MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Physical Science & Biophysics Journal
ISSN: 2641-9165

Prediction of Heart Disease Using Machine Learning with Data Mining

Tanyildizi Kökkülünk H*

Radiotherapy Program, Vocational School of Health Sciences, Altınbaş University, Turkey

***Corresponding author:** Handan Tanyildizi-Kökkülünk, Radiotherapy Program, Vocational School of Health Sciences, Altınbaş University, Incirli st., No:11, Bakırkoy, Istanbul, Turkey, Tel: +902126040100; Email: handan.kokkulunk@altinbas.edu.tr

Research Article

Volume 7 Issue 1

Received Date: December 20, 2022

Published Date: January 12, 2023

DOI: 10.23880/psbj-16000228

Abstract

Aim: In this study, it was aimed to make a categorical estimation of the absent/presence of heart disease by using some parameters (age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thalach) of healthy and heart disease individuals.

Material and Methods: The classification was obtained with multiple linear regression (MLR) of machine learning in the R Studio program. Machine learning has been improved by selecting parameters that have a high contribution to the prediction by using the Akaike information criterion.

Results: The classification was performed using the biomarkers from glm.fit.1, which produced the lowest AIC value (237.48). The accuracy of the MLR model used was 88%, the precision was 93%, the sensitivity was 86%, and the specificity was 91%. It was found that age data from biomarkers contributed little to the prediction.

Conclusion: MLR is a preferable method for categorical disease classification.

Keywords: Heart; AIC criterion; Machine learning; Prediction

Introduction

Machine learning offers strategies, tactics, and resources that can assist in resolving diagnostic and prognostic issues in a range of medical specialties. The significance of clinical indicators and their combinations for prognosis is examined using machine learning. It has developed into a method that is often used to gather medical data for things like planning treatments, outcome studies, and estimating illness progression. Additionally, machine learning is employed for data analysis in the form of smart alerts, continuous data interpretation in intensive care units, and the replication of inaccurate or missing data based on pattern discovery in the available data. It is well recognized that when machine learning techniques are successfully deployed, they aid in the integration of computer-based healthcare systems, present chances to facilitate and enhance the work of medical professionals, and ultimately improve the effectiveness and caliber of medical care [1]. Machine learning is being used in the healthcare industry not to replace doctors, but to reduce their workload and give patients feedback more quickly and efficiently.

Machine learning has several important applications in the field of medical diagnosis [2]. In this approach, doctor-based techniques are used to create hypotheses from patient data. In order to do this, the system is enhanced with symbolic learning techniques and knowledge

management capabilities that are appropriate for the doctor's interpretation of the case. As a result, the forecast is generated using straightforward rules or, most often, a decision tree. The reporting of a medical imaging as a certain radiologist using machine learning is an illustration of this.

Biomedical signal processing is an additional application area [3]. With the use of machine learning techniques, it is feasible to model the linear or non-linear relationships that exist between the data and find the fundamental features and information that are concealed in physiological signals or that are likely to be disregarded. Additionally, machine learning is employed in radiography, magnetic resonance imaging, endoscopy, confocal microscopy, computer tomography, and other imaging techniques, particularly for the detection of cancerous regions. In addition to all of these uses, the most typical application of machine learning is to forecast diseases using categorical classification algorithms and patient data [4-6].

In addition to increasing daily, cardiovascular disorders make up a sizable share of non- communicable diseases. More than a third of all deaths are caused by cardiovascular illnesses, which constitute the leading cause of death worldwide [7]. Coronary heart disease, cerebrovascular disease, hypertension, congenital heart disease, peripheral artery disease, rheumatic heart disease, and inflammatory heart disease are all examples of cardiovascular illness. Tobacco use, physical inactivity, a poor diet, and abusing alcohol are the main causes of cardiovascular disease [8]. Rapid diagnosis has become more crucial due to the rising disease incidence and death.

The goal of this study was to use some demographic factors and biomarkers in combination with machine learning to categorize those with no history of heart disease (healthy) and those who have been diagnosed with heart disease (patients).

