2021 44th International Conference on Telecommunications and Signal Processing (TSP) | 978-1-6654-2933-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/TSP52935.2021.9522628

Obesity Level Estimation based on Machine Learning Methods and Artificial Neural Networks

Yaren Celik[a] , Selda Guney[b] , Berna Dengiz[a]
[a] Department of Industrial Engineering, Baskent University, Ankara, Turkey
[b] Department of Electrical and Electronics Engineering, Baskent University, Ankara, Turkey

Abstract – **Obesity is a growing societal and public health problem starting from 1980 that needs more attention. For this reason, new studies are emerging day by day, including those looking for obesity in children, especially the impact factors, and how to predict the emergence of the situation under these factors. In this study, different classification methods were applied for the estimation of obesity levels. Based on the evaluation criteria, the results were compared for different machine learning methods. When the Cubic SVM method was applied by selecting the appropriate features specific to the problem, 97.8% accuracy was obtained.**

Keywords - Machine Learning; Feature Selection; Classification; Obesity Prediction; Artificial Neural Network; Support Vector Machine.

I. INTRODUCTION

The World Health Organization (WHO) (OMS, 2016), describes obesity and overweight as excessive fat accumulation in certain body areas that can be harmful for health, the number of people that suffers from obesity has doubled since 1980 and also in 2014 more than 1900 million adults, 18 years old or older, are suffering from alteration of their weight. Obesity can be caused by biological hazard factors such as hereditary background. Besides, there are other risk factors as social, psychological, and eating habits. On the other side, authors also propose other determining factors for obesity such as "being only child, family conflicts as divorce, depression and anxiety" [1].

Based on the previous statements and the literature you can find in many studies working the obesity influence factors, they have implemented several data mining techniques. Several authors have studies to analyze the disease and generate web tools to calculate the obesity level of a person, nevertheless such tools are limited to the calculation of the body mass index, omitting relevant factors such as family background and time dedicated to. Based on this, the authors considered an intelligent tool was needed to be able to detect obesity levels on people more efficiently [1].

This study had the objective of implementing several machine learning techniques to determine if one person suffers from obesity. The organization of the study is as follows: the relevant studies are given in the second part. In the third part, the data set and its properties are given. The application of classification methods is given in the fourth part. The results and discussion in the last part are given.

II. PREVIOUS WORK

Obesity has become an area of interest for research and many studies can be found working with the factors that produce the disease. Next, you can find a brief review of works proposed by different authors who implement data mining techniques on datasets with features related with this health issue.

Adnan et al. used data mining to predict children obesity in their study [2]. The purpose of their survey was to provide the necessary knowledge for the obesity problem, to introduce data mining for prediction, describe the current efforts in that area and show the benefits and weaknesses of each technique used. The techniques involved were Neural Networks, Naïve Bayes and Decision Trees.

In [3], they had an initial approach to the study of predicting children obesity, collecting information from primary sources: Parents, children, and caretakers. The authors identified risk factors such as: Obesity and level of education of the parents, lifestyle and habits of the children and influence of environment. The proposed framework uses a hybrid technique of Naïve Bayes and decision trees called NBTree. A framework was presented with a hybrid approach, based on Naïve Bayes for prediction and genetic algorithms for parameter optimization, applied to the problem of predicting children obesity, with a low rate of negative samples compared to positive samples [4]. As result, they obtained 19 parameters to be implemented in prediction with a precision of 75%.

In another study, the authors presented MyHealthyKids, an intervention system for primary schools with the goal of handling and reducing children obesity problems [5]. The system was composed of three modules: Obesity prediction, persuasion and recipe suggestion. The prediction module was based on Naïve Bayes and tests showed that the system had a precision of 73.3%.

Dugan et al. [6] generated a predictive study of children obesity with subjects older than 2 years old. In this study, the methods analyzed included: RandomTree, RandomForest, J48, ID3,

Naïve Bayes and Bayes. Their results showed that ID3 had better behavior with 85% in precision and 89% in sensibility.

III. MATERIALS AND METHODS

A. Data Set

This paper contains data for the estimation of obesity levels in people from the countries of Mexico,Peru and Colombia, with ages between 14 and 61 and diverse eating habits and physical condition as mentioned, data was collected using a web platform with a survey where anonymous users answered each question, then the information was processed obtaining 16 features and 2111 records, after a balancing process described. The data contains numerical and continous data, so it can be used for analysis based on algorithms of classification, prediction, segmentation and association [7].

