Comparitive Analysis of Heart Disease Prediction

Kundrapu Supriya
IIIT Naya Raipur, India
Email :kundrapu20100@iiitnr.edu.in

Jai Vardhan
IIIT Naya Raipur, India
Email : jai20100@iiitnr.edu.in

K Sai Swetha
IIIT Naya Raipur, India
Email : kothapalli20101@iiitnr.edu.in

Abstract—**Globally, heart disease is the leading cause of mortality. Many modern technologies are utilised to treat cardiac disease. The most prevalent problem at medical centres is that many medical workers lack the necessary knowledge and skills to treat their patients, resulting in poor outcomes and, in some cases, death. Machine learning algorithms are being used to anticipate cardiac illness to tackle these problems, making automated diagnosis in hospitals easier. This project is implemented on a public heart disease prediction dataset, which is bascially combined using 4 different datasets. The dataset is divided into train and test sets with a ratio of 70:30 and different algorithms such as Logistic Regression, Decision Tree and K Neighbor Classifiers have been used to predict heart disease. Also, various evaluation metrics such as accuracy, precision, recall and f1 score have been used to analyse the performance of the models. Furthermore, the log-likelihood value (goodness of fit) has been calculated for logistic regression model and nested models have been implemented. Then the models have been compared using lr test.**

I. Introduction

Health is regarded as a gift that is dependent on good care and several other elements. Modern living has a significant impact on health, which is increasing day by day [1]. Among all other health issues, heart disease is becoming a serious issues, since many people are losing their lives as a result of it. According to the WHO, cardiac illnesses are responsible for taking the lives of 17.7 million people each year, accounting for 31 % of all deaths worldwide [4]. According to the Indian Heart Association, half of all heart strokes occur in people under the age of 50, and a quarter of all heart strokes occur in those under the age of 40 [2]. Other than change in lifestyle, several elements such as obesity, high blood pressure, diabetes, and others contribute to the chance of developing a heart disease [3]. As a variety of condition affects the heart and blood arteries, diagnosis of heart disease is very crucial and complex task.

Medical diagnosis is seen as a critical yet difficult process that must be completed correctly and swiftly. This system's automation would be incredibly beneficial [5]. Unfortunately, not all doctors are experts in every field, and there is a lack of skilled people in some areas. Medical professionals have accumulated a large amount of medical data that can be reviewed and relevant information retrieved. Techniques for extracting meaningful and hidden facts from large volumes of data are known as data mining techniques. In a healthcare system, the majority of the data is discrete. As a result, making decisions based on discrete facts becomes a difficult undertaking [6]. Therefore, the main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis. As a result, an automatic medical diagnosing system would most likely be quite benificial. Machine Learning, a subset of data mining, excels at dealing with large, well-structured datasets. In the medical field, machine learning can be used to diagnose, detect, and forecast a variety of ailments [7] . The study's main goal is to give clinicians a tool for detecting heart problems.In recent researches many machine learning and deep learning algorithms such as KNN [8], ANN, Naives Bayesian, Decision Tree [9], Random Forest [10], etc have been used for predicting heart disease. In this project, we proposed a method based on Logistic Regression. The significant contributions of this report are as follows:

- Combining four different dataset taken from four different hospitals.
- Training the model using Logistic Regression, Random Forest and KNeighbors Classifier.
- Implementing nested models and comparing the models using lr test.
- Using different evaluation metrics for evaluating the performance of the model.

The following sections of this report are organized chronologically. Section II provides an overview of the recent researches based on heart disease prediction. Section III methodology used for predicting the heart disease. Section IV provides a in detail explanation of the experimental analysis of the model. Finally, Section V concludes this project.

