Using Machine Learning Method for classification
Body Mass Index for clinical decision

Firouz Amani (firouzamani2019@gmail.com)
Ardabil University of Medical Sciences

Alireza Mohammadnia
Ardabil University of Medical Sciences

Paniz Amani
University of Tabriz

Research Article

Keywords: Body Mass Index, Machine Learning, Classification, Algorithms

Posted Date: February 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-250227/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

**Background:** Body mass index (BMI) is a good method for measuring overweight and obesity among people. The aim of this study was to develop a machine learning method to classify BMI for clinical application.

**Methods:** In this study, we used the dataset of 1316 people who were selected randomly from all areas of Ardabil city. The dataset included demographic and anthropometric data. Classification algorithms such as Random Forest (RF), Gaussian Naive Bayes (GNB), Decision Tree (DT), Support-Vector Machines (SVM), Multi-layer Perceptron (MLP), K-nearest neighbors (KNN) and Logistic Regression (LR) were used for classification of people based on BMI data. The performance of these algorithms were evaluated using Precision, Recall, Mean Squared Errors (MSE) and Accuracy. All programming was done in Python 3.7 in Jupyter Notebook.

**Results:** According to BMI, 603 (45.8%) of all samples were normal and 713 (54.2%) were at-risk. The precision of RF, GNB, DT, SVM, MLP, KNN and LR for people at risk was 0.93, 0.86, 0.99, 0.82, 1.00, 0.82 and 0.99 respectively. Also, the accuracy of RF, GNB, DT, SVM, MLP, KNN and LR were 95%, 83%, 100%, 82%, 100%, 82% and 100%.

**Conclusion:** In comparison, classification algorithms results showed that the LR, MLP and DT had the higher full accuracy than other algorithms in detecting people at risk.

Background

Obesity and overweight are a complex, multifactorial, and major public health problem worldwide which could affect people in all age groups and increase the risk of many diseases among people [1–2]. This index is used to measure the obesity and overweight and detection of people at risk of obesity and overweight [3]. BMI is defined as a person's weight in kilograms divided by the square of his height in meters (kg/m2). According to the WHO reports, BMI less than 25 was considered normal, and more than 25 as at-risk of overweight and obesity. By 2030, approximately 38% of the world’s elderly population will be obese [2–4].

Recently, obesity and being overweight are increasing rapidly in the developed and developing countries and it is estimated that by 2030 due to many factors, up to 57.8% of the world’s elderly people would suffer from being overweight or obese [5–6].

The aim of this study was to investigate the classification of BMI by using several machine learning algorithms.

Methods

1. Data collection method and dataset
In this study, the used dataset is from the data used in the obesity and overweight research which approved before by Ardabil University of Medical Science and part of data published in a paper by Amani et al [7].

The used dataset included the information BMI of 1316 people of Ardabil city at year 2019.

The detailed clarification about this dataset is given in Table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sample Size</th>
<th>Feature size including class label</th>
<th>Classes</th>
<th>Presence of missing attribute</th>
<th>Presence of noisy attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>1316</td>
<td>8</td>
<td>2</td>
<td>NO</td>
<td>NO</td>
</tr>
</tbody>
</table>

**Machine learning strategies**

The aim of the study was to classification of people based on BMI data by using seven classification machine learning algorithms such as RF, GNB, DT, SVM, MLP, KNN and LR.

**Logistic Regression (LR)** is a machine learning technique for regression and classification problems which to assign observations to a discrete set of classes.

**Gaussian Naive Bayes classifier (GNB)** is a group of simple classifiers based on probabilities created assuming the independence of random variables and based on Bayes theorem.

**Decision Tree (DT)**

A decision tree is a map of the possible results of a series of related choices or options so that it allows an individual or organization to weigh possible actions in terms of costs, opportunities, and benefits.

**Support vector machine (SVM)** is classified as a pattern recognition algorithm. The SVM algorithm can be used wherever there is a need to identify patterns or classify objects in specific classes.

**Multi-layer Perceptron (MLP)**

The artificial neural network creates a structure similar to the biological structure of the human brain and neural network to be able to learn to generalize and make the decision.

**Random Forest (RF)** is a combined learning method for regression classification, which works on the training time and class output (classification) or for predictions of each tree separately, based on a structure consisting of a large number of decision trees.

**K-nearest neighbors (KNN)**
In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951, and later expanded by Thomas Cover. It is used for classification and regression.

**Data preprocessing**

Data preprocessing is necessary to prepare the BMI data in a manner that a machine learning model can accept. Separating the training and testing datasets ensures that the model learns only from the training data and tests its performance with the testing data. The dataset was divided into training and test data. The training data contain 80% of the total dataset, and the test and validation data contain 20% each.

