Original Article

Accuracy Analysis of Heart Disease Prediction using Logistic Regression in Comparison with the Linear Regression Algorithm

B. Manoj Kumar[1], Uma Priyadarsini[2]

[1]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, pin: 602105
[2]Project Guide, Corresponding Author, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, pin: 602105

Abstract

Aim: The main objective of this research article is to employ the detection of heart disease by using the Logistic Regression (LR) classifier in comparison with the Linear regression (LR) model. **Materials & Methods:** The dataset used in this paper was collected from the UCI machine learning repository database. The sample size for the detection of heart disease was sample 60 (Group 1=30 and Group 2 =30) and calculation was performed utilizing G-power 0.8 with alpha and beta qualities of 0.05, 0.2 with a confidence interval of 95%. The detection of heart disease is performed by the Logistic Regression (LR) classifier with a number of samples (N=30) and Linear regression (LR) model with a number of samples (N=30). **Results:** The Logistic Regression (LR) classifier has a 88.68 percent higher accuracy rate when compared to the accuracy rate of the Linear regression (LR) model is 78.56 percent. The study has a significance value of p=0.025. **Conclusion:** Logistic Regression (LR) classifier provides better outcomes in accuracy rate when compared to Linear regression (LR) model for detection of heart disease.

Keywords: Image processing, Novel Logistic Regression (LR) classifier, Linear Regression model, accuracy rate, Heart Disease, Segmentation.

https://doi.org/10.47750/pnr.2022.13.S04.199

INTRODUCTION

Heart problems are a prominent risk factor for mortality, and they have piqued the scientific community's interest (Finegold, Asaria, and Francis 2013). There are many different types of heart disorders, such as coronary artery disease, congestive heart failure, and abnormal heart rhythms, to name a few. There are a large number of people who suffer from heart disease. Before they attack patients, heart disorders may or may not exhibit symptoms (Glick and Greenberg 2005). As a result, we must be able to forecast whether or not a person will be afflicted by heart disease. Healthcare experts that operate in the field of cardiovascular problems have their own limitations and are unable to predict the possibility of cardiac illness with great accuracy. This research uses the Logistic Regression (LR) machine learning model to optimize heart disease classification accuracy using a health care dataset that identifies individuals as having or not having heart disease based on information in their medical records (Khanna et al. 2015). Image processing has numerous applications in biomedical, forensics, remote sensing, farming, food control, cars, and communications (Párraga Álava 2015).

There are numerous studies in the literature that use machine learning algorithms to diagnose vehicle heart problems. IEEE Xplore distributed 97 examination papers, and Google Scholar tracked down 133 articles. In (Reddy and Khare 2017), an effective hybrid model is created to more effectively diagnose cardiac disease. A model is constructed in this study by merging the effective output characteristics of Decision trees, Naive Bayes, Neural Networks, and Support Vector Machines, and it exhibits remarkable precision and effectiveness in the treatment of heart disorders. Researchers created a technique to analyze the likelihood of coronary disease in (Mythili et al. 2013). The designed automated tool is easy to use and accepts basic patient information as input. A decision support system for cardiac disease classification was provided by Mrudula Gudadhe et al. (Rajakumar and George 2013). The two main methods employed in this system were support vector machine (SVM) and artificial neural network (ANN). Sairabi H. Mujawar et al. (Mujawar and Devale 2015) used modified k-means

