International Journal of Industrial Engineering & Production Research March 2023 Vol. 34, No. 1: 1-15
DOI: 10.22068/ijiepr.34.1.5

RESEARCH PAPER

Prediction of Heart Diseases Using Logistic Regression and Likelihood Ratios

Mohammad Yaseliani[1] & Majid Khedmati[2*]

Received 30 August 2022; Revised 28 Novamber 2022; Accepted 12 December 2022;

ABSTRACT

Diagnosis of diseases is a critical problem that can help for more accurate decision-making regarding the patients' health and required treatments. Machine learning is a solution to detect and understand the symptoms related to heart disease. In this paper, a logistic regression model is proposed to predict heart disease based on a dataset with 299 people and 13 variables and to evaluate the impact of different predictors on the outcome. In this regard, at first, the effect of each predictor on the precise prediction of the outcome has been evaluated and analyzed by statistical measurements such as AIC scores and p-values. The logit models of different predictors have also been analyzed and compared to select the predictors with the highest impact on heart disease. Then, the combined model that best fits the dataset has been determined using two statistical approaches. Based on the results, the proposed model predicts heart disease with a sensitivity and specificity of 84.21% and 90.38%, respectively. Finally, using normal probability density curves, the likelihood ratios have been established based on classes 1 and 0. The results show that the likelihood ratio classifier performs as satisfactorily as the logistic regression model.

KEYWORDS: *Logistic regression; Heart disease; Likelihood ratio; Receiver operating characteristic (ROC); Akaike information criterion (AIC).*

1. Introduction

According to WHO, cardiovascular diseases (CVDs) are the number one cause of death all over the world, taking about 17.9 million lives each year. Some factors of CVDs can be easily measured and the people at risk can be treated. Most CVDs may be avoided by implementing population-wide programs to address behavioral risk factors such as cigarette use, poor nutrition, obesity, and physical inactivity [1]. According to WHO, 80% of heart attacks and strokes are preventable, and controlling high blood pressure, high cholesterol, and high blood sugar plays an important role in preventing from heart disease [2]. It should be noted that CVDs are considered multifactor diseases which are caused by a range of genetic, environmental, and nutritional factors [3].

* **Corresponding author:** ***Majid Khedmati***
khedmati@sharif.edu

1. *Department of Industrial Engineering, Isfahan University of Technology, Isfahan, Iran.*
2. *Department of Industrial Engineering, Sharif University of Technology, Azadi Ave., Tehran 1458889694 Iran.*

Machine learning is the process of developing and applying algorithms to detect automatically the patterns in the data. Machine learning is a branch of applied statistics that involves creating computer models that rely on inference and pattern recognition rather than explicit rules [5]. The machine learning algorithms range from simple linear regression to complicated multilayer neural networks [4] and they can be utilized, along with statistical models, to create predictive models [6]. Recently, machine learning has been used in machine learning-based smart healthcare, maintenance prioritization in healthcare facilities, multiple cancer diagnosis frameworks, and duration of hospitalization in COVID-19 patients [7-10]. In the case of heart diseases, machine learning algorithms predict the likelihood of a diagnosis, analyzing some factors and predictors [11-14].

Data mining and machine learning approaches turn the extensive collection of raw healthcare data into information and knowledge to make informed decisions and predictions [15]. There exist different machine learning approaches including discriminant function analysis (DFA)

and logistic regression to handle and predict categorical variables [16-18]. Among them, logistic regression is a statistical tool for estimating the association between a binary outcome and one or more covariates where the covariates may be either discrete or continuous [16]. Compared to other methods, logistic regression does not require normally distributed predictors and is not sensitive to outliers. Furthermore, logistic regression can abide by differences in the two classes [19]. Besides, although some approaches such as DFA may be superior to logistic regression when the normality criteria are satisfied, the differences between these methods become insignificant, when the sample size is high enough (50 observations or more) [20].

The likelihood ratio is the probability of a clinical finding in patients with disease divided by observing that in patients without the disease [21]. As an example, the likelihood ratio compares the chances of properly predicting cancer against the chances of wrongly predicting cancer. The likelihood ratio indicates how much a diagnostic test result will increase or decrease the suspected disease's pre-test likelihood. This method has recently been used in the prediction of sex from a set of continuous variables [17]. Accordingly, it is utilized in this paper to discern how close it is to the predictions of the logistic regression model developed for heart disease prediction.

During the past few years, various research studies have developed machine learning-based methods for the prediciton of heart disease. The following studies have used the UCI heart disease dataset which is described in the next section. Magar et al. [22] applied logistic regression, support vector machines (SVM), naïve bayes (NB), and decision tree (DT) classifiers UCI heart disease dataset and achieved an accuracy of 83% using the logistic regression model. Shah et al. [23] used data transformation technqiues and deployed a k-nearest neighbors (KNN), NB, random forest (RF), and DT classifier. They achieved an accuracy of approximately 90% using their k-nearest neighbors model. Pandita et al. [24] used logistic regression, KNN, SVM, NB, and RF to predict heart diseases. They achieved more than 89% accuracy using KNN and created a web application based on this classifier. Akella et al. [25] developed an artificial neural network (ANN) model to predict heart diseases with 93% accuracy. They compared the performance of this model with other methods, including logistic regression, NB, fuzzy KNN, and K-means clustering, which all were inferior to their ANN model in terms of accuracy. There have also been other studies which have developed state-of-the-art machine learning-based classifiers to classify heart diseases on other heart disease datasets [26].

Although many studies have developed special classifiers for the prediction of heart disease, a logistic regression model based on two variable selection approaches (i.e., forward selection and backward elimination) is proposed in this paper for creating an accuracte final model. In addition, the likelihood ratio analysis is utilized to get the predictions and compare its performance with a machine learning model for the prediction of heart diseases. To the best of authors' knowledge, this approach has not been previously considered in the literature by other researchers.

In this paper, it has been tried to provide a model for estimation of heart disease based on a set of variables such as age, sex, serum cholesterol, etc. These predictors are compared through logistic regression models, and then, two feature selection methods, including forward selection and backward elimination approaches are employed and compared to select a set of variables that have the highest predictive power. Finally, a likelihood ratio analysis is performed to discern if they can be utilized as a classifier and alternative to the logistic regression model to predict heart disease.

The rest of the paper is organized as follows. The materials and methods are presented in section 2. The results are presented and discussed in detail in section 3. Finally, the paper is concluded in section 4.

