Middle Black Sea Journal of Health Science / Mid Blac Sea J Health Sci **November 2022;8(4):582-591**

RESEARCH ARTICLE **DOI:** 10.19127/mbsjohs.1142542

Estimation of Risk Factors Related to Heart Attack with Xgboost That Machine Learning Model

Onural Özhan[1](ID), Zeynep Küçükakçali[2](ID)

[1]Department of Medical Pharmacology, Inonu University, Medical Faculty, Malatya, Turkey.
[2]Department of Biostatistics and Medical Informatics, Inonu University Faculty of Medicine, Malatya, Turkey.

Received: 05 September 2022, Accepted: 08 July August 2022, Published online: 30 November 2022

Abstract

Objective: The objective of this work is to classify heart attack cases using the open-access heart attack dataset and one of the machine learning techniques called XGBoost. Another aim is to reveal the risk factors associated with having a heart attack as a result of the modeling and to associate these factors with heart attack.

Methods: In the study, modeling was done with the XGBoost method using an open access data set including the factors associated with heart attack. Model results were evaluated with accuracy, balanced accuracy, specificity, positive predictive value, negative predictive value, and F1-score performance metrics. In addition, 10-fold cross-validation method was used in the modeling phase. Finally, variable importance values were obtained by modeling.

Results: Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score from by XGBoost modeling were 89.4%, 89.4%, 88.4%, 90.3%, 88.4%, 90.3%, and 88.4%, respectively. According to the variable importance values obtained for the input variables in the data set examined in this study, thal2, oldpeak, thal3, ca1, and exang1 were obtained as the most important variables associated with heart attack.

Conclusions: With the machine learning model used, the heart attack dataset was classified quite successfully, and the associated risk factors were revealed. Machine learning models can be used as clinical decision support systems for early diagnosis and treatment.

Keywords: Heart attack, machine learning, XGBoost, modelling, variable importance

Suggested Citation: Özhan O, Küçükakçalı Z. Estimation of Risk Factors Related to Heart Attack with Xgboost That Machine Learning Model Mid Blac Sea J Health Sci, 2022;8(4):582-591.

Address for correspondence/reprints:
Name and Surname: Onural Özhan
Telephone number: +90 (535) 932 33 44
E-mail: onural.ozhan@inonu.edu.tr

INTRODUCTION

Deaths caused by cardiovascular diseases, which are included in the "Non-Communicable Diseases 2013-2020 Action Plan" of the World Health Organization (WHO), are in the first place in the world and in our country. Especially nowadays, it is the most common cause of mortality and morbidity in developed and western countries, and the possibility of the disease in developing countries is increasing day by day. Cardiovascular diseases such as peripheral vascular disease, coronary artery disease, heart failure, dyslipidemia, and hypertension (HT) affect 400 million people worldwide, representing a diverse range of races, ages, and genders. Studies show that between 1990 and 2020, the death rate from cardiovascular diseases will increase from 28.9% to 36.3% all over the world (1, 2). Cardiovascular diseases are more likely to affect people with metabolic problems like insulin resistance, glucose intolerance, abdominal obesity, HT, hypertriglyceridemia, high low-density lipoprotein (LDL) and low high-density lipoprotein (HDL). Cardiovascular diseases are generally characterized by atherosclerosis, thrombosis, and vascular dysfunction resulting from high blood pressure (3).

Myocardial infarction (MI), usually known as a heart attack, is the most prevalent cardiovascular disease (4). MI is a condition in which the cardiac muscle cells suffer damage from a lack of oxygen because the necessary amount of blood does not flow because a portion of the heart's blood supply has deteriorated. Additionally, if the heart muscle goes for an extended period of time without oxygen, death may result. Within the first hour, 50% of MI-related deaths occur, and within the first 24 hours, this rate rises to 80% (5). MI is a significant public health issue that regularly affects society's productive age group, results in serious issues because of post-acute consequences, and can even be fatal. It is one of the most significant causes of morbidity and mortality in our nation and industrialized countries, despite recent improvements in diagnosis and treatment (6). The importance of early disease diagnosis in disease prevention and treatment cannot be overstated. The fact that heart diseases, particularly MI, are the leading cause of death in people of all ages is the most important reason for long-term cardiovascular disease research (3).