and Naive Bayes to predict cardiac disease. Using WEKA, Jaymin Patel et al. (Patel, TejalUpadhyay, and Patel 2015) compared different Decision tree classification algorithms for better performance in heart disease detection. They compared the J48 algorithm, logistic model tree, and random forest algorithms. Shadab Adam Pattekari et al. (Pattekari and Parveen 2012) used Naive Bayes, Decision trees, and neural networks to create a prototype of a heart disease prediction system. It's done through a web application. K. Polaraju et al. (Polaraju, Prasad, and Others 2017) suggested a multiple regression model for predicting heart disease risk, demonstrating that multiple linear regression is acceptable for predicting heart disease risk. Soleimani and Neshati (Soleimani and Neshati 2015) used 711 data from patients with symptoms like chest infections, backache, cold shivers, breathlessness, nausea, and vomiting to create three logistic regression models with 28 factors to predict heart disease risk.
Our institution is passionate about high quality evidence based research and has excelled in various fields (Devarajan et al., 2021; Dhanraj & Rajeshkumar, 2021; Kamath et al., 2020; Nandhini et al., 2020; Parakh et al., 2020; Perumal et al., 2021; Pham et al., 2021; Sathiyamoorthi et al., 2021; Tesfaye Jule et al., 2021; Uganya et al., 2021). The current technique's main disadvantage is that it can exhibit poor run speed if the training set is big; accuracy is dependent on data quality. In comparison to the Linear Regression method, this research provides an automation strategy for heart disease identification using the innovative Logistic Regression (LR) method. The goal of the LR algorithm is to improve the accuracy of heart disease diagnosis. The proposed cardiovascular disease categorization outperforms the existing linear technique according to the performance analysis.

MATERIALS AND METHODS

Dataset Description
This work was carried out in the Image Processing Laboratory, Department of Computer Science and Engineering, Saveetha School of Engineering. In this study, the dataset was collected from the UCI machine learning repository. 60 sample images were taken. The database is divided by the amount of 75% training and 25% testing. Two sets are taken and 30 data samples for each set, total number of samples considered are 60. Group 1 was a Linear regression algorithm and Group 2 was a novel Logistic Regression (LR) algorithm. The output is obtained by using Matlab software for the prediction of heart disease. The calculation is performed utilizing G-power 0.8 with alpha and beta qualities 0.05, 0.2 with a confidence interval at 95% (Polaraju, Prasad, and Others 2017).

Linear Regression
Linear regression is a machine training algorithm that establishes a direct relationship between a dependent variable and at least one independent variable. Simple linear regression is known as linear regression with one independent variable, while multiple linear regressions is known as linear regression with more than one free element. Logical regression changes its result using the logistic sigmoid function to yield a probability value that may then be translated to two or more discrete classes, whereas linear regression uses continuous numeric output. The model is built with 75% of the data and tested with the remaining data.

$$Y = X0 + X1a \quad (1)$$

$$Error = \sum (actual\ output - predicted\ output) \quad (2)$$

$$X0 = \underline{m} - X1\underline{n} \quad (3)$$

When X0 is determined, then, at that point X1 is determined utilizing the accompanying formula.

$$X1 = \sum_{i=1}^{n} (x_i - \underline{x})(y_i - \underline{y}) \quad (4)$$

Pseudocode
Input: Heart disease_Input Features
Assign Training and Testing Dataset of Heart disease
Output: Prediction of Heart disease
Step 1: Import the Heart Disease Training and Testing Dataset
Step 2: Prepare the data that has been imported.
Step 3: Make a decision on a categorization algorithm.
Step 4: Determine the linear regression algorithm to use.
Step 5: Find the Starting Vector
Step 8: Calculate bias-connected estimates for the first and second moment estimations.
Step 9: Results of the classification.

Logistic Regression
Logistic regression is one of the machine learning classification techniques for assessing a dataset with one or more relationships between the independent variable (IVs) that determine a result and a categorical dependent variable (DV). The Logistic Regression (LR) Classifier uses a set of indicators to predict the outcome of a dependent variable. When the dependent factor is all out and dichotomous, and the autonomous variables are all

out, constant, or blended, it is suitable. In calculated relapse, the dependent variable has a value of 1 (one) for the likelihood of achieving an occasion and 0 (zero) for the likelihood of failing to achieve an occasion. Demonstrating with logistic regression can be done in two ways:
1. Stepwise Regression - A strategy in which autonomous components are entered one by one and the model's importance is assessed.
2. Stepwise Regression in Reverse – A approach in which all autonomous elements are used and irrelevant factors are removed step by step in an ongoing interaction to ensure that the model adequately fits the data.

