International Journal of Computer and Communication Technology

Volume 9
Issue 1 *Research on Computing and Communication Sciences.*

Article 9

July 2023

Predicting Accurate Heart Attacks Using Logistic Regression

Vishal Baral
Carelon Global Solution, vishal.baral19@gmail.com

Pranati Palai
RNIASE, Cuttack, pranatipalai@gmail.com

Soumen Nayak
IIT (ISM),Dhanbad, soumen.nayak@gmail.com

Recommended Citation

Baral, Vishal; Palai, Pranati; and Nayak, Soumen (2023) "Predicting Accurate Heart Attacks Using Logistic Regression," *International Journal of Computer and Communication Technology*: Vol. 9: Iss. 1, Article 9.
DOI: 10.47893/IJCCT.2023.1444
Available at: https://www.interscience.in/ijcct/vol9/iss1/9

Predicting Accurate Heart Attacks Using Logistic Regression

Vishal Baral[1],Pranati Palai[2], Soumen Nayak[3*]

[1]Software Engineer, Carelon Global Solution, India

[2]Department of Teacher Education, RNIASE, Cuttack, India

[3]Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

E-mail-vishal.baral19@gmail.com, pranatipalai@gmail.com, *soumen.nayak@gmail.com

Abstract.

A heart attack is one of the leading causes of death today. According to a large data population used as a training set for the algorithm for machine learning, classification is a technique for predicting the target class from input data. A difficulty in clinical data analytics is predicting heart attacks with greater precision. The focal point of this work is to analyze the heart attack dataset (Kaggle repository) to find a Machine learning classifier technique that predicts if a person is prone to heart attack with maximum accuracy based on various health factors. The efficacy of the three classifiers, namely Logistic Regression, Random Forest, and Decision Tree, is demonstrated for predicting heart attack. This work compares the three classification algorithms among various factors. Logistic Regression outperforms all for predicting the values from the dataset accurately.

Keywords: Heart attack predictions, machine learning, logistic regression, classification algorithms

1 Introduction

The emergence in medical cases has led up to a hike in the size or number of medical databases that accumulate a considerable number of data to be examined for medical research, which makes it even more essential to evoke important and required information from this data conducive to improve the standard of the medical facility. Machine learning has become a crucial tool for data extraction and value prediction in medical technology. Diagnosing heart disease is challenging due to several risk factors, such as diabetes, high cholesterol, high blood pressure (BP), irregular heartbeat, etc. Diverse techniques have been employed in machine learning, and data mining is used to assess the magnitude of human heart attacks.

The main objective of this research is to enhance heart attack prediction performance accuracy and find a suitable ML classifier. Several studies have resulted in constraints on the selection of classifiers for algorithmic use. This paper analyzes the heart attack dataset to find a better classification technique that predicts if a body is prone to have a heart attack with maximum accuracy based on various health factors. For classification, three machine learning classifiers are used, namely ***Decision Tree (***DeT) Classifier, ***Random Forest (***RaF) Classifier, and ***Logistic Regression (***LoR) [12] DeT algorithm classifier is a Supervised Machine Learning (SML) Technique, a classification algorithm in which the data is continually split according to certain specifications.

RaF Classifier cluster of many DeT, which individually develop a prediction. The prediction with the most votes becomes the final prediction. LoR is a classification algorithm that provides discrete values (Boolean values, Binary values) according to a provided set of independent variables. LoR is a classification algorithm that provides discrete values (Boolean values, Binary values) according to a provided set of independent variables. This work compares these three classification algorithms among various factors to develop the most efficient algorithm to predict the values from the dataset more accurately.

The paper is arranged in multiple sections as follows—section 2 addresses work related to the heart, current approaches, and available techniques. In Section 3, the details of the classifiers have been mentioned. Section 4 deals with the experimental result and discussion. It also reveals how the experiment was carried out and the outcomes obtained. Finally, section 5 exhibits the conclusion with the future scope of the work.

