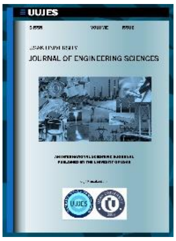

Usak University

Journal of Engineering Sciences

An international e-journal published by the University of Usak

Journal homepage: dergipark.org.tr/uujes

Research article

COMPARISON OF CLASSIFIERS FOR THE RISK OF OBESITY PREDICTION AMONG HIGH SCHOOL STUDENTS

Emel Kuruoğlu Kandemir[1], Çağın Kandemir Çavaş[1], Ayça Efe[2]*

[1]Dokuz Eylül University, Faculty of Science, Department of Computer Science, Turkey
[2]Dokuz Eylül University, The Graduate School of Natural and Applied Sciences, Turkey

Received: 28 Nov 2018 Revised: 21 Feb 2019 Accepted: 27 Feb 2019 Online available: 30 June 2019
Handling Editor: Kemal Mazanoğlu

Abstract

Obesity, which negatively affects human health, is a chronic disease due to genetic and living conditions. In this study, it was aimed to examine the observations with three main techniques: logistic regression, artificial neural networks and Naive Bayes, where the response variable was two categories of obese/not obese. Obesity questionnaire data, that was answered by 504 senior students in three randomly selected high schools in Gaziemir, Izmir, were analysed, and the predictive competences of the results of the three methods were evaluated. It was found that obesity is affected by the mother and father's being obese and eating too much fruit. In addition, gender and diet status were significantly related with the obesity risk.

In the artificial neural network, backward propagation learning algorithm was used as the learning rule in the adjustment of the connection weights according to the output. With the Naive Bayes method, a classification based on the probability values of the data was performed. The logistic regression model coefficient values were determined, using the maximum likelihood method. According to obesity questionnaire data, it was determined whether the relationship of each obesity risk factor with the response variable was statistically significant. The Naive Bayes method has the highest accuracy in prediction obesity compared to the other two methods.

Keywords: *Logistic regression; artificial neural networks; Naive Bayes classification; obesity.*

1. Introduction

Today, obesity is the one of the problems encountered in medicine and it is very important to diagnose it as a disease. The determination of risk factors affecting obesity

*Corresponding author: Çağın Kandemir Çavaş (ORCID ID: 0000-0003-2241-3546)
E-mail: cagin.kandemir@deu.edu.tr

development can be evaluated in terms of measures and diagnosis. Several studies, some of which are below, have been carried out.

Obesity is a major public health problem and is increasing rapidly all over the world [1-3]. Yıldırım and Uskun [4] examined the risk factors affecting the development of obesity in high school students with a population-based case control study. In the study, it was determined that the presence of obesity in the family / near family was significant risk factors in terms of obesity by going to school in a vehicle, being asked to stay at the weight of the friend with the same sex. Mbakwa [5] aimed to examine the intestinal microbiota composition of school-aged children in association with overweight [BMI ≥ 85th percentile]. Aswathikutty et al. [6] found no evidence of any association of obesity and physical activity with traumatic dental injuries (TDI) among adolescents from East London. Ferenci and Kovács [7] predicted body fat percentage from anthropometric and laboratory measurements using artificial neural networks. Efe [8] predicted the obesity among high school students by using logistic regression and artificial neural network.

Data mining is the popular technique that can solve the complex problems in all scientific areas. The Naive Bayes classifier is one of these techniques that can find solutions with high accuracy rates and it was used in the health field too. The method was used by Orphanou et al. [9] for the diagnosis of coronary heart disease, Cui et al. [10] used the method of predicted osteonecrosis of the femoral head with cannulated screw fixation. Zhang et al. [11] predicted the carcinogenicity of chemicals by using Naive Bayes classifier. Wang and Zhang [12] detected engine imagery EEG signals using the same method.

The aim of this study is to determine the risk factors and to predict the obesity among high school students by using logistic regression, artificial neural network and Naive Bayes classifier models.

2. Methods

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. There are three basic tests which are Likelihood ratio test, Wald test and score test to determine the significance of the variables in the logistic model. Likelihood ratio is a significance test based on the likelihood function. It analyses whether a current model is as good as the saturated model. Wald test analyses whether an independent variable has a significant relationship with the dependent variable. The Wald test statistic is obtained by using maximum likelihood estimate of the slope parameter divided by its standard error. In logistic regression, the coefficient expresses the change in the logit for one unit change in independent variable X.