2. Materials and Methods

2.1. Logistic regression

Logistic regression is a statistical method for binary classification which can be generalized to multiclass classification [27]. Prediction of acute kidney injury, survival prediction of hepatocellular carcinoma, prediction of low-velocity impact damage in composite structures, fog computing, early diagnosis of Alzheimer's disease, prediction of deforestation, prediction of tuberculosis, and prediction of anxiety disorders are some of the recent applications of logistic regression models in various fields [28-35]. In addition, logistic regression has a great potential of being combined with other algorithms such as genetic algorithms, evolutionary generalized radial basis function, and non-parametric models to make accurate predictions [29, 36, 37]. It has been shown that the combination of logistic regression classifier with genetic algorithms can improve the prediction accuracy, significantly

[29]. The essence of the problem considered in this paper is a binary classification problem. The probability of a binary output, given a set of variables, is represented as follows [38].

$$p(y|x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m}} \quad (1)$$

where, y is the outcome variable, x is the set of predictors, and, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients. The odds ratio (OR) is one of the statistical measurements used in clinical research and decision-making. It is particularly important since it informs the doctors clearly and directly about which therapy method is most likely to help the patient [39]. The odds ratio is defined as the ratio of the probability of occurrence to non-occurrence of an event [40].

$$\text{odds} = \frac{p(y|x)}{1 - p(y|x)} \quad (2)$$

The logit function is represented as the natural logarithm of the odds ratio [40].

$$\text{logit}(x) = \ln(\text{odds}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m \quad (3)$$

In these formulas, p(y|x) represents the probability of class 1, given a set of variables x [40].

In this paper, the coefficients of logit models are calculated for each predictor. Additionally, the odds ratios and probabilities are utilized to compare the results of logit models and likelihood ratio classifiers, which have been described in the next sections.

2.2. Description of the dataset

The dataset used in this paper is the "Heart Disease Dataset" of the UCI Machine Learning Repository. The dataset consisted of a sample of 299 people, aged between 29 and 78 years old. The number of class 1 is 134 against 165 instances representing class 0 [41]. There have been four principal investigators responsible for the data collection at four institutions:

Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

The traditional risk factors causing heart diseases that are reported in the Framingham risk score are age, sex, blood pressure, serum total cholesterol level, low-density lipoprotein or high-density lipoprotein cholesterol level, cigarette smoking, and diabetes [42]. In this paper, a subset of 13 features is used to create a model relevant to clinical conditions related to heart disease. The collected data consisted of 13 predictors, including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, the slope of ST segment in the peak of exercise, number of major vessels colored by fluoroscopy, and types of thalassemia [41].

2.3. Preprocessing information and cleansing considerations

In this paper, the dataset is divided into training and test set. The training set includes 70% of the dataset, and the test set is comprised of 30% of the dataset. The analysis is performed using Python software (version 3.9) to get the results based on the training set, and evaluations are performed based on the test set. In this regard, the categorical variables are changed and transformed into binary variables to be used in the logit model. For categorical predictors with two possible values including sex, fasting blood sugar, and exercise-induced angina, one of the possible states is utilized to input these particular predictors in their corresponding logit model where, in this paper, the states of "male" for sex, "yes" for fasting blood sugar, and "yes" for exercise-induced angina are used as the input of these variables. Other categorical variables including the chest pain type, resting electrocardiographic results, thalassemia, and the slope of the ST segment at the peak of exercise have more than two possible states and hence, they are divided into 3 or 4 different predictors according to their possible states. There are four types of chest pain, three types of resting electrocardiographic results, three types of thalassemia, and three types of the slope of the ST segment at the peak of exercise. In addition, there are six continuous variables including age, serum cholesterol, resting blood pressure, maximum heart rate achieved, ST depression induced by exercise relative to rest, and the number of major vessels colored by fluoroscopy. In Table 1, the types of all the predictors as well as their indicators in the logit models are presented. Applying the logistic regression function in Python, the results are obtained for all

predictors and the combined models are developed. In Figure 1, these six continuous variables have been analyzed using box plots, separately, for both classes in which, classes 1 and 0 represent having and not having heart disease, respectively. In Figure 1, the positive class stands for class 1, and the negative class stands for class 0.

Tab. 1. Types of predictors

Variable	Type
age (x1)	numerical
male (x2)	binary
resting blood pressure (x3)	numerical
serum cholesterol (x4)	numerical
maximum heart rate achieved (x5)	numerical
asymptomatic for chest pain (x6)	binary
atypical angina for chest pain (x7)	binary
non-anginal for chest pain (x8)	binary
typical angina for chest pain (x9)	binary
left ventricular hypertrophy for resting electrocardiographic results (x10)	binary
ST-T wave for resting electrocardiographic results (x11)	binary
normal for resting electrocardiographic results (x12)	binary
fixed for thalassemia (x13)	binary
normal for thalassemia (x14)	binary
reversible defect for thalassemia (x15)	binary
down sloping for the slope of ST segment in the peak of exercise (x16)	binary
flat for the slope of ST segment in the peak of exercise (x17)	binary
up sloping for the slope of ST segment in the peak of exercise (x18)	binary
number of major vessels colored by fluoroscopy (x19)	numerical
yes for fasting blood sugar (x20)	binary
ST depression induced by exercise relative to rest (x21)	numerical
yes for exercise-induced angina (x22)	binary

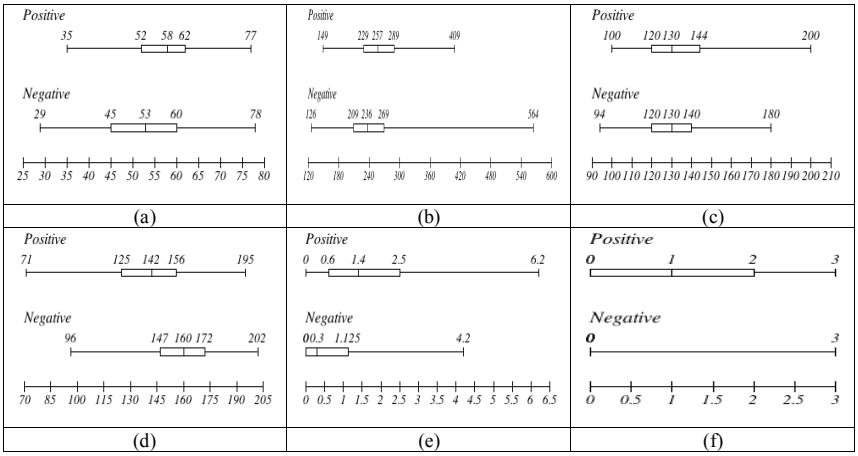


Fig. 1. Box-plots of various variables for positive and negative test result a) age, b) serum cholesterol, c) resting blood pressure, d) maximum heart rate achieved, e) ST depression induced by exercise relative to rest, f) number of major vessels colored by fluoroscopy

3. Results and Discussion

3.1. Description of logit models and their statistical measurements

In this section, the logistic regression models are developed for various predictors and the results are presented Table 2. For categorical predictors such as male, number 1 means being a male and 0 means being a female. Similarly, for fasting blood sugar, 1 means having fasting blood sugar, and 0 represents not having blood sugar. Other binary preditors are interpreted in the same way. To train and test the models, we used 90 samples to calculate the accuracy and 209 samples to train the models, which corresponds to 30% and 70% of the dataset, respectively.