Material and Methods

Dataset and Preparation of Machine Learning

The UCI machine learning repository was used to get the heart disease dataset [9]. In order to divide the dataset consisting of 303 people exactly as training and test in the ratio of 80:20, 3 people were randomly removed. A total of 162 cases with heart disease and 138 instances without it were included in the heart disease dataset, and each instance was given 13 parameters to define it. The detailed descriptions of the parameters used were given in Table 1. Data types and some statistical values were given in Table 2.

age	age in years
sex	sex [1 = male, 0 = female]
cp	chest pain type [Value 0: typical angina, Value 1: atypical angina, Value 2: non-anginal pain, Value 3: asymptomatic]
trestbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	(fasting blood sugar > 120 mg/dl) [1 = true; 0 = false]
restecg	resting electrocardiographic results [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]
thalach	maximum heart rate achieved
exang	exercise induced angina [1 = yes, 0 = no]
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment [Value 0: upsloping, Value 1: flat, Value 2: downsloping]
ca	number of major vessels (0-3) colored by flourosopy
thalach	[0 = error (in the original dataset 0 maps to NaN's), 1 = fixed defect, 2 = normal, 3 = reversable defect]
target	(the lable): [0 = no disease, 1 = disease]

Table 1: Detailed information of parameters.

Name of Parameters	Minimum	Maximum	Mean
age	29	77	54.44
trestbps	94	200	131.59
chol	126	564	246.43
thalach	71	202	149.44
oldpeak	0	6.2	1.02
	Number of Person		
sex	0= 95, 1= 205		
cp	0=143, 1= 49, 2= 86, 3= 22		
fbs	0=256, 1= 44		
restecg	0=145, 1= 151, 2=4		
exang	0= 201, 1= 99		
slope	0=19, 1= 140, 2=141		
ca	0=172, 1= 65, 2= 38, 3= 20, 4=5		
thal	0=2, 1= 17, 2= 164, 3= 117		

Table 2: Data types and some statistical values for the entire data set.

The data collection with the.xlsx extension was used in the study to program machine learning algorithms using the R Studio software. 20% of the data set is used for testing, and the remaining 80% for training. The data set reserved for the test was used to check the accuracy of the classification and to evaluate the performance.

Classification Algorithm

Machine learning includes a variety of categorization and regression estimation techniques. Multiple linear regression (MLR) was chosen among machine learning regression methods for estimation since the dependent variable in the data set utilized in this study must be estimated as a categorical data type.

Multiple Linear regression (MLR)

One technique for determining the relationship between multiple independent variables (x1–xn) and a dependent variable, y, is known as multiple linear regression (MLR). According to the statement, MLR is a popular technique for estimating an unknown variable's value from the known values of two or more other variables [10]. The following Equation 1, which is created for n independent variables in a linear or linearizable way, often serves as the expression for this method [11].

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots\ldots + b_n x_n \quad (1)$$

Ten alternative fit processes were carried out by altering the biomarkers that were employed in the study. The goal of these ten separate methods is to identify the biomarkers that have the greatest impact on prediction. Each fit operation's Akaike information criterion (AIC) was determined [12]. The AIC is a single numerical value that can be used to identify the best model for a given dataset among the various models. It is well known that a model that has a lower AIC value than the rest classifies more accurately.

Performance Evaluation Metrics

Utilizing criteria like accuracy, precision, sensitivity, and specificity, the machine learning classification algorithm's efficiency is evaluated [13]. Following the estimation, the confusion matrix for this was established. It displays the numbers with TP true positive, TN true negative, FP false positive, and FN false negative in order to depict the positive healthy people and the negative heart disease patients in the confusion matrix and metrics. As a result, the calculations for accuracy, precision (P), sensitivity, and specificity are displayed below at Figure 1.