Logistic regression is used to obtain odds' ratio in the presence of independent variables. The result is the impact of each variable on the odds ratio of the observed event of interest [13]. The logistic model serves to predict the risk to be any value between 0 and 1. In other words, there is no risk above 1 or below 0 [14].

The first artificial neural network model was carried out in 1943 by Walter Pitts, a mathematician with Warren McCulloch, a neurosurgeon [15]. Artificial neural networks are computer systems that are able to learn how to react to events from the environment by learning examples using human-generated examples - examples of actual brain functions. Similar to the functional properties of the human brain, learning, association, classification, generalisation, identifying feature and optimisation are applied. A neural network is a two-stage regression or classification model, typically represented by a network diagram [16].

The main operational principle of artificial neural networks puts forth that an input set is received then transformed into an output set. In the process, the network needs to be trained to generate the proper outputs for the presented inputs. The input set have to be transformed into a vector. This vector is presented to the network and network generates the required output vector for this particular vector. In every single iteration weight connections of network are regulated to generate the proper output. Figure 1 demonstrates a graphical illustration of artificial neuron model.

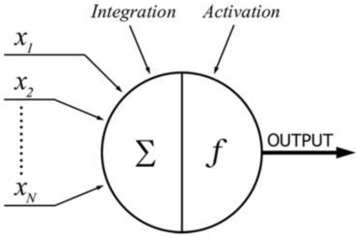

Fig. 1 An artificial neuron.

The general artificial neuron model consists of five basic elements as input, connection weights, net function, activation function and output. Input is external information sent to an artificial neural cell from outside. Inputs are determined by the samples demanded to be learnt by network. Connection weights are show the strength of information received by the cell and its effect on the cell. The weights can take positive or negative values. Net function computes weighted inputs received by the cell. Each input is multiplied with its own weight. The net function is comprised of sum of weighted inputs as given below:

$$\mathrm{net}(w, x) = w_1x_1 + w_2x_2 + \ldots + w_nx_n \tag{1}$$

where w is the matrix of weights, x is the input matrix, and n is the number of inputs.

Activation function processes weighted inputs are received by the neuron to determine the output that shall be generated by the neuron in response to this input. In the simplest case, the output y is computed as follows:

$$y = f(\mathrm{net}) = \begin{cases} 1 & \sum w_i x_i \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where θ is a threshold-level.

As in the case for summing function in the computation of output value, too different activation functions can be employed. In multilayer perceptron network model, which is widely used in modelling as activation function, sigmoid function is the most widely selected one. Typically, the activation function is chosen by the designer and then the weight and threshold values will be adjusted some learning rule.

The Naive Bayes classifier predicts the conditional probability of the random variable Y when the random variable X is given, using the Bayes rule given below:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \qquad X = \{X_1, \ldots, X_N\} \tag{3}$$

When the independent variable *X* is categorical, Naive Bayes' main objective is to estimate the class of the variable *Y* over its knowledge. It can be explained by the expression in Eq.4.

$$P(Y = y_i \mid X = x_k) = \frac{P(X = x_k \mid Y = y_i)P(Y = y_i)}{\sum_j P(X = x_k \mid Y = y_j)P(Y = y_j)} \tag{4}$$

The assumption that X_j has a conditional independent property when *Y* is given, is as follows:

$$P(X_1 \ldots X_N \mid Y) = P(X_1 \mid Y) \ldots P(X_N \mid Y) = \prod_{j=1}^{N} P(X_j \mid Y) \tag{5}$$

Under the assumption, *P*(*Xj*|*Y*) is estimated from relative frequencies in the data as follows:

$$P(X_j \mid Y) = \frac{\sum (X_j = x_{jk} \wedge Y = y_i)}{\sum (Y = y_i)} \tag{6}$$

If the parametric distribution assumption for continuous distribution is not possible, Kernel density estimation can be used as the non-parametric method.

3. Application and results

The obesity questionnaire data, which was answered by 504 senior students in three randomly selected high schools in Gaziemir, İzmir, were analysed and three methods were compared.

