International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 03 March 2024
Page No: 485 - 490
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0070

Heart Disease Prediction Using Machine Learning

Patricia Rufes[1], J Sorna Jenita[2], Mabel Rakshitha T[3], Arthika Infanta A[4], Divya M S[5]*
[1, 2, 3, 4, 5] UG Biomedical Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India.
Emails: *patriciarufes.work@gmail.com[1], jjenita006@gmail.com[2], mabeldavid1218@gmail.com[3], arthika.1622@gmail.com[4], divyams389@gmail.com[5]*
****Corresponding Author Orcid ID:*** *https://orcid.org/0009-0007-3672-7438*

Abstract

Over the past few decades, heart-related diseases, or cardiovascular diseases (CVDs) have emerged as the leading cause of death not only in India but worldwide. They are the main cause of many diseases in the world. Machine learning algorithms and techniques have been applied to various clinical trials to automate the analysis of large and complex data. Recently, many scientists have been using AI techniques to support life and experts to detect heart-related diseases in the care industry. Compared to the brain, which is the largest organ in the human body, the heart is the next largest organ. Blood is pumped and sent to all the organs of the body. Cardiovascular disease prediction plays an important role in clinical practice. Informative studies will be useful in anticipating more data that can help predict various diseases and focus on treatment. There is a lot of data about the patients that are seen each month. The stored data can be a source for predicting future disease outbreaks. Some methods of data mining and artificial intelligence are used to predict heart diseases, such as artificial neural networks (ANN), animation, etc. To reduce the number of people who die from heart disease, we need to do it quickly and have a way to detect it. Data mining techniques and artificial intelligence calculations help experts make better cancer predictions and diagnoses. The main goal of this research project is to use AI statistics to predict coronary heart disease in patients.

Keywords: *Machine Learning, Heart Disease, Risk Prediction, Feature Selection*

1. Introduction

Machine learning techniques around us are compared and used for analysis for a variety of data science applications. The main motivation of this research project is to explore the methods of feature selection, data preparation, and post-processing of training models in machine learning. [1-4] The challenge we face today with sophisticated models and libraries is the large amount of data and larger accuracy gaps that can be encountered during training, testing, and validation in the non-mature learner world. Therefore, this project is carried out with the motivation to explore the antecedents of the model and also learn from the data obtained through the Implementation of a logistic regression model. Also, as all machine learning inspires the development of appropriate computer and decision support systems that help in the early detection of heart disease, in this project it will be decided whether the patient will die in 10 years or not. An example to classify whether a person will die of a heart attack. It is not based on different conditions (eg. risk factors for heart disease). Therefore, an early diagnosis of cancer can help high-risk patients make decisions about lifestyle changes and ultimately reduce complications, which is an important indication for the medical field. [5]

2. Methodology and Algorithms Used

Predicting the likelihood of developing heart disease in the future was the primary goal of system design.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 03 March 2024
Page No: 485 - 490
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0070

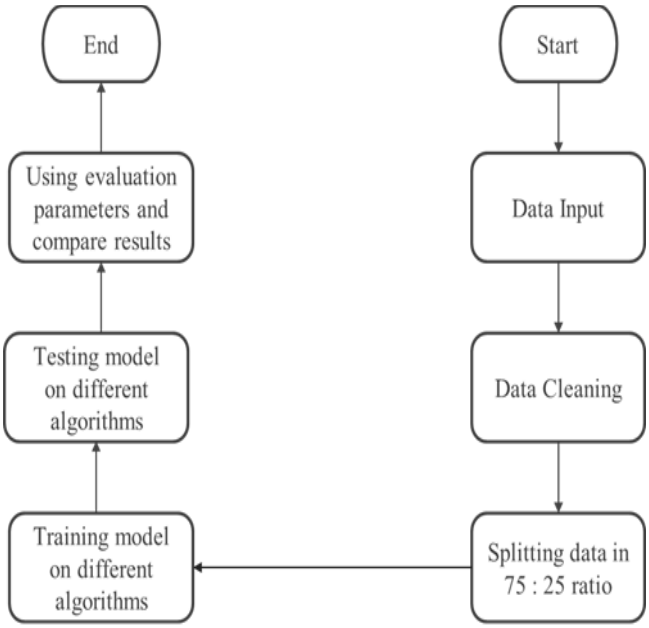

Figure 1 Flow Diagram of the Model Implemented

Our system has been trained using a variety of feature selection algorithms, including Naive Bayes, Support Vector Machines, and many more, as well as logistic regression as a machine-learning methodology. These algorithms are covered in full below. [6]

Logistic Regression

An algorithm for supervised classification is called logistic regression. It is an algorithm for predictive analysis built on the idea of probability. The underlying logistic function is used to estimate probabilities, which is how the connection between the dependent variable and one or more independent variables is measured. The logistic regression hypothesis is constrained between 0 and 1 by the use of the sigmoid function as a cost function. The correct presentation of data is crucial to the success of logistic regression. Therefore, key features from the available data set are chosen utilizing recursive and backward elimination strategies to increase the model's power. [7]

Random Forest

Random forest is the most potent and popular algorithm in machine learning. It's a part of machine learning under supervision. In machine learning, it is applied to both regression and classification issues. The way random forest works is as follows: [8]

- It gathers data.
- It averages the decision trees and constructs decision trees using various samples.