2 Related Work

Various projects have predicted if a person with specific input values might be the prey of a particular disease. In the project where, by using four algorithms, kidney disease is predicted by Jena et al. [1], it was concluded that the Multilayer Perceptron was the one with the highest accuracy with the output among Multilayer Perceptron, Decision Table, J48, and Naïve Bayes. Three classifiers were used (Decision Table, Naive Bayes, and J48) to precisely predict the risk in chronic disease by examining the composite usage of the classification algorithms by Jena et al. [2]. The J48 algorithm performance was observed to be better than other algorithms, with high accuracy of 99 percent in prediction. Patra et al. [3] designed FRESNN, which is a hybrid characteristic collection method for the typical selection of gene expression data feature selection mechanism that promoted evidence that is more valuable for discrimination, where the distinct hybrid selection outline resulted in the enhancement of the stability being one among other factors.

When it comes to the applications of the same, problems related to the heart are researched by various individuals and groups like Mohan et al. [4], who suggested a hybrid HRFLM approach used to combine the features of the linear method and random forest. Finally, the concluding HRFLM is to be accurate in heart disease prediction.

Latha and Jeeva [5] used various ensemble algorithms for experiments stacking, bagging, majority voting, and boosting to analyze the prediction accuracy of heart disease. A final comparison of results showed that the majority voted with an accuracy of 85.48%, which was the highest. Muthuvel et al. [6] used the Naïve Bayes, DeT, K- nearest neighbor, and Support Vector Machine. In the end, the highest calculated accuracy using each of the model's algorithms was that of KNN, with 83.60% accuracy. A framework for Decision Support System was designed by Mhatre and Verma [7] and used to analyze to predict the severity of cardiovascular disease with a genetic-based neural network approach. The data were classified into five classes using the Backpropagation Algorithm. The attributes, after taking into consideration, gave an accuracy of 78.763%. Ranga and Rohila [8] found the accuracy to be the highest in the case of the support vector, which was nearly equal to 82.6400% when they used KNN, SVM, artificial neural network, DeT, and RaF classifiers. Mishra et al. [13] have used ensemble techniques that enhance the accuracy of prediction of cardiac arrest.

3 Materials and Methods

Data Analysis is a method of inspecting and refining data and applying statistical or logical techniques to evaluate and use data for decision-making / predicting future values using the data. The data analysis is done in two forms classification and prediction. They perform a train-test split to generate a model and predict the future of the data.

Classification is a method/ process of separating a given set of data into different classes. This approach is a part of the supervised learning of Machine Learning, where the systems learn from the input data to develop new observations. The process involved predicting a particular output based on the given input in classification. The data that is readily available is fetched to expect the results. The records are classified based on this data itself. The further categorization of the data into a training set and a test set is done. The reference used for classification is the data contained by the training set, which has been classified beforehand [12]. Classification predictive modeling is the task assigned to approximate the mapping function from the given data set to give the most appropriate output.

Prediction in Machine Learning: An algorithm that has been trained using a training dataset and then used the same methodology that the model learned from the training dataset will be used to predict the output of a given collection of data.

The three classifiers used for the classification of the heart attack dataset that has been taken from the UCI dataset are as follows:

Logistic Regression: This classification algorithm is used only when the value of the target/output variable is categorical. In other words, LoR is mainly used when the output of the given data is binary (0 or 1). As mentioned by this analysis, the input data belong to the generalized linear model class. They are categorized by their response distribution, which helps transfer the average value to a scale to define them as linear and additive concerning the background variables [9].

Random Forest: RaF constitutes various distinct DeT that function as a group. Each tree in any RaF gives a prediction of class, while the model's prediction is selected based on which category has the maximum number of votes. RAF is the collection of un-snipped regression or classification trees from the irregularly chosen samples from the training data set. In the induction process, all the random characteristics are selected. The prediction is made by combining (the majority vote and averaging for classification and regression, respectively) [10].