The goal of machine learning, a subset of artificial intelligence, is to predict new data as it is presented to it through data-driven learning. The researchers' goal is to teach computers to detect complex patterns and make data-driven decisions (7). In recent years, one of the technologies that have seen widespread usage in the diagnosis of diseases and clinical decision support systems is machine learning methods. These approaches have a wide range of application areas and have been increasingly

popular in recent years. Machine learning techniques often carry out the classification process in the disease prediction process. Machine learning, which has a widespread application area in the field of health, constitutes the fundamental infrastructure of applications in the determination of genetic diseases, early diagnosis of cancer diseases and chronic diseases, and the identification of patterns in medical imaging. In the last decade, with more computing power, ML methods have achieved very high performances in the field of health (8,9). Extreme Gradient Boosting (XGBoost), one of the machine learning methods, is one of the most effective supervised learning algorithms and its basic structure is based on gradient boosting and decision tree algorithms boost is an ensemble method that uses boosting to combine a set of weak classifiers to create a strong classifier. Starting with a basic learner, the strong learner is trained iteratively (10,11).

The purpose of this study is to use the machine learning technique XGBoost on the open-access heart attack dataset to classify instances of heart attacks and identify the factors associated with them.

METHODS

Deaths caused by cardiovascular diseases, which are included in the "Non-Communicable Diseases 2013-2020 Action Plan" of the World Health Organization (WHO), are in the first place in the world and in our country. Especially nowadays, it is the most common cause of mortality and morbidity in developed and western countries, and the possibility of the disease in developing countries is increasing day by day. Cardiovascular diseases such as peripheral vascular disease, coronary artery disease, heart failure, dyslipidemia, and hypertension (HT) affect 400 million people worldwide, representing a diverse range of races, ages, and genders. Studies show that between 1990 and 2020, the death rate from cardiovascular diseases will increase from 28.9% to 36.3% all over the world (1,2). Cardiovascular diseases are more likely to affect people with metabolic problems like insulin resistance, glucose intolerance, abdominal obesity, HT, hypertriglyceridemia, high low-density lipoprotein (LDL), and low high-density lipoprotein (HDL). Cardiovascular diseases are generally characterized by atherosclerosis, thrombosis, and vascular dysfunction resulting from high blood pressure (3).

Myocardial infarction (MI), usually known as a heart attack, is the most prevalent cardiovascular disease (4). MI is a condition in which the cardiac muscle cells suffer damage from a lack of oxygen because the necessary amount of blood does not flow because a portion of the heart's blood supply has deteriorated. Additionally, if the heart muscle goes for an extended period of time without oxygen, death may result. Within the first hour,

50% of MI-related deaths occur, and within the first 24 hours, this rate rises to 80% (5). MI is a significant public health issue that regularly affects society's productive age group, results in serious issues because of post-acute consequences, and can even be fatal. It is one of the most significant causes of morbidity and mortality in our nation and industrialized countries, despite recent improvements in diagnosis and treatment (6). The importance of early disease diagnosis in disease prevention and treatment cannot be overstated. The fact that heart diseases, particularly MI, are the leading cause of death in people of all ages is the most important reason for long-term cardiovascular disease research (3).