215 (43%) male and 289 (57%) female students participated in the survey. The mean age was 18.2 years. 76 of the 504 students are obese. Dependent variable is whether students are obese. The independent variables were 35 items obtained from the questionnaire. The final model includes the meaningful variables such as gender (SEX- [Male, Female]), obesity among parents (OAP- [None (OAP), one of the parents is obese (OAP (1), both parents are obese (OAP (2))]]), whether they are currently dieting (DIET-Dieting (No, Yes)), fruit intake (FIN- [eating low level fruit(FIN), eating less fruit-Medium level (day/week) (FIN (1)), eating a lot of fruit (day/week) (FIN (2))]]) were taken into the model (Table 1).

Since the presence and absence of obesity is taken into consideration and due to the high number of variables, logistic regression model has been established with backward elimination method. The validity of the model has been tested and determined. Odds ratio (OR) was used for the interpretation of variables as a result of the model.

95% confidence interval for the odds ratio of SEX variable does not include an odds ratio of 1 and the Wald statistic of this variable is greater than 2. Thus, we would conclude that the SEX (Male, Female) variable is significantly related with the obesity risk. Based on the coefficients in Table 1 the odds of being obese for females are 34.2% (odds ratio=0.342) of the odds for males for being obese or another explanation the odds of being obese are 65% (0.342-1=0.658) lower for females than for males.

95% confidence interval for the odds ratio of OAP variable does not include 1. Also the Wald statistic of OAP variable is greater than value of 2. For this reason, we may conclude that the OAP variable is significant. For OAP variable, the odds ratio is 2.106 for the first group which is the group one of the parents of whom is obese (OAP (1)). Thus we would say that having an obese parent has 2.106 times more risk factor than the case; neither of the parents are obese. The odds ratio for the second group, which is the group that both parents are obese (OAP (2)), is 4.095. Thus we would say that having parents both obese has 4.095 times more risk factor than having parents both non-obese.

The odds ratio is 4.514 for the DIET variable. Since the 95% confidence interval does not include 1 and also the Wald statistic of this variable greater than 2, the relationship is found to be significant. Thus we would say that ones going on a diet have the 4.514 times more risk of becoming obese than ones who are not going on a diet.

Although the odds ratio for the first group is 0.651, the confidence interval of one of the groups for FIN (1) variable contains value of 1 and also the Wald statistic value is smaller than 2 and thus it is not significant. Besides, the odds ratio for the second group (FIN (2)) is 2.009 which means that the risk of being obese is about 2 times higher for a person who eats a lot of fruit than one who eats a little fruit (Table 1).

According to the logistic regression model, the obesity of the parents, gender, fruit intake and the diet of the students indicate that obesity should be suspected at first.

Table 1. Results (LogR model).

			95%CI		
Variable	**Wald**	**OR**	**Lower**	**Upper**	**p-value**
SEX(1)	14.748	0.342	0.197	0.591	0.000
OAP	9.988				0.007
OAP(1)	4.921	2.106	1.091	4.068	0.027
OAP(2)	6.311	4.095	1.363	12.298	0.012
DIET(1)	16.248	4.514	2.169	9.394	0.000
FIN	11.029				0.004
FIN(1)	1.355	0.651	0.316	1.341	0.244
FIN(2)	5.129	1.977	1.096	3.566	0.024
Constant	43.608				0.000

The neural network that is used in this study is a feed-forward back-propagation neural network with three layers: an input layer, three hidden layers and an output layer. Sigmoidal function is used as the activation function. In predicting outcome variable in logistic regression the whole data set has been used. For neural networks 3/4 and 1/4 data set are divided as training and testing, respectively. The learning rate and momentum constant for network training were chosen as 0.25 and 0.5, respectively.

For the Naive Bayes classifier, 25% of the total data was separated for test, 75% as training data, and 35 attributes were created as two-class model as obese / not obese. The modelling study was carried out by MATLAB R2013b Toolbox. The test data on the trained model is correctly classified as 90.4%.

In the study, the accuracy rate for obese / not obese classification was found as 86.5% in the logistic regression model, 86.3% in artificial neural network method and 90.4% in Naive Bayes method as seen in Table 2.

Table 2. Accuracy rates.