The 16 features related with eating habits are: Frequent consumption of high caloric food (FAVC), frequency of consumption of vegetables(FCVC), number of main meals (NCP), consumption of food between meals (CAEC), consumption ofwater daily (CH20), and consumption of alcohol (CALC). The features related with the physicalcondition are: Calories consumption monitoring (SCC), physical activity frequency (FAF), time usingtechnology devices (TUE), transportation used (MTRANS), other variables obtained are: Gender, age,height and weight. Finally, all data was labeled and the NObesity class variable was divided into 7 classes, whose data numbers are shown in Figure 1. These are: Insufficient weight, normal weight, overweight level I, overweight level II, obesity type I, obesity type II and obesity type III [7].

After all the data has been collected, the data has been preprocessed so that it can be used for different data mining techniques. The number of records is 485 and the data is labeled using Equation (1) [7].

$$mass\ body\ index = \frac{weight}{(weight * height)} \quad (1)$$

After all calculations were made to obtain the body mass index for each individual, the results were compared with data provided by the WHO and Mexican Normativity:

- Underweight less than 18.5
- Normal 18.5 to 24.9
- Overweight 25.0 to 29.9
- Obesity I 30.0 to 34.9
- Obesity II 35.0 to 39.9
- Obesity III higher than 40

When designing the system, there are stages of first training the system and then testing the trained system. Some of the data obtained is real, some of it is synthetic data. The reason for using synthetic data is data imbalance between classes. In this way, a total of 2111 data were used during the study [7]. It has been decided to perform k-cross-validation (k=10) in classification applications. In the artificial neural network application, it was decided to allocate 20% of the data set for testing and 80% for training.

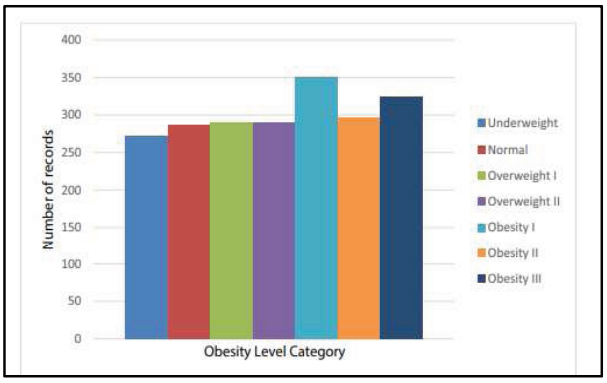

Fig. 1. Balanced Distribution of data regarding the obesity levels category.

In this study, 8 of 16 Features are numeric, the other 8 features are categorical. For the use of categorical data in algorithms with numerical characteristics, the data must be converted into numerical data. There are many methods used for this conversion (such as ordinal encoder, label encoder and one hot encoder). Label Encoder is widely used in such studies for ease of use. For this reason, it was preferred in this study. Label Encoder, which is an important pre-processing step for structured data sets in supervised learning, refers to converting categorical values into digital form. For example, the observation values of the gender variable in a data set with Label Encoder; It has been converted to numerical form as 0: Female and 1: Male.

B. Methods

In this study, Decision Tree, Support Vector Machine and Artificial Neural Network are used to classify obesity levels.

1) Decision Tree

Decision trees are one of the methods often used in classification problems. It is simpler to implement and understand than other machine learning methods. A decision tree is created using a data set when classifying, the data is applied to the tree one by one, and the necessary rules for forecasting are revealed [8]. Algorithms such as ID3, C4.5, C 5, CART, SLIQ, CHAID, Random Forest are used in decision tree applications. The main purpose of these algorithms is created a decision tree from the available data set by minimizing the generalization error [9].

2) Artificial Neural Network (ANN)

Artificial neural networks (ANN) are an information processing system consisting of many nerve cells that have the ability to perform calculations simultaneously. This system can also be called a mathematical modeling technique that matches, classifies, or uses the information entered according to the selected algorithm with the output for predictive purposes [10]. The basic elements being make up ANN as follows:

- It consists of many basic units (Architectural structure) called neurons, where information processing takes place.
- Signals are transmitted along connecting lines between neurons. (Algorithm, training or learning).
- Each connection line has a numerical weight. This weight is multiplied by the signal value.

- Each neuron uses the sum of its input signals as an input to a non-linear activation function, usually producing an output signal (Activation function).