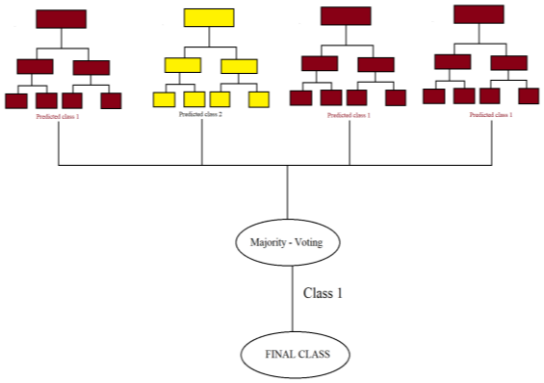


Fig 1: Structural flow of Random Forest Tree

***Decision Tree*:** It is a part of Supervised Machine Learning and is a logic-based algorithm [11] where the data is in continuous order. The tree mainly consists of two entities, namely called leaves and nodes. The leaves are the final output of the data. The idea is similar to that of the structure of an average tree in nature which consists of a single root and various branches, nodes, and leaves. Likewise, a DeT is built from nodes and the branches that are the segments connecting the nodes [10]. A single node of a DeT symbolizes a feature in a case to be classified, and a branch represents value as the assumptions for the node. Starting at the root node, the classification of the instances begins, and then their feature values become the basis of sorting later.

4Results and Discussion

The demonstration is done on the Heart Attack dataset, which consists of 340 instances and 14 characteristic columns. The merits that were considered as the input were

1. Age (in whole numbers),
2. Sex or gender (0 for male or 1 for female and 2 for others),
3. Chest pain type (consists of 4 values, i.e., 0, 1, 2, 3),
4. Fasting blood sugar > 120 mg/dl,
5. Resting blood pressure,
6. Resting Electro-Cardio graphic results (values 0,1,2),
7. Max heart rate achieved,
8. Serum cholesterol in mg/dl,
9. Number of major vessels (0-3) colored by fluoroscopy,
10. Exercise-induced angina,
11. Old peak = ST depression induced by exercise relative to rest,
12. Slope of the peak exercise ST segment,
13. Thal (with values as 0 = normal, 1 = fixed defect, 2 = reversible defect).

The "target" field refers to the presence of heart disease in the patient. It is an integer – value denoted with binary outcomes that are 0 and 1, which refers to no/fewer chances of heart attack and more chance of getting a heart attack, respectively.

0 = no/less probability of heart attack
1 = more likelihood of heart attack

The model used for testing the modulus is split into 85% as train and 15% as testing. The result of different classifications has different performance values with varying accuracy values. All the performance values are displayed in Table 1 and Table 2.

Classification	**Accuracy**
Logistic Regression	0.895
Random Forest Tree	0.802
Decision Tree	0.709

Table1:AveragePrecision (Avg. Pre.),Recall, and Precision valuesofClassification

Classification	**AveragePrecision (Avg. Pre.)**	**Precision**	**Recall**
LoR	0.918	0.806	0.894
RaF	0.898	0.826	0.804
DeT	0.689	0.753	0.697

Precision measured by the fraction of relevant instances among the retrievedinstances Eq (1).

$$\text{Precision} = \frac{TPositive}{TPositive+FPositive} \quad (1)$$

where T Positive is True positive, and F Positive is False positive.

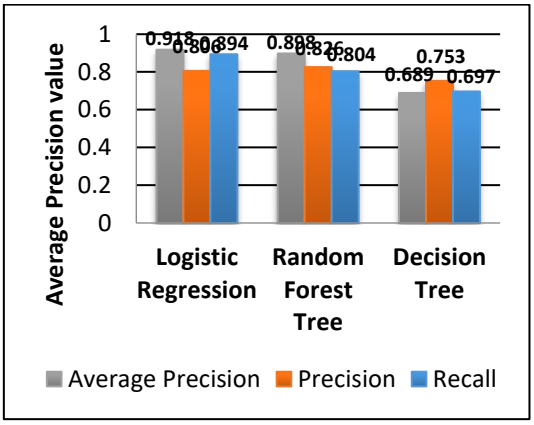
Recall measured on the fraction of the total amount of relevant instances that were actually retrieved.

$$\text{Recall} = \frac{TPositive}{TPositive+FNegative} \quad (2)$$

where FN is False Negative.