The goal of machine learning, a subset of artificial intelligence, is to predict new data as it is presented to it through data-driven learning. The researchers' goal is to teach computers to detect complex patterns and make data-driven decisions (7). In recent years, one of the technologies that has seen widespread usage in the diagnosis of diseases and clinical decision support systems is machine learning methods. These approaches have a wide range of application areas and have been increasingly popular in recent years. Machine learning techniques often carry out the classification process in the disease prediction process. Machine learning, which has a widespread application area in the field of health, constitutes the fundamental infrastructure of applications in the determination of genetic diseases, early diagnosis of cancer diseases and chronic diseases, and the identification of patterns in medical imaging. In the last decade, with more computing power, ML methods have achieved very high performances in the field of health (8,9). Extreme Gradient Boosting (XGBoost), one of the machine learning methods, is one of the most effective supervised learning algorithms and its basic structure is based on gradient boosting and decision tree algorithms XGBoost is an ensemble method that uses boosting to combine a set of weak classifiers to create a strong classifier. Starting with a basic learner, the strong learner is trained iteratively (10,11).

The purpose of this study is to use the machine learning technique XGBoost on the open-access heart attack dataset to classify instances of heart attacks and identify the factors associated with them.

XGBoost METHOD

Gradient Boost is a powerful machine learning technique that is regularly used for regression and classification problems where weak prediction models frequently generate ensemble forms of decision trees. Gradient Boost is typically applied in situations where these problems arise. Using the boosting method, attempts to generate a large number of weak learners sequentially and incorporate them into a complex model (11, 12). XGBoost is a robust machine learning model that utilizes

gradient boosting and decision tree methods. In terms of speed and performance, it has a major edge over other machine learning algorithms, with the potential to process nearly ten times faster. It also has a variety of regularizations that enhance overall performance while reducing overfitting and over-learning. XGBoost is an ensemble method for creating a robust classifier by combining a set of weak classifiers with reinforcement. XGBoost can achieve better performance than other methods by using different regularization techniques to control the complexity of the trees (13,14).

Tablo 1. Explanations of the Variables in the Data Set and Their Characteristics

Variable	Explanations of The Variables	Variable Type	Variable Role
target	target: 0= less chance of heart attack 1= more chance of heart attack	Qualitative	Output
age	age	Quantitative	Predictor
sex	Sex of the patient (0=female;1=male)	Qualitative	Predictor
trestbps	resting blood pressure	Quantitative	Predictor
chol	serum cholestoral in mg/dl	Quantitative	Predictor
fbs	fasting blood sugar > 120 mg/dl 1 = true; 0 = false	Qualitative	Predictor
cp	Chest pain type 0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic	Qualitative	Predictor
thalach	maximum heart rate achieved	Quantitative	Predictor
exang	exercise induced angına 1 = yes; 0 = no	Qualitative	Predictor
restecg	resting electrocardiographic results (values 0,1,2) ○ Value 0: normal ○ Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) ○ Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria	Qualitative	Predictor
slope	the slope of the peak exercise ST segment 0 = unsloping;1 = flat;2 = downsloping	Qualitative	Predictor
ca	number of major vessels (0-3) colored by flourosopy	Qualitative	Predictor
oldpeak	oldpeak = ST depression induced by exercise relative to rest	Quantitative	Predictor
thal	Thalium Stress Test result (0,3)	Qualitative	Predictor

Statistical analysis

The median (minimum-maximum) is used to summarize quantitative data, whereas qualitative factors are presented as numbers and percentages. Using the Kolmogorov-Smirnov test, a normal distribution was determined. The Pearson Chi-square test, Mann-Whitney U test, and Yates' correction chi-square test were used to determine whether there was a statistically significant difference between the output variable ("less probability of heart attack") and input variables. $p<0.05$ value was considered statistically significant. In all analyzes, IBM SPSS Statistics 26.0 for the Windows package program was used.

Modelling

XGBoost was used in the modeling. The n-fold cross-validation method was used for the analyses. The data set was divided 80:20 as a training and test dataset. The data is divided into n parts in the n-fold cross-validation method, and the model is applied to n parts. One of the n components is used for testing, while the remaining n-1 components are used to train the model. The modeling process in this study was carried out using 5-fold cross-validation. As performance evaluation criteria, accuracy, balanced accuracy, sensitivity, selectivity, positive predictive value, negative predictive value, and F1-score were used. In addition, variable importances were calculated, which gives information about how much the input variables explain to the output variable.