Pseudocode

Input: heart disease_Input Features
Assign training and testing dataset for heart disease
Output: Classification of heart disease as a result of the output
Step 1: Logistic Regression (Input Matrix M, Input Features I=[1.....n]) is the first step.
Step 2: Matrix M is required for Input features I.
Step 3: Select Regressor as a parameter.
Step 4: Set the Class Label Vector C= [1… N]
Step 5: Perform while (conditioning)
Step 6: create a binary vector B
Step 7: If it belongs to the target class C, and B=1, it belongs to the class label C.
Step 8: If B does not correspond to the class label C, set it to 0.
Step 9: Use Logistic Regression to discover the parameter Regressor for the Matrix M.
Step ten: Finish while
Step 11: Retain the results of the classification.

Statistical Analysis

Matlab software is used to generate the results (Gilat 2004). A monitor with a resolution of 1024x768 pixels was required to train these datasets (10th gen, i5, 12GB RAM, 500 GB HDD). The software programme IBM SPSS is employed in this study for statistical analysis (Yockey 2017). The independent sample t test was used to determine the mean, standard deviation, and standard error mean statistical significance between the groups, and then the two groups were compared using SPSS software to obtain accurate values for the two different s, which were then used with the graph to calculate the significant value with maximum accuracy (88.68 percent), mean value (97 percent), and standard deviation value (1.50124). Accuracy is a dependent variable, while SVM and LR are independent variables.

RESULTS

The accuracy rate of the Logistic Regression (LR) classifier is compared to the linear classifier in Figure 1. The LR classifier has a higher accuracy rate of 88.68 when compared to the linear classifier, which has 78.56 respectively. The LR classifier is significantly different from the linear classifier ($p<0.05$ independent sample test). Logistic and linear accuracy rates are plotted on the X-axis. Y-axis: Mean accuracy rate for keyword identification, ± 1 SD with 95 percent confidence interval.

Table 1 presents the evaluation metrics for the comparison of the LR classifier with the linear classifier. The LR classifier has a 88.68 accuracy rate, whereas the linear classifier has 78.56, respectively. In all parameters, the LR classifier outperforms the linear in the classification of heart disease, with a higher accuracy rate.

Table 2 displays the statistical computations for LR, linear classifier, such as mean, standard deviation, and standard error mean. In the t-test, the accuracy rate parameter is used. The LR classifier has a mean accuracy rate of 88.68, while the linear classifier has 78.56, respectively. The LR classifier has a standard deviation of 1.50124, while linear has a standard deviation of 2.68312 respectively. The LR classifier has a Standard Error Mean of 0.58932, while linear has a Standard Error Mean of 1.47823 respectively.

Table 3 shows the statistical computations for independent samples of LR compared to the linear regression classifier. The accuracy rate has a significance level of 0.025. The LR classifier is compared to a linear classifier using an Independent samples T-test with a confidence interval of 95 percent and a threshold of significance of 0.37283. The significance level is 0.001, the significance level is two-tailed, the mean difference, the standard error difference, and the lower and upper interval difference are all included in this independent sample test.

DISCUSSION

This study used two alternative algorithms and prediction models to evaluate heart disease classification. In terms of accuracy, the investigations show that Logistic Regression (LR) (88.68 percent) outperforms Linear Regression (78.56 percent).

