International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

PREDICTION OF HEART DISEASE USING LOGISTIC REGRESSION

Kavya Sreehari[1], Devika Santhosh Kumar[2], Muhammed Shameem S[3], Sumeesh S[4], Bismi M[5]

[1],[2],[3],[4] UG Scholar, Department of Computer Science and Engineering,
[5]Asst. Prof, Department of Computer Science and Engineering,
UKF College of Engineering and Technology, Kerala, India

Abstract: *Heart disease symptoms are caused by abnormal heartbeats and diseased heart muscle. There are two cases of prediction of which the correct prediction can help to prevent threats whereas incorrect prediction can lead to fatal. Here, the paper is based on heart disease prediction using logistic regression model. Machine Learning is one the trending technology in which various researches around the world is used for predicting diseases. Nowadays it's important for the early detection and for its treatment.The dataset consists of 14 attributes used for performing the analysis. Accuracy is validated and promising results are achieved. Heart disease dataset analysed is used to predict the result whether the patient has heart disease or not i.e., using logistic regression technique. This prediction gives the result in the form of logistic representations which produces efficient and accurate results in healthcare sectors.*

Keywords: ***Heart disease, Machine learning, Logistic regression***

1. INTRODUCTION

According to WHO, 17.9 million people year die due to heart related diseases. There are different types of heart disease, some are preventable. Heart disease refers to any condition affecting the heart and blood vessels. There are different types of heart disease affecting the heart and blood vessels in different ways. Early detection of cardiovascular disease may help to avoid complications. Unhealthy lifestyle habits are reason for increase in heart related diseases diet high in saturated fats, trans fats, and cholesterol has been linked to heart disease and related conditions, such as atherosclerosis. Excess salt in the diet can increase blood pressure. Some risk factors for heart disease can be controlled whereas some cannot be controlled, for example family history.

One of the main risk factor for heart disease is high blood pressure. It is often called a "silent killer" as it usually shows no symptoms. Early detection of heart diseases is required to reduce the health complications. In healthcare sector machine learning has been used in diagnosing and predicting various diseases using different models. The study intends to find the most important predicators of heart diseases and predicting the overall risk by using logistic regression. Healthcare expenses are overwhelming national and corporate budgets due to asymptomatic diseases including heart diseases. Therefore, there is an immediate need to detect and treat such diseases.

Coronary heart disease is also known as ischemic heart disease. Plaque build-up in the walls of the arteries leads to coronary heart disease. Some of the symptoms of heart disease include pain in the chest, breath shortness, and discomfort in the arms, back, neck, jaw, or stomach.

2. RELATED WORKS

Machine learning is widely used in almost all fields including healthcare sector. Most machine learning algorithms alive concerned with discovering interrelationship between datasets. Once Machine Learning Algorithms can identify on certain correlations, the model can either use these relationships to predict future observations or generalize the info to reveal interesting patterns. In Machine Learning there are various types of algorithms like Regression, Linear Regression, , Naive Bayes Classifier, Bayes theorem, Logistic Regression KNN (K-Nearest Neighbour Classifier), Decision Tress, Entropy, ID3, SVM (Support Vector Machines), K-means Algorithm, Random Forest .Lots of research work have been done for assessment of the classification accuracies of different machine learning algorithms by using the Cleveland heart disease database which is uninhibitedly accessible at an online data mining repository of the UCI. Authors Bayu Adhi Tama, Afriyan Firdaus, Rodiyatul

FS in their work suggested a research related to the identification of diabetes malady with utilization of ML procedures. This disease was viewed as incredibly a thrust area of ML. Roughly 285 million individuals around the globe were experiencing diabetes as per a study directed by International Diabetes Federation (IDF). As a matter of fact, detection of type 2 diabetes at beginning phase isn't a simple undertaking, yet research done by the authors, in which data mining was used on the grounds that it gives the best results, helped in the disclosure of information from accessible data. In their research, they utilized SVMs for the mining of related information of various patients from the previous records..