**Machine learning model selection**

Seven classification algorithms such as RF, GNB, DT, SVM, MLP, KNN and LR were applied and used to trained and evaluated training datasets.

**Variable selection**

Features selection for classification model attempts to select minimally sized subset according to following criteria: (1) The classification accuracy should have to increase; (2) The values for the selected features should have to close as possible to the original class distribution. All features are listed in the Table 2. The response variable in dataset of our study was BMI of patients which divided into two classes: Normal (18.5 ≤ BMI < 25) and At-risk (25 ≤ BMI).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Class Type</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>2 class [1, 2]</td>
<td>Integer</td>
</tr>
<tr>
<td>Age</td>
<td>20−49</td>
<td>Integer</td>
</tr>
<tr>
<td>WHR</td>
<td>0−2.27</td>
<td>Integer</td>
</tr>
<tr>
<td>WC</td>
<td>45−160</td>
<td>Integer</td>
</tr>
<tr>
<td>HC</td>
<td>37−160</td>
<td>Integer</td>
</tr>
<tr>
<td>Height</td>
<td>110−194</td>
<td>Integer</td>
</tr>
<tr>
<td>Weight</td>
<td>43.5−111</td>
<td>Integer</td>
</tr>
<tr>
<td>BMI</td>
<td>2class[1, 2, 3]</td>
<td>Integer</td>
</tr>
</tbody>
</table>

**Model assessment**

The confusion matrix which included TP, FP, FN and TN has been used to determine the relationship between the actual values and predicted values [16]. Table 3 shows the structure of confusion matrix.
We compared the classification performance of all ML algorithms by using Accuracy, Precision, Recall (Sensitivity), F1-score and MSE.

<table>
<thead>
<tr>
<th>Actual values</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative</td>
<td>Positives</td>
</tr>
<tr>
<td><strong>Predicted values</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>True Positives (TP)</td>
<td>False Positives (FP)</td>
</tr>
<tr>
<td>Positives</td>
<td>False Negatives (FN)</td>
<td>True Negatives (TN)</td>
</tr>
</tbody>
</table>

TP and TN represent the number of true positive or true negative samples. Accuracy is a statistical measure which is defined as the quotient of correct predictions (both True positives (TP) and True negatives (TN)) made by a classifier divided by the sum of all predictions made by the positive cases, i.e. the correctly and the incorrectly cases predicted as positive. Recall, also known as sensitivity, is the ratio of the correctly identified positive cases to all the actual positive cases, which is the sum of the "False Negatives" and "True classifier, including false positives (FP) and False negatives (FN). Therefore, the formula for quantifying binary accuracy is: Precision is the ratio of the correctly identified positive cases to all the predicted Positives".

**Results**

**Patient's characteristics:**

Of all people, 686 (52.1%) were men and 630 (47.9%) were women. The mean age of participants was 28.5 ± 7.4 years (range 20 to 49).

All of participants were from urban population who lived in Ardabil city. (Table 4)
Table 4
Demographic characteristics of the participants (n = 1316)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Groups</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years (28.5 ± 7.4, Range: 20–49)</td>
<td>20–30</td>
<td>831</td>
<td>63.1</td>
</tr>
<tr>
<td></td>
<td>30–40</td>
<td>343</td>
<td>26.1</td>
</tr>
<tr>
<td></td>
<td>&gt; 40</td>
<td>142</td>
<td>10.8</td>
</tr>
<tr>
<td>Gender</td>
<td>Female</td>
<td>630</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>686</td>
<td>52.1</td>
</tr>
<tr>
<td>WHR</td>
<td>Healthy</td>
<td>984</td>
<td>74.8</td>
</tr>
<tr>
<td></td>
<td>At-Risk</td>
<td>332</td>
<td>25.2</td>
</tr>
<tr>
<td>BMI, k/m$^2$ (26.1 ± 4.5, range:18.5–51.2)</td>
<td>Normal</td>
<td>603</td>
<td>45.8</td>
</tr>
<tr>
<td></td>
<td>Overweight</td>
<td>468</td>
<td>35.6</td>
</tr>
<tr>
<td></td>
<td>Obesity</td>
<td>245</td>
<td>18.6</td>
</tr>
</tbody>
</table>

**Performance of machine learning algorithms**

In this ML model, we predicted the performance on BMI dataset by Accuracy, Precision (Positive Predictive Value), Recall (Sensitivity) and F1-score. Figure 1 shows the performance of the predictive model using different data mining algorithm techniques. As shown in Fig. 1, the LR and MLP with 100% and RF with 97% had the highest sensitivity than other algorithms. Also the algorithms DT, LR and MLP with 100% had the highest accuracy rate than others in the classification of people based on BMI data.