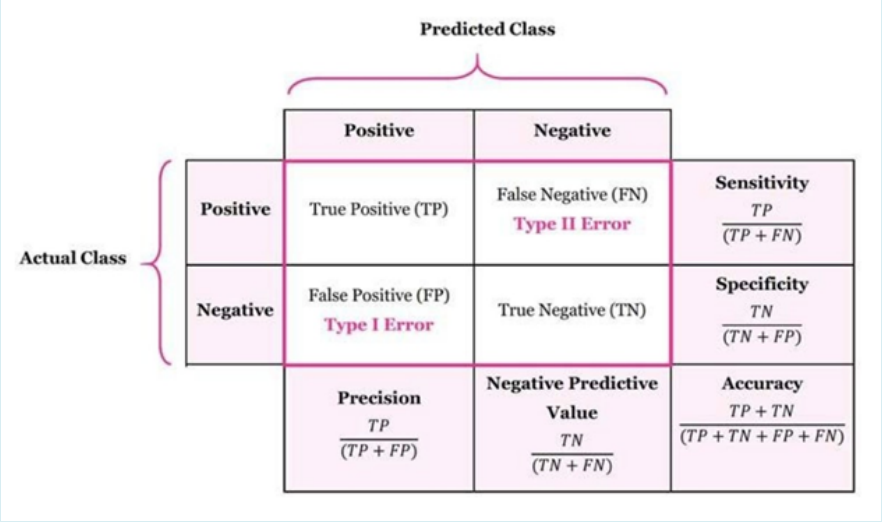

Figure 1: Numerical data in the form of TP, TN, FP, and FN are used to measure performance. Based on these findings, accuracy, precision, specificity, and sensitivity are computed. The percentage of positive cases constitutes sensitivity, the percentage of negative cases is specificity, and the percentage of correctly classified cases constitutes accuracy [14].

Results and Discussion

The study examined age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, and thalach data to investigate the presence of heart disease using certain biomarkers. While these parameters are within the typical reference range for healthy people, they do not match the reference range for sick people. Adults who are healthy are supposed to have no cp complaints, fbs values under 120 mg/dl [15], resting heart rates between 60 and 100 [16], no exang symptoms, and chol values under 200 mg/dl [17]. While there is a significant difference in age, sex, cp, trestbps, restecg, thalach, exang, oldpeak, slope, ca, and thalach values between healthy and sick persons in the data set used in the study ($p<0.05$), there is no significant difference in the chol and fbs values ($p>0.05$).

Ten different randomly generated fitts were explored in order to identify the biomarkers that contribute significantly to the AIC value for disease prediction. The glm.fit method is employed in this step to enhance machine learning. The biomarkers and AIC data included in each foot are displayed in Table 3.

Number	Fit names	Biomarkers	AIC
1	glm.fit.0	age+sex+cp+trestbps+chol+fbs+restecg+thalach+exang+oldpeak+slope+ca+thal	283.51
2	glm.fit.1	sex+cp+trestbps+chol+fbs+restecg+thalach+exang+oldpeak+slope+ca+thal	237.48
3	glm.fit.2	cp+trestbps+chol+fbs+restecg+thalach+exang+oldpeak+slope+ca+thal	251.85
4	glm.fit.3	trestbps+chol+fbs+restecg+thalach+exang+oldpeak+slope+ca+thal	272.43
5	glm.fit.4	chol+fbs+restecg+thalach+exang+oldpeak+slope+ca+thal	271.74
6	glm.fit.5	fbs+restecg+thalach+exang+oldpeak+slope+ca+thal	269.93
7	glm.fit.6	restecg+thalach+exang+oldpeak+slope+ca+thal	268.04
8	glm.fit.7	thalach+exang+oldpeak+slope+ca+thal	269.11
9	glm.fit.8	exang+oldpeak+slope+ca+thal	276.41
10	glm.fit.9	oldpeak+slope+ca+thal	301.29

Table 3: The biomarkers and AIC results for each fit group.

It is advised to choose the model with the lowest AIC value for predicting the disease. Due to this, the estimation was performed using the biomarkers from glm.fit.1, which produced the lowest AIC value (237.48). It was found that

age data from biomarkers contributed little to the prediction.

Machine learning was carried out with MLR in the form of healthy/patient prediction using sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, and thal biomarkers. Table 4 contains the complexity matrix that was produced as a result of the estimation.

		Estimation	
		Positive	Negative
Real	Positive	31	5
	Negative	2	22

Table 4: Confusion matrix for healthy/patient estimation with MLR.

The classification of healthy and heart sick people using the MLR algorithm of machine learning was found to be 88% successful. The precision value, which shows how many of the values we predicted as positive are actually positive, was calculated as 93%. The sensitivity and specificity were found to be 86% and 91%, respectively. Table 5 displays a comparison of studies that classified heart disease using various methods and the same data set.

Literature	Method	Accuracy (%)
Chaurasia, et al. [7]	Classification and Regression Tree	83.49
Mohan, et al. [18]	HRFLM classification	88.4
Taşçı, et al. [19]	Naive Bayes	88.52
In this study	MLR	88

Table 5: The literature comparison.