3) Support Vector Machines (SVM)

Support Vector Machine (SVM) is a machine learning method in which data is divided into two classes with the help of a correct plane, or hyperplane. This method, which is often used for linearly decomposable data, can also be used for data that cannot be linearly decomposed. It can make the data linearly decomposable with the help of kernel functions. The main goal in this method is to determine the bracket that will minimize the square of errors [11],[12].

IV. RESULTS AND DISCUSSION

There are many methods in machine learning techniques. First, these methods are divided into classification, clustering, and link analysis models according to the purpose of the problem. Among them, according to the calculation methods, there are differences such as rule-based, statistical distance-based and deterministic. Since this study will be applied to a classification problem, classification performances were examined using methods based on each basis.

In this study, data set is composed of 7 classes and a total of 2111 samples. Each sample in the data set represents a separate observation. While 16 basic features measured are independent variables, "class" indicates a dependent variable. Therefore, independent variables were accepted as input variables, and the classification of obesity levels was made from these data. For this purpose, the system must be trained first. After the system is trained and tested with another test dataset, it is accepted that the system becomes ready to predict according to new data if the accuracy values are sufficient.

In this study, a computer with 2GB Nvidia Geforce GT940 graphics card, 8GB RAM, 6th generation Intel Core i5 processor and Matlab R2021a program were used.

In this study, 8 of 16 features and class values are numeric and 8 features are categorical. K-cross validation is applied to examine the test results. K value was determined as 10. Then feature selection is applied to the dataset. First of all, Backward elimination method was applied for the feature selection process to reduce the number of features and eliminate irrelevant data. The results were evaluated by extracting individual features and the points where improvements occurred were recorded. By using backward feature elimination method, the “CALC” "MTRANS", “FCVC” and “FAMILY_HISTORY” features are removed from the 16 features.

When the results were examined, there was no improvement when any feature was removed after 12 features. Therefore, the best accuracy result was obtained as 97.8% with using 12 features.

Then the next stage is classification. For the application of an artificial neural network, 12 features were used, which were decided in the previous stage. The number of hidden layer was selected as 2. Various trials have been made for different numbers of neurons. When the results were examined, the best results were obtained when 10 neurons were used in each hidden layer. Also tansig function is used as a transfer function. The accuracy value was calculated as 96.52%.

Within the scope of this study, the results of classification methods were examined and recorded in Table 1. It is generally observed that most methods of support vector machines work better than Decision Trees algorithms. The best decision tree was obtained as Bagged Trees with 95.4% accuracy and the best support vector machine was obtained as Cubic SVM with 97.8% accuracy.

TABLE 1 Accuracy values obtained for different classification methods using features

k=10 (by using 12 features)	
Fine Tree	93,1
Medium Tree	81,9
Coarse Tree	60,4
Linear SVM	96,0
Quadratic SVM	97,2
Cubic SVM	**97,8**
Fine Gaussian SVM	85,6
Medium Gaussian SVM	95,4
Coarse Gaussian SVM	88,4
Boosted Trees	90,4
Bagged Trees	95,4
RUSBoosted Trees	83,2
Artificial Neural Network	96,5

As for the application of an artificial neural network, the result was above 90%. When all the results within the scope of the study were evaluated, the best result was obtained with Cubic SVM. Also, the obtained different success results are given in Table 2 when the cubic SVM method is applied.

TABLE 2 Results obtained for different success criteria for the most successful classification method

Categories	**TP**	**FP**	**FN**	**precision**	**TP rate**
insufficient	267	5	5	98,16%	0,98162
normal	271	10	16	96,44%	0,94425
obesity I	345	3	6	99,14%	0,98291
obesity II	295	4	2	98,66%	0,99327
obesity III	323	0	1	100,00%	0,99691
overweight I	279	16	11	94,58%	0,96207
overweight II	284	9	6	96,93%	0,97931
	Average			97,70%	97,72%

When the studies on obesity and studies using the same data set are examined in the literature, it is seen from Table 3 that better results are obtained than other studies. Different techniques were used to obtain the best precision rates to detect obesity in the literature. Accordingly, The Decision Trees method in the study achieved 97.4% precision levels for classifying users with the disease, in addition, the technique showed a True Positive (TP) rate of 97.8%, which is stated by the researchers to

guarantee a high percentage of success for classifying the data [1]. The proposed method is also better results than literature which have 75% precision [4], 85% precision [5], and 73.3% precision [6]. When the results obtained in this study and the results of the studies used the same data set in the literature are examined, it is observed that results similar to the literature, even better than some, are obtained. These results are shown in Table 3. The contribution of this study is to obtain higher or similar accuracy values by using less features than the literature. In future studies, it is also a guide in terms of considering the 12 features that affect the obesity level very much.