It is slower than a single decision tree, but it can handle datasets with categorical variables. Does not deal with null values. When compared to a single decision tree, Random Forest offers the primary benefit of greater accuracy and lower variation. The number of trees directly affects the accuracy of the outcome; hence a higher number of trees would yield a higher accuracy. The data is divided into 25% testing data and 75% training data in the suggested model. Considering this, the data is evaluated and trained in every conceivable combination before producing the optimal model. Now, the Random Forest Classifier is used to train the model. A predicted result is obtained from each tree, and the entropy of each result is determined separately. [9] Flow Diagram of the Model Implements shown in Figure 1.

Naive Bayes

A probabilistic classifier known as a "naive Bayes" classifier uses strong (naive) independence assumptions between the characteristics to apply the Bayes theorem. A Naive Bayesian model is particularly helpful in the field of medical science for diagnosing heart patients since it is simple to construct and does not require complex iterative parameter estimates. The Naive Bayesian classifier is popular because it frequently outperforms more complex classification techniques, even if it is incredibly simple. Using P(c), P(x), and P (x|c), one may get the posterior probability, P (c|x), using the Bayes theorem. The naive bayes classifier assumes that a predictor's (x) value has an independent impact on a given class (c) regardless of the values of other predictors. We refer to this presumption as class conditional independence. The equation below represents the Baye's theorem equation. [10]

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 03 March 2024
Page No: 485 - 490
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0070

$$P(c \mid x) = (P(x \mid c)\, P(c))/P(x) \quad (1)$$
$$P(c \mid x) = P(x1 \mid c)*P(x2 \mid c)*\ldots*P(xn \mid c)*P(c) \quad (2)$$

Equation 1 Baye's Theorem Equation

Where P (c|x) is the posterior probability of class (target) given predictor (attribute).
P(c) is the prior probability of class.
P (x|c) is the likelihood which is the probability of the predictor given class.
P(x) is the prior probability of the predictor. [11]

K – Nearest Neighbor (KNN)

KNN looks for correlations between values in the dataset and predictions. Since there is no specific set of parameters associated with a given functional form, it employs a non-parametric approach. It doesn't make any assumptions about the dataset's properties or output. [12] Because KNN memorizes the training data rather than precisely learning and fixing the weights, it is often referred to as a lazy classifier. Therefore, rather than during training, most of the computational work is done during classification. Typically, KNN finds the class that the new feature is closest to by attempting to determine which class it is. The algorithm's ease of implementation is one of KNN's advantages. Due to the k-nearest neighbor average, it is also highly successful when dealing with noisy data. Large data sets are easily handled by it, and the resulting decision boundaries can take on any shape. Divide the dataset into training and testing sets using the train-test-split function. The testing set will be used to evaluate how well our model performs on "unseen data," also referred to as real-world data, while the training set will be used to train the machine learning model. [13]

Gradient Boosting Machines (GBM)

Gradient Boosting Machines are a potent type of machine learning algorithms that perform exceptionally well in regression and classification problems involving predictive modeling. GBM is an ensemble learning technique that builds a powerful predictive model by aggregating the predictions of several weak learners, usually decision trees. The goal of GBM is to gradually enhance the forecasts by fixing the mistakes made by the prior model. Gradient Boosting Machines have several benefits, such as robustness against overfitting when hyperparameters are appropriately tuned, feature importance analysis, and the capacity to handle complex relationships in data. To attain the best performance, GBM requires careful tuning and can be computationally expensive. [14]

Cat Boost

A machine learning technique called CatBoost focuses on gradient boosting for tasks including regression and classification. Because of its unique design, which eliminates the need for significant preprocessing, it is an effective tool for managing mixed data types found in real-world datasets. It can handle categorical features with ease. The acronym CatBoost represents "Category Boosting. CatBoost is freely accessible for use via command-line interfaces and Python libraries. It works with popular machine learning frameworks such as XGBoost and scikit-learn. The library is a useful tool for data scientists and machine learning practitioners because of its robust regularization, automatic management of missing values, and effective handling of categorical features. [15]