Chaurasia, et al. [7] performed heart disease classification with popular algorithms such as CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) and reported that they achieved 83.49%, 72.93%, and 82.50% accuracy rates, respectively. By classifying using deep learning with artificial neural networks, the hybrid random forest with a linear model, and many other models, Mohan, et al. [18] recommended the adoption of the hybrid random forest with a linear model (HRFLM), with the greatest accuracy rate of 88.4. Taşçı, et al. [19] classified cardiac disease in their experiments using well-known algorithms as Naive Bayes, Random Forest, Decision Tree, Logistic Regression, and SVM. They claimed that the Naive Bayes algorithm stood out among all of these techniques with an accuracy of 88.52%.

The MLR algorithm was used in this study to classify cardiac disease, and the results were found to be 88% accurate and consistent with the literature.

Conclusion

In this study, instead of using complex or multiple algorithms as in the literature, the biomarkers used in classification were rearranged. In this way, only the markers with a high contribution to the classification were included in the study and 88% accuracy was obtained by using a single algorithm called MLR. Considering the clinical intensity, it has been seen that the MLR algorithm, which is easy to apply, provides fast and highly accurate results, will be useful in estimating heart disease.

References

1. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, et al. (2017) Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2(4): 230-243.

2. Stausberg J, Person M (1999) A process model of diagnostic reasoning in medicine. International Journal of Medical Informatics 54(1): 9-23.

3. Choubey A, Bhargava Choubey S, Subba Rao SPV (2021) 7 A Machine Learning Algorithm for Biomedical Signal Processing Application. Machine Learning Methods for Signal, Image and Speech Processing, River Publishers, pp: 149-168.

4. Srinivas S (2020) A Machine Learning-Based Approach for Predicting Patient Punctuality in Ambulatory Care Centers. Int J Environ Res Public Health 17(10): 3703.

5. Anusuya V, Gomathi V (2021) An Efficient Technique for Disease Prediction by Using Enhanced Machine Learning Algorithms for Categorical Medical Dataset. Information Technology and Control 50(1): 102-122.

6. Uddin S, Khan A, Hossain ME, Moni MA (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making 19(1): 281.

7. Chaurasia V, Pal S (2013) Early Prediction of Heart Diseases Using Data Mining Techniques. Caribbean Journal of Science and Technology 1: 208-217.

8. Srinivas K, Rao G, Govardhan D (2010) Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. ICCSE 2010 5th International Conference on Computer Science and Education, China.

9. UCI Machine Learning Repository (2022) Heart Disease Data Set.

10. Yee MM, Aung EE, Khaing YM (2022) Forecasting Stock Market using Multiple Linear Regression. International Journal of Trend in Scientific Research and Development 3(5): 2174-2176.

11. Giacomino A, Abollino O, Malandrino M, Mentasti E (2011) The role of chemometrics in single and sequential extraction assays: A Review. Part II. Cluster analysis, multiple linear regression, mixture resolution, experimental design and other techniques. Analytica chimica acta 688(2): 122-139.

12. Khalid A, Sarwat AI (2021) Unified Univariate-Neural Network Models for Lithium-Ion Battery State-of-Charge Forecasting Using Minimized Akaike Information Criterion Algorithm. IEEE Access 9: 39154-39170.

13. Hasan M, Islam M, Zarif MII, Hashem MMA (2019) Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. Internet of Things 7: 100059.

14. Shajihan N (2020) Classification of stages of Diabetic Retinopathy using Deep Learning.

15. Ding HJ, Shiau YC, Wang JJ, Ho ST, Kao A (2002) The influences of blood glucose and duration of fasting on myocardial glucose uptake of [18F] fluoro-2-deoxy-D-glucose. Nucl Med Commun 23(10): 961-965.

16. Li SJ, Sartipy U, Lund LH, Dahlström U, Adiels, M, et al. (2015) Prognostic Significance of Resting Heart Rate and Use of β-Blockers in Atrial Fibrillation and Sinus Rhythm in Patients With Heart Failure and Reduced Ejection Fraction. Circulation: Heart Failure 8(5): 871-879.

17. Ma H (2004) Cholesterol and Human Health. Nature and Science 2(4): 17-21.

18. Mohan S, Thirumalai C, Srivastava G (2019) Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 7: 81542-81554.

19. Taşçi ME, Şamli R (2020) Veri Madenciliği İle Kalp Hastalığı Teşhisi. European Journal of Science and Technology, pp: 88-95.