Method	**Accuracy Rate**
Logistic Regression	%86.5
Artificial Neural Network	%86.3
Naive Bayes	%90.4

In the logistic regression model, it was found that the mother and father were obese and eating too much fruit was effective factors in obesity. Also, gender and dieting status variables are significantly related with the obesity risk.

The learning rate and momentum constant for network training were chosen as 0.25 and 0.5, respectively. If the number of neurons is very low, false estimates are obtained, and in case of very high number of neurons, the network data is memorized and transformed into a model that produces results. There are no precise methods for determining the optimal number of neurons in the intermediate layer. For this reason, the number of neurons in the hidden layer of the network was determined to be the lowest of the error of the network by experimenting with different numbers of neurons.

As a result, the true classification rate for logistic regression, artificial neural network and Naive Bayes have been found %86.5, %86.3 and %90.4 respectively.

4. Conclusions

In this study, three successful data mining techniques such as logistic regression, artificial neural network and Naive Bayes were applied in the healthcare area. Useful results were obtained in prediction of obesity risk factors. Naive Bayes method has the highest accuracy in classification when compared to artificial neural network and logistic regression methods. This result can be referred as Naive Bayes method is an effective probability-based classification method that can provide robust prediction for the risk of obesity.

References

1. Marques CDF, Silva RCR, Machado MEC, de Santana MLP, Cairo RCA, de Jesus Pinto E, et al. The prevalence of overweight and obesity in adolescents in Bahia, Brazil. Nutricion Hospitalaria, 2013;28(2):491 – 496.
2. Cirulli ET, Guo L, Swisher CL, Shah N, Huang L, Napier LA, et al. Profound perturbation of the human metabolome by obesity. bioRxiv, 2018; 298224:1-30.
3. Bookman JS, Schwarzkopf R, Rathod P, Iorio R and Deshmukh AJ. Obesity: the modifiable risk factor in total joint arthroplasty. Orthopedic Clinics of North America, 2018;49(3):291-296.
4. Yıldırım S and Uskun E. Risk factors affecting obesity development in high school students: a community based case-control study. Türk Pediatri Ars, 2018;53(3):155 – 162.
5. Mbakwa CA, Hermes GD, Penders J, Savelkoul PH, Thijs C, Dagnelie PC and Arts IC. Gut microbiota and body weight in school-aged children: the KOALA birth cohort Study. Obesity, 2018;26(11):1767-1776.
6. Aswathikutty A, Marcenes W, Stansfeld SA and Bernabé E. Obesity, physical activity and traumatic dental injuries in adolescents from East London. Dental Traumatology, 2017;33(2):137 – 142.
7. Ferenci T and Kovács L. Predicting body fat percentage from anthropometric and laboratory measurements using artificial neural networks. Applied Soft Computing, 2018; 67:834 – 839.
8. Efe A. Evaluation of obesity risk factors using logistic regression and artificial neural networks, Master Thesis in Statistics, Dokuz Eylül University, İzmir, Turkey, 2012.
9. Orphanou K, Dagliati A, Sacchi L, Stassopoulou A, Keravnou E and Bellazzi R. Incorporating repeating temporal association rules in Naïve Bayes classifiers for coronary heart disease diagnosis. Journal of Biomedical Informatics, 2018;81:74 – 82.

10. Cui S, Zhao L, Wang, Y, Dong Q, Ma J, Wang Y et al. Using Naive Bayes Classifier to predict osteonecrosis of the femoral head with cannulated screw fixation. Injury, 2018;49(10):1865 – 1870.
11. Zhang H, Cao ZX, Li M, Li YZ and Peng C. Novel naive Bayes classification models for predicting the carcinogenicity of chemicals. Food and Chemical Toxicology, 2016;97: 141 – 149.
12. Wang H and Zhang Y. Detection of motor imagery EEG signals employing Naïve Bayes based learning process. Measurement, 2016;86:148 – 158.
13. Sperandei S. Understanding logistic regression analysis. Biochemia Medica, 2014;24(1):12 – 18.
14. Hosmer D and Lemeshow S. Applied logistic regression. New York: John Wiley&Sons; 1989.
15. Elmas Ç. Yapay sinir ağları. Ankara: Seçkin Yayıncılık; 2003.
16. Hastie T, Tibshirani R and Friedman J. The elements of statistical learning. 2nd edition. New York: Springer; 2008.