II. Related Works

M. Tarawneh et al. [11] proposed a Intelligent Heart Disease Prediction Systems (IHDPS) have been built to forecast various heart illnesses utilising Decision Tree, Nave Bayes, and Neural Network (NN). Using 15 distinct medical vital signs, such as age, sex, and chest pain, the system has extracted the hidden knowledge (patterns and correlations) connected with heart disorders. The prediction accuracy of the Naive Bayes, decision tree, and neural network algorithms was 95 percent, 94.93 percent, and 93.54 percent, respectively. J. L. Castillo et al. [12] presented a study that might be used to diagnose heart disease, AIDS, brain cancer, diabetes, dengue fever, and hepatitis C. To map the patients onto distinct classifications of the aforementioned disorders, machine-learning methods such as Naive Bayes, J48, K-Nearest Neighbors, and C4.5 were used. When the results of the harnessed classifiers were compared, C4.5 came out on top with an 83.6 percent prediction accuracy, followed by J48 and Naive Bayes classifiers

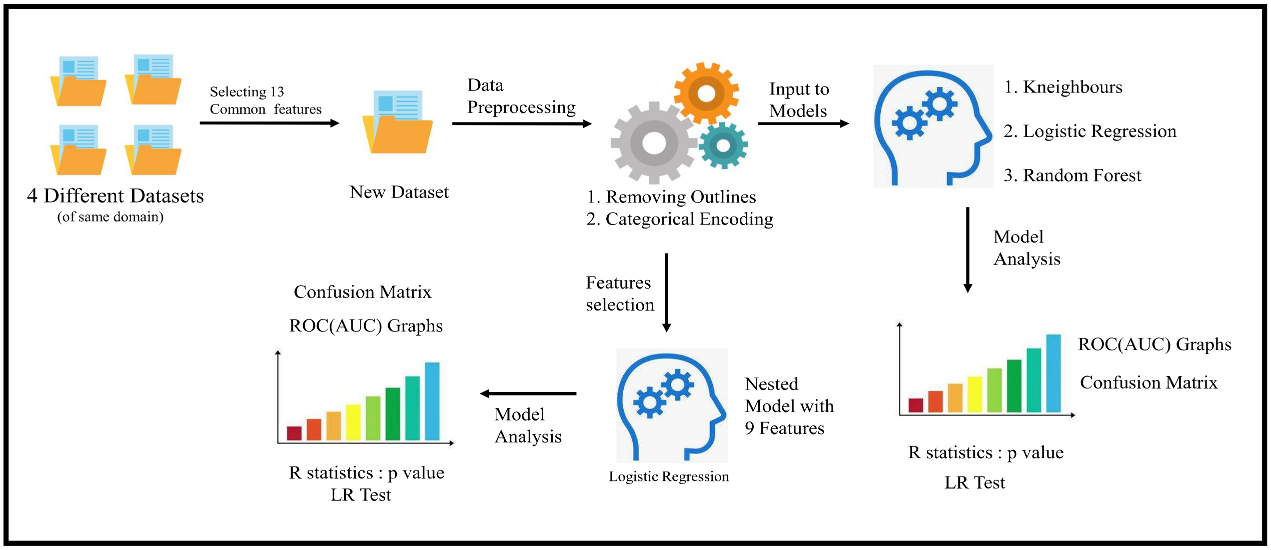


Fig. 1. Overview of proposed Solution

with 81.1 percent and 75.97 percent prediction accuracy, respectively. T. Bhardwaj et al. [13] used different algorithms such as Dichotomiser3 (ID3), Classification and Regression Tree (CART), and Decision Tree (DT) were used to diagnose coronary heart disorders. The CART, DT, and ID3 had accurate prediction rates of 83.49 percent, 82.50 percent, and 72.93 percent, respectively.