Many papers from the last three to five years are less accurate in predicting cardiac disease than the demands of today. Sharma Purshottam et al (Sharma and Saxena 2017) released an article in 2015 called "Efficient detection method for heart disease using decision tree." They employed D-Tree classifiers as their method and achieved an accuracy of 86.3 percent. Similarly, Sairabi H Mujawar et al. published 'Prediction of cardiovascular disease using modified K-means and nave bayes'. This research was completed in 2015 and published in a peer-reviewed journal. Their detection accuracy for cardiac illness was 85 percent, and their undetection accuracy was 82 percent (Párraga Álava 2015). This demonstrates that the accuracy percentage is dependent on the technique used. In (Kim 2014), in addition to the 13 commonly used qualities such as gender, heart rate, lipid, and so on, two more attributes such as obesity and smoking are included for the prediction of coronary illnesses. The accuracy of Neural Networks, D-Tree, and NB is 93 percent, 92.62 percent, and 90.74 percent, respectively, according to this study. According to the results of this study, Neural Networks surpassed the other two methodologies in terms of heart disease prediction accuracy. A comparison of several data mining algorithms for the prediction of cardiac illnesses is undertaken in (Anggoro and Kurnia 2020). SVM, KNN and Logistic Regression were used to create and validate the models. Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighborhood (KNN) models had accuracy levels of 94.02 percent, 91.23 percent, and 85.05 percent, respectively.
The proposed study's fundamental flaw is that it is highly sensitive to interference and overfitting. If the number of data is smaller than the number of features, logistic regression should not be utilized; otherwise, overfitting may occur. In the future, an intelligent platform may be developed that may help a patient diagnosed with chronic disease choose the best treatment options. The paper's next focus will be on predicting cardiac illnesses utilizing advanced methodologies and algorithms that are less time consuming.

CONCLUSION

The proposed model includes the Logistic model and Linear regression model, in which the Logistic model has the highest accuracy values. The accuracy rate of the Logistic model is 88.68% higher compared with the Linear regression model that has an accuracy rate of 78.56% in the detection of heart disease with an improved accuracy rate.

DECLARATION

Conflicts of Interest

No conflict of interest in this manuscript

Authors Contributions

Author name BMK was involved in data collection, data analysis & manuscript writing. Author guide name UP was involved in conceptualization, data validation, and critical review of manuscripts.

Acknowledgment

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (Formerly known as Saveetha University) for successfully carrying out this work.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Saveetha University
2. Saveetha Institute of Medical And Technical Sciences
3. Saveetha School of Engineering

REFERENCES

1. Anggoro, Dimas Aryo, and Naqshauliza Devi Kurnia. 2020. “Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease.” *Aquatic Microbial Ecology: International Journal* 8 (5). https://www.academia.edu/download/63554423/ijeter3285202020200607-113398-1w5oaj3.pdf.
2. Finegold, Judith A., Perviz Asaria, and Darrel P. Francis. 2013. “Mortality from Ischaemic Heart Disease by Country, Region, and Age: Statistics from World Health Organisation and United Nations.” *International Journal of Cardiology* 168 (2): 934–45.
3. Gilat, Amos. 2004. *Matlab: An Introduction With Applications*. John Wiley & Sons.
4. Glick, Michael, and Barbara L. Greenberg. 2005. “The Potential Role of Dentists in Identifying Patients’ Risk of Experiencing Coronary Heart Disease Events.” *Journal of the American Dental Association* 136 (11): 1541–46.
5. Khanna, Divyansh, Rohan Sahu, Veeky Baths, and Bharat Deshpande. 2015. “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* 5 (5): 414.
6. Kim, Hae-Young. 2014. “Analysis of Variance (ANOVA) Comparing Means of More than Two Groups.” *Restorative Dentistry & Endodontics* 39 (1): 74–77.
7. Mujawar, Sairabi H., and P. R. Devale. 2015. “Prediction of Heart Disease Using Modified K-Means and by Using Naive Bayes.” *International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified*