TABLE 3 Comparison of the results with the studies in the literature

STUDY	DATA SET		APPLIED METHODS AND RESULTS		
	Number of Data	**Number of Feature**	**Method**	**Result**	
Obesity level estimation software based on decision trees.	712	18	J48	Precision %97,4 TP rate %97,8	
			Naive Bayes	Precision %90,1 TP rate %91,1	
			Simple Logistic	Precision %90,4 TP rate %91,6	
This study	2111	12	**Cubic SVM (Matlab Clasification Learner)**	**Precision %97,7 TP rate %97,7 Accuracy %97,8**	**Prediction Time: 3700 obs/sec Training Time: 12.211 sec**
			Artificial Neural Network (Matlab nntoolbox)	Accuracy %96,5	Time: 00:00:02

V. CONCLUSION

It is important to determine the obesity level, which is one of the important diseases of our age, with the right questions. The effects of the used feature on classification success were examined. Thus, the features that affect the obesity level the most have been revealed. By reducing the number of 16 features to 12, the account load has also been reduced. The successes of different classification methods were compared using these features. Among the methods used, the most successful result was obtained with Cubic SVM as 97.8%. When the results are examined, it is seen that faster results are obtained with fewer features. So, it is saved data collection time and working time. In future studies, the success of artificial neural networks can be increased by fine tuning. Higher precision results can be achieved. Also, the categorical and numerical data in the data set complicates this study. It can be investigated whether better results can be obtained with different encoding methods.

REFERENCES

[1] E. De-La-Hoz-Correa, F. E. Mendoza-Palechor, A. De-La-Hoz-Manotas, R. C. Morales-Ortega, and S. H. B. Adriana, "Obesity level estimation software based on decision Trees", J. Comput. Sci., pp:67-77, 2019.

[2] M.H.B.M. Adnan, W. Husain, and F. Damanhoori, "A survey on utilization of data mining for childhood obesity prediction", IEEE Xplore Press, pp: 1-6, 2010.

[3] M.H.M. Adnan, and W. Husain, "A framework for childhood obesity classifications and predictions using NBtree", IEEE Xplore Press, pp: 1-6, 2011.

[4] M.H.B.M. Adnan, and W. Husain, "A hybrid approach using Naïve Bayes and genetic algorithm for childhood obesity prediction", IEEE Xplore Press, pp: 281-285, 2012.

[5] W. Husain, M.H.M. Adnan, L.K. Ping, J. Poh, and L.K. Meng, "My healthy kids: Intelligent obesity intervention system for primary school children", The Society of Digital Information and Wireless Communication, pp: 627-633, 2013.

[6] T.M. Dugan, S. Mukhopadhyay, A. Carroll, and S. Downs, "Machine learning techniques for prediction of early childhood obesity", Applied Clin. Inform., pp: 506-520, 2015.

[7] F. M. Palechor, and A. De La Hoz Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico", Data in Brief, 2019.

[8] G. Silahtaroğlu, "Veri Madenciliği Kavram ve Algoritmaları", Papatya Yayıncılık Eğitim, pp:46-59, 2013.

[9] L. Rokach, and O. Maimon, "Decision Trees, Data Mining and Knowledge Discovery Handbook", Springer, pp:165-192, 2005.

[10] F. Altiparmak , B. Dengiz, and A. A. Bulgak, "Buffer allocation and performance modeling in asynchronous assembly system operations: An artificial neural network metamodeling approach", Applied Soft Computing, pp:946-956, 2006

[11] A. Haltaş, and A. Alkan, "Medlıne Veritabanı Üzerinde Bulunan Tıbbi Dökümanların Kanser Türlerine Göre Otomatik Sınıflandırılması", Bilişim Teknolojileri Dergisi, pp: 181-186, 2016.

[12] A. R. Bagasta1 , Z. Rustam, J. Pandelaki , and W. A. Nugroho, "Comparison of Cubic SVM with Gaussian SVM: Classification of Infarction for detecting Ischemic Stroke", IOP Conference Series: Materials Science and Engineering, pp:546-556, 2019.