The percentage of people with heart disease who are accurately diagnosed by the test is known as sensitivity while the percentage of those without heart disease who are accurately excluded by the test is known as specificity. A confusion matrix is a prominent tool that is used in classification problems and it can be used to solve multiclass as well as binary classification problems. This matrix is used to demonstrate the counts based on expected and actual values. The output "TN" stands for True Negative and displays the number of correctly identified negative cases. Similarly, "TP" stands for True Positive, which denotes the number of correctly identified positive cases. The terms "FP" and "FN" stand for False Positive and False Negative, respectively. "FP" is the number of actual negative cases categorized as positive, and "FN" shows the number of actual positive cases, classified as negative [43]. At first, the total accuracy is calculated and then, the sensitivity and specificity are obtained. Equations (4)-(6) represent the formulation of sensitivity, specificity, and accuracy measures [44].

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{specificity} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

The Akaike Information Criterion (AIC) was established by Akaike as a measure to compare the alternative models. The model selection is critical, because an under-fitted model may not represent the real nature of variability in the outcome variable, whereas an overfitted model loses generality. The AIC is then used to choose the model that best balances these problems. Akaike showed that the selection of the best model is determined by an AIC score formula, as follows:

$$\text{AIC} = ٢ * K - ٢ * \ln(L) \tag{7}$$

where L represents the likelihood function, and K denotes the number of estimated parameters (degrees of freedom) [45].

Tab. 2. Logistic regression models

Logit Model	AIC	BIC
-1.857+0.030*x1	404.59	411.99
-1.243+1.526*x2	389.50	396.90
-1.918+0.013*x3	409.62	417.02
-0.756+0.002*x4	410.80	418.21
5.251-0.036*x5	367.12	374.52
-1.027+1.766*x6	342.45	349.85
0.068-1.644*x7	397.23	404.63
0.106-1.011*x8	388.41	395.81
-0.134-0.375*x9	412.77	420.17
-0.635+0.923*x10	404.10	411.50
-0.163 (no coefficient)	415.26	422.66
0.287-0.923*x12	403.92	411.32
-0.184+0.338*x13	413.81	421.21
0.923-2.045*x14	340.11	347.51
-0.966+2.032*x15	347.22	354.62
-0.165+0.031*x16	414.39	421.79
-0.843+1.443*x17	374.57	381.97
0.496-1.463*x18	369.85	377.25
-0.915+1.100*x19	344.65	352.05
-0.125-0.230*x20	415.10	422.50
-1.042+0.828*x21	357.56	364.96
-0.714+1.772*x22	371.01	378.41

Tab. 3. The accuracies of logistic regression models

Variable	Accuracy (%)	Sensitivity (%)	Specificity (%)
age (x1)	60.00	31.57	80.76
male (x2)	57.77	73.68	46.15
resting blood pressure (x3)	66.67	34.21	90.38

serum cholesterol (x4)	58.88	13.15	92.30
maximum heart rate achieved (x5)	70.00	60.52	76.92
asymptomatic for chest pain (x6)	82.22	86.84	78.84
atypical angina for chest pain (x7)	50.00	94.73	17.30
non-anginal for chest pain (x8)	68.88	92.10	51.92
typical angina for chest pain (x9)	57.77	0.00	100.00
left ventricular hypertrophy for resting electrocardiographic results (x10)	55.55	57.89	53.84
ST-T wave for resting electrocardiographic results (x11)	57.77	0.00	100.00
normal for resting electrocardiographic results (x12)	55.55	60.52	51.92
fixed for thalassemia (x13)	60.00	5.26	100
normal for thalassemia (x14)	77.77	71.05	82.69
reversible defect for thalassemia (x15)	75.55	65.78	82.69
down sloping for the slope of ST segment in the peak of exercise (x16)	57.77	0.00	100.00
flat for the slope of ST segment in the peak of exercise (x17)	71.11	68.42	73.07
up sloping for the slope of ST segment in the peak of exercise (x18)	74.44	76.31	71.15
number of major vessels colored by fluoroscopy (x19)	80.00	68.42	88.46
yes for fasting blood sugar (x20)	57.77	0.00	100.00
ST depression induced by exercise relative to rest (x21)	72.22	55.26	84.61
yes for exercise-induced angina (x22)	70.00	55.26	80.76

Bayesian Information Criterion (BIC) is another model selection criterion that assesses the trade-off between model fit and model complexity which is presented in the following equation:

$$\mathrm{BIC} = -۲ * \ln(L) + ۲ * \ln(N) * K \quad (8)$$

where L represents the likelihood function, N denotes the number of measurements, and K stands for the number of estimated parameters [45]. A better fit is indicated by a lower AIC or BIC value [45]. These measurements, as well as the p-values, are used in this paper for a comparison of the models.

In Tables 2 and 3, the logit models and their corresponding statistical measurements have been presented.

3.2. Comparison of logit models

As mentioned previously, less AIC score means the model is better at predicting the output [45, 46]. Based on the results, all the predictors are significant, by a p-value less than 0.05, except the variables down sloping for the slope of ST segment in the peak of exercise, fixed for thalassemia, ST-T wave for resting electrocardiographic results, typical angina for chest pain, and yes for fasting blood sugar. According to the AIC scores, it is concluded that the normal (thalassemia) is the best predictor of heart disease. The AIC of the logit model with normal (thalassemia) is 340.11 and the total accuracy is equal to 77.77%. After normal thalassemia, asymptomatic chest pain with an AIC of 342.45, and the number of major vessels colored by fluoroscopy with an AIC of 344.65 is selected as the best predictors of heart disease. The ranking of other predictors from lowest to highest AIC is the reversible defect, ST depression induced by exercise relative to rest, male, maximum heart rate achieved, up sloping for the slope of ST segment in the peak of exercise, yes for exercise-induced angina, flat for the slope of ST segment in the peak of exercise, non-anginal for chest pain, atypical angina for chest pain, normal for resting electrocardiographic results, left ventricular hypertrophy for resting electrocardiographic results, age, resting blood pressure, serum cholesterol, typical angina for chest pain, fixed for thalassemia, down sloping for the slope of ST segment in the peak of exercise, yes for fasting blood sugar, and ST-T wave for resting electrocardiographic results. In real-world applications, the results obtained based on the predictors with lower AIC are more reliable. However, for more analysis, the models are evaluated based on BIC scores in which, normal (thalassemia), asymptomatic chest pain, and the number of major vessels colored by fluoroscopy are selected, again, as the best three predictors of

heart disease. However, the ranking of other predictors from the lowest to the highest BIC led to different results. If a range of model sizes are being compared and no finite size provides the exact parametric model, AIC will do better than BIC [47]. Thus, AIC is selected as the main factor in comparing different predictors. A combined model is also constructed to be used as the predictive model. In the next section, this combined model has been introduced.