The data in [5] comes from the UCI data repository. Jyoti Soni et al. proposed a method using WEKA software to forecast cardiac disease using KStar, J48, SMO, and Bayes Net, as well as Multilayer perceptron. Based on the results of many factors Using k-fold cross validation, SMO (89 percent accuracy) and Bayes Net (87 percent accuracy) procedures outperform KStar, Multilayer Perceptron, and J48 techniques. The precision of the algorithms is still unsatisfactory. Using data from the UCI repository, paper [14] compares the performance of various machine learning algorithms such as Naive Bayes, KNN, Decision Tree, and ANN. ANN had the best accuracy of 85.3 percent among them. While Nave Bayes and KNN accounted for around 78 percent of the total, Decision Tree accounted for 80 percent. Similarly, a variety of research use machine learning or data mining to make medical science predictions . We believe that foreseeing future events could greatly assist medical professionals in taking urgent safeguards and devising an effective treatment approach in a timely manner.Our research solves these problems by predicting patients' conditions by forecasting vital signs using various regression models.

III. Proposed Methodology

The overview of the proposed solution is shown in the 1. The brief description of the different models shown in the figure are given in the subsections below:

A. Different Models

1) Logistic Regression: The method of modelling the probability of a discrete result given an input variable is known as logistic regression. The most frequent logistic regression models have a binary outcome, which might be true or false, yes or no, and so forth. The equation (1) represents the logistic equation.

$$logisticequation = \frac{1}{1+e^{-z}} \tag{1}$$

2) Random Forest: The Random Forest classifier combines the results of numerous decision trees applied to different subsets of a dataset to increase the dataset's projected accuracy. Random forest collects forecasts from each tree and predicts the final output based on the majority votes of predictions rather than relying on a single decision tree.

3) KNeighbors Classifiers: A k-nearest-neighbor algorithm, abbreviated as k-nn, is a data categorization method that calculates how probable a data point is to belong to one of two groups based on which group the data points closest to it belong to.

4) XGBoost Classifier: XGBoost is a gradient boosting algorithm that uses a decision-tree-based ensemble Machine Learning algorithm. Artificial neural networks tend to outperform all other algorithms or frameworks in prediction problems involving unstructured data (images, text, etc.). Decision tree-based algorithms, on the other hand, are currently considered best-in-class for small-to-medium structured/tabular data.

B. Log Likelihood

A likelihood technique defines how well a parameter $(theta)$ explains the observed data and is used to determine how well a model fits the data. Consider the case below: You've got a collection of the observations x1, x2,...xn, each with its unique probability density function fX. (x). fX1, X2...

is the function of their joint density. Xn(x1, x2,...xn) = fX(x1) * fX(x2) *... * fX(xn) = fX(x1) * fX(x2) *... * fX(xn) = fX(x1) * fX(x2) *... * fX(xn)

We may rewrite the above equation using summation instead of products because the sum of the logs of the multiplied components equals the log of a product as shown in the equation eqrefeq2 below.

$$\ln[f_X(x_1) * f_X(x_2) * ... * f_X(x_n)] = \sum[\ln[f_X(x_i)] \quad (2)$$

C. Nested Models

A nested model is a regression model that include a portion (or) subset of independent variables to form a different regression model.Basically nested models are used to know if a set of full independent variable fit a regression model better than a subset of independent variables. In order to know if the if a nested model is better than a full set of independent variable we perform likelihood-ratio test. The example of a nested model is given below: eq (3) represents a model with full set of independent variables i.e, Model A and eq (4) represents a model with subset of variables i.e, Model B. Therefore we can say that Model B is nested model in Model A.

$$points = \beta0 + \beta1(age) + \beta3(heartbeat) + \beta4() + \epsilon \quad (3)$$

$$points = \beta0 + \beta1(age) + \beta3(heartbeat) + \epsilon \quad (4)$$

D. Lr Test

Log-Likelihood Test uses null and alternative hypothesis to compare nested models. The outcome of the log-likelihood test is the test statistic and chi-square p-value. We can say that a full model is better than nested model if the p-value is greater than the significance value i.e, 0.05 and we will reject the null hypothesis.

