IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

Clinical Support System for Predicting Heart Diseases using Machine Learning Techniques

Prof. Vikram Chavan[1], Arpit Bisane[2], Aishwarya Jadhav[3], Bholeshwar Choudhary[4]
Professor, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2,3,4]
Sinhgad Institute of Technology, Lonavala, Maharashtra, India

Abstract: *Heart disease remains a leading global cause of mortality, and accurate prediction poses challenges for clinicians due to its complexity and cost. In this study, we propose a clinical support system for heart disease prediction to aid clinicians in diagnostics and decision-making. Machine learning algorithms, including Logistic Regression, Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine, are applied to risk factor data obtained from medical records. Through experiments conducted on the UCI dataset, Logistic Regression demonstrated superior performance with an accuracy of 91.2% using train-test split techniques. Furthermore, we recommend future validation of our proposed system using prospectively collected data. The findings of this study have the potential to improve heart disease prediction and contribute to more informed clinical decision-making.*

Keywords: Heart Disease, Machine Learning, Logistic Regression, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine

I. INTRODUCTION

According to the world health organization (WHO), cardiovascular diseases (CVDs) are the first cause of death worldwide, with more than 17.9 million people died in 2016[1]. CVDs are a group of syndromes affecting blood vessels and heart; they include heart disease (HD), which is often expressed as coronary heart disease [2]. However, HD can be prevented by observing a healthy lifestyle and avoiding risk factors. Thus, understanding what is contributing to these alarming factors may help for the prevention and prediction of HD. Typical, angiography is the primary diagnosis method; it is used to determine the localization of heart vessels' stenosis. Being costly, time-consuming, and invasive had motivated researchers to develop automatic systems based on information gathered through a set of medical data, such as data from past treatment outcomes as well as the latest medical research results and databases [3].

II. LITERATURE SURVEY

Most research have used the UCI heart disease data set due to its availability [4]. This data set contains four sub data set and 76 attributes; the number of selected attributes and common features used in each study is ranging from 76 to 8, including the class attribute.

Various prediction models were built using well-known ML techniques. The author [5] suggested a predictive model using C4.5 and fast decision tree algorithms applied on the four collected and separated UCI data sets; this model achieved an accuracy of 78.06% and 75.48% for C4.5 and fast decision tree respectively using only Cleveland data set. The author [7] predicted HD using the meta-algorithm Ad boost on Cleveland data set and suggested reducing the number of attributes from 76 to 28 to provide higher accuracy of 80.14%.

Despite a substantial research output, no gold-standard model is available to predict HD. Hence, there is still a need for improvement. Also, many parameters impact the construction of the HD prediction model; these include the data set of choice, the number of attributes and the output class, and the algorithm used.

The objective of this project is to develop a clinical decision support system for predicting the risk level of heart disease (HD) using the UCI Cleveland data set. A classification model is proposed to identify patterns in the data of existing HD patients. The following section outlines our methodology, including a brief description of the data set utilized.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

III. PROPOSED METHOD

Figure 1 illustrates the proposed architecture of the model for predicting the presence of HD. The initial step of the process involves pre-processing to handle missing values. Machine learning algorithms are then utilized to predict whether a patient has HD or not. The performance of the algorithms is evaluated using performance measures based on the confusion matrix.

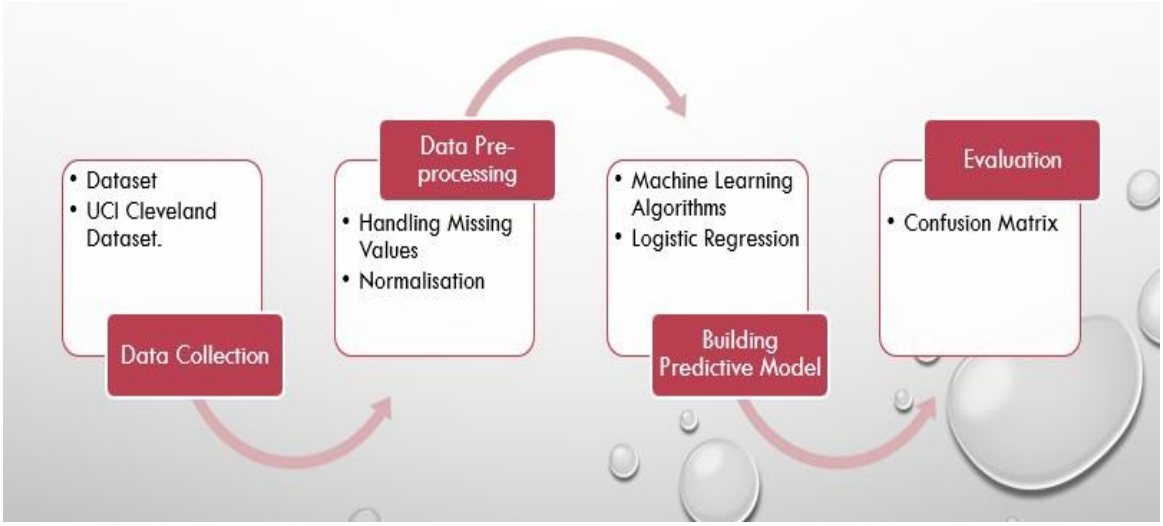

Figure 1 Flow Chart of our model for prediction of HD

IV. DATA PREPARATION

Data preparation, as a critical first step in any predictive model, is carried out to transform data into a format that enhances model efficiency. Medical data are often incomplete, with missing attribute values and noisy due to the presence of outliers or irrelevant data.