3.3. Building a combined model

There are two approaches to building a combined model: forward selection and backward elimination approaches [48]. In this paper, the backward elimination and forward selection approaches are used to get the best-combined model. In the backward elimination, the model starts containing all the predictors and if removing a predictor leads to a decrease in the significance level or increase in the AIC score, the predictor is retained in the model. Otherwise, the predictor is dropped. Applying the backward elimination approach, the best combined model is determined as -4.079 + 1.753x2 + 1.224x6 – 0.752x12 + 1.609x15 + 0.888x17 + 1.092x19 + 0.589x21. All these predictors are significant, with a p-value less than 0.05. The sensitivity of this model is 84.21%, and the specificity is equal to 90.38%. Also, the total accuracy of this model is obtained as 87.77%. In addition, the AIC score and p-value of the combined model are 232.184 and 1.252 * 10-38, respectively.

In the forward selection approach, one starts with a model without any predictors and if adding a new predictor does not decrease the significance level but the AIC score decreases, the predictor is added to the model.

Tab. 4. Comparison of the accuracy with other studies

Author(s)	Method	Accuracy	Dataset
Magar et al. [22]	Logistic regression, SVM, NB, DT	82.89%	-
Shah et al. [23]	NB, KNN, RF, DT	90.79%	-
Akella et al. [25]	DT, RF, SVM, ANN, KNN	93.03%	-
Zhang et al. [49]	Logistic Regression	85.86%	-
Prasad et al. [50]	Logistic Regression	86.89%	-
Khanna et al. [51]	Logistic Regression	84.80%	50% training, 50% testing
Khanna et al. [51]	SVM (linear)	87.60%	50% training, 50% testing
Kodati et al. [52]	Naïve Bayes	83.70%	-
Latha & Jeeva [53]	Majority vote with NB, BN, RF, MP	85.48%	-
This paper	Logistic Regression	87.77%	70% training, 30% testing

Otherwise, it is removed from the model. Using forward selection approach, the best combined model is determined as -4.131 + 2.011x2 + 1.419x6 + 0.974x17 + 1.037x19 + 0.614x21. The AIC of this model is 244.999 and the p-value is equal to 1.245 * 10-36.

According to AIC scores, the combined model resulting from the backward elimination approach is selected as the best model. In addition, the p-value of the model obtained from the backward elimination approach is slightly more significant.

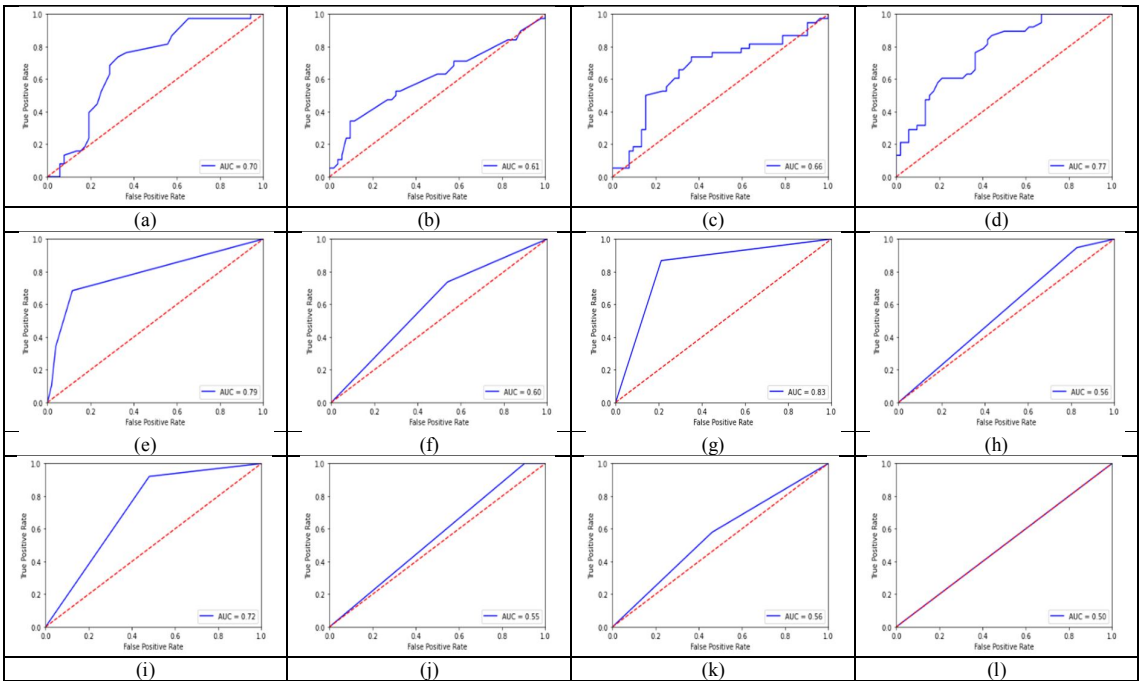
In Table 4, the accuracy of the proposed model is compared to the accuracy of the competing models in the literature applied to the heart disease dataset. Based on the results, the proposed logistic regression model has superior accuracy, compared to other models. Comparing the logistic regression model proposed in this paper to the ones proposed by Zhang et al. [49], Prasad et al. [50], and Khanna et al. [51], the proposed model resulted in, respectively, 1.91%, 0.88%, and 2.97% increase in the accuracy. Although the accuracy of the ANN classifier proposed by Akella et al. [25] and the KNN classifier proposed by Shah et al. [23] are higher than that of our logit model, it should be noted that the ANN classifier is black-box in nature and the decsription of the results obtained by this classifier is a chellenging task. Moreover, the KNN classifier is highly dependent on the training set and works based on the majority voting of the results. However, the logistic regression model can explain the contribution of each variable to the output and provides better insight into how the model works. Therefore, despite the lower total accuracy of our logistic regression model, it is easily interpretable and the selection of variables has been made based on various statistical analyses which is the advantage of our model compared to other studies. Overall, the logistic regression model proposed in the

current study has been created based on the best predictive variables and has a significant performance on the UCI dataset compared to other studies and fair interpretability to make more informed medical decisions.

