Intelligent Methods in Engineering Sciences 2(1): 001-004, 2023

IMIENS

INTELLIGENT METHODS
IN ENGINEERING SCIENCES

https://www.imiens.org

International
Open Access

March, 2023

e-ISSN 2979-9236

Research Article

https://doi.org/10.58190/imiens.2023.6

Heart Attack Risk Analysis and Estimation Using Machine Learning Methods

Hadice Okay [a,*], ***Abidin Çalışkan*** [b]

[a] *Batman University, Graduate School of Education, Department of Electrical and Electronics Engineering Batman, Turkiye*
[b] *Batman University, Faculty of Engineering and Architecture, Department of Computer Engineering, Batman, Turkiye*

ARTICLE INFO

Article history:
Received 12 December 2022
Accepted 1 January 2023

Keywords:
Analysis
Artificial intelligence
Classification
Heart disease
Logistic regression
Visualization

ABSTRACT

Heart disease is a disease that is difficult to diagnose and leaves serious damage to individuals like many other diseases today. It is not known whether the risk of this disease is carried or not, and it is observed that there is an increase in the number of individuals at risk today. This increase; It requires accelerating the diagnosis of the disease to humanity by making early intervention and risk analysis together with developing technologies. Machine learning methods are developing rapidly in this field, facilitating early diagnosis in medicine. Diagnosing the disease with the developed methods provides a great advantage in terms of time cost. With the developments made, the diagnosis of diseases related to more than one parameter is carried out in a very short and reliable way. In this study; with the dataset consisting of the parameters and values of carrying the risk of heart attack, the classification of the risk of heart attack with high / low probability was made using Logistic Regression, which is one of the machine learning methods. By referring to what the parameters are, the distribution and values of these parameters on the dataset are determined. Obtained values; The effect of the parameters on the result status was analyzed using visualization methods. The main purpose of these analyzes is to determine the need for corrections on the dataset before training the network. As a result of the experimental analysis, 97% overall accuracy was achieved with the proposed approach.

1. Introduction

Heart attack; It is a condition in which the coronary vessels that feed the heart are blocked or narrowed due to a lack of oxygen or nutrients [1]. As a result of this narrowing and blockage, blood flow to the heart muscle slows down and if the heart is not fed enough, it can cause heart attack and permanent damage with heart attack. Considering the current data of the World Health Organization, it is seen that more than 31% of the death rates in Turkey and in the world are caused by a heart attack [2]. Although this rate is increasing day by day, it has a high potential to cause different damages even in rescued patients. Heart attack; Although it is a disease that will raise suspicion in cases such as dizziness, shortness of breath, indigestion, pain in the chest-arm-neck and fatigue, it is a disease that is diagnosed by looking at the values of many causes and conditions in medicine [3]. The process of making the diagnosis requires a long process depending on the parameter and intensity examined.

With the introduction of computers into our lives, operations performed with the human brain in daily life are performed by machines in a very short time, providing faster results. Studies for machines to perform operations quickly and to produce accurate results are increasing day by day [4]. These developments, together with machine learning methods, offer incredible convenience to human life. In addition to providing convenience, the time cost is quite high in the results of the human brain and daily working power. Developed machine learning methods manage to reduce this cost to milliseconds [5].

In this study; using Logistic Regression (LR), one of the machine learning methods, the heart attack risk status was classified as high/low. By giving information about the parameters that trigger the heart attack on the dataset used, the distribution of these parameters on the dataset and the situations in which they are related to each other were examined with different analysis methods. After the examinations, it is aimed to produce faster results by training with the dataset using the LR method and to break new ground by making it available in the field of medicine. Time and security of the trained network have been tested and there has been no obstacle to the dissemination of its use.

Other parts of the article are summarized as follows; Literature review is given in the second part. Material and method are detailed in the third section. The conclusion part is in the last part.

* **Corresponding Author:** okayhdc2016@gmail.com

2. Literature Review

With the use and spread of machine learning methods, studies with various success results in the field of heart diseases have been carried out and continue to be carried out. These studies; It has been developed and implemented by researchers by working with various algorithms such as K-Nearest Neighbor Algorithm (KNN), Support Vector Machines (DVM), Naïve Bayes (NB), Decision Trees (DT), LR and Artificial Neural Networks (ANN). Using the 10x cross validation method to classify the data using the DVM algorithm and further improve its performance, Yılmaz and friends. Using different datasets such as Statlog, SPECT Heart and Pima Indians, they found 97%, 98% and 96% accuracy values, respectively [6]. Kahramanli and Allahverdi developed a hybrid neural network including ANN and Fuzzy Neural Networks (BSA) and achieved an accuracy of 87.4% when applied to the Cleveland dataset [7]. Using the same algorithm (LR) as our study, Detrano and friends found an accuracy of 77% [8]. Gudadhe and friends They obtained 80% ,41 % accuracy by using DVM algorithms [9].