Another method to do log-likelihood ratio test is by knowing both the models log-likelihood values and taking the ratio of log-likelihood of nested model to the log-likelihood of the full model.

$$LRTest = -2\ln_e(\frac{L_n(\theta)}{L_f(\theta)}) \quad (5)$$

where $L_n(\theta)$ represents log-likelihood of nested model and $L_f(\theta)$ represents log-likelihood of full model. The equation of the loglikelihood ratio test is shown in the equation (5)

IV. Experimental Analysis

A. Dataset Description

The dataset that has been used in this project consists of data from four different databases (Hungary, Cleveland, Long Beach V, and Switzerland). It contains 75 independent variables and 1 dependent variable. But for predicting whether the heart disease is present or not, we initially use only 14 attributes including dependent variable. Hence, the shape the dataset us (1025,14)

B. Data Preprocessing

Different data preprocessing techniques such as missing values imputation and outlier detection have been used to transform the raw data into useful and effective format for traning the model. There is not need of missing values imputations as there are no missing values present in the data. Outlier are detected using box plots and are removed using Inter-Quantile Range (IQR) Method. The box plots of column trestbps (resting blood pressure) before and after outlier removal are shown in the Fig. 2

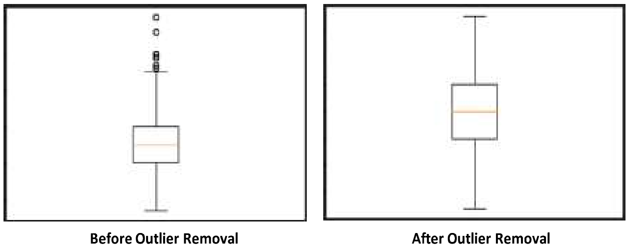


Fig. 2. Box plots before and after removal of outliers

C. Feature Selection

From the given dataset important features are selected using feature importance graph and correlation matrix. The feature importance graph and Correlation Plots are shown in Fig. 3 and Fig. 4 respectively.

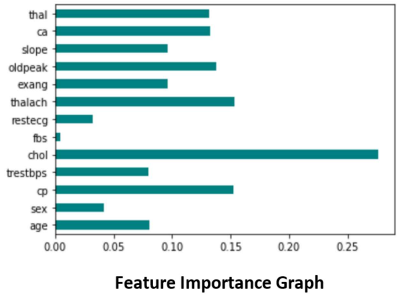


Fig. 3. Correlation Plot

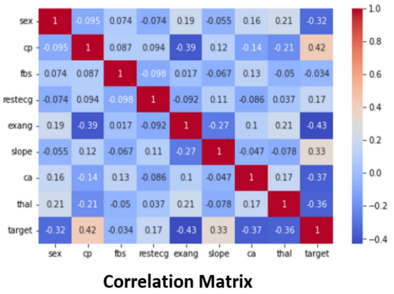


Fig. 4. Correlation Plot

D. Train-Test Split

The dataset containing 1025 points is split into two independent datasets for training and testing the machine learning model. The dataset is divided in a 70:30 ratio, with training data accounting for 70% of the original data and testing data accounting for 30% of the original data.

E. Logistic Regression

Using 13 features logistic regression model and then predicting the heart disease. Analyzing the performance using accuracy metrics such as accuracy score, cross-validation score, roc and auc curves, etc.

The precision, recall, F1-score are estimated based on the following eq (6), (7),(8) respectively.

$$Precision = \frac{T_P}{T_P + F_P} \tag{6}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{7}$$

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

where T_P ,F_N ,F_p, T_N are taken from the confusion matrix. The calculated values of precision, recall, F1-score are shown in the Table II.

TABLE I
EVALUATION METRICS SCORES FOR LOGISTIC REGRESSION

Evaluation Metrics	Value (%)
Accuracy	87.2
Precision	87.3
Recall	89.6
F1	88.4

F. Comparing different models

The Random Forest, Logistic Regression and KNegihbors algorithm are used in order to predict the existence of heart disease. The performances of all the algorithm used in the project are listed in II. In terms of both accuracy and cross validation accuracy scores XGBoost Classifier achieved an higher accuracy of 99.6 % and 99.3% respectively and than other methods.