RESULTS

The mean age of the patients used in the current study was 54.37±9.08 years. Of the patients, 96 were female and 207 were male. Tables 2 and 3 contain descriptive statistics pertaining to the target variable that this study looked at. In terms of variables other than the "fbs" variable, there is a statistically significant difference between the dependent variable classes.

Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score obtained from the XGboost model were 89.4%, 89.4%, 88.4%, 90.3%, 88.4%, and 90.3%, and 88.4%, respectively.

Table 2. Descriptive statistics for Quantitative Input variables

Variables	Predicted Class		p*
	More Chance Of Heart Attack	Less Chance Of Heart Attack	
	Median (minimum-maximum)	**Median (minium-maximum)**	
age	52(29-76)	58(35-77)	<0.001*
trestbps	130(94-180)	130(100-200)	0.035*
chol	234(126-564)	249(131-409)	0.036*
thalach	161(96-202)	142(71-195)	<0.001*
oldpeak	0.2(0-4.2)	1.4(0-6.2)	<0.001*

*** Mann Whitney U test**

In Figure 2, the values of performance metrics are plotted for the XGboost model.

Table 4 shows the values of the performance criteria of the XGBoost model used in this study to classify cases of a heart attacks.

The graph of the variables associated with the output varies according to the variable importance obtained as a result of the modeling is given in figure 2.

DISCUSSION

Each year, cardiovascular disease kills 17,9 million people, accounting for 31 percent of all deaths globally. Cardiovascular diseases include coronary heart disease, cerebrovascular disease, rheumatic heart disease, and various heart and blood vessel diseases. Ischemic heart diseases are involved in the pathophysiology of most deaths due to cardiovascular diseases. Ischemic heart diseases cause mortality and morbidity worldwide (15,16). One of these diseases is MI, which is characterized as myocardial cell damage caused by persistent

ischemia. A heart attack is a physiological disorder that causes significant chest discomfort as a result of insufficiency caused by a defect in the coronary arteries of the heart and is fatal. A heart attack happens as a result of oxygen deprivation caused by a sudden decrease or halt in blood flow in the arteries that feed the heart for a variety of reasons. It can cause varying degrees of damage to the heart muscle fed by the blocked channel, as well as tissue death (17). Heart attack is a major health concern that is most prevalent in industrialized countries and is becoming more prevalent in emerging countries.

MI is a significant public health issue that occurs frequently in the productive age group of society, creates substantial problems owing to post-acute complications, and can result in mortality. According to World Health Organization (WHO) figures, 16.7 million people die each year as a result of heart attacks. This figure reflects one-third of all deaths worldwide (18). Machine learning is a subfield of computer science that focuses on the development and application of algorithms that give computers the ability to learn based on the types of data they are given.

Tablo 3. Descriptive statistics for Qualitative Input variables

Variables		Predicted Class		**p** value**
		more chance of heart attack	less chance of heart attack	
sex	0	72(43.6%)	24(17.4%)	**<0.001****
	1	93(56.4%)	114(82.6%)	
cp	0	39(23.6%)	104(75.4%)	**<0.001***
	1	41(24.8%)	9(6.5%)	
	2	69(41.8%)	18(13.0%)	
	3	16(9.7%)	7(5.1%)	
fbs	0	142(86.1%)	116(84.1%)	0.744
	1	23(13.9%)	22(15.9%)	
restecg	0	68(41.2%)	79(57.2%)	**0.007***
	1	96(58.2%)	56(40.6%)	
	2	1(0.6%)	3(2.2%)	
exang	0	142(86.1%)	62(44.9%)	**<0.001****
	1	23(13.9%)	76(55.1%)	
slope	0	9(5.5%)	12(8.7%)	**<0.001***
	1	49(29.7%)	91(65.9%)	
	2	107(64.8%)	35(25.4%)	
ca	0	130(78.8%)	45(32.6%)	**<0.001***
	1	21(12.7%)	44(31.9%)	
	2	7(4.2%)	31(22.5%)	
	3	3(1.8%)	17(12.3%)	
	4	4(2.4%)	1(0.7%)	
thal	0	1(0.6%)	1(0.7%)	**<0.001***
	1	6(3.6%)	12(8.7%)	
	2	130(78.8%)	36(26.1%)	
	3	28(17.0%)	89(64.5%)	