3. Materials and Methods

In this study; using the LR algorithm from machine learning algorithms, it was classified whether the risk of having a heart attack was carried or not. The dataset used in the classification process includes the symptoms of 303 different individuals that may pose a risk of heart attack [10]. 90% of this dataset was used for training the network and the remaining 10% was used for testing the network.

Evaluation and conclusion were made by following the steps mentioned below and making the necessary analyzes.

- Exploratory Data Analysis
- Data Values Analysis
- Standardization
- Correlation Chart Analysis
- LR

3.1. Exploratory Data Analysis

In the first stage of the study, the Exploratory Data Analysis data set was examined and the problem was discussed. The number of data, the parameters of the data and the effect of these parameters on the result are emphasized [11]. The data set used consists of the values of the parameters that include the risk of having a heart attack of 303 individuals. Some of these parameters and their states are given below.

The next stage of the study; The distribution of the parameters that make up the data set is examined and the weight values are determined by the machine learning code that will be written according to the importance of these parameters in terms of heart attack risk.

Table 1. Attribute Information of Data Set Parameters

Variable	Attribute	Explanation
age	Age	Patient's Age
sex	Gender	0:Woman 1 :Man
cp	Chest Pain Type	0-1-2-3 Values.
trestbps	Blood Pressure	Numerical Value In Mercury (Mm)
chol	Cholesterol Value	Its Value In Mg/Dl
fbs	Fasting Blood Sugar (Higher Than 120 Mg/Dl?)	0: No 1: Yes
restecg	ECG Results At Rest	0: Indicates Probable Or Definite Left Ventricular Hypertrophy According To Estes Criteria. 1: Normal 2: ST-T Wave Abnormality
thalach	Max Heart Rate	Numerical Value
exang	Does The Patient Have Angina During Exercise?	0: No 1: Yes
oldpeak	Numerical Value Measured In Depression	Value
slope	The Slope Of The Hill Exercise	0: Slope Down 1: Straight 2: Slope
ca	Number Of Major Blood Vessels	0-1-2-3 Values
thal	Blood Flow Rate	1: Fixed Defect 2: Normal 3: Reversible Defect
target	Conclusion	0: Patient 1: Not Sick

3.2. Data Values Analysis

The main characteristics of the data in the data set are given in Table 2 [12]. Information is given that there are 14 parameters and whether these parameters have values or not. The type of values of the parameters are shown. The expression

“303 non-null”; it shows that all parameters in the data set containing 303 data have their values, that is, the parameter is not left blank. Values of parameters other than “oldpeak” are “int64” 64-bit integers; It shows that the values of the “oldpeak” parameter are “float64” 64-bit rational numbers.

Table 2. General characteristics of data set parameters

Title	Parameters	Non-Null Value	Datatype
0	age	303 non-null	int64
1	sex	303 non-null	int64
2	cp	303 non-null	int64
3	trestbps	303 non-null	int64
4	chol	303 non-null	int64
5	fbs	303 non-null	int64
6	restecg	303 non-null	int64
7	thalach	303 non-null	int64
8	exang	303 non-null	int64
9	oldpeak	303 non-null	float64
10	slope	303 non-null	int64
11	ca	303 non-null	int64
12	thal	303 non-null	int64
13	target	303 non-null	int64

Lost Data Analysis; In other words, missing value analysis analyzes whether the values corresponding to the parameters in the data set are empty or full [13]. The purpose of this analysis method; is to ensure that all parameters correspond to a value. Parameters that cannot be found; It prevents data loss in the data set by giving the median or average value of the relevant parameter. The fullness of the data is very important in terms of training the network, visualizing it and giving the closest prediction value to the truth. The missing data analysis results of the data set used are given in Table 3 [12]. In the figure, the sum of the values left blank along the column for each parameter is given. As seen in the figure, there is no parameter with an empty value in the data set used.

Table 3. Missing data analysis results of the data set

Parameters	Null Value
age	0
sex	0
cp	0
trestbps	0
chol	0
fbs	0
restecg	0
thalach	0
exang	0
oldpeak	0
slope	0
ca	0
thal	0
target	0

3.3. Standardization

Data in the standardization process; it was made by using the method that we will obtain zero mean standard deviation over the values of the parameters. The values found in the standardization process are made ready for use in training and visualization of the network [14]. Finding the mean value is performed just before the visualization on the dataset, enabling meaningful conclusions to be drawn from the data. The standardization process is a very necessary step in order to train the network correctly and to make accurate predictions [15].

3.4. Correlation Chart Analysis

Correlation Chart Analysis; analyzes the relationship of all parameters with each other and with the output value [16]. If the values between the parameters are directly proportional, the value is positive; In case of an inverse ratio, the value is negative.

For example, it is seen in Figure 1; inverse proportion between “slope” and “olpeak” values; Parameters «age» and «fbs» are directly proportional to each other. Parameters that are directly proportional to the output value are; Parameters are «slope, cp ..» .