TABLE II
ACCURACY AND CROSS-VALIDATION SCORES FOR DIFFERENT METHODS USED

Method	Accuracy Score (%)	Cross Validation(%)
Random Forest	70	95.8
Logistic Regression	87.2	85.6
KNeighbors	95.4	88.6
XGBoost	99.6	99.3

G. Comparision between logistic regression and logistic regression nested models

A nested logistic regression model has been implemented by taking a subset of features from the logistic model. And both are compared using log-likelihood test. The loglikelihood values of both the models are calculated and then lr test is done using the equation (4). The log-likelihood ratio is 3.2. And from that we got to know the logistic regression model is better than the nested logistic regression model. The correlation matrix of the of the nested models is shown below in fig. 5

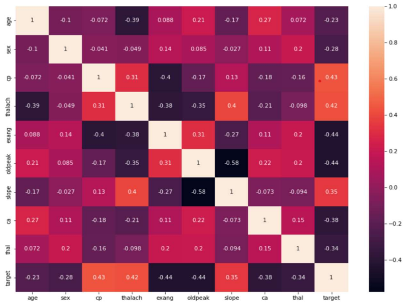
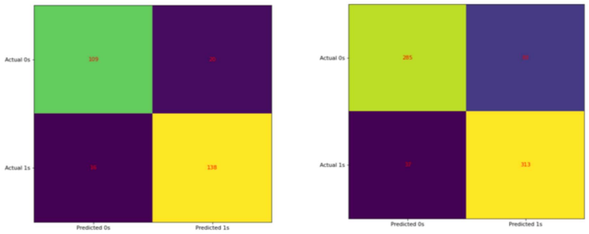


Fig. 5. Correlation Matrix

Confusion matrix for both logistic regression and logistic regression nested nodel is shown in the Fig. 6



Confusion Matrix for logistic Regression Model

Confusion Matrix for logistic Regression Nested Model

Fig. 6. Confusion Matrix

The comparision of the accuracy metrics such as precision, recall and f1 score are listed in Table III. From the table we can say that the LR models has better accuracy, precision, recall and f1 score than Nested LR model.

TABLE III
ACCURACY METRICS COMPARISION FOR LR AND NESTED LR

Accuracy Metrics	LR	Nested LR
Accuracy	87.2	83.4
Precision	87.3	79.0
Recall	89.6	89.4
f1 score	88.4	83.9

H. Comparision between existing models and proposed models

The Proposed models have been compared with the existing models in the table IV. Out of all the methods XGBoost has achieved an higher accuracy of 99.6 %.

TABLE IV
COMPARISION OF PROPOSED MODELS WITH THE EXISITING MODELS

References	**Methods Used in various papers**	**Accuracy (%)**
T.Islam et al. [1]	Random Forest	70
	Logistic Regression	87.2
	KNeighbors	95.4
	Naive Bayes	95
N.Caball'e e-Cervig ey al. [12]	Decision Tree	94.93
	Neural Networks	93.54
	C4.5 algorithm	83.6
T.Bhardwaj et al.b13	J48	81.1
	Naive Bayes	75.97
	CART	83.49
Dangare et al. [14]	Decision Tree	82.50
	ID3	72.93
	ANN	85.3
Proposed Methods	**Random Forest**	**70**
	Logistic Regression	**87.2**
	KNeighbors	**95.4**
	XGBoost	**99.6**

V. CONCLUSION

This paper demonstrates different machine learning algorithms for the predicting heart disease. And made different conclusions based on the heart disease dataset with several machine learning algorithms. Initially, different data preprocessing techniqes like outlier detection and removal have been performed in order increase the accuracy. After that the model was evaluated using Logistic Regression, Random Forest and KNeighbors.Out of which KNeighbors obtained a better accuracy of 95.4 % as compared to other two algorithms. And Logistic Regression nested model is trained and compared with the logistic regression model using lr test. From the log likelihood test we got to know that logistic regression is better than logistic regression nested model. Along with this different evaluation metrics have been used to compare these models. And the proposed models have been compared with the existing models and it has shown that XGBoost Classifier is better than other classifiers and regression models.

