Volume 9, Issue 3, March – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2016

Heart Disease Prediction Using ML

M.Sangeetha
Assistant professor
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil, Virudhnagar, India

S.Arun Kumar
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil, Virudhnagar, India

K. Pazhani Bharathi
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil, Virudhnagar, India

P .Kumara Guru
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil, Virudhnagar , India

P.Bhuvan Prakash Reddy
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil, Virudhnagar, India

Abstract:- Machine Learning and artificial intelligence have found valuable on variety of disciplines during their growth, particularly in the light of massive increase in data in recent years. It has the potential to be more dependable in terms of producing quicker and more accurate illness prediction judgments. Therefore, the use of machine learning algorithms to forecast different diseases is growing. Building a model can also aid in the visualization and analysis of diseases to increase the accuracy and consistency of reporting. This article has looked into using several machine learning algorithms to identify cardiac disease. This article's study has demonstrated a step procedure. In a dataset on heart disease initially prepared in the format needed to run machine learning algorithms. The UCI is the source of patient medical records and other data. The presence are absence of heart disease in patients is then ascertained using the heart disease dataset. Second, this paper presents a number of noteworthy findings. The confusion matrix is used to validate the accuracy rate of machine learning methods, including Gradient Boosting Classifier, Support Vector Machine, and Logistic Regression. According to recent research, the Logistic Regression method outperforms other algorithms in terms of accuracy, yielding a high 95% rate. It also outperforms the other four algorithms in terms of recall, precision, and f1-score correctness. The difficult and future research component of this project will be raising the accuracy rates of the machine learning algorithms to between 97% and 100%.

Keywords:- *Machine Learning, Artificial Intelligence, Heart Disease , logistic Regression, KNN, Support Vector Machine.*

I. INTRODUCTION

Artificial intelligence (AI) includes machine learning (ML), which makes it possible for a software program to increase its prediction accuracy without having to be explicitly coded. Machine learning algorithms employ historical data as input to forecast new output values . The area of machine learning is vast and diverse, with daily advancements in both its use and breadth. This is why many firms now view machine learning as a critical differentiator in the competition. To forecast and determine the correctness of a dataset, machine learning uses ensemble learning, unsupervised learning, and supervised learning classifiers. To reach a conclusion or make a prediction, machine learning algorithms can create a model using train data, which is a sample of data. The current study examines the application of machine learning techniques in the medical field, with a primary focus on imitating certain human behaviors or mental processes and diagnosing illnesses based on a range of inputs. A collection of disorders affecting the heart are collectively referred to as "heart diseases." Globally, cardiovascular illnesses account for 18.9 million deaths annually, according to data from the World Health Organization .

To identify cardiac disorders, numerous kinds of study have been investigated and carried out using different machine learning algorithms. Ghumbre et al. claim that the UCI dataset is used to forecast heart disorders using machine learning and deep learning technicqes. The authors came to the conclusion that this analysis was the better served by machine learning methods. The article by Rohit Bharti et al. on machine learning techniques for heart disease prediction came to the conclusion that distinct data mining and neural systems should be employed to determine the severity of HD among patients . The application of a predictive data mining technique on the same dataset has been the subject of some

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2016

investigation . In this article, a machine learning predictive model will be developed to help analyze heart disease in connection to medical history. Medical records and patient details are obtained from the UCI repository. This dataset will be utilized to determine if the patients have heart disease. In order to diagnose the HD dataset, this article considers 14 patient factors. The biggest problem facing the medical sector today is providing higher-level infrastructure and facilities to diagnose diseases early and treat patients promptly in order to enhance patient outcomes and quality of life. Worldwide, heart disease accounts for about 32% of deaths. Lack of infrastructure, doctors, and technology in emerging and underdeveloped nations makes it difficult to anticipate diseases in their early stages, prevent complications, and lower mortality rates. The development of information and communication technology has helped patients of all income levels by giving them access to real-time information at a lower cost of diagnosis and health monitoring. This has significantly increased the patients' detailed medical records. Researchers can access the extensive medical records. It helps diagnose diseases with less medical procedures by classifying the disease's existence or absence . The study's patient characteristics, such as age, sex, blood pressure, serum cholesterol, exang, etc., are taken into consideration in this article. The primary cause of death worldwide is heart disease, also referred to as cardiovascular disease (CVD). Reducing death rates and minimizing major consequences need early detection and management. Here's where machine learning (ML) comes into play, providing a formidable method to forecast heart disease risk and maybe save lives. In the medical field, machine learning is vital. Machine learning allows us to identify, track, and forecast a wide range of illnesses. Predicting the chance of contracting specific diseases through data mining and machine learning approaches has gained popularity recently. Applications of data mining techniques for disease prediction are found in the already published work. While some research has tried to forecast the likelihood that the disease would proceed in the future, they have not yet produced reliable findings. This paper's primary objective is to provide precise predictions about the likelihood that a person may develop heart disease.