Support Vector Machines (SVM)

Support Vector Machines are a class of supervised machine learning algorithms used for both classification and regression tasks. SVMs are particularly well-suited for problems with complex decision boundaries and relatively small datasets. The primary goal of SVM is to find a hyperplane that best separates data points of different classes in feature space. [16] Examine the attached image, which shows the existence of both positive and negative hyperplanes in addition to the important ideas of support vectors and the maximum margin hyperplane. Figure 2 illustrates the application of the hyperplane system to the data and the usage of support vector machines for data classification. Examine the complex relationships that exist between two of these support vectors. Explain and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 03 March 2024
Page No: 485 - 490
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0070

talk about the importance of these connections in relation to support vector machines (SVMs) and machine learning.

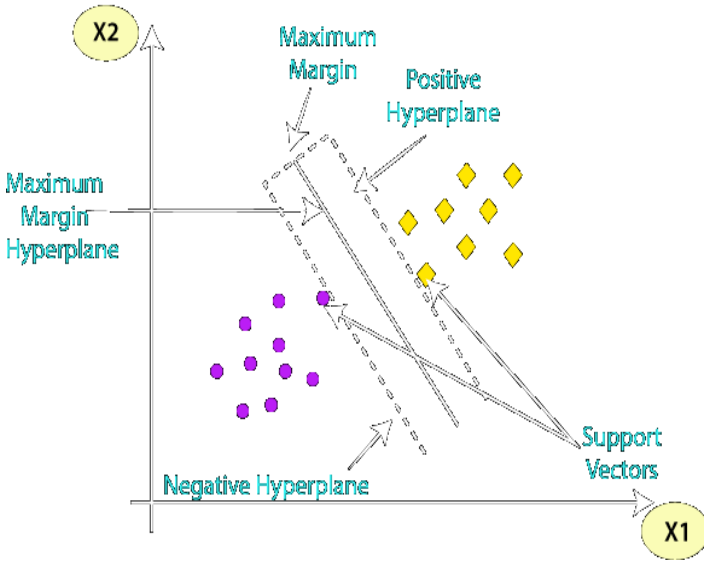

Figure 2 Two Categories Classified Using Hyperplane

3. Results and Discussion

3.1 Evaluation Metrics

For the evaluation of the output from the training data, the accuracy must be analyzed through a confusion matrix. A confusion matrix, also known as an error matrix, is a table that is commonly used to describe the performance of a classification model (or "classifier") on a set of test data whose real values are known. It allows you to visualize the performance of an algorithm. It makes it easy to identify class confusion, such as when one class is widely mislabelled as the other. The confusion matrix's essential feature is that the number of accurate and incorrect predictions are summarized with count values and broken down by class, rather than simply the number of errors committed. [17]

Accuracy

The equation used to measure the accuracy of the data used is explained below:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

Equation 2 Accuracy Equation

Where True Positive (TP) – observation is positive and predicted to be positive
False Positive (FP) – Observation is negative but predicted to be positive
True Negative (TN) – Observation is positive and is predicted to be negative
False Negative (FN) – Observation is negative but predicted to be negative

Recall

Recall is defined as the ratio of the total number of correctly categorized positive examples divided by the total number of positive examples. High recall suggests that the class was correctly recognized. Equation 5.2 below calculates recall as follows:

$$\text{Recall} = TP/(TP+FN)$$

Equation 3 Recall Equation

Precision

To calculate the value of precision, divide the total number of successfully classified positive examples by the total number of anticipated positive examples. High Precision shows that an example labeled as positive is positive (a low number of FP). Precision is calculated as:

$$\text{Precision} = TP/(TP+FP)$$

Equation 4 Precision Equation

3.2 Result

After comparing several feature selection approaches, the Logistic regression algorithm outperforms the other algorithms in terms of accuracy. The accuracy was calculated using the confusion matrix of each algorithm, where the number of true positives, true negatives, false positives, and false negatives is given, and the value is calculated using the accuracy equation, and it is concluded that Support Vector Machines are the best with 94.28% accuracy, as shown in the comparison below. Table 1 below presents the accuracy, precision, recall, and F-score of all algorithms used to predict heart disease.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 03 March 2024
Page No: 485 - 490
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0070

Table 1 Algorithms and the Accuracy, Precision, Recall, & F-Score

Algorithm	Accuracy	Precision Score	Recall	F-Score
Logistic Regression	80.33 %	0.85	0.65	0.74
Naïve Bayes	78.69%	0.78	0.69	0.73
K-Nearest Neighbors	67.21%	0.69	0.42	0.52
Random Forest	73.77%	0.72	0.62	0.67
Gradient Boosting Machines	92.46%	0.90	0.74	0.82
CatBoost	87.15%	0.84	0.61	0.71
Support Vector Machines	94.28%	0.92	0.75	0.83

Conclusion

This research presents a survey of machine learning-based heart disease diagnosis algorithms. This survey examines seven techniques of using machine learning models to identify cardiac disease. According to the results analysis, the accuracy, precision, recall, and F1-measure parameters for the SVM algorithm-based heart disease detection are high, whereas the KNN method has poor values. The current survey report provides the best insight into several machine learning-based heart disease diagnosis approaches. This study can be expanded in the future by including more attributes to the heart disease dataset and making it more interactive for users. It can also be carried out as a mobile application, requiring less computational time and complexity.