3.4. Comparison of ROC curves

ROC curve is another concept that is used to assist with data interpretation [54]. The performance of the classifiers can be evaluated and compared based on receiver operating characteristic (ROC) curves. It should be noted that ROC curves are extensively used in medical decision-making and have been increasingly popular in machine learning and data mining research, in recent years [55]. For instance, this concept has recently been used in assessing image quality in digital mammography and evaluating the performance of clinical prediction models [56, 57]. The model with the largest ROC area would best classify the output.

The ROC curves of different predictors are shown in Figure 2. Most of the classifiers consider equal weights for false negative and false positive rates, while in real applications, this is not true [58]. For instance, in the medical diagnosis of cancer, the false negative error is much more severe than the false positive error, which means the prediction of having cancer, while not having cancer [59]. Although ROC curves appear to be straightforward, they are subject to several typical misunderstandings and problems, when used in reality [58]. In this regard, ROC curves assume that the misclassification costs are the same [60]. Since the nature of this study is related to the diagnosis of a disease, the false negative error comes at a much greater cost than the false positive error. Therefore, more emphasis is concentrated on the AIC score as a proper criterion for comparing different models. According to AUCs, the asymptomatic, the number of major vessels colored by fluoroscopy, and normal (thalassemia) have the largest area under the curve (AUC). This conclusion is in accordance with the one obtained based on the AIC scores. The maximum heart rate achieved has also an AUC equal to normal (thalassemia). These predictors are followed by ST depression induced by exercise relative to rest, reversible defect, up sloping for the slope of ST segment in the peak of exercise, non-anginal for chest pain, flat for the slope of ST segment in the peak of exercise, age, yes for exercise-induced angina, serum cholesterol, resting blood pressure, male, atypical angina for chest pain, left ventricular hypertrophy for resting electrocardiographic results, normal for resting electrocardiographic results, typical angina for chest pain, down sloping for the slope of ST segment in the peak of exercise, fixed for thalassemia, ST-T wave for resting electrocardiographic results, and yes for fasting blood sugar ranked from highest to lowest AUC. In Figure 3, the AUC of the combined model has been presented. The AUC of this model is 0.91, which is expectable, due to its higher prediction accuracy compared to other models' performances.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

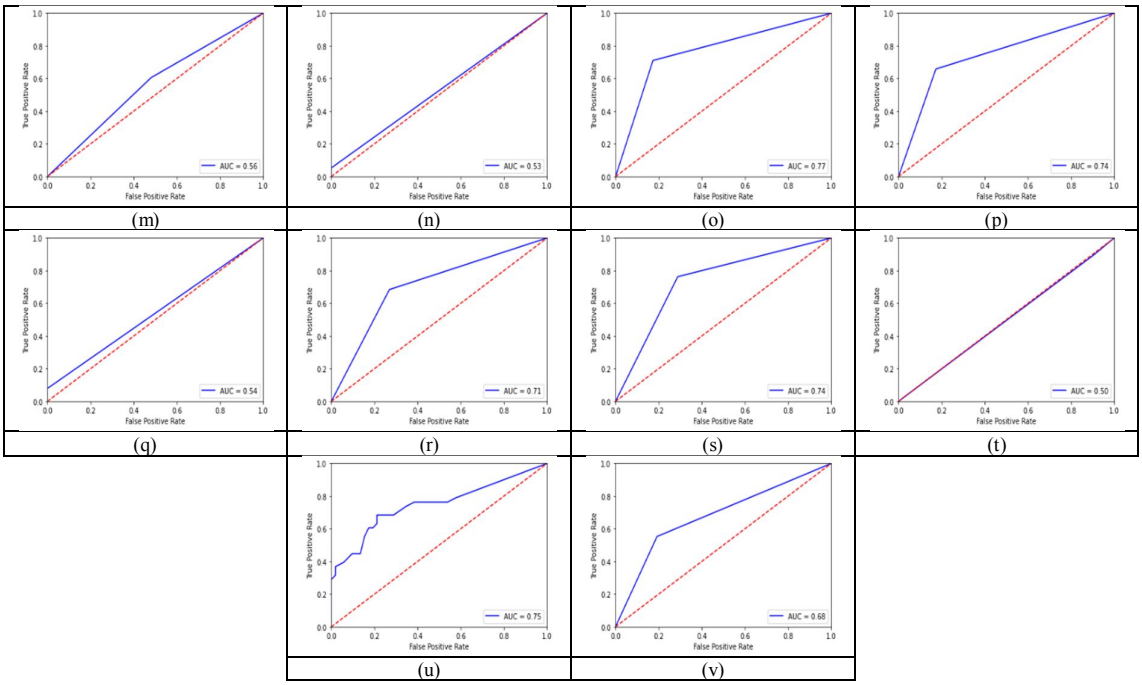


Fig. 2. ROC curves for various models a) age, b) resting blood pressure, c) serum cholesterol, d) maximum heart rate achieved, e) number of major vessels colored by fluoroscopy, f) male, g) asymptomatic, h) atypical angina, i) non-anginal, j) typical angina, k) left ventricular hypertrophy, l) ST-T, m) normal (resting electrocardiographic results), n) fixed, o) normal (thalassemia) p) reversible defect, q) down sloping, r) flat, s) up sloping, t) yes (fasting blood sugar), u) ST depression induced by exercise relative to rest, v) yes (exercise-induced angina)

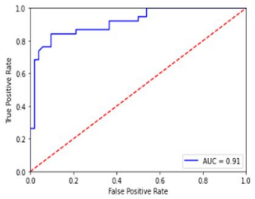


Fig. 3. ROC curve for the combined model

3.5. Likelihood ratio analysis

The normal distribution is an underlying assumption of many statistical processes. There exist a number of normality tests including the Shapiro-Wilk (SW) test, KS test, LL test, CVM test, AD test, CSQ test, JB test, and DP omnibus test. It has been shown that the Shapiro-Wilk test outperforms other normality tests [61]. In addition, it provides better power than the KS test even after Lilliefors correction [62, 63]. Hence, in this paper, the SW normality test is utilized to detect whether the numerical predictors are normally distributed to be used in the likelihood ratio analysis. According to the results presented in Tables 6 and 8, none of the numerical predictors are normally distributed for both classes 0 and 1.