Average Precision(AP) is calculated based on the area under the precision-recall curve. The value of AP is always within 0 and 1 as the value of precision and recall value have a range of [0,1].

$$\text{AP} = \int_0^1 p(r)dr \quad (3)$$



Graph 1: Based on Average Precision, Precision and Recall of each classification

The average precision of Logistic Regression is the highest (0.918) among all three regressions used as Random forest tree has value of 0.898 and Decision trees has that of 0.689, that shows Logistic Regression has a better average precision on heart attack data set.

Table2: Accuracy of each Classification

The key parameter for assessing classification models is **accuracy**. Informally, accuracy is the ratio of the model's correct predictions to the total number of predictions that it was asked to make. The following is the official definition of accuracy:

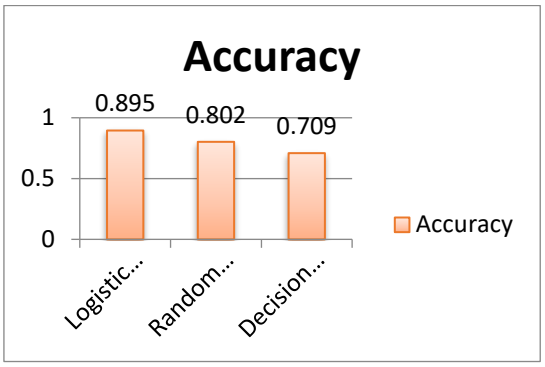
$$\text{Accuracy} = \frac{\text{Number\_of\_correct\_predictions}}{\text{Total\_number\_of\_predictions}} \quad (4)$$

For binary outcomes of data-set, the accuracy can also be calculated in terms of positives (correct positive prediction of heart attack entries) and negatives (correct negative prediction of heart attack entries) as follows:

$$\text{Accuracy} = \frac{\text{TPositive+TNegative}}{\text{TPositive+TNegative+FPositive+FNegative}} \quad (5)$$

T Positive = True Positives, *T Negative* = True Negatives, *F Positive* = False Positives, and *F Negative* = False Negatives.

According to Graph 2, the higher accuracy of predicting the correct outcomes for some given data is Logistic Regression whereas the value of logistic regression prediction is more compared to any other regression prediction.



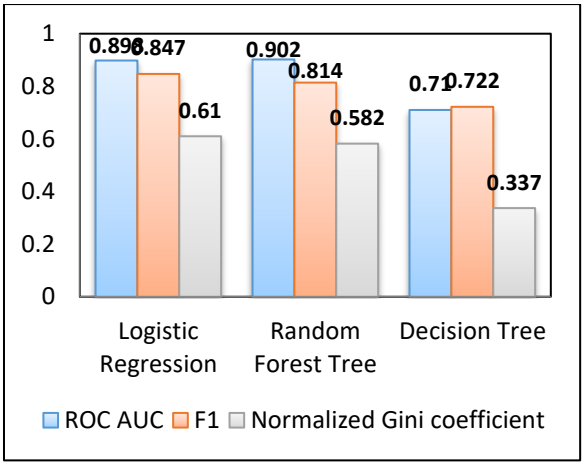
Graph 2: The Accuracy values of the modulus created different classification

The other factors of the classifier prediction models are F_1, ROC AUC, Normalized Gini Coefficient, Log Loss, and Model building time (in secs) as given in Table 3.

Table 3: Other performance factors of the classifications

Classification	F_1	Log loss	ROC AUC	Normalized Gini coefficient	Build Time(sec)
Logistic Regression	0.847	0.441	0.898	0.610	1
Random Forest Tree	0.814	0.707	0.902	0.582	4
Decision Tree	0.722	10.047	0.710	0.337	1

In the above data, building the model using different classification is represented by build time as Logistic Regression and Decision tree have the same build time,i.e., 1 second, which is faster than the Random Forest tree. On comparing another factor like Area Under the Curve of Receiver Characteristic Operator(ROC AUC), F_1, Normalized Gini coefficient Logistic Regression have a higher value in F_1 and normalized Gini Coefficient whereas ROC-AUC of Random Forest Tree has higher values.