References

[1]. Droz˙dz˙ , K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman,W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022, 21, 240. [CrossRef] [PubMed]

[2]. 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].

[3]. Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked 2021, 26, 100655. [CrossRef]

[4]. Mienye I. D., Sun Y. Improved heart disease Prediction using particle swarm optimization based stacked sparse autoencoder. Electronics. 2021; 10(19): p. 2347. Doi: 10.3390/electronics10192347. [CrossRef] [Google Scholar].

[5]. Kaur J., Khehra B. S. Fuzzy logic and hybrid-based approaches for the risk of heart disease detection: state-of-the-art review. Journal of the Institution of Engineers: Series B. 2021 doi: 10.1007/s40031-021-00644-z. [CrossRef] [Google Scholar].

[6]. Shorewala V. Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked. 2021; 26 doi: 10.1016/j.imu.2021.100655.100655 [CrossRef] [Google Scholar]

[7]. Biswal A. K., Singh D., Pattanayak B. K., Samanta D., Chaudhry S. A., Irshad A. Adaptive fault-tolerant system and optimal power allocation for smart vehicles in smart cities using controller area network. Security and Communication Networks. 2021; 2021 doi: 10.1155/2021/2147958.e2147958

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 03 March 2024
Page No: 485 - 490
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0070

[CrossRef] [Google Scholar].

[8]. Zhenya Q., Zhang Z. A hybrid cost-sensitive ensemble for heart disease prediction. BMC Medical Informatics and Decision Making. 2021; 21(1): p. 73. Doi: 10.1186/s12911-021-01436-7. [PMC free article] [PubMed] [CrossRef] [Google Scholar].

[9]. Spencer R., Thabtah F., Abdelhamid N., Thompson M. Exploring feature selection and classification methods for predicting heart disease. Digital Health. 2020; 6 doi: 10.1177/2055207620914777.205520762091 4777 [PMC free article] [PubMed] [CrossRef] [Google Scholar].

[10]. Tama B. A., Im S., Lee S. Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble. BioMed Research International. 2020; 2020 doi: 10.1155/2020/9816142.9816142 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[11]. Fitriyani N. L., Syafrudin M., Alfian G., Rhee J. HDPM: an effective heart disease prediction model for a clinical decision support system. IEEE Access. 2020; 8 doi: 10.1109/access.2020.3010511.133034 [CrossRef] [Google Scholar].

[12]. Nourmohammadi-Khiarak J., Feizi-Derakhshi M.-R., Behrouzi K., Mazaheri S., Zamani-Harghalani Y., Tayebi R. M. New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. Health Technology. 2020; 10(3):667–678. Doi: 10.1007/s12553-019-00396-3. [CrossRef] [Google Scholar].

[13]. Sultan Bin Habib A.-Z., Tasnim T., Billah M. M. A study on coronary disease prediction using boosting-based ensemble machine learning approaches. Proceedings of the 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET); December 2019; Dhaka, Bangladesh. pp. 1–6. [CrossRef] [Google Scholar].

[14]. Alarsan F. I., Younes M. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. Journal of Big Data. 2019; 6(1): p. 81. Doi: 10.1186/s40537-019-0244-x. [CrossRef] [Google Scholar].

[15]. Tao R., Zhang S., Huang X., et al. Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods. IEEE Transactions on Biomedical Engineering. 2019; 66(6):1658–1667. Doi: 10.1109/tbme.2018.2877649. [PubMed] [CrossRef] [Google Scholar].

[16]. H. Li, K. Ota, and M. Dong, ‘‘Learning IoT in edge: Deep learning for the Internet of Things with edge computing,’’ IEEE Netw., vol. 32, no. 1, pp. 96–101, Jan. /Feb. 2018.

[17]. F. Sabahi, ‘‘Bimodal fuzzy analytic hierarchy process (BFAHP) for coronary heart disease risk assessment,’’ J. Biomed. Informat. vol. 83, pp. 204–216, Jul. 2018. Doi: 10.1016/j.jbi.2018.03.016.