Bulletin of Environment, Pharmacology and Life Sciences
Bull. Env. Pharmacol. Life Sci., Special Issue [1]2022 : 909-914

Online ISSN 2277-1808
Journal's URL:http://www.bepls.com
CODEN: BEPLAD

ORIGINAL ARTICLE **OPEN ACCESS**

Heart Disease Prediction Using Real Dataset via Effective Machine Learning Technique

Seira Shinde*, Tanay Baban Kapse, Mansi Ramwani, Sunidhi Singh, Suman Sahu
Computer Science & Engineering Department of college BIT,Raipur.
Email: seira.tak@bitraipur.ac.in

ABSTRACT

The heart is indeed the most important internal part of the body. Cardiovascular disease is one of the leading causes of death all over the world. This occurs when the heart is unable to adequately pump blood to all regions of the body. The researcher needs to look whether there is a relation among the existence or non-existence of cardiovascular disease and other factors like age, trestbps, chol, thalach, gender, cp, restecg, fbs, oldpeak, slope, ca, and thal in this study. The information presented pertains to a total of 303 patients. The data on cardiac disease was analysed using the binary logistic regression (blr) model. {x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13} = {patient's age, gender, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal} respectively are the variables in the model. This type of research aims to raise public information of the most critical element which can lead to cardiovascular disease such that individuals can act quickly to prevent it.

Keywords: *Heart disease binary logistic regression machine learning data visualization accuracy.*

Received 21.02.2022 Revised 24.03.2022 Accepted 22.04.2022

INTRODUCTION

The human body is made up of several organs, each of which serves a specific purpose. As it pumps blood into our lungs, the cardiac is perhaps the most essential organ in the human body. The rib cage protects the heart, whereas the Pericardium tissue layer protects the skin. It is divided into 4 chambers by oxygenated and deoxygenated blood. Plaque accumulation inside the arteries, veins as well as capillaries causes heart disease. Plaque is indeed a lustrous substance that contains cholesterol, lipid molecular particles, as well as minerals [1]. The inner lining of an artery is damaged by increased blood pressure, smoking tobacco, and excessive cholesterol or fats. Medical data models are used to identify and forecast the presence of illnesses using machine learning. Logistic regression is a machine learning technique that is less well-known. Heart disease is the world's biggest cause of death. Widely, 17.7 million people are predicted to die from cardiovascular illnesses in the next two decades. Heart disease affects both men and women equally [2]. To test hypotheses concerning the links between result and predictor variables, logistic regression was utilised. In contrast to discriminant analysis, logistic regression does not require regularly distributed data. The use of logistic regression to predict a discrete result from a set of factors is very useful [3]. Group size, for example, might be constant, discrete, dichotomous, or mixed. The pattern of antidepressants at a tertiary care institution was predicted using a binary logistic regression technique. They discovered that female patients suffer from depression at a higher rate than male ones. The goal of this research was to figure out what variables affect whether or not someone gets cardiovascular disease. The strategy that will be employed in this investigation is a binary logistic regression model [4]. The aim of the research is to investigate if a patient's medical traits, like gender, age, chest pain, fasting blood sugar level, and other factors, suggest that they could be diagnosed with cardiovascular disease. Kaggle is used to choose a dataset containing the patient's medical history and features. This information might be used to predict whether or not the patient would develop cardiovascular disease. In order to determine if a patient is at risk of developing cardiovascular disease, we categorise them depending on 13 medical factors [5]. A lot of effort and study has gone into developing better and more accurate models for the Heart Disease Dataset in recent years. Python and machine learning libraries were used in our research. Many researchers have utilised a 10-fold cross validation on the complete data and published the result for illness detection in the case of medical data diagnosis, while others have not used this approach for heart disease prediction. We employed the test-train split concept in our work.

MATERIAL AND METHODS

Data set and data sources

The heart disease dataset, which comprised of 303 samples, was collected via the Kaggle website. The research had a dependent factor and thirteen independent factors, with the response variable implying that a value of '0' or '1' shows the non-existence or existence of cardiovascular disease, respectively. x1 (age), x2 (gender), x3 (cp), x4 (trestbps), x5 (chol), x6 (fbs), x7 (restecg), x8 (thalach), x9 (exang), x10 (oldpeak), x11 (slope), x12(ca), and x13 (thal) and the dependent variable y (existence of cardiovascular disease). The research must pinpoint the fundamental elements that influence the start of cardiovascular disease and forecast the risk of developing it [6].