Tab. 5. Statistical parameters for various numerical variables with class 0

Variable	Min	Max	Median	Mean	Standard Deviation
age	29	78	56	53.393	9.764
resting blood pressure	94	180	130	129.163	16.397
serum cholesterol	16	564	236	245.533	58.775
maximum heart rate achieved	96	202	160	157.339	20.171
ST depression induced by exercise relative to rest	0	4.200	0.300	0.618	0.782
number of major vessels colored by fluoroscopy	0	3	0	0.284	0.642

Tab. 6. Shapiro-Wilk test for class 0

Variable	Shapiro-Wilk test (p-value)
age	0.095
resting blood pressure	0.017
serum cholesterol	8.793*10-12
maximum heart rate achieved	3.369*10-4
ST depression induced by exercise relative to rest	4.729*10-14
number of major vessels colored by fluoroscopy	1.110*10-16

Tab. 7. Statistical parameters for various numerical variables with class 1

Variable	Min	Max	Median	Mean	Standard Deviation
age	35	77	58	56.850	8.055
resting blood pressure	100	200	130	133.992	18.557
serum cholesterol	149	409	257	259.037	49.829
maximum heart rate achieved	71	195	142	139.343	22.535
ST depression induced by exercise relative to rest	0	6.200	1.400	1.573	1.277
number of major vessels colored by fluoroscopy	0	3	1	1.164	1.027

Tab. 8. Shapiro-Wilk test for class 1

Variable	Shapiro-Wilk test (p-value)
age	0.004
resting blood pressure	6.493*10-5
serum cholesterol	0.096
maximum heart rate achieved	0.271
ST depression induced by exercise relative to rest	4.871*10-6
number of major vessels colored by fluoroscopy	2.932*10-10

To calculate the likelihood ratios, we used the maximum heart rate achieved as an instance to describe the concept of likelihood ratios. These curves are drawn by Python based on the mean and variance of maximum heart rate achieved for class 1 and class 0, separately. In Tables 5 and 7, these statistical parameters have been presented. To calculate the likelihood ratio for say, the maximum heart rate achieved of 120, two hypotheses have been considered; i) the patient's test is positive, versus ii) the patient's test is negative. By putting 120 in the logit function of maximum heart rate achieved, the dependent variable of the logistic response function is obtained as 2.537, and accordingly, the probability of class 1 or having heart disease for a person with a maximum heart rate of 120 is 0.7172. Analyzing the probability densities in Figure 4.b, reveals that the probability of class 1 to the probability of class 0 is greater than 1.

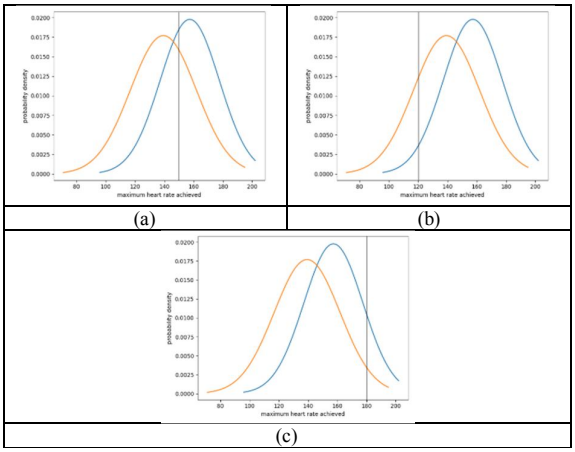


Fig. 4. Probability density plots of maximum heart rate achieved, the red curve for having heart disease and blue curve for not having heart disease a) maximum heart rate of 150, b) maximum heart rate of 120, c) maximum heart rate of 180.

Hence, the likelihood ratio analysis verifies the results of scanning the logit model and it is concluded that the patient's test is positive or belongs to class 1. Considering the maximum heart rate achieved of 180 and putting it in the logit model, we received the odds ratio of 0.2925, which indicates that the probability of having heart disease given the input of 180 is 29.25% and hence, the patient's test is negative or belongs to class 0. It was also supported by analyzing Figure 4.c, which proves that the probability of having heart disease to the probability of not having heart disease is less than 1 and consequently, the patient's test is negative. As extra evidence, analyzing the point where the likelihood ratios are equal, demonstrates that a maximum heart rate between 140 and 160 would give us equal probabilities for both classes. By assuming equal probabilities in the logit model, we also reached a maximum heart rate achieved of about 146, which was a satisfactory result, consistent with the interpretation of the likelihood ratios. It is notable that in this special case, the standard deviations are so high, which makes the probability densities so low; but the number of our samples in the dataset is above 100 for each class and thus, this problem cannot affect the quality of our interpretations.

4. Conclusions

This paper presents the estimation of having heart disease by logistic regression. According to AIC scores, it is concluded that normal thalassemia has the highest impact on heart disease prediction, followed by asymptomatic chest pain and the number of major vessels colored by fluoroscopy. These results were confirmed by comparing the AUCs which resulted in the same predictors, as well. A combined regression model has also been proposed that can predict heart disease with a sensitivity of 84.21%, a specificity of 90.38%, and a total accuracy of 87.77%. Besides, using the normal probability densities, the likelihood ratios have been calculated and used as a classifier. The likelihood ratio classifier proposed in this paper resulted in a similar performance to the logit model in the prediction of heart disease. However, more studies are needed to prove that their accuracy is as satisfactory as a logistic regression model.

Acknowledgment

The dataset used in this paper is the "Heart Disease Dataset"
it can be find in UCI Machine Learning Repository site:
https://github.com/mohammadaiai/Prediction-of-Heart-Disease/blob/main/Dataset/HeartDataset.csv

References

[1] "Cardiovascular Diseases." Accessed May 28, (2021). https://www.who.int/westernpacific/health topics/cardiovascular-diseases.

[2] "Cardiovascular Diseases: Avoiding Heart Attacks and Strokes." Accessed May 28, 2021. https://www.who.int/news-room/q-a-detail/cardiovascular-diseases-avoiding-heart-attacks-and-strokes.

[3] M. Ruiz Canela, A. Hruby, C. B. Clish, L. Liang, M. A. Martínez González, F. B. Hu, "Comprehensive Metabolomic Profiling and Incident Cardiovascular Disease: A Systematic Review." Journal of the American Heart Association, Vol. 6, No. 10 (2017).

[4] D. W. T. Jackson, A. D. Short, "Introduction to Beach Morphodynamics." Sandy Beach Morphodynamics, (2020).

[5] J. S. Dramsch, "70 Years of Machine Learning in Geoscience in Review." Advances in Geophysics, Vol. 61, (2020), pp. 1-55.

[6] D. Mpanya, C. Turgay, E. Klug, H. Ntsinjana, "Machine Learning and Statistical Methods for Predicting Mortality in Heart Failure." Heart Failure Reviews, Vol. 26, No. 3, (2021), pp. 545-552.

[7] M. Aazam, S. Zeadally, E. F. Flushing, "Task Offloading in Edge Computing for Machine Learning-Based Smart Healthcare." Computer Networks, Vol. 191, (2021): p. 108019.