Figure 1. Correlation Chart Analysis

3.5. Logistic Regression

LR is a statistical model that uses the probability of a particular class or event depending on the binary dependent variable to model [17]. LR, which is one of the machine learning methods; During its use and implementation, step-by-step analyzes were made and the best value level estimation was ensured after the operations to prevent the wrong learning and memorization of the network on the dataset. First of all, 90% of the data was used for training the network to find the weight values on the parameters. Considering the effect of the estimation result of the data, appropriate transformations were made on the parameter values, and the network was provided to determine the weight value for the parameters. To test the weight values found, 10% data was used in the dataset. With the data allocated for the test, the success of the network has resulted in an accurate prediction value of 97%.

4. Conclusion

In this study; it was aimed to analyze and predict the risk of having a heart attack with the LR method, which is one of the machine learning methods. In the first stage of training the network, the data were examined and the weight values on the result were calculated. In the dataset used, 12 different parameter values belonging to 303 different individuals were analyzed and some transformations were made in order for the network to learn correctly and not memorize it.

The risk of heart attack was classified as high/low probability by using the data set consisting of parameters and values for carrying the risk of heart attack. An overall accuracy of 97% was obtained in the experimental analyzes of this study.

Even if the correct prediction value obtained is at an acceptable level, in the continuation of the studies, it will be tried to decide on the algorithm that finds the best correct prediction rate by comparing it with different algorithms. In future work, new approaches to heart attack risk analysis and estimation will be designed using

different types of datasets. In addition, different classification algorithms will be designed and added to the approach in order to highlight more efficient features in feature selection.

Author's Note

Abstract version of this paper was presented at 6th International Conference on Engineering Technologies (ICENTE'22), 17-19 November 2022, Konya, Turkey.

References

[1] R. S. Paffenbarger Jr, A. L. Wing, and R. T. Hyde, "Physical activity as an index of heart attack risk in college alumni," *American Journal of epidemiology*, 108(3), 161-175, 1978.

[2] R. Katarya, and S. K. Meena, "Machine learning techniques for heart disease prediction: a comparative study and analysis," *Health and Technology*, 11(1), 87-97, 2021.

[3] Bashir , Saba, *et al*. "Improving heart disease prediction using feature selection approaches," *In: 2019 16th international bhurban conference on applied sciences and technology (IBCAST)*. IEEE, 2019. p. 619-623.

[4] A. Çalışkan, "A New Ensemble Approach for Congestive Heart Failure and Arrhythmia Classification Using Shifted One-Dimensional Local Binary Patterns with Long Short-Term Memory," *The Computer Journal*, 65: 9, pp. 2535-2546, 2022.

[5] A. Çalışkan, "Detecting human activity types from 3D posture data using deep learning models," *Biomedical Signal Processing and Control*, 81, 104479, 2023.

[6] N.Yılmaz, O.Inan and M.S. Uzer, "A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases," *Journal of medical systems*, 38:48-59, 2014.

[7] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Systems with Applications*, 35.1-2:82-89, 2008.

[8] R. Detrano, A. Janosi, and W. Steinbrunn, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American Journal of Cardiology*, 64.5:304-310, 1989.

[9] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," *In: 2010 International Conference on Computer and Communication Technology (ICCCT)*. IEEE, pp. 741-745, September 2010.

[10] E. M. Senan, I. Abunadi, M. E. Jadhav, and, S. M. Fati , "Score and Correlation Coefficient-Based Feature Selection for Predicting Heart Failure Diagnosis by Using Machine Learning Algorithms," *Computational and Mathematical Methods in Medicine*, 2021.

[11] UCI Machine Learning Repository, Heart disease data set, Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease [Accessed: 25-Aug-2022].

[12] R. Mogot, "Heart Attact Analysis," Kaggle Web. (2022). https://www.kaggle.com/code/roderikmogot/heart-attack-analysis

[13] M. L. Vigni, C. Durante and M. Cocchi "Exploratory data analysis," *In Data Handling in Science and Technology*, vol:28, pp. 55 – 126.

[14] T. Ahmad, and M.N. Aziz, "Data preprocessing and feature selection for machine learning intrusion detection systems," *ICIC Express Lett*, 13.2:, 93-101, 2019.

[15] A. Tunc, I. Ülger, "Application of Normalization to Financial Values with Binning and Five Number Summary Methods for Feature Selection in Data Mining Applications," *Conf.: XVIII. Academic Informatics, Adnan Menderes University*, 30:3, 2016.

[16] A. Mehmood, M. Iqbal, Z. Mehmood, A. Irtaza, M. Nawaz, T. Nazir, and M. Masood "Prediction of heart disease using deep convolutional neural networks," *Arabian Journal for Science and Engineering*, 46(4), 3409-3422, 2021.

[17] J. Tolles and W. J. Meurer, "Logistic regression: relating patient characteristics to outcomes", 316(5), pp. 533-534, 2016.