Table 1. Data description of existence or non-existence of cardiovascular disease

Variables	Representation of Variables	Type
Y	Existence or non-existence of cardiovascular disease: 1 means existence 0 means non-existence	Dependent
x1	Age (Patient's Age)	Independent
x2	Gender: 1 means male; 0 means female	Independent
x3	Cp	Independent
x4	Trestbps	Independent
x5	Chol	Independent
x6	fbs: 1 means true; 0 means false	Independent
x7	Restecg	Independent
x8	Thalach	Independent
x9	Exang: 0 means no; 1 means yes	Independent
x10	Oldpeak	Independent
x11	Slope	Independent
x12	Ca	Independent
x13	Thalassemia	Independent

Training and Testing

The training phase pulls features (independent variables) from the dataset, whereas the testing phase (which contains dependent variables) determines how the suitable model performs in terms of prediction. The dataset has been separated into two parts. These are the stages of training and testing. We divided the dataset into two parts: 80 percent training and 20 percent testing. And we've assigned a number to the random state: 1. We employ the random state option to initialise the fixed internal random number generator, which determines how data is divided into train and test indices. Setting the randomized state to a fixed value guarantees that almost every moment the procedure is performed, the very similar series of non-specific numbers is generated [7].

Methods

This study uses the Logistic Regression machine learning technique, which can help practitioners or medical analysts effectively detect Heart Disease. This approach includes looking through journals, published articles, and fresh statistics on heart disease. The technique of the provided model serves as a framework. The method consists of a set of methods that convert original data to easily understandable data patterns for users. Depending on the approach employed, data preprocessing deals with missing values. Finally, we put the suggested model to the test, evaluating it for accuracy and performance using a variety of performance indicators. An effective Heart Disease Prediction System (EHDPS) has been created in this model. For prediction, this model employs 13 medical characteristics including chest pain, fasting sugar, blood pressure, cholesterol, age, and gender [8-10]. A statistical analysis technique for predicting a data value based on past data set observations is known as logistic regression. On the other hand, the binary response variable contradicts the normality constraints of traditional regression models. A logistic regression model states that the fitted probability of occurrence is an appropriate function of a linear function of the observed figures of the relevant explanatory variables. This method's key advantage is that it can produce a simple probabilistic categorization formula. The difficulty of LR to deal effectively with non-linear and interaction outcomes of descriptive variables is one of its shortcomings. LR comes in helpful when you need to anticipate the existence or non-existence of a feature or outcome depending on the figures of a set of independent features.

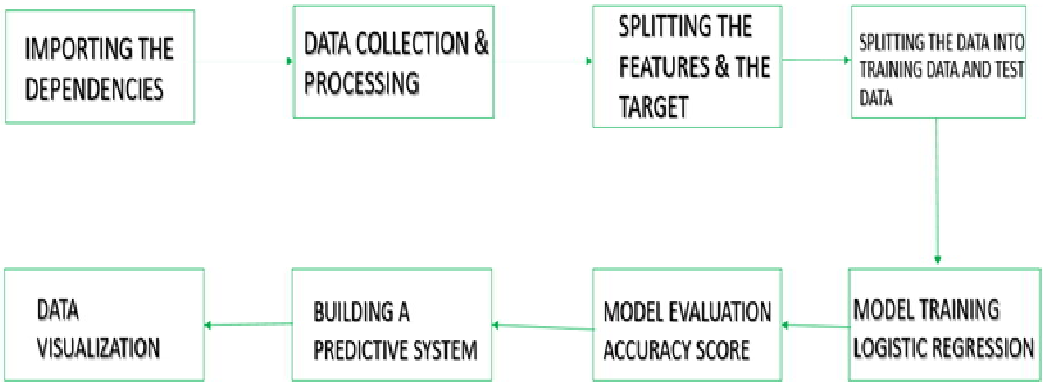

Fig.1 Flowchart of the methodology followed

The binary logistic regression analysis technique will be used in this investigation. At its most simplest form, logistic regression is a statistical model that describes a reliant variable using a logistic sigmoid function in the form of 0 and 1, however there are many more complicated forms. It looks into the link between a categorical variable and a group of independent factors. When analysing data with a reliant variable vs. a self-reliant variable, logistic regression is commonly utilised. The broad explanatory variable model proposed by

$$logi(y) = \ln(p/1-p) = \alpha + b1x1 + b2x2 + b3x3 + \ldots + e$$

The total % accuracy can be used to compare the method's accuracy. The formula for calculating the accuracy % is given below