Organization) 3 (10): 10265–73.
8. Mythili, T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu. 2013. "A Heart Disease Prediction Model Using SVM-Decision Trees-Logistic Regression (SDL)." *International Journal of Computer Applications in Technology* 68 (16). https://www.researchgate.net/profile/Mythili-Thirugnanam/publication/273261237_A_Heart_Disease_Prediction_Model_using_SVM-Decision_Trees-Logistic_Regression_SDL/links/5cd3c7bf92851c4eab8c563b/A-Heart-Disease-Prediction-Model-using-SVM-Decision-Trees-Logistic-Regression-SDL.pdf.
9. Párraga Álava, Jorge Antonio. 2015. "Computer Vision and Medical Image Processing: A Brief Survey of Application Areas." In *Argentine Symposium on Artificial Intelligence (ASAI 2015)-JAIIO 44 (Rosario, 2015)*. sedici.unlp.edu.ar. http://sedici.unlp.edu.ar/handle/10915/52114.
10. Patel, Jaymin, Dr TejalUpadhyay, and Samir Patel. 2015. "Heart Disease Prediction Using Machine Learning and Data Mining Technique." *Heart Disease* 7 (1): 129–37.
11. Pattekari, Shadab Adam, and Asma Parveen. 2012. "Prediction System for Heart Disease Using Naïve Bayes." *International Journal of Advanced Computer and Mathematical Sciences* 3 (3): 290–94.
12. Polaraju, K., D. Durga Prasad, and Others. 2017. "Prediction of Heart Disease Using Multiple Linear Regression Model." *International Journal of Engineering Development and Research* 5 (4): 2321–9939.
13. Rajakumar, B. R., and Aloysius George. 2013. "On Hybridizing Fuzzy Min Max Neural Network and Firefly Algorithm for Automated Heart Disease Diagnosis." In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 1–5.
14. Reddy, G. Thippa, and Neelu Khare. 2017. "An Efficient System for Heart Disease Prediction Using Hybrid OF CAT with Rule-Based Fuzzy Logic Model." *Journal of Circuits, Systems and Computers* 26 (04): 1750061.
15. Sharma, Purushottam, and Kanak Saxena. 2017. "Application of Fuzzy Logic and Genetic Algorithm in Heart Disease Risk Level Prediction." *International Journal of System Assurance Engineering and Management* 8 (2): 1109–25.
16. Soleimani, Paria, and Arezoo Neshati. 2015. "Applying the Regression Technique for Prediction of the Acute Heart Attack." *International Journal of Biomedical and Biological Engineering* 9 (11): 767–71.
17. Yockey, Ronald D. 2017. *SPSS® Demystified: A Simple Guide and Reference*. Routledge.

TABLES AND FIGURES

Table 1. The evaluation metrics of the LR classifier with the Linear classifier has been calculated. The LR classifier has a 88.68 accuracy rate, whereas the Linear classifier has 78.56, respectively. In all parameters, the LR classifier outperforms the Linear in the classification of heart disease, with a higher accuracy rate.

SI.No.	Test Size	ACCURACY RATE	
		Logistic regression classifier	Linear regression model
1	Test1	85.23	76.10
2	Test2	85.54	76.23
3	Test3	85.36	76.19
4	Test4	86.34	77.92
5	Test5	86.12	77.92
6	Test6	87.56	77.01
7	Test7	88.35	77.85
8	Test8	88.36	78.28
9	Test9	88.45	78.58
10	Test10	88.54	78.34

Average Test Results	88.68	78.56

Table 2. The statistical calculation such as Mean, standard deviation and standard error Mean for LR classifier and Linear regression model. The accuracy rate parameter used in the t-test. The mean accuracy rate of LR classifier is 88.68 and the Linear model is 78.56. The Standard Deviation of LR classifier is 1.50124 and Linear model is 2.68312. The Standard Error Mean of LR classifier is 0.58932 and Linear model is 1.47823.

Group		**N**	**Mean**	**Standard Deviation**	**Standard Error Mean**
Accuracy rate	Linear regression model	10	78.56	2.68312	1.47823
	Logistic regression (LR) model	10	88.68	1.50124	0.58932

Table 3. The statistical calculations for independent samples test between LR classifier and Linear regression model. The sig. for accuracy rate is 0.025. Independent samples T-test is applied for comparison of LR classifier and Linear regression model with the confidence interval as 95% and level of significance as 0.34097. This independent sample test consists of significance as 0.001, significance (2-tailed), Mean difference, standard error difference, and lower and upper interval difference.