3. PROPOSED SYSTEM

The proposed system has datasets consist of 14 main attributes for the prediction of heart disease. The system is based on logistic regression method in which efficient algorithms are utilized for heart disease prediction which helps to prevent at the earlier stage. In this system, the data is collected and then converted it into knowledge by data analysis. This proposed system consists of data which determines whether the patient has heart diseases or not with the help of some parameters. Here, we use the sources for datasets from Kaggle. The proposed system consisting of sklearn library which helps in testing and training of the datasets. sklearn also known as Scikit-learn is probably most useful library for implementing machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modelling including classification, regression, clustering and also dimensionality reduction. Logistic Regression is a type of regression analysis which is used in statistics to predict the outcome of a categorical dependent variable from a collection of predictors or independent variables. In logistic regression the dependent variable always represented in binary. Logistic regression method mainly in prediction and calculation of the probability of success. The proposed system splits the data into two parts one for testing and the other for training using sklearn library functions imported. The logistic regression model to call the labels by logistic model and use the accuracy function to predict the labels and find the accuracy of the model. Out of the 14 attributes 13 are of integer data type and the other 1 is of floating data type. Finally, analyses the results with the help of various Comparing Models. The data we are having, is classified into different structured data based on the features of the patient's heart. From the available data, we need to create a model that predicts the patient's disease using a logistic regression algorithm. First, we have to import datasets and then read the datasets; the data should contain different variables such as age, gender, sex, chest pain, slope, target, resting blood pressure, thalach etc. The data need to be explored so that the information gets verified. After that create a temporary variable and also build a model for logistic regression. By using logistic regression, the accuracy is increased as compared to the other works done on the existing system.

3.1. LOGISTIC REGRESSION

Regression is a method for finding the relationship between independent variables and a dependent variable. It is used as a method for predictive modelling in which an algorithm is used to predict continuous outcomes.

The curve from the logistic function indicates the likelihood of something like a mouse is obese or not based on its weight, the cells are cancerous or not, etc.

Types of linear regression:

a) Binary logistic regression

b) Multinomial logistic regression

c) Ordinal logistic regression

A logistic regression model can consider different input criteria. The logistic function can take into consideration different factors such as the student's grade point average, SAT score and number of extracurricular activities. It then scores new cases on their chances of two outcome categories.

Logistic regression has become a vital tool in machine learning. It enables the algorithms that are used in machine learning applications to classify data's based on data history. As additional data comes in, the algorithms improve at predicting classifications within data sets.

This model has good accuracy for simple datasets and it performs well when the dataset is linearly separable. It requires less training. It is also less inclined to over-fitting but it can over-fit in high dimensional datasets.