Percentage of predictive frequency = (number of accurate data / total data) * 100

- IMPORTING THE DEPENDENCIES: Numpy (for numerical datasets), Pandas (for dataframes), and the sklearn model are all dependencies (for train-test split, logistic regression, and finding accuracy score).
- COLLECTION AND PROCESSING OF DATA: The csv data is being loaded into a Panda dataframe.
- SPLITTING THE TARGET AND THE FEATURES: The input and output characteristics are separated.
- SEPARATION OF THE INFORMATION INTO TRAINING AND TESTING DATA
- LOGISTIC REGRESSION MODEL TRAINING: At its simplest form, logistic regression is a statistical model that represents a reliant variable using a logistic sigmoid function. Using training data to train the Logistic Regression model.
- EVALUATION OF THE MODEL: The Correctness SCORE indicates the model's accuracy.
- BUILDING A PREDICTIVE SYSTEM: We enter a patient's data and forecast the patient's outcome.
- DATA VISUALIZATION: It depicts the statistical analysis of different patients based on their age, gender, cholesterol level, blood pressure, and other factors

RESULTS AND DISCUSSION

Multiple Patients Heart Disease Prediction

We used a live real-time dataset of individuals with cardiac disease. By producing a.csv file, we created a model to figure out the cardiac status of several patients at once. This method will be beneficial in hospitals. We employed data visualization to conduct a statistical study of the model's inputs.

Visualization of Data

The process of translating big data sets and observations into plots, scatter plots, and other visual depiction is known as data visualization. The resulting graphical representation of data makes it simpler to spot and convey genuine patterns, anomalies, and fresh insights into the data's content. A dashboard is a tool for visualizing data. It shows data on several sheets and panels in order to keep track of events rapidly. A visual tab provides contemporaneous data by collecting complicated data points from enormous data sets, as opposed to a visual representation, which shows a constant graphical image. An dynamic dashboard allows you to quickly organize, categorize, and delve into a range of data. Data science tools may assist you in swiftly determining what is occurring, why it is happening, and what will happen next. More people are using data visualisation tools on their PCs and mobile devices to acquire

insights as the volume of big data expands. Dashboards are used by business people, data analysts, and data scientists to make strategic business decisions.

Accuracy

Accuracy % result in Logistic Regression.

