Heart Disease Prediction using Machine Learning

N. Saranya, P. Kaviyarasu, A. Keerthana, C. Oveya,

Abstract*: Deriving the methodologies to detect heart issues at an earlier stage and intimating the patient to improve their health. To resolve this problem, we will use Machine Learning techniques to predict the incidence at an earlier stage. We have a tendency to use sure parameters like age, sex, height, weight, case history, smoking and alcohol consumption and test like pressure ,cholesterol, diabetes, ECG, ECHO for prediction. In machine learning there are many algorithms which will be used to solve this issue. The algorithms include K-Nearest Neighbour, Support vector classifier, decision tree classifier, logistic regression and Random Forest classifier. Using these parameters and algorithms we need to predict whether or not the patient has heart disease or not and recommend the patient to improve his/her health.

Keywords: Classifier, Heart disease, K nearest neighbor , Prediction, Random forest.

I. INTRODUCTION

The contents of this paper primarily concentrate on predicting heart disease using machine learning techniques. If the heart is not functioning properly, this may affect other parts of the body such as the brain, kidney, etc. Heart disease is a condition which affects the heart's functioning. Different individuals will show different symptoms of heart disease which can vary consequently. They often have back pain, jaw pain, neck pain, abdominal disorders, and breath weakness, chest pain, pain to the arms, and pain to the shoulders. There are a number of common heart diseases including heart failure and stroke and coronary artery disease. In today’s era heart problems are the primary reason for deaths. Some heart diseases are heart failure and coronary cardiomyopathy. Among varied serious diseases, cardiomyopathy acts as an excellent deal of attention in medical analysis. It's tough for doctors to predict the heart attack because it could be an advanced task that needs a lot of expertise and high information. The diagnosis of heart can give machine-driven prediction regarding the heart condition of patients at earliest. It's necessary to appear at the signs, symptoms and physical examination of the patient. There are several factors that increase the possibility of heart condition, like smoking habits, body cholesterin level and case history, obesity, high force per unit area and lack of physical exertion.

A big challenge faced by health care organisations, like hospitals and medical centers, is that the supply of quality services. The information information consists of numerical and categorical data. Before the extra method of the dataset, improvement and filtering are applied on these records therefore on filter the impertinent data from the information. Records or knowledge of case history is extremely massive, however these are from several dissimilar foundations. The interpretations that are done by physicians are essential elements of this knowledge. The information within the world could be clamant, incomplete and inconsistent, therefore knowledge preprocessing is going to be needed in order to fill the omitted

values within the database. Once preprocessing is completed, we have a tendency to train and test the information. we have a tendency to train 80% of the information and test 20% of the data. The accuracy of the result should be among 0 to 1. Suggestions are given to the patients to boost their health. The planned system can make sure bound info that's patterns and relationships involving cardiopathy from historical heart disease databases. It will even answer the queries for designation cardiopathy, therefore, it should be helpful to doctors to make intelligent clinical decisions.

II. LITERATURE SURVEY

[3] Decision Tree is an approach for early detection of heart condition by utilizing a variety of options. Authors have gotten potency up to 95% with the assistance of trained neural network observation the success of neural network researchers operating within the domain of SVM have used this method to classify and win higher lead to case wherever the feature vector are multidimensional and non linear these method defeated all different existing quantum up to date techniques as a result of it's the aptitude to figure underneath dataset of high spatial property.Deep learning arising space of computer science showed some promising leads to different fields of diagnosing with high accuracy.Naive Thomas Bayes classifier it's quite simple to coach the classifier on tiny dataset if there exist high biases and low variance provides it a significant advantage over the classifier with low biases and high variance like KNN as a result of later classifier can suffer downside of overfitting.

[5] Three completely different machine-supervised learning algorithms i.e. Naive Bayes, K-NN, decision list formulae are used for data set analysis. The experiment is carried out using a training knowledge collection consisting of 3000 instances with fourteen attributes which are completely different. The dataset is divided into 2 components that are used for coaching 70 percent of the knowledge and used for testing 30 percent. In support of the experimental findings, it is clear that the classification accuracy of the Naive Bayes method (52.33%) is higher than that of the List method (52%) and the K-NN algorithm (45.67%).