*:Pearson chi square test; **: Chi-square test with Yates correction

Table 4. Performance metrics of the XGboost model

Metric	Value (%) (95% CI)
Accuracy	89.4 (86-92.9)
Balanced Accuracy	89.4 (85.9-92.8)
Sensitivity	88.4 (81.9-93.2)
Specificity	90.3(84.7-94.4)
Positive predictive value	88.4 (81.9-93.2)
Negative predictive value	90.3 (84.7-94.4)
F1 score	88.4 (88.4-92)

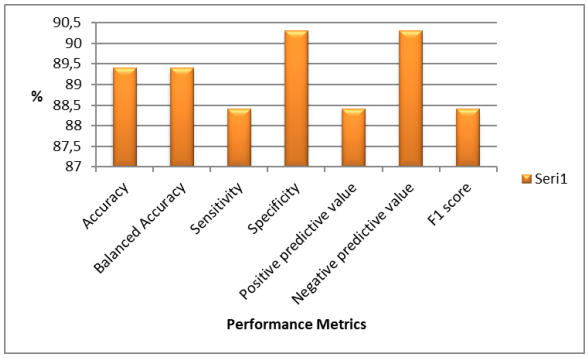

Figure 1. Graph of values for performance metrics for XGboost model

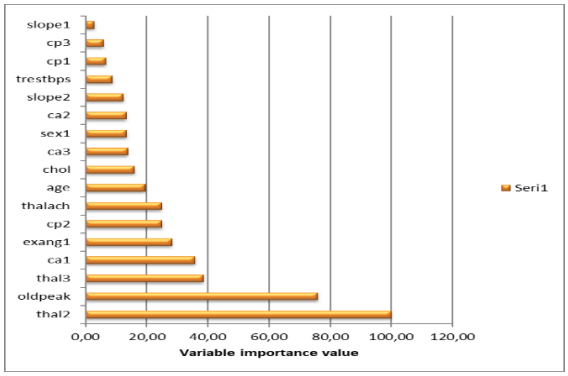

Figure 2. Variable importance graph

Not only is machine learning a database problem, but it is also a branch of artificial intelligence that models future events based on historical data and makes predictions about such events (19). In recent years, machine learning methods have been widely used in disease diagnosis and clinical decision support systems. Early disease detection and identification of disease-causing factors are made possible by machine learning methods, which are widely used in the field of health (20,21).

In the study, an open-source data set consisting of MI patients' data were classified using the XGBoost method, and factors associated with a case of more chance of heart attack, which is among the categories of the target variable, were determined. Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score from the performance criteria obtained by modeling were 89.4%, 89.4%, 88.4%, 90.3%, 88.4%, 90.3%, and 88.4%, respectively.

XGBoost method gave successful estimation results in the classification of heart attack status according to the values of performance metrics obtained from the study. In addition, risk factors for heart attack were also obtained with the variable importance values calculated as a result of the model. Thal2, old peak, thal3, ca1, and exang1 are among the most significant risk factors connected with having a heart attack. In research utilizing the same data set, the deep learning technique yielded an 81.4% accuracy rate. In addition, that, age, ca, old peak and exang variables were found to be associated with a heart attack in the study (22).

CONCLUSION

The XGBoost machine learning model utilized in the study correctly categorized the

state of having a heart attack. In addition, the study's results highlighted the risk factors for having a heart attack. Lastly, machine learning technology can be advantageous for medical data accessibility and early diagnosis.