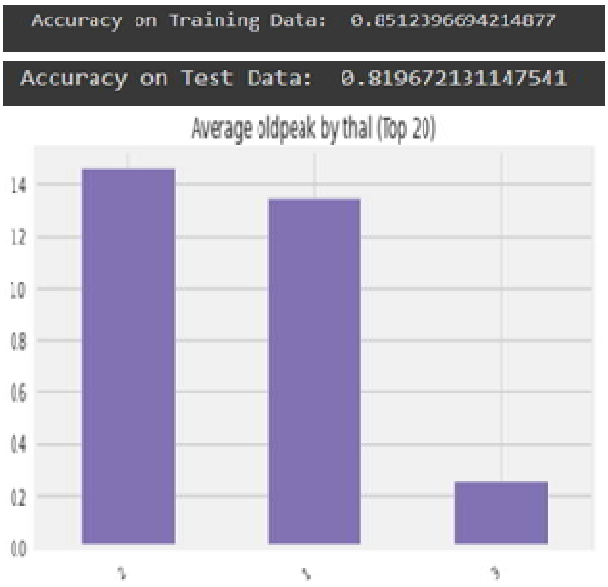

Figure 2. Average oldpeak by thal

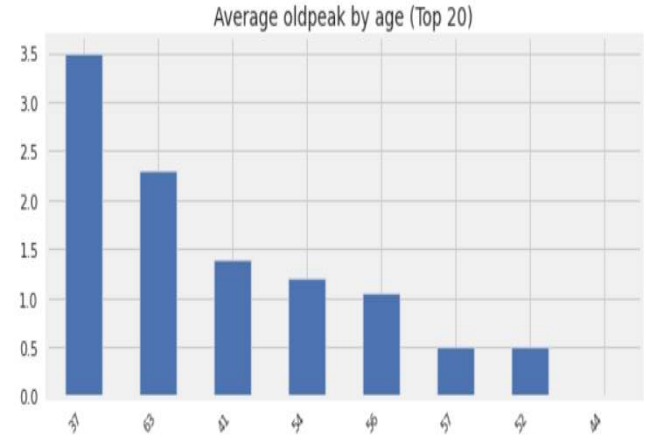

Figure 3. Average oldpeak by age

Figure 4. Accuracy comparison

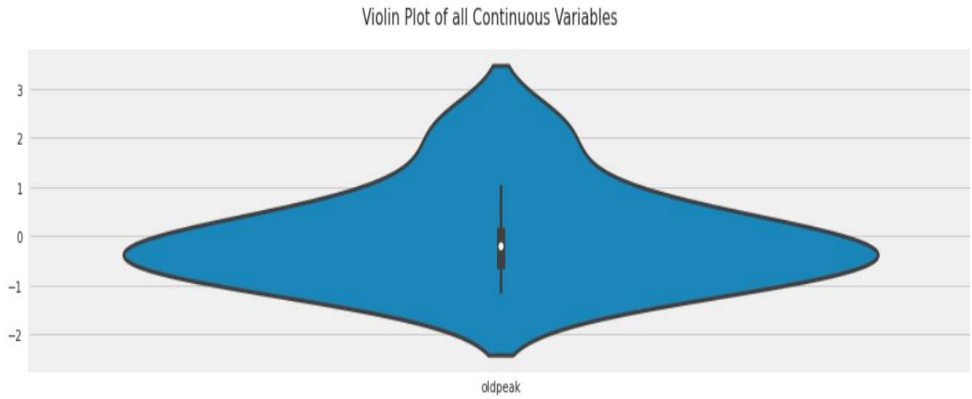

Figure 5. Violin Plot of all Continuous variables

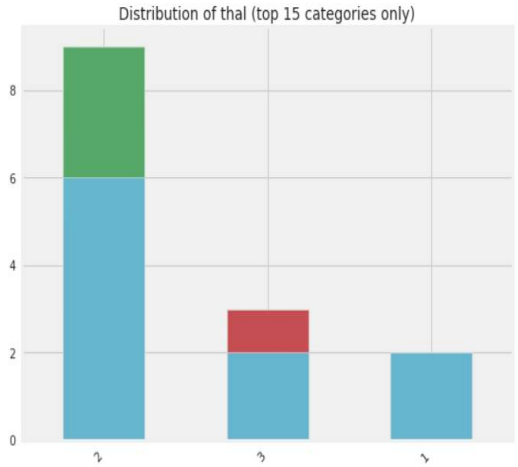

Figure 6. Distribution of thal

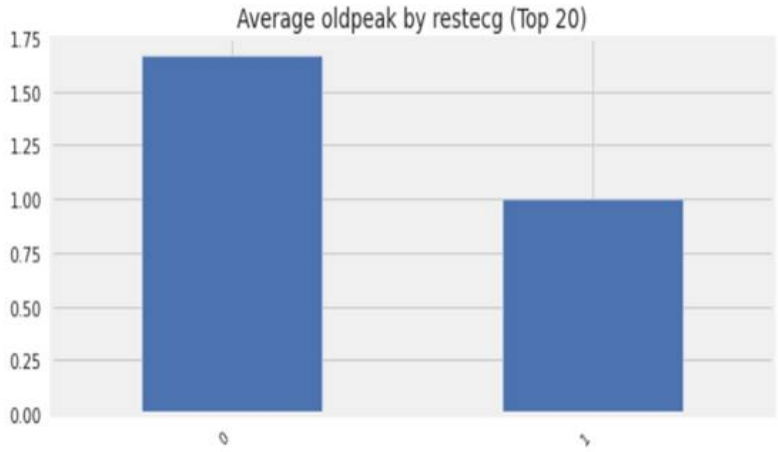

Figure 7. Average oldpeak by restecg

CONCLUSION

Finally, the study reveals that cardiovascular disease is directly proportional to x2 (gender), x3 (cp), x4 (trestbps), x5 (chol), x7 (restecg), x9 (exang), x10 (oldpeak), x12 (ca), and x13 (thal). According to the percentage accuracy value, the binary logistic model has a percentage accuracy of 85.12 percent. In hospital administration, this binary logistic regression model will be used to predict cardiac sickness in the coming time ahead. The great bulk of data in present scenario is digitised, distributed, as well as underutilised. We can also analyse the provided data to hunt for new trends. The study's main purpose is to come up with a mechanism for properly forecasting heart problems. To predict heart disease, we may use the logistic regression approach, often known as sklearn in machine learning. The paper's long-term goal is to use new techniques and algorithms to predict cardiac diseases in a timely way.

ACKNOWLEDGMENTS

We would like to convey our heartfelt thanks to Prof Seira Shinde, our research supervisor, for her invaluable counsel, patience, support, and direction over the course of our study. The research work is supported by Bhilai Institute of Technology, Raipur, Chhattisgarh.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. Nor Fatihah Zulkiflee, Mohd Saifullah Rusiman (2021). *Heart Disease Prediction Using Logistic Regression.* Enhanced knowledge in Sciences and Technology Vol. 1 No. 2 (2021) 177-184.
2. A Rajdhan, M Sai, A Agarwal, D Ravi (2020). Heart Disease Prediction using Machine Learning. International Journal of Engineering Research and Technology. DOI:10.17577/IJERTV9IS110259