[8] R. Ahmed, F. Nasiri, T. Zayed, "A Novel Neutrosophic-Based Machine Learning Approach for Maintenance Prioritization in Healthcare Facilities." Journal of Building Engineering, Vol. 42, (2021), p. 102480.

[9] C. H. Hsu, X. Chen, W. Lin, C. Jiang, Y. Zhang, Z. Hao, Y. C. Chung, "Effective

Multiple Cancer Disease Diagnosis Frameworks for Improved Healthcare Using Machine Learning." Measurement, Vol. 175, (2021), p. 109145.

[10] J. Ebinger, M. Wells, D. Ouyang, T. Davis, N. Kaufman, S. Cheng, S. Chugh, "A Machine Learning Algorithm Predicts Duration of Hospitalization in COVID-19 Patients." Intelligence-Based Medicine, Vol. 5, (2021), p. 100035.

[11] P. M. Naidu, C. Rajendra, "Detection of Health Care using Data Mining Concepts Through Web." International Journal of Advanced Research in Computer Engineering and Technology, Vol. 1, No. 4, (2012), pp. 45-50.

[12] G. Antonogeorgos, D. B. Panagiotakos, K. N. Priftis, A. Tzonou, "Logistic Regression and Linear Discriminant Analyses in Evaluating Factors Associated with Asthma Prevalence among 10- to 12-Years-Old Children: Divergence and Similarity of the Two Statistical Methods." International Journal of Pediatrics, Vol. 2009, (2009), pp. 1-6.

[13] K. H. Miao, J. H. Miao, G. J. Miao, "Diagnosing Coronary Heart Disease Using Ensemble Machine Learning." International Journal of Advanced Computer Science and Applications, Vol. 7, No. 10, (2016), pp. 30-39.

[14] L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour, S. A. C. Bukhari, "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure." IEEE Access, Vol. 7, (2019), pp. 54007-54014.

[15] M. S. Amin, Y. K. Chiam, K. D. Varathan, "Identification of Significant Features and Data Mining Techniques in Predicting Heart Disease." Telematics and Informatics, Vol. 36, (2019), pp. 82-93.

[16] R. H. Riffenburgh, D. L. Gillen, "Logistic Regression for Binary Outcomes." in Statistics in Medicine, (2020), pp. 437-457.

[17] R. Verma, K. Krishan, D. Rani, A. Kumar, V. Sharma, R. Shrestha, T. Kanchan, "Estimation of Sex in Forensic Examinations Using Logistic Regression and Likelihood Ratios." Forensic Science International: Reports, Vol. 2, (2020), p. 100118.

[18] M. A. Bidmos, A. A. Adebesin, P. Mazengenya, O. I. Olateju, O. Adegboye, "Estimation of Sex from Metatarsals Using Discriminant Function and Logistic Regression Analyses." Australian Journal of Forensic Sciences, (2020).

[19] E. A. DiGangi, M. K. Moore, Research Methods in Human Skeletal Biology. Oxford; Waltham, MA: Academic Press, (2013).

[20] R. H. Riffenburgh, D. L. Gillen, "Tests on Categorical Data." in Statistics in Medicine, (2006), pp. 241-79.

[21] David. Smith, FRCS General Surgery: 500 SBAs and EMIs. Jaypee Brothers Medical Publishers (P) Ltd., (2013).

[22] R. Magar, R. Memane, S. Raut, "Heart Disease Prediction Using Machine Learning." International Journal of Emerging Technologies and Innovative Research, Vol. 7, (2020), pp. 2081-2085.

[23] D. Shah, S. Patel and S. K. Bharti, "Heart Disease Prediction Using Machine Learning Techniques." SN Computer Science, Vol. 1, No. 345, (2020).

[24] A. Pandita, S. Vashisht, A. Tyagi, S. Yadav, "Prediction of Heart Disease Using Machine Learning Techniques." International Journal for Research in Applied Science and Engineering Technology, Vol. 9, (2021), pp. 2422-2429.

[25] A. Akella, S. Akella, "Machine learning algorithms for predicting coronary artery disease: efforts toward an open-source

solution." Future Science OA, Vol. 7, No. 6, FSO698, (2021).

[26] M. M. Ali, B. K. Paul, K. Ahmed, F.M. Bui, J. M. W. Quinn, M. A. Moni, "Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis and Comparison." Computers in Biology and Medicine, Vol. 136, (2021), p. 104672.

[27] A. Subasi, Practical Machine Learning for Data Analysis Using Python. 1st ed. San Diego: Elsevier Inc, (2020).

[28] X. Song, X. Liu, F. Liu, C. Wang, "Comparison of Machine Learning and Logistic Regression Models in Predicting Acute Kidney Injury: A Systematic Review and Meta-Analysis." International Journal of Medical Informatics, Vol. 151, (2021), p. 104484.

[29] W. Książek, M. Gandor, P. Pławiak, "Comparison of Various Approaches to Combine Logistic Regression with Genetic Algorithms in Survival Prediction of Hepatocellular Carcinoma." Computers in Biology and Medicine, Vol. 134, (2021).

[30] F. Jiang, Z. Guan, Z. Li, X. Wang, "A Method of Predicting Visual Detectability of Low-Velocity Impact Damage in Composite Structures Based on Logistic Regression Model." Chinese Journal of Aeronautics, Vol. 34, No. 1, (2021), pp. 296-308.

[31] R. Priyadarshini, N. Malarvizhi, P. Karthikeyan, "Estimation of Trust Using Logistic Regression in Fog Computing." Microprocessors and Microsystems, (2021), p. 104026.

[32] R. Xiao, X. Cui, H. Qiao, X. Zheng, Y. Zhang, C. Zhang, X. Liu, "Early Diagnosis Model of Alzheimer's Disease Based on Sparse Logistic Regression with the Generalized Elastic Net." Biomedical Signal Processing and Control, Vol. 66, (2021), p. 102362.

[33] B. Bera, S. Saha, S. Bhattacharjee, "Forest Cover Dynamics (1998 to 2019) and Prediction of Deforestation Probability Using Binary Logistic Regression (BLR) Model of Silabati Watershed, India." Trees, Forests and People, Vol. 2, (2020), p. 100034.

[34] K. Ghazvini, M. Yousefi, F. Firoozeh, S. Mansouri, "Predictors of Tuberculosis: Application of a Logistic Regression Model." Gene Reports, Vol. 17, (2019), p. 100527.

[35] W. A. van Eeden, C. Luo, A. M. van Hemert, I. V. E. Carlier, B. W. Penninx, K. J. Wardenaar, H. Hoos, E. J. Giltay, "Predicting the 9-Year Course of Mood and Anxiety Disorders with Automated Machine Learning: A Comparison between Auto-Sklearn, Naïve Bayes Classifier, and Traditional Logistic Regression." Psychiatry Research, Vol. 299, (2021).