**Discussion**

The main goal of this study was presentation of the efficacy of ML algorithms and techniques in BMI data which we used various machine learning (ML) algorithms to improve the classification of at-risk people based on BMI data which could be provided significant insights compared with traditional statistical models.

Among all ML models, DT, LR and MLP showed higher performance than others. Similar to this study, Wu et al in a study on fatty liver disease by using machine learning algorithms showed that among studied algorithms, the random forest model showed higher performance than other classification models which have some difference with our study results [8].
To our knowledge, this is the first population based study attempted to classification of at-risk people based on BMI data by using various machine learning algorithms. There are many kind of machine learning algorithms have been developed along with the most popular Bayesian algorithm and logistic regression, it is hard to make a proper algorithm for clinical decision making and clinical practices [9]. Therefore, the performances of different algorithms could provide the most important consideration, along with the easy to use and the interpretation of the models. However, our model could effectively detect the at-risk people based on BMI data without using advanced methods. In addition, the model could provide an easy, fast, low cost, and noninvasive method to accurately detection of people with normal and abnormal BMI [10]. By considering the increasing health issues related to obesity and overweight has in daily reports, machine learning allow massive amounts of data to be analyzed rapidly [11]. Therefore, it is the opportunity to apply machine learning algorithms to the classification of individual patients in medical practice and treatment and control of future related problems for people in term of their health and life style. By using various machine learning prediction models, physicians and health staff could be able to extract the minimum data necessary to make a prediction decision about people with normal and non-normal BMI [12].

Lee et al in a study showed that accuracy of ML method ranged from 60.4–73.8% which was lower than our study results because in our study the accuracy of ML algorithms ranged from 82–100% [13].

Uddin et al in a study entitled “Comparing different supervised machine learning algorithms for disease prediction” showed that of all ML algorithms, the algorithm RF had high accuracy in compare with other algorithms which was not in line with our study results because in this study we resulted that the best accuracy related to the algorithms such as DT, LR and MLP each with 100% [14].

Bastin Takhti et al in a study entitled “A model for diagnosis of kidney disease using machine learning techniques” showed that similar to our study on BMI data, the results showed that machine learning techniques could be effective in the diagnosis of kidney disease and of all algorithms, the most accuracy was related to the SVM with 0.97 and recall was for DT with 0.96 and most precision was related to the MLP with 0.99. In our study the most accuracy, recall and precision of BMI classification was related to the DT, LR and MLP but the accuracy of SVM was 0.82 which was lower than Bastin Takhti study rate [15].

**Conclusion**

In this study, seven machine learning techniques were used to classification of healthy people from at-risk people based on BMI data. All the algorithms worked with a reasonable accuracy and speed. However, the DT, LR and MLP algorithms showed maximum precision and minimum errors among all algorithms and also, these algorithms showed better performance than other ML classification techniques. This prediction outcome has the potential to help clinicians and health system staff to make more precise and meaningful decisions about people at-risk of overweight and obesity to provide the prediction program for decreasing their risk of diseases and change their bad life style in compare with healthy people.
Declarations

Ethics approval and consent to participate: The used dataset is used from another study about BMI which published by the first author in JBE. The original study was approved by Ethical committee of Ardabil University of Medical Sciences.

Consent for publication: Yes

Availability of data and materials: Yes

Competing interests: no

Funding: no

Authors' contributions: FA write the draft of article, sampling, data collection, statistical analysis. AM help in ML programming, data analysis and review Article. PA help in complete data and data entry, review article and writing some section of manuscript.

We confirm that all methods included in this study were carried out in accordance with relevant guidelines and regulations.

We confirm that all experimental protocols in this study were approved by Ethical committee of Ardabil University of Medical Sciences.

We confirm that the Informed consent in original data study was obtained from all subjects or, if subjects are under 18, from a parent and/or legal guardian orally.

References


Figures
Figure 1

Comparison performance of different machine learning algorithms on classification of people based on BMI data. *Note: RF= Random Forest, SVM= Support Vector Machine, Multi-layer Perceptron (MLP), K-nearest neighbors (KNN), LR= Logistic Regression, DT= Decision Tree Classifier, GNB= Gaussian Naïve Bayes, Accuracy, Precision, Sensitivity (Recall) and F1-score

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BMI.useddataset.xlsx