Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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 measure the overweight and obesity among people. The aim of this study was to develop a machine learning method to classification of BMI for clinical application.

Methods: In this study we used the dataset of 1316 people who selected randomly from all area of Ardabil city. Dataset included demographic and anthropometric data. Classification algorithms such as Random forest (RF), Gaussian Naïve 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 algorithms were evaluated with Precision, Recall, Mean Squared Errors (MSE) and Accuracy. All programing 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, 100, 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 compare classification algorithms results showed that, the LR , MLP and DT had the higher full accuracy than other algorithms in detection of people at-risk.

Background

Obesity and overweight are a complex, multifactorial, and major public health problem world-wide which could be affects people in all age groups and increased the risk of many diseases among people [1–2]. This index is used for measure of 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 the WHO reports, BMI less than 25 was considered as normal as 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 1
Description of the BMI datasets.

Dataset	Sample Size	Feature size including class label	Classes	Presence of missing attribute	Presence of noisy attributes
BMI	1316	8	2	NO	NO

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 ($18.5 \leq BMI < 25$) and At-risk ($25 \leq BMI$).

Table 2
Features of obesity type dataset

Feature	Class	Type
Gender	2 class [1, 2]	Integer
Age	20–49	Integer
WHR	0–2.27	Integer
WC	45–160	Integer
HC	37–160	Integer
Height	110–194	Integer
Weight	43.5–111	Integer
BMI	2class[1, 2, 3]	Integer

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 3
structure of confusion matrix

		Actual values	
		Negative	Positives
Predicted values	Negative	**True Positives(TP)**	False Positives(FP)
	Positives	False Negatives(FN)	**True Negatives(TN)**

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)

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

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 claasification 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 whit 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

1. -Jafari-Adli Sh, Jouyandeh Z & Qorbani M (2014). Prevalence of obesity and overweight in adults and children in Iran; a systematic review. J Diabetes Metab Disord 13: 121.
2. -Hruby A & Hu FB (2015). The epidemiology of Obesity: A big Picture. Pharmacoeconomics 33: 673–689.
3. -Yang L, Liu J, Xing Y, Du L, Chen J, Liu X & Hao J (2017). Correlation of Body Mass Index and Waist-Hip Ratio with Severity and Complications of Hyperlipidemic Acute Pancreatitis in Chinese Patients. Gastroenterol Res Pract 2017: 6757805.
4. -WHO (2018). Obesity and overweight – Fact Sheets 16 February 2018. From: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.
5. -Misra A & Khurana L (2008). Obesity and the metabolic syndrome in developing countries. J Clin Endocrinol Metab 93: S9–S30.
6. -Kelly T, Yang W, Chen CS, Reynolds K & He J (2008). Global burden of obesity in 2005 and projections to 2030. Int J Obes (Lond) 32:1431–1437.

7. -Amani F, Tabrizian S, Zakeri A, Pirzadeh A, Zeynizadeh S. Using Multinomial Logistic Regression for Modeling Obesity and Overweight Among People in Urban Area of Ardabil City, Ardabil, Iran. jbe. 6(3):170–178.
8. -Fattahzadeh-Ardalani G, Masoumi R, Amani F & Zakeri A (2017). Prevalence of overweight and obesity among high school girls in Ardabil, Iran. Int J Adv Med 4:486–489.
9. -Yaghoobi M, Rimaz Sh, Arbabisarjou A, Liaghat S & Salehinia H (2015). The prevalence of Obesity and overweight in Iranian Women: a study in Zahedan (Southeast of Iran). Scholars Journal of Aplied Medical Sciences 3: 1411–1415.
10. -Hosseinpanah F, Barzin M, Eskandary PS, Mirmiran P & Azizi F (2009). Trends of obesity and abdominal obesity in Tehranian adults: A cohort study. BMC Public Health 9: 426.
11. -Doustjalali SR, Gujjar KR, Sharma R & Shafiei-Sabet N (2016). Correlation between Body Mass Index (BMI) and Waist to Hip Ratio (WHR) among Undergraduate Students. Pakistan Journal of Nutrition 15: 618–624.
12. - Amirul Syafiq Mohd Ghazali, Zalila Ali, Norlida Mohd Noor, and Adam Baharum. Multinomial logistic regression modelling of obesity and overweight among primary school students in a rural area of Negeri Sembilan. AIP Conference Proceedings 1682, 050006 (2015); https://doi.org/10.1063/1.4932497.
13. -Bum Ju Lee, Boncho Ku, Jun-Su Jang, Jong Yeol Kim, "A Novel Method for Classifying Body Mass Index on the Basis of Speech Signals for Future Clinical Applications: A Pilot Study", Evidence-Based Complementary and Alternative Medicine, vol. 2013,Article ID 150265, 10 pages, 2013. https://doi.org/10.1155/2013/150265
14. Uddin S, Khan A, Hossain MDE, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making 2019;19:281.
15. Bastin Takhti S, Firouzi Jahantigh F. A model for diagnosis of kidney disease using machine learning techniques. Razi J Med Sci. 2019;26(8):14–22.