Graph 3 Representing some of the other factors of the classification

After observing all the data and the graphs the result of Logistic Regression on the heart attack Dataset gave the highest accuracy value than other classifications taken into consideration here (Random Forest Tree, Decision Tree). As the accuracy value of Logistic Regression is 0.895 (89.5%) which is greater than Random Forest trees 0.802 (80.2%) and Decision Trees 0.709 (70.9%). This shows that for any binary classification problem (target data output is 0 or 1).

5 Conclusion

This paper uses three classification algorithms, LoR, RaF, and DeT, to check which algorithm predicts heart attacks using various parameters with maximum accuracy. The LoR classification algorithm has outflanked the other two classification algorithms regarding the accuracy, average precision, log loss, and build time. Therefore, LoR is more accurate than the classification algorithms used in our paper to predict heart-attack using the dataset. In the future, the heterogeneous classifiers will be used to improve the result's accuracy and minimize the weaknesses in the individual classifiers.

References

1. Jena L., Patra B., Nayak S., Mishra S., Tripathy S. (2021) Risk Prediction of Kidney Disease Using Machine Learning Strategies. In: Mishra D., Buyya R., Mohapatra P., Patnaik S. (eds) Intelligent and Cloud Computing. Smart Innovation, Systems and Technologies, vol 153. Springer, Singapore.
2. Jena L., Nayak S., Swain R. (2020) Chronic Disease Risk (CDR) Prediction in Biomedical Data Using Machine Learning Approach. In: Mohanty M., Das S. (eds) Advances in Intelligent Computing and Communication. Lecture Notes in Networks and Systems, vol 109. Springer, Singapore.
3. Patra B., Jena L., Bhutia S., Nayak S. (2021) Evolutionary Hybrid Feature Selection for Cancer Diagnosis. In: Mishra D., Buyya R., Mohapatra P.,

Patnaik S. (eds) Intelligent and Cloud Computing. Smart Innovation, Systems and Technologies, vol 153. Springer, Singapore

4. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707
5. C. Beulah Christalin Latha, S. Carolin Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics in Medicine Unlocked, Volume 16, 2019, 100203, ISSN 2352-9148.
6. Muthuvel, Marimuthu&Sivaraju, Deivarani& Ramamoorthy, Gayathri. (2019). Analysis of Heart Disease Prediction using Various Machine Learning Techniques.
7. TejaliMhatre ,Satishkumar Varma, 2019, Heart Disease Prediction using Evolutionary based Artificial Neural Network, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 08, Issue 08 (August 2019)
8. Ranga, Virender &Rohila, D.. (2018). Parametric Analysis of Heart Attack Prediction Using Machine Learning Techniques. International Journal of Grid and Distributed Computing. 11. 37-48
9. Dalgaard, P.: Logistic regression. In: Introductory Statistics with R.StatisticsandComputing.Springer,NewYork,NY,2008.
10. Ali, J., Khan, R., Ahmad, N., Maqsood, I.: Random Forests andDecisionTrees.InternationalJournalofComputerScienceIssues(IJCSI),vol.9,2012.
11. Kotsiantis,S.B.,ZaharakisI.,PintelasP.:Supervisedmachinelearning:Areviewofclassificationtechniques.Emergingartificialintelligence applications in computer engineering,vol. 160(1), pp. 3-24,2007.
12. Umadevi,S.,Marseline,K.S.J.:Asurveyondataminingclassificationalgorithms,2017InternationalConferenceonSignalProcessingandCommunication(ICSPC),Coimbatore,pp.264-268,2017.
13. Mishra, N., Desai, N. P., Wadhwani, A., Baluch, M. F.: Visual Analysis of Cardiac Arrest Prediction Using Machine Learning Algorithms: A Health Education Awareness Initiative. In Handbook of Research on Instructional Technologies in Health Education and Allied Disciplines, IGI Global, pp. 331-363, 2023.