Ethics Committee Approval: Ethics committee approval is not required in this study.
Peer-review: Externally peer-reviewed.
Author Contributions: Concept- OÖ, ZK Design- OÖ, ZK Literature Review- OÖ, ZK Critical Review- OÖ, ZK
Conflict of Interest: The authors have no interests to declare
Financial Disclosure: There are no financial supports.

REFERENCES

1- Abanonu G. Major risk factors for coronary artery disease and evaluation of C-Reactive protein. Published Specialization Thesis Istanbul. 2005.
2- House W. Follow-up to the political declaration of the high-level meeting of the general assembly on the prevention and control of non-communicable diseases. World Health Organization. 2013.
3- Lee CH, Kim J-H. A review on the medicinal potentials of ginseng and ginsenosides on cardiovascular diseases. J Ginseng Res. 2014;38(3):161-6.
4- Halıcı Z, Yasin Bayır HS, Çadırcı E, Keleş MS, Bayram E. Investigation of the Effects of Amiodarone on Erythropoietin Levels in Isoproterenol-induced Acute and Chronic Myocardial Infarction Model in Rats. The Eurasian Journal of Medicine. 2002;38:68-72
5- Storrow AB, Gibler WB. Chest pain centers: diagnosis of acute coronary syndromes. Ann Emerg Med. 2000;35(5):449-61.
6- Şentürk S. Investigation of the effect of l-lysine on total sialic acid levels in rats with myocardial infarction with isoproterenol. Trakya University Institute of Health Sciences Department of Biochemistry Master's Program Erzurum, 2008.
7- Polikar R. Ensemble learning. Ensemble machine learning: Springer; 2012. p. 1-34.
8- Akman M, Genç Y, Ankarali H. Random Forests Yöntemi ve Saglik Alaninda Bir Uygulama/Random Forests Methods and an Application in Health Science. Turkey Clinics Biostatistics. 2011;3(1):36.
9- Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record. 2002;31(1):76-7.
10- Dikker J. Boosted tree learning for balanced item recommendation in online retail. Master thesis. 2017.
11- Patrous ZS. Evaluating XGBoost For User Classification by Using Behavioral Features Extracted from Smartphone Sensors. [Master Thesis]: KTH Royal Institute of Technology, School of Computer Science and Communication, Sweden.; 2018. Access link: https://www.diva-portal.org/smash/get/diva2:1240595/FULLTEXT01.pdf
12- Wang J, Li P, Ran R, Che Y, Zhou Y. A short-term photovoltaic power prediction model based on the gradient boost decision tree. Appl Sci. 2018;8(5):689.
13- Ogunleye A, Wang Q-G. XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Trans Comput Biol Bioinform. 2019;17(6):2131-40.
14- Li W, Yin Y, Quan X, Zhang H. Gene expression value prediction based on XGBoost algorithm. Frontiers in Genetics. 2019;10:1077.
15- Organization WH. Hearts: technical package for cardiovascular disease management in primary health care. 2016. Access link: https://apps.who.int/iris/bitstream/handle/10665/252661/9789241511377-eng.pdf
16- Thippeswamy B, Thakker S, Tubachi S, Kalyani G, Netra M, Patil U, et al. Cardioprotective effect of Cucumis trigonus Roxb on isoproterenol-induced myocardial infarction in rat. Am J Pharmacol Toxicol 2009;4(2):29-37.
17- Ateş S. Determining the Most Appropriate Ambulance Locations for Heart Attack Cases

with Geographic Information Systems: Graduate School of Sciences; 2010.
18- Upaganlawar A, Gandhi H, Balaraman R. Isoproterenol induced myocardial infarction: protective role of natural products. J Pharmacol Toxicol. 2011;6(1):1-17.
19- Alpaydin E. Introduction to machine learning: MIT press; 2020.
20- Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94.
21- Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat. Biomed. Eng. 2018;2(10):719-31.
22- Zeynep T, İpek BC, Guldogan E. Performance evaluation of the deep learning models in the classification of heart attack and determination of related factors. J. Cogn. Sci. 2020;5(2):99-103.