Figures

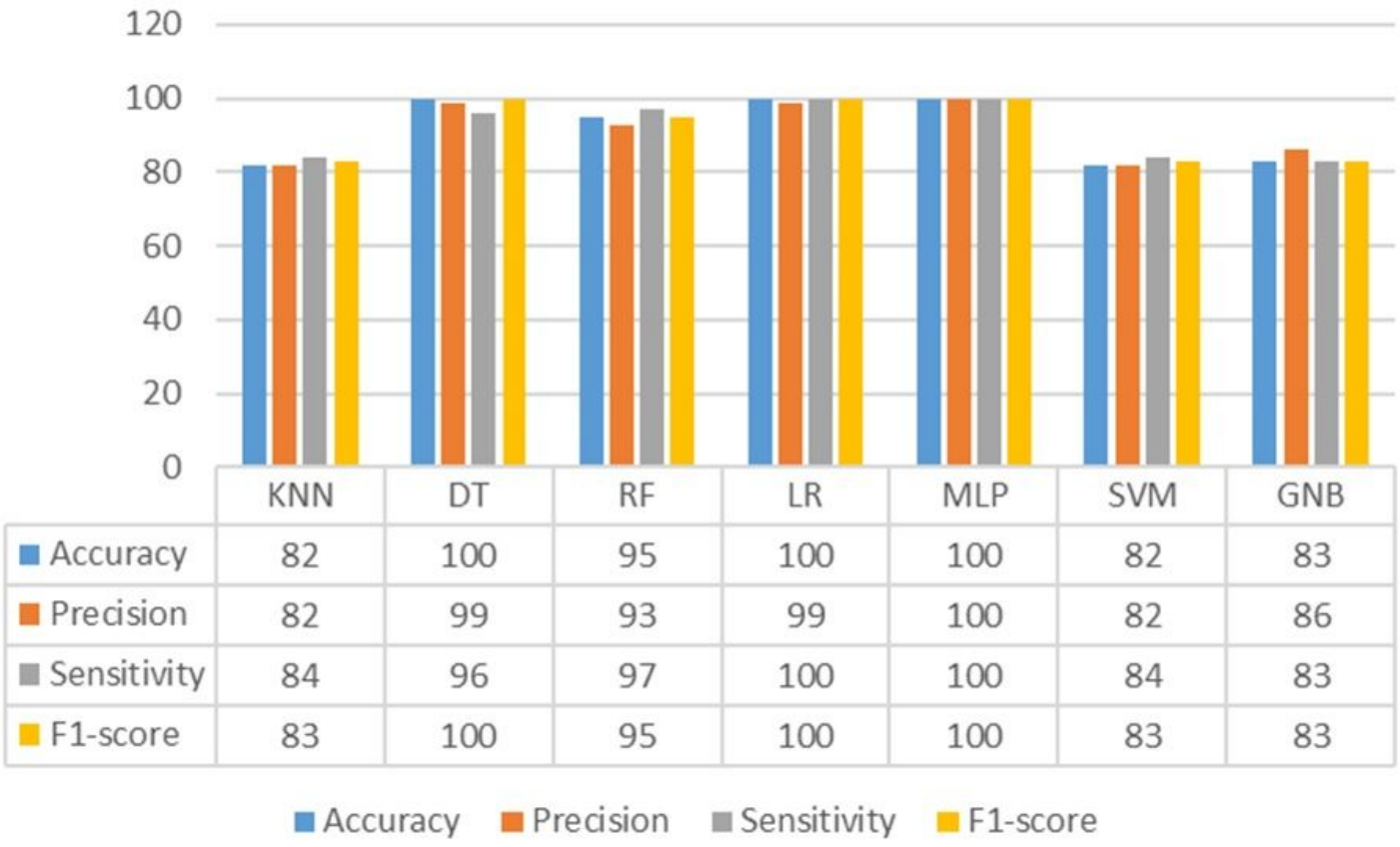

Figure 1

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