II. METHODOLOGY

The methodology and analysis used in this research project are explained in this section. First and foremost, the study's first steps involve gathering data and choosing pertinent qualities. Subsequently, the pertinent data is preprocessed into the necessary format. Next, the provided data is divided into training and testing dataset categories. The model is then trained using the provided data and the algorithms. The testing data is used to determine this model's correctness. Many modules, including data collecting, attribute selection, pre-processing, data balancing, and disease prediction, are used to load the operations of this investigation. Furthermore, because it is simpler to comprehend and analyse the relationship between the input variables and the output classes, binning—which turns continuous information into categorical input—can also help improve the interpretability of the results. However, using continuous input, such numerical values, in classification algorithms can be more challenging because the algorithm may need to make assumptions about where to draw the boundaries between various classes or categories. Feature Selection and Reduction. To improve the interpretability and effectiveness of classification algorithms, we suggest using binning as a technique for turning continuous input, such age, into categorical input. Based on particular values of the input variables, the algorithm can distinguish between different classes of data by grouping continuous input into discrete groups or bins. A classification algorithm can use this information, for example, if the input variable is "Age Group" and the potential values are "Young," "Middle-aged," and "Elderly," to divide the data into several classes or categories depending on the age group of the persons in the dataset. In addition, other continuous-valued parameters like weight, ap hi, ap lo, and height were also transformed into category values. The study's findings show that classification algorithms might function better and be easier to understand when continuous data is binned into categories. The patients' comprehensive medical records have expanded dramatically as a result. The vast medical records are available for researchers to access. By categorising an illness's presence or absence, it facilitates disease diagnosis with fewer medical procedures . This article takes into account the study's patient characteristics, including age, sex, blood pressure, serum cholesterol, exang, etc. Heart disease, often known as cardiovascular disease (CVD), is the leading cause of death globally. Early detection and management are necessary to lower death rates and lessen serious consequences. This is where machine learning (ML) enters the picture, offering a powerful way to predict the risk of heart disease and perhaps save lives. Participants in this extensive study of US citizens without a history of clinical cardiovascular disease (CVD) had a significant lifetime risk of the disease, which was further increased for those who were overweight or obese. Obese adults were found to have a worse overall survival rate, a larger proportion of life spent with CVD morbidity (unhealthy life years), and an earlier onset of incident CVD compared to those with a normal BMI . This implies that the characteristics of height

A. Data Collection

The datasets used in this paper are gathered from the UCI repository and was taken into consideration by several authors throughout research analysis. In order to forecast heart disease, the initial stage involves organizing the information from the UCI repository and splitting it into two sections training and testing. In the training dataset comprises 80% of the data, whereas the testing dataset is used for analysis.

B. Analyzing Vast Data

A vast amount of patient data, including lab results, radiology scans, demographics, lifestyle factors, and medical history, may be analyzed by ML algorithms. This data has hidden patterns that conventional approaches might overlook.

C. Data Set and Attributes

A dataset's attributes are its characteristics that are crucial to examine and forecast in relation to our issue. To anticipate diseases, a number of patient characteristics are

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2016

taken into account, including gender, chest discomfort, fasting blood pressure, exang, etc. On the other hand, selecting attributes for a model can be done using the correlation matrix.

D. Preprocessing of Data

To get precise and flawless findings, we must clean and eliminate the noise or missing values from the dataset—a process called data cleaning. Python 3.8 offers several standard methods that we can use to fill in missing and noisy values, as shown in . Next, we must modify our dataset by taking into account its aggregation, generalization, smoothing, and normalizing. Several factors are taken into consideration in order to integrate, which is one of the most important stages in the pre-processing of data. The dataset can occasionally be more complicated or challenging to comprehend. The best course of action in this situation is to reduce the dataset to the necessary format.