3. M Diwakar, A Tripathi, K Joshi, M Memoria, P Singh, N Kumar (2020). Latest Trends on Heart Disease Prediction using Machine Learning and Image Fusion. Material and Today Proceeding Volume 37, Part 2, 2021, Pages 3213-3218
4. S K Mohan, Chandrasegar Thirumalai, Gautam Srivastava (2019). *Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques.* IEEE Access. **DOI:** 10.1109/ACCESS.2019.2923707
5. P Singh, Sanjay Singh, Gayatri H Pandi-Jain (2018). Effective Heart Disease Prediction System using Data Mining Techniques. International Journal of Nanomedicine. 13(T-NANO 2014 Abstracts) Pages 121—124
6. J. Vijayashree, N. Ch. Sriman Narayana Iyengar (2016). Heart Disease Prediction System using Data Mining and Hybrid Intelligent Techniques. International Journal of Bio-science and Bio-technology. Vol.8, No.4 (2016), pp. 139-148
7. Y Khourdifi, M Bahaj (2018). Heart Disease Prediction and Classification using Machine Learning algorithms optimized by Particle Swarm Optimization and Ant Colony Optimization. International Journal of Intelligent Engineering & Systems. DOI: 10.22266/ijies2019.0228.24
8. Maryam I. Al-Janabi, Mahmoud H. Qutqut, Mohammad Hijjawi (2018). *Machine Learning Classification Techniques for Heart Disease Prediction: A Review*. International Journal of Engineering and Technology, 7(4) 5373-5379.
9. Ea Maini, B Venkateswarlu, B Maini, Dj Marwaha (2021). Machine Learning-based heart disease prediction system for Indian population: An exploratory study done in South India. Elsevier. Med J Armed Forces India ;77(3):302-311doi: 10.1016/j.mjafi.2020.10.013.
10. M Ahmad, M Alfayad, S Aftab, M Adnan Khan, Areej Fatima, B Shoaib, M Sh. Daoud, Nouh Sabri Elmitwally (2021). *Data and Machine Learning Fusion Architecture for Cardiovascular Disease Prediction.* Computers, Materials and Continua. 69(2):2717-2731, DOI:10.32604/cmc.2021.019013

CITATION OF THIS ARTICLE

S Shinde, T B Kapse, M Ramwani, S Singh, S Sahu. Heart Disease Prediction Using Real Dataset via Effective Machine Learning Technique. Bull. Env.Pharmacol. Life Sci., Spl Issue [1] 2022 : 909-914