabio, Classification and features selection method for obesity level prediction. J. Theor. Appl. Inf. Technol. **99**(11), 1992–8645 (2021)