Group		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval (Lower)	95% Confidence Interval (Upper)
Accuracy	Equal variances assumed	0.435	0.025	15.824	15	.001	12.84574	0.84563	12.68213	12.63218
	Equal variances not assumed			13.103	13.830	.001	12.00923	0.28129	12.21987	11.02463

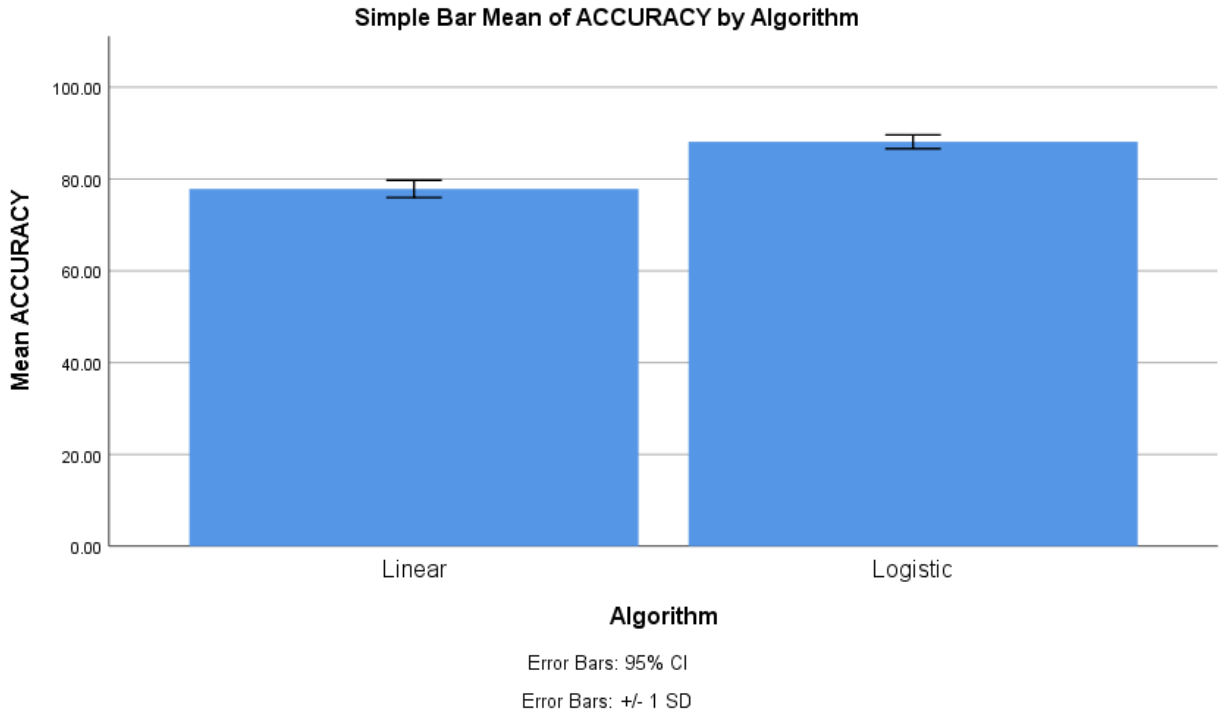

Fig. 1. Simple Bar graph for LR classifier accuracy rate is compared with Linear regression model. The LR classifier is higher in terms of accuracy rate 88.68 when compared with Linear regression model 78.56. Variable results with its standard deviation ranging from 100 lower to 150 higher in LR classifier where Linear regression model standard deviation ranging from 200 lower to 300 higher. There is a significant difference between LR classifier and Linear regression model ($p<0.05$ Independent sample test). X-axis: Linear regression model accuracy rate vs LR classifier Y-axis: Mean of accuracy rate, for identification of keywords ± 1 SD with 95 % CI.