ACKNOWLEDGMENT

To begin, we would like to offer our heartfelt gratitude to Prof. Anurag Singh, our project guide, for his direction, assistance, and encouragement during our project work. We appreciate his research guidance, creative ideas, hard labour, and work dedication. Without his tireless labour, this project would not have been possible. He is more than a guide; he is our mentor in creating our personal and professional lives, and we would not be where we are now if it weren't for him. We owe him our deepest gratitude for his unwavering support in every way.

REFERENCES

[1] M. T. Islam, S. R. Rafa and M. G. Kibria, ”Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means,” 2020 23rd International Conference on Computer and Information Technology (ICCIT), 2020, pp. 1-6, doi: 10.1109/ICCIT51783.2020.9392655.

[2] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, ”Prediction of Heart Disease Using Machine Learning,” 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.

[3] M. Chakarverti, S. Yadav and R. Rajan, ”Classification Technique for Heart Disease Prediction in Data Mining,” 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2019, pp. 1578-1582, doi: 10.1109/ICICICT46008.2019.8993191.

[4] Ramalingam, V V Dandapath, Ayantan Raja, M. (2018). Heart disease prediction using machine learning techniques: A survey. International Journal of Engineering Technology. 7. 684. 10.14419/ijet.v7i2.8.10557.

[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni ,” Predictive Data Mining for Medical Diagnosis: An Overviewof Heart Disease Prediction”, International Journal of Computer Applications, Volume 17, No 8, 2011

[6] Apurb Rajdhan, Milan Sai, Avi Agarwal, Dundigalla Ravi, ”Heart Disease Prediction using Machine Learning”, International Journal of Engineering Research Technology (IJERT), Vol. 9 Issue 04, April-2020.

[7] A. N. Repaka, S. D. Ravikanti and R. G. Franklin, ”Design And Implementing Heart Disease Prediction Using Naives Bayesian,” 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 292-297, doi: 10.1109/ICOEI.2019.8862604.

[8] Nida Khateeb and Muhammad Usman. 2017. Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique. In Proceedings of the International Conference on Big Data and Internet of Thing (BDIOT2017). Association for Computing Machinery, New York, NY, USA, 21–26. https://doi.org/10.1145/3175684.3175703

[9] Purushottam, K. Saxena and R. Sharma, ”Efficient heart disease prediction system using decision tree,” International Conference on Computing, Communication Automation, 2015, pp. 72-77, doi: 10.1109/CCAA.2015.7148346.

[10] Yang, L., Wu, H., Jin, X. et al. Study of cardiovascular disease prediction model based on random forest in eastern China. Sci Rep 10, 5245 (2020). https://doi.org/10.1038/s41598-020-62133-5

[11] M. Tarawneh and O. Embarak, “Hybrid approach for heart disease prediction using data mining techniques,” in Proceedings of the International Conference on Emerg- ing Internetworking, Data Web Technologies, pp. 447–454, Springer, Fujairah Campus, UAE, February 2019.

[12] N. Caball'e-Cervig'on, J. L. Castillo-Sequera, J. A. G'omez-Pulido, J. M. G'omez-Pulido, and M. L. Polo-Luque, “Machine learning applied to diag- nosis of human diseases: a systematic review,” Applied Sciences, vol. 10, no. 15, p. 5135, 2020.

[13]] T. Bhardwaj and S. C. Sharma, “Cloud-wban: an experimental framework for cloud-enabled wireless body area network with efficient virtual resource utilization,” Sustainable Computing: Informatics and Systems, vol. 20, pp. 14–33, 2018.

[14] Dangare, Chaitrali Apte, Sulabha. (2012). A Data Mining Approach for Prediction of Heart Disease Using Neural Networks. 3.