III. LITERATURE ANALYSIS

Through skin electrodes, the ECG records from the electrical activity of the human heart in a variety of wave shapes. It is a noninvasive method of identifying heart disease that takes cardiac health, heart rate, and pulse into account. The human body's cell count is not in direct contact with the environment. Additionally, they rely on the circulatory system to provide them with transportation. Two types of fluids circulate through the cardiovascular system. The first kind of fluid is blood. The heart and blood vessels are formed here by the circulatory system. The second kind of fluid is lymph. Lymph nodes and lymphatic veins make up the lymphatic system's structural elements. Create HDPM in order to offer high prediction accuracy, the existence or absence of heart disease, and the patient's current status. A flow chart depicting the HDPM development process is displayed in Figure 4. First, compile the heart disease datasets. Second, data preparation is done in order to change the data. Third, use the DBSCAN technique, which is based on outlier identification, to identify the outlier data that the ideal parameters supply . Fourth, eliminate the identified outlier data from the training dataset. Fifth, use the SMOTE-ENN based data balancing technique to balance the training dataset. Sixth, to use MLA based on XGBoost to generate HDPM and learn from the training dataset. Finally, performance measures are offered for assessing the model performance that has been presented. HDPM is generated within the CDSS . To prevent overfitting, a 10-fold cross-validation method is used in this work. Cross-validation enables the models to learn from several training datasets by means of iterative sampling. Consequently, data maximizing is applied to validation, aiding in the prevention of overfitting. The 10-cross field validation technique will be employed to preserve the bias variance trade-off, which eventually yields a generalized model and guards against overfitting, as previous research has shown .Figure displays the suggested heart disease prediction model that makes use of XGBoost.

IV. MACHINE LEARNING ANALYSIS

In the following analysis will compare between 4 different Classifications models Logistic Regression, KNN, SVM, and XGBOOSTR in terms of prediction the Heart Disease. Where I am going to use the following techniques to help me in developing robust models . Standard Scaling, Cross-validation method, Grid Search, Metric measurements ssuch as accuracy, precision, F1 score etc. All the models provide very good prediction results. Based on these models which model give the accurate prediction with the less time taken that model is chosen and train the data using that model.

In trems of simplicity we can say Logistic Regression provided high predictive result and at the same time it is the simplest and fastest model in terms of parameters and training but if we lock to other models like KNN it providing the best result but it slower in terms of prediction process because it is required to calculate the distance between all the points in the dataset to classify every single point.

V. K-NEAREST NEIGHBORS (K-NN)

The technique that uses extreme vectors is referred to as a support vector machine, and the extreme vectors themselves are termed support vectors. This SVM graphic below uses decision boundaries or hyperplanes to classify two distinct categories. (x2, x1) is training sample dataset, where x2 was the target vector and x1was the x-axis vector. The simplest classification algorithm, K-NN, is based on supervised learning methods. Though it is mostly utilized for classification, the K-Nearest Neighbour approach were used for regression. The K-NN technique is used to classify incoming data point based on how similar the stored old data is. It suggests that when new data enters a relevant category, the K-NN algorithm can classify it swiftly.

VI. CONCLUSION

Here, the horizontal x-axis and vertical y-axis are independent and dependent variables of a function, respectively. Figure 4 is a simple example of the K-NN classification algorithm. The test sample (Yellow Square with what symbol) should be classified as either a green triangle or a red star in this algorithm. When k=3 is considered in a small dash circle, the yellow square would be a green triangle because the majority number in this region is green triangles, not red stars. Now, if we consider k=7, which is in a large dash circle, then the yellow square would be red stars because the number of red stars is four and the green triangles are 3. So, It can conclude that the majority vote in a specific region is important here Although the heart is an essential organ in the human body, heart illness becoming more and more common worldwide, which is a serious worry. Thus if we have a model that can forecast the early stages of cardiac disease, we can manage this illness. Therefore we need to develop a machine learning model that are more precise and aid in the cost-and doubt-free diagnosis of cardiac disease. It may serve as the main method for determining the state of the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2016

heart. This is the reason that this article concentrates on the accuracy rate for the confusion matrix in predicting heart disease. In accordance with this notion, the statistics of the provided algorithms are utilized to validate the statistics among the machine learning algorithms and estimates accuracy rate of the confusion matrix. After comparing the five methods, it is discovered that the Logistic Regression algorithm performs well in terms of accuracy rates . The 95% accuracy of the Logistic Regression model suggests that in the near future, machine learning algorithms will be regarded as predetermined tools for the detection of cardiac problems. For Logistic Regression, additional statistics including the f1-score, recall, and precision rate have been computed as 95%, 95%, and 95%, respectively. These estimated numbers point to this algorithm's maximum accuracy. These results imply that machine learning algorithms are capable of learning about the disease predictions in an efficient manner. This type of research may be expanded to diagnose additional illnesses.

