Global Transitions Proceedings 3 (2022) 127–130

Contents lists available at ScienceDirect

Global Transitions Proceedings

journal homepage: http://www.keaipublishing.com/en/journals/global-transitions-proceedings/

Logistic regression technique for prediction of cardiovascular disease

Ambrish G*, Bharathi Ganesh, Anitha Ganesh, Chetana Srinivas, Dhanraj, Kiran Mensinkal

East West Institute of Technology, Anjana Nagar, Bangalore 560091, India

ARTICLE INFO

Keywords:
Cardiovascular disease
Feature selection
Logistic regression
Machine learning
UCI dataset

ABSTRACT

One of the most life-threatening disease is cardiovascular disease. Its high mortality rate contributes to nearly 17 million deaths all over the world. Early diagnosis helps to treat the disease in timely manner to prevent mortality. There are several machine and deep learning techniques available to classify the presence and absence of the disease. In this research, Logistic Regression (LR) techniques is applied to UCI dataset to classify the cardiac disease. To improve the performance of the model, pre-processing of data by Cleaning the dataset, finding the missing values are done and features selection were performed by correlation with the target value for all the feature. The highly positive correlated features were selected. Then classification is performed by dividing the dataset into training. testing in the ratio of 90:10, 80:20, 70:30, 40:60 and 50:50. The splitting ratio of 90:10 gives best accuracy as listed below. The LR model obtained 87.10% accuracy.

1. Introduction

Today the greatest challenge to medical industry to provide higher level facility to health infrastructure to diagnose the disease in the initial day and give timely treatment to improve the quality of life through quality of service. Around 31% of mortality occurs world due to cardiac disease [1,2]. The developing and under developing countries lacks in infrastructure and technologies, infrastructure and doctors to predict the disease in early stage to avoid complications reduce mortality. The growth of Information and telecommunication technology has benefited from rich to poor patients by providing real time information to the patients with lower cost of diagnose and monitoring the patients' health. This has increase in detail health records of the patients dramatically. The vast medical records are available to the research. The medical industry faces enormous challenges in using the huge medical data. The vast amount of data is transformed to obtain valuable and accurate information speedily by machine. Thus, machine learning is the important area. The highly useful machine learning models used to discover the hidden pattern and correlation among features in the dataset [3,4]. The medical dataset is inconsistent and redundant, appropriate preprocessing is pivot step [5]. Various researcher has included risk of different feature the most prevalent are 14 features. Since the feature selection become an important part of the study, based on the feature selection the model increases or decrease the prediction accuracy [6]. The cardiac disease can be predicted with the help of machine learning with greater accuracy will help healthcare to diagnose and treat patient in early stage supporting many patients to diagnose disease in short period of time. Thus, saving millions of lives.

In this study, Logistic Regression classifier model are applied. Results are compared with existing studies.

2. Problem statement

Presently, the major challenge of the medical industry is to predict the cardio vascular disease with less expensive and more reliable method to avoid the compounding effect of the disease in low income or developing countries. The early detection not only reduce the cost but also improves the quality of life.

3. Research aim and scope of the research

The aim of this research is to develop an efficient way to predict the presence of the cardiovascular disease. The steps as mentioned below.

(1) The UCI dataset is used to predict the disease.
(2) The Features are selected based on high positive correlation values with the target and used random order of data (without sorting).
(3) The performance of the model is evaluated by Five different training and testing ratio of dataset.
(4) To check the behavior of the model with low to high training and testing data.

* Corresponding author.
E-mail address: ambrishmgcse@gmail.com (A. G).

https://doi.org/10.1016/j.gltp.2022.04.008

Available online 3 April 2022

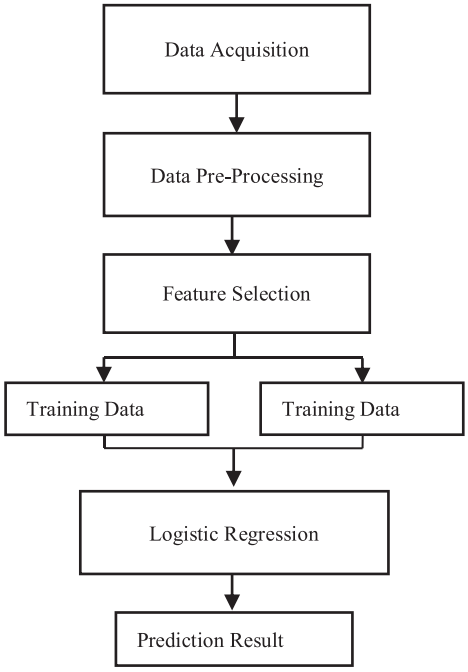

Fig. 1. Flow chart of logistic regression model.