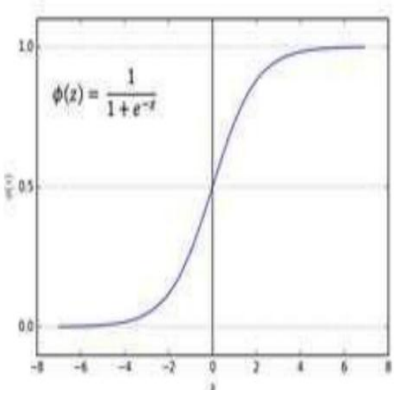

FIGURE 1: Logistic Regression

3.2 WORKFLOW

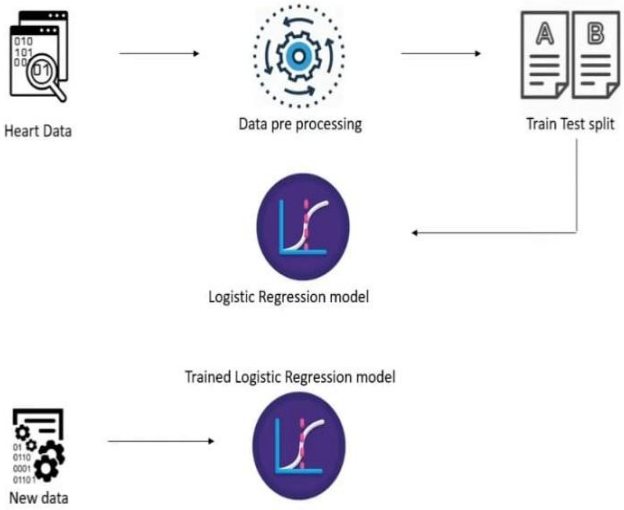

FIGURE 2: Workflow

The workflow first consists of a several health parameters in which datasets are available for the 14 attributes. Then we need to process this dataset using machine learning algorithms since pre-processing is need as it helps in cleaning the data. Then the datasets are splitted into two. They are training data and testing data. We use 80-20 split, in which 80% of data is used for training and 20% is used for testing the data. Here we use, x_train and y_train which contains the training data and their corresponding labels. Similarly, it contains x_test and y_test for testing the data and their labels which is done through sklearn library. Since the model is based on logistic regression which is uses binary classification, it predicts the output as either 1 or 0. The predicted output that represents 0 shows that the person has no diseases. Whereas the output showing 1 represents that the person is affected by heart disease. Here that is represented by a target table. Every dataset consists of some values i.e., it will not represent null values. The pre-processing stage plays an important role when passing the data for prediction. The distribution of various parameters like age and sex, cholesterol and fasting blood, ECG resting and thalach, exang and old peak, slope and ca etc are well analyzed. This analysis helps in correct prediction of heart disease as it results either 0 or 1.

Then this training data is passed to a machine learning model, here the model is logistic regression model in which this model uses binary classification. The model is used to classify whether a person having heart diseases or not. Once we train the logistic regression model, then some evaluations are done to check its performance. Then a trained logistic regression model is obtained. To this model we feed new data and performance is evaluated i.e., whether it predicts heart disease or not.

4. RESULTS AND ANALYSIS

The Logistics Regression increase its accuracy with increasing training by 50% to 90% and 90% training and 10% testing provides highest accuracy of 81%. because the behavior of Logistic regression is as training increases the accuracy of prediction also increased. The result obtained is quite promising. There are several techniques and methods are present for prediction of disorder. to enhance 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 concentrate on the behavior of the algorithm.

5. CONCLUSION

The aim of this study is to predict the heart disease with the help of logistic regression using 14 attributes. These attributes are selected after the backward elimination process. The logistic regression model analysis depicts that men are more suspectable to heart diseases than women. The total cholesterol level and glucose level has no significant changes. Further the accuracy of the model is also evaluated and it produces efficient accuracy rate which is satisfying. Heart diseases are severe and every year lots of people are dying due to the late detection of heart diseases. Any non-medical employee is capable of using this software and predicts the heart disease and reduce the time complexity. This is an open domain for predicting the heart disease with satisfying accuracy. The model is more specific than sensitive. Logistic regression is a good performance machine learning model to predict the risk of major chronic diseases with simple clinical predictors.

REFERENCE

[1] V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," 2020, 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842.

[2] Mozaffarian, D., Benjamin, E., Go, A., Arnett, D., Blaha, M. Cushman, M. et al. (2015).Heart Disease and Stroke Statistics—2015, Update. Circulation, 131(4). doi:10.1161/cir.0000000000000152.

[3] Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of Artificial Intelligence inMachine Learning: Review and Prospect. International Journal of Computer Applications,115(9), 31-41. doi: 10.5120/20182-2402.

[4] Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. (2019). Applications of ArtificialIntelligence in Transport: An Overview. Sustainability, 11(1), 189. doi: 10.3390/su11010189.

[5] Strecht, Pedro & Cruz, Luís & Soares, Carlos & Moreira, João & Abreu, Rui. (2015). A

Comparative Study of Classification and Regression Algorithms for Modelling Students'AcademicPerformance.https://www.researchgate.net/publication/278030689_A_Comparative_Study_of_Classification_and_Regression_Algorithms_for_Modelling_Students'_Academic_Performance, viewed:10th June 2019

[6] Avinash Golande, Pavan Kumar T," Heart Disease Prediction Using Effective MachineLearning Techniques", International Journal of Recent Technology and Engineering, Vol 8,pp.944-950,2019.

[7] T. Nagamani, S. Logeswari, B. Gomathy," Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and ExploringEngineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[8] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict HeartFailure Disease", (IJACSA) International Journal of Advanced Computer Science andApplications, Vol. 10, No. 6, 2019.

[9] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin," Design andImplementation Heart Disease Prediction Using Naives Bayesian", International Conferenceon Trends in Electronics and Information (ICOEI 2019).