Additionally, we might examine data from the past and incorporate additional machine learning techniques. Additional potential uses of this research could encompass the following: forecasting cardiovascular disease, diabetes, breast cancer, tumors, and multiple illness scenarios.

REFERENCES

[1]. Wikipedia contributors. (2022, June 22). Machine learning. In Wikipedia, The Free Encyclopedia. Retrieved 06:31, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=1094363111

[2]. Victor Chang, Vallabhanent Rupa Bhavani, Ariel Qianwen Xu, MA Hossain. An artificial intellegence model for heart disease detection using machine learning. Healthcare Analytics, volume 2, November 2022, 100016. https://doi.org/10.1016/j.health.2022.100016

[3]. Rohit Bharti, Aditya Khamparia, Mohammed Shabaz, Gaurav Dhiman, Sagar pande, and Parneet Singh. Prediction of Heart Disease Using a combination of Machine Learning and Deep learning. Hindawi Computational Intelligence and Neuroscience, Volume 2021, Article ID 8387680, 11 pages. https://doi.org/10.1155/2021/8387680.

[4]. Khaled Mohamed Almustafa. Prediction of heart disease and classifiers sensitivity analysis. Almustafa BMC Bioinfirmatics (2020) 21: 278. https://doi.org/10.1186/s12859-020-03626-y.

[5]. World Health Organization and J. Dostupno, cardiovascular diseases: key facts, vol. 13, no. 2016, p. 6, 2016. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[6]. Yuan X., Chen J., Zhang K., Wu Y., Yang T. A stable ai-based binary and multiple class heart disease prediction model for IoMT. IEEE Transactions on Industrial Informatics . 2022;18(3):2032–2040. doi: 10.1109/tii.2021.3098306. [CrossRef] [Google Scholar]

[7]. Rob Stocker, Tim Turner, and Mai Shouman. Heart disease patient diagnosis using k-Nearest Neighbour. International Journal of Information and Education Technology, vol. 2, No. 3, June 2012.

[8]. IOP Conference Series: Materials Science and Engineering, Volume 1022, Rajpura, India, October 24, 2020 First International Conference on Computational Research and Data Analytics (ICCRDA 2020)

[9]. Tyagi A., Mehra R. Intellectual heartbeats classification model for diagnosis of heart disease from ECG signal using hybrid convolutional neural network with Goa. SN Applied Sciences . 2021;3(2):p. 265. doi: 10.1007/s42452-021-04185-4. [CrossRef] [Google Scholar]

[10]. Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32. [Google Scholar] [CrossRef]

[11]. Alotaibi, F.S. Implementation of Machine Learning Model to Predict Heart Failure Disease. .[Google Scholar] [CrossRef]

[12]. Benjamin, E.J.; Muntner, P.; Alonso, A.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Das, S.R.; et al. Heart disease and stroke statistics—2019 update: A report from the American heart association. *Circulation* 2019, *139*, e56–e528. .[Google Scholar] [CrossRef]

[13]. Estes, C.; Bellentani, S.; Caballeria, J.; Colombo, M.; Craxi, A.; Crespo, J.; Day, C.P.; Anstee, Q.M.; Arias-Loste, M.T.; Bantel, H.; et al. estimating the disease burden from NAFLD in the United States, the United Kingdom, China, France, Germany, Italy, Japan, and Spain between 2016 and 2030. In 2018, J. Hepatol. 69, 896–904. [Scholar Google] [Cross Reference] [PubMed]

[14]. Alotaibi, F.S. Machine Learning Model Implementation for Heart Failure Disease Prediction. 2019, 10, 261–268. Int. J. Adv. Comput. Sci. Appl. [Scholar Google] [Cross Reference]