The predicted attribute (target) in the original dataset contains two values, where a value of 0 indicates the absence of HD and a value of 1 represents the presence of HD.

4.1 Data Classification

Predictive modeling involves building a model that can make accurate predictions based on observed data in a training dataset. In this study, we have a specific target variable that indicates the presence or absence of heart disease, and therefore a supervised learning classification algorithm is appropriate for training the data. We have utilized the following machine learning algorithms (Logistic Regression, Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine) to build our proposed model.

Logistic Regression is used to obtain odds ratios when dealing with multiple explanatory variables. The procedure is similar to multiple linear regression, but the response variable is binomial [11].

Naïve Bayes (NB) is a probabilistic classifier that applies Bayes' Law and makes the assumption of naïve conditional independence. This means that the presence or absence of one particular feature in a class is assumed to be unrelated to the presence or absence of any other feature. Thus, the status of one feature does not affect the status of another feature. NB is known for its robustness and is commonly used for classification purposes [8].

K-Nearest Neighbor (KNN), also a supervised learning model, is used to classify test data based on the training samples directly. It is a method that classifies objects based on the closest training data in the feature space. The class of a new sample is predicted based on distance metrics, such as Euclidean distance. In practical steps, KNN calculates the k (number of nearest neighbors), finds the distance between the training data, and then sorts the distances. The class label for the test data is assigned based on majority voting.

Support Vector Machine (SVM) is a widely used supervised learning model for classification tasks. It operates in finite-dimensional vector spaces, where each dimension represents a feature of a sample. SVM aims to find the best

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

hyperplane with the highest margin that separates the two classes. SVM has shown effectiveness in handling high-dimensional datasets due to its computational efficiency [8].

4.2 Performance Measurement

To assess the validity of the predictive model, several performance measurements such as sensitivity, specificity, accuracy, and precision can be calculated using a confusion matrix. Specificity measures the proportion of negatives that are correctly identified, while sensitivity measures the percentage of true positives that are correctly identified. These measures can be mathematically expressed using the following formulas, where TP, TN, FP, and FN represent True Positive (the number of positive data correctly labeled by the classifier), True Negative (the number of negative data correctly labeled by the classifier), False Positive (the number of negative data incorrectly labeled as positive), and False Negative (the number of positive data mislabeled as negative), respectively.

4.3 Training and Testing

Finally, this resulting data split into 70% train and 30% test data, which was further passed to the Logistic Regression model to fit, predict and score the model.

V. CONCLUSION

This study has been conducted to help clinicians to produce an accurate and efficient predictive system; the model validation is conducted with the train-test split of data. The results showed that LR achieved the highest accuracy compared to other algorithms. We believe that this technique is the best in our model since the used dataset is not large, so the process didn't take a long time.

The results were significant, and we believe that the achieved results using our predictive model based on ML algorithms could improve the knowledge on the prediction of heart disease risk through better diagnoses and interpretation; therefore, appropriate clinical decisions.

REFERENCES

[1]. URL:http://who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
[2]. URL: http://nhlbi.nih.gov. National heart, lung, and blood institute
[3]. N. Mishra and S. Silakari "Predictive Analytics: A Survey, Trends, Application, Opportunities and Challenges," International Journal of computer science and information technologies, vol 3(3), pp. 4434-4438, 2012.
[4]. H. Alharti. "Healthcare predictive analytics: An overview with a focus on Saudi Arabia," Journal of Infection and Public Health, vol 11(6), pp. 749-756, 2018..
[5]. R. El-Bialy, M. A. Salamay, O. H.Karam, & M.E. Khalifa. "Feature Analysis of Coronary Artery Heart Disease Data Sets". International Conference on Communication, Management and Information Technology. Procedia Computer Science, vol 65, pp. 459-468, 2015
[6]. L. M. Hung, D. T. Toan, & V. T. Lang. "Automatic Heart Disease Prediction Using Feature Selection and Data Mining Technique,". Journal of Computer Science and Cybernetics, vol 34(1), pp. 33-47, 2018.
[7]. K. H., Miao, J. H. Miao & G. Miao. "Diagnosing Coronary Heart Disease Using Ensemble Machine Learning," International Journal of Advanced Computer Science and Applications, vol 7(10), 2016.
[8]. Kononenko; " Inductive and Bayesian learning in medical diagnosis," Applied Artificial Intelligence, vol 7(4), pp. 317-337, 1993.
[9]. N. S. Altman. "An introduction to kernel and nearest-neighbor nonparametric regression," The American Statistician, vol 46(3), pp. 175–185, 1992.
[10]. C. Cortes & V. Vapnik. "Support-vector networks," Machine Learning, vol 20(3), pp. 273–297, 1995.
[11]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/