[36] A. Castaño, F. Fernández-Navarro, P. A. Gutiérrez, C. Hervás-Martínez, "Permanent Disability Classification by Combining Evolutionary Generalized Radial Basis Function and Logistic Regression Methods." Expert Systems with Applications, Vol. 39, No. 9, (2012), pp. 8350-8355.

[37] P. M. Kuhnert, K. A. Do, R. McClure, "Combining Non-Parametric Models with Logistic Regression: An Application to Motor Vehicle Injury Data." Computational Statistics & Data Analysis, Vol. 34, No. 3, (2000), pp. 371-386.

[38] J. I. E. Hoffman, "Logistic Regression." in Biostatistics for Medical and Biomedical Practitioners, (2015), pp. 601-611.

[39] M. L. McHugh, "The Odds Ratio: Calculation, Usage, and Interpretation." Biochemia Medica, Vol. 19, No. 2, (2009), pp. 120-126.

[40] L. Liu, "Biostatistical Basis of Inference in Heart Failure Study." in Heart Failure: Epidemiology and Research Methods, (2018), pp. 43-82.

[41] "UCI Machine Learning Repository: Data Set." Accessed May 28, (2021). https://archive.ics.uci.edu/ml/datasets/heart+disease.

[42] R. Chou, B. Arora, T. Dana, R. Fu, M. Walker, L. Humphrey, "Screening Asymptomatic Adults with Resting or Exercise Electrocardiography: A Review of the Evidence for the U.S. Preventive Services Task Force." Annals of Internal Medicine, Vol. 155, No. 6, (2011), p. 375.

[43] A. Kulkarni, D. Chong, F. A. Batarseh, "Foundations of Data Imbalance and Solutions for a Data Democracy." In Data Democracy, (2020), pp. 83-106.

[44] D. Boyce, "Evaluation of Medical Laboratory Tests." In Orthopaedic Physical Therapy Secrets, Vol. 336, No. 7644, (2008), pp. 125-134.

[45] E. A. Mohammed, C. Naugler, B. H. Far, "Emerging Business Intelligence Framework for a Clinical Laboratory Through Big Data Analytics." in Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology, pp. (2015), pp. 577-602.

[46] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer Series in Statistics. New York, NY: Springer, (2009).

[47] J. J. Dziak, D. L. Coffman, S. T. Lanza, R. Li, L. S. Jermiin, "Sensitivity and Specificity of Information Criteria." Briefings in Bioinformatics, Vol. 21, No. 2, (2020), pp. 553-565.

[48] D. M. Diez, M. Çetinkaya-Rundel, C. D. Barr, Openintro Statistics. 4th ed. OpenIntro, Inc, (2015).

[49] Y. Zhang, L. Diao, L. Ma, "Logistic Regression Models in Predicting Heart Disease." Journal of Physics: Conference Series, Vol. 1769, (2021), p. 012024.

[50] R. Prasad, P. Anjali, S. Adil, N. Deepa, "Heart Disease Prediction Using Logistic Regression Algorithm Using Machine Learning." International Journal of Engineering and Advanced Technology, Vol. 8, No. 3S, (2019), pp. 659-662.

[51] Di. Khanna, R. Sahu, V. Baths, B. Deshpande, "Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease." International Journal of Machine Learning and Computing, Vol. 5, No. 5, (2015), pp. 414-419.

[52] S. Kodati, R. Vivekanandam, "Analysis of Heart Disease using in Data Mining Tools Orange and Weka." Global journal of computer science and technology, Vol. 18, No. 1, (2018), pp. 17-21.

[53] C. B. C. Latha, S. C. Jeeva, "Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques." Informatics in Medicine Unlocked, Vol. 16, (2019), p. 100203.

[54] R. Jain, M. K. Camarillo, W. T. Stringfellow, Drinking Water Security for Engineers, Planners, and Managers. Integrated Water Security Series. Amsterdam: Elsevier, Butterworth-Heinemann, (2014).

[55] T. Fawcett, "An Introduction to ROC Analysis." Pattern Recognition Letters, Vol. 27, No. 8, (2006), pp. 861-874.

[56] J. Boita, R. E. van Engen, A. Mackenzie, A. Tingberg, H. Bosmans, A. Bolejko, S. Zackrisson, et al., "Validation of a Candidate Instrument to Assess Image Quality in Digital Mammography Using ROC Analysis." European Journal of Radiology, Vol. 139, (2021), p. 109686.

[57] J. Y. Verbakel, E. W. Steyerberg, H. Uno, B. D. Cock, L. Wynants, G. S. Collins, B. N. Calster, "Erratum to 'ROC Curves for Clinical Prediction Models Part 1. ROC Plots Showed No Added Value above the AUC When Evaluating the Performance of Clinical Prediction Models' [J Clin Epidemiol. 126C (2020): 207-16]."Journal

of Clinical Epidemiology, Vol. 130, (2021), pp. 171-173.

[58] Institute of Electrical and Electronics Engineers, ed. The 2010 International Joint Conference on Neural Networks: (IJCNN 2010); Barcelona, Spain, 18 - 23 July 2010; [Associated with the 2010 IEEE World Congress on Computational Intelligence (IEEE WCCI 2010)]. Piscataway, NJ: IEEE, (2010).

[59] C. Sammut, G. I. Webb, Encyclopedia of Machine Learning and Data Mining. Second edition. Springer Reference. New York, NY: Springer, (2017).

[60] S. Halligan, D. G. Altman, S. Mallett, "Disadvantages of Using the Area under the Receiver Operating Characteristic Curve to Assess Imaging Tests: A Discussion and Proposal for an Alternative Approach." European Radiology, Vol. 25, No. 4, (2015), pp. 932-939.

[61] B. W. Yap, C. H. Sim, "Comparisons of Various Types of Normality Tests." Journal of Statistical Computation and Simulation, Vol. 81, No. 12, (2011), pp. 2141-2155.

[62] H. C. Thode, Testing for Normality. CRC Press, (2002).

[63] D. J. Steinskog, D. B. Tjøstheim, N. G. Kvamstø, "A Cautionary Note on the Use of the Kolmogorov–Smirnov Test for Normality." Monthly Weather Review, Vol. 135, No. 3, (2007), pp. 1151-1157.

Follow this article at the following site:

Mohammad Yaseliani & Majid Khedmati. Prediction of Heart Diseases Using Logistic Regression and Likelihood Ratios. IJIEPR 2023; 34 (1) :1-15
URL: http://ijiepr.iust.ac.ir/article-1-1590-en.html