Table 1
Feature selection using correlation.

Features	Correlation
Exang	0.436757
Cp	0.433798
Oldpeak	0.430696
Thalach	0.421741
Ca	0.391724
Slope	0.344029

Table 2
Split percentage of training and test set.

Serial Number	Training Set	Test Set
1	50%	50%
2	60%	40%
3	70%	30%
4	80%	20%
5	90%	10%

4. Related work

Firda Anindita Latifah et.al., proposed comparative study of machine learning model namely, logistical regression and random forest for classification of heart disease. The research done on Framingham dataset with 3656 records and training to testing ratio of 70:30. The accuracy of 85.04% was achieved by the model [7].

Zameer Khan et.al., proposed empirical study of several ML algorithms namely, logistic regression, KNN Classifier, RF, SVM, Decision Tree, Gauss Naïve Bayesian used for classification of cardiovascular disease. The research done on UCI Cleveland dataset. The logistic regression achieved accuracy of 85.71 [8].

Thanuja Nishadi A S et al., proposed logistic regression model for classification of heart disease on Framingham dataset with 4238 rrecords. The logistic regression achieved the accuracy of 86.66% [9].

Montu Saw et al., proposed logistic regression model to classify the cardiac disease. The research uses Framingham datasets and logistics regression achieved accuracy of 87.02% [10]. Detrano et.al., proposed logistic regression model and obtained 77.0% accuracy [11].

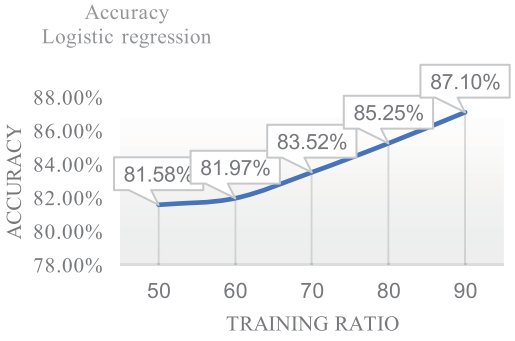

Fig. 2. Logistic regression accuracy for training and testing ratio.

Table 3
Accuracy result of logistic regression classifier.

Training and Testing				
90:10	80:20	70:30	60:40	50:50
87.10%	85.25	83.52%	81.97%	81.58%

Table 4
Classification report of logistic regression classifier.

Confusion Matrix			Classification Report			
	Pos	Neg	Precision	Recall	F1-Score	Accuracy
Pos	15	2	0.857	0.857	0.857	87.10
Neg	2	12				

Saba Bashir et al., proposed several ML algorithms namely, logistic regression, decision tree, logistics regression support vector machine etc. The research uses UCI dataset and the logistic regression achieved accuracy of 82.56% and logistic regression support vector machine achieved accuracy of 84.85% [12].

Kannan, R et al., proposed various ML algorithms namely, LR, SVM, RF Stochastic Gradient Boosting using UCI Cleveland dataset. The logistic regression outperformed other models with best accuracy of 86.51% [13].

Ganesan M et al., proposed four different ML algorithm, SVM, J48, Logistic Regression, Multilayer perceptron using UCI dataset to classify cardiac disease and obtained best accuracy. The logistic regression achieved an accuracy of 83.70% [14].

Dinesh Kumar G et al., proposed five machine learning algorithm using UCI dataset to classify the cardiac disease. The Logistic Regression achieved overall accuracy of 86.51% [15].

5. Proposed work

In the proposed system, the analysis of the cardiac disease UCI dataset is carried out using suitable data acquisition, preprocessing by cleaning the data, then using selects all the features which have high correlation with the target function. Then logistics regression model was trained and tested for predicting the cardiac disease is present or not. The Fig. 1 shows the workflow to build logistic regression cardiac disease classification model.

The machine learning model was implemented in python 3.7 environment and used ML libraries sklearn, pandas and matplotlib. Jupyter Notebook to run the code.

5.1. Data acquisition

The cardiac disease dataset obtained from the UCI ML repository. It contains 13 features and 303 records.

Table 5
Comparative result of logistic regression classifier.

	Year	Author	Tool/ Techniques	Logistic Regression
UCI	2019	[8]	Rapid Miner (Logistic Regression)	82.56%
DATASET	2019	[8]	Rapid Miner (Logistic Regression Support Vector Machine)	84.85%
	2020	[1]	Python	85.04%
	2020	[2]	Python	85.71%
	2021	This Paper	Python sklearn (90:10)	87.10%

5.2. Data pre-processing

Cardiovascular disease UCI dataset is first loaded and then data cleaning and finding missing values was performed on all records. The dataset contains complete information. The attributes of the dataset are multiclass variable in characteristics with double classification.

5.3. Feature selection

The patient record is identified uniquely by two features of the dataset by sex and age from 13 attributes of the dataset and assign individual ids. The rest of the features consists of medical information. The medical information are vital attributes predicting heart disease. The correlation performed on all 13 attributes with the target value to select the features with high and positive correlation feature as shown in Table 1.

5.4. Splitting dataset

The Table 2 below shows the splitting of the dataset in the following ratios of training and testing set in percentile.

5.5. Classification

One of the Simplest and best ML classification algorithm is Logistic Regression. The LR is the supervised ML binary classification algorithm widely used in most application. It works on categorical dependent variable the result can be discrete or binary categorical variable 0 or 1. The sigmoid function is used as a cost function. Sigmoid function maps a predicted real value to a probabilistic value between '0' and '1'.

Logistic Sigmoid function:

$$P(x) = 1/(1 + e \wedge (-x)) \quad (1)$$

Where, P(x) is probability estimation function a value between 0 and 1, x is input to the probability function (algorithm's prediction value), the mathematical constant e is Euler's number and its value is approximately equal to 2.71828 as shown in Eq. (1).

To predict the cardiac disease logistic regression ML model is used, firstly the LR model are trained with five splitting condition and tested with test data for prediction to get the best accuracy and to find the models behavior. The algorithm results category of 1 and 0 for presence and absences of cardiac disease.

The Logistics Regression Model is described in Pseudocode 1 is used in both training and testing the data instance.

Pseudocode 1 Logistic Regression
1: **Input:** Feature selected data
2: **Output:** Best classification
3: **Algorithm:**
4: For i ← 1 to k
5: For each training & testing data instance d_i:
6: Set the target value for the regression to $Z \leftarrow \frac{y_j - P(1-d_j)}{[P(1-d_j)\cdot(1-p(1-d_j))]}$
7: Initialize the weight of instance d_j to P(1|d_j). (1-P). (1|d_j)
8: Finalize a f(j) to the data with class value (z_j) & weights (w_j) Classification Label Decision
9. Assign (class label:1) if P (1|dj) >0.5, otherwise (class label:2)

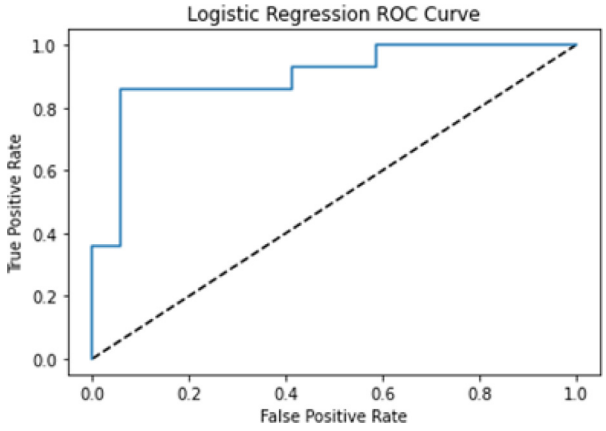

Fig. 3. ROC curve.

6. Results and discussion

The logistical regression is tested with UCI dataset with five different ratios and their accuracy as shown in the table below. The accuracy of 87.10% obtained by logistical regression for split ratio of training and testing is 90:10. The increasing accuracy of the model by increasing the training is shown in Fig. 2 and accuracy of result in Table 3.

The Logistics Regression increase its accuracy with increasing training by 50% to 90% and 90% training and 10% testing provides highest accuracy of 87.10%.

The Table 4 shows classification report, precision, recall, f1-score and accuracy of LR classifier for UCI dataset with 90% training and 10% testing. The model has precision of 0.857, recall 0.857, F1-score 0.857 and accuracy of 87.10%.

The ROC (Receiver Operator Characteristics) curve as shown in the Fig. 3 is used to further investigation in to the model. The performance of the model is visualized by ROC Curve and the tradeoff between TPR (True Positive Rate) and FPR (False Positive Rate). It ranges from 0 to 1 and the area under it signifies the capabilities of distinguish the class of ML model. The ROC curve as near to one it is more capable of classifying.

The Table 5 represents the various previous research work carried on Logistic Regression using rapid minor and python on UCI Dataset from the year 2019 to 2021 with accuracy of prediction.

7. Conclusion

One of the important areas in industry of medical is prediction of cardiovascular disease, with the available data of the patient to predict the absence and presence of cardia disease. There are several techniques and methods are present for prediction of cardiovascular disease. In this research, Logistic Regression supervised ML algorithm are used to classify the heart disease. To improve the performance, pre-processing of corpus like Cleaning, finding the missing values are done. The vital part is feature selection, which increase the accuracy of algorithm and even focus on the behaviour of the algorithm. As the behaviour of Logistic regression is as training increases the accuracy of prediction also increased. The LR classifier achieved 87.10% of accuracy with training

90% and testing 10%. The results outperformed compared to previous research work. The limitation is only UCI dataset is used in the study and future work try to implement on multiple datasets.

References

[1] World Health Organization and J. Dostupno, cardiovascular diseases: key facts, vol. 13, no. 2016, p. 6, 2016. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] K. Uyar, A. Ilhan, Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks, Proced. Comput. Sci. 120 (2017) 588–593.

[3] N. Kausar, S. Palaniappan, B.B. Samir, A. Abdullah, N. Dey, Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients, in: in Applications of Intelligent Optimization in Biology and Medicine, Cham, Switzerland: Springer, 2016, pp. 217–231.

[4] M. Shouman, T. Turner, R. Stocker, Integrating clustering with different data mining techniques in the diagnosis of heart disease, J. Comput. Sci. Eng. 20 (1) (2013) 1–10.

[5] M.S. Amin, Y.K. Chiam, K.D. Varathan, Identification of significant features and data mining techniques in predicting heart disease, Telemat. Inf. 36 (2019) 82–93 Mar..

[6] D. Singh, J.S. Samagh, A comprehensive review of heart disease prediction using machine learning, J. Crit. Rev. 7 (12) (2020) 2020.

[7] F.A. Latifah, I. Slamet, Comparison of heart disease classification with logistic regression algorithm and random forest algorithm, in: Proceedings of the AIP Conference, 2020, p. 2296, doi:10.1063/5.0030579.

[8] Z. Khan, D.K. Mishra, V. Sharma, A. Sharma, Empirical study of various classification techniques for heart disease prediction, in: Proceedings of the IEEE 5th International Conference on Computing Communication and Automation (ICCCA), 2020, pp. 57–62, doi:10.1109/ICCCA49541.2020.9250852.

[9] Nishadi, A.S.T. (n.d.). International journal of advanced research and publications predicting heart diseases in logistic regression of machine learning algorithms by python jupyterlab. https://www.kaggle.com

[10] M. Saw, T. Saxena, S. Kaithwas, R. Yadav and N. Lal, Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning, in: saw (Ed.), 2020 International Conference on Computer Communication and Informatics (ICCCI), IEEE, Coimbatore, India, 2020, pp. 1–6. January 22-24.

[11] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K.H., Lee, S., & Froelicher, V. (n.d.). International application of a new probability algorithm for the diagnosis of coronary artery disease.

[12] S. Bashir, Z.S. Khan, F. Hassan Khan, A. Anjum, K. Bashir, Improving heart disease prediction using feature selection approaches, in: Proceedings of the 16th International Bhurban3 Conference on Applied Sciences and Technology (IBCAST), 2019, pp. 619–623, doi:10.1109/IBCAST.2019.8667106.

[13] R. Kannan, V. Vasanthi, Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease, in: Springer Briefs in Applied Sciences and Technology, Springer Verlag, 2019, pp. 63–72, doi:10.1007/978-981-13-0059-2_8.

[14] M. Ganesan, N. Sivakumar, IoT based heart disease prediction and diagnosis model for healthcare using machine learning models, in: Proceedings of the IEEE (ICSCAN), 2019, pp. 1–5, doi:10.1109/ICSCAN.2019.8878850.

[15] K.G. Dinesh, K. Arumugaraj, K.D. Santhosh, V. Mareeswari, Prediction of cardiovascular disease using machine learning algorithms, in: Proceedings of the International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018, pp. 1–7, doi:10.1109/ICCTCT.2018.8550857.