Manuscript received on April 02, 2020.
Revised Manuscript received on April 15, 2020.
Manuscript published on May 30, 2020.
***** Correspondence Author

Mrs. N. Saranya*, Assistant Professor**,** Department of Computer Science and Engineering,Sri Shakthi Institute of Engineering and Technology, Coimbatore

P. Kaviyarasu, UG Student**,** Department of Computer Science and Engineering,Sri Shakthi Institute of Engineering and Technology, Coimbatore

A. Keerthana, UG Student **,** Department of Computer Science and Engineering,Sri Shakthi Institute of Engineering and Technology, Coimbatore

C. Oveya, UG Student**,** Department of Computer Science and Engineering,Sri Shakthi Institute of Engineering and Technology, Coimbatore

[6] In this paper, we aim to apply machine-learning algorithms to the dataset of heart disease to predict heart disease, supporting the patient's knowledge about each attribute. Our aim was to test completely different rating models and outline the most economical one.Either cross-validation, grid search, operation and option is employed or not, entirely different algorithms perform higher depending on the situation.As a consequence, performance metrics are also a customary method of evaluating algorithms when selecting a function, parameter calibration, and calibration to compare the dataset.The average accuracy price of the simplest results while not improving at 83.6 percent for SVM and NB compared to 81.35 percent for RF. These SVM and NB shows function on average, and when configured by FCBF, PSO and ACO, we discover that the simplest one is 99.6% PA-RF.

III. OBJECTIVE

A. Straightforward to use: The key detachment of this project is the creation of a platform that is in a position to be transparent and easy to use, as here one can give the medical details of the patient and support the choices derived from the algorithmic law, we can then note the problem.Since this algorithmic rule can perform the task, a well-trained model can be a smaller amount required to produce errors in predicting cardiovascular disease and thus increase precision in a descriptive manner, thus saving time and promoting what is more like patients for doctors to determine whether or not they are vulnerable to some form of heart attack or not, otherwise we prefer to be rigorous in undertaking and doing while not requiring a specialist.

B. No human intervention needed: In order to find the center unwellness one can give medical information such as age, steroid, etc. and here the algorithmic rule will provide the results backed by the choices derived and then here the chances of error are very low because there is no human involvement and it also saves the patients or doctors plenty of time whether they can continue for treatment or absolutely. This is also clearly the case because findings are presented to them more quickly. This will create in-turn the precautionary / preventive technique of loads of heart care faster as it saves doctors and patients the critical time, so that they can undergo more therapies and measures to be taken to minimize the impact of this cardiovascular disease.

C. Detective work cardiovascular disease and conjointly counsel precautions: Through this project, our goal is to search and predict cardiovascular disease involvement and provide the precautions to reduce the impact of heart disease. Receiving recommendations on the measures to be taken would promote the progress of doctors and patients primarily towards further steps in their care.

D. Economical use of accessible annotated knowledge samples: Through this project, our goal is to search and predict cardiovascular disease involvement and provide the precautions to reduce the impact of heart disease. Receiving recommendations on the measures to be taken would promote the progress of doctors and patients primarily towards further steps in their care.

IV. PROPOSED SYSTEM

Prediction of heart diseases can be a web-based application of machine learning, informed by a true dataset. The user will input their specific medical details to urge that user to predict cardiovascular disease. The algorithmic rule will measure the likelihood that cardiovascular disease is present.The result shows itself on the net tab. Therefore reducing the price and time needed to foresee the unwellness. Information format plays a critical half throughout this procedure. When the user information application is uploaded, it can check its proper file format.

Our system are implementing the subsequent 2 algorithms:

Random Forest Classifier

●K Nearest Neighbour

The algorithms are qualified victimization of the information set obtained from one of Coimbatore's hospitals. 80 percent of the entries in the knowledge collection are used for instruction, and the remaining 20 percent are used to check the algorithmic rule for accuracy. In addition, some steps are taken to refine the algorithms and improve the accuracy thereby. These steps include cleaning up of the information set and pre-processing of data. The algorithms were judged to support their accuracy and the Random Forest (100 percent) is calculated to be the most accurate than KNN (91.36 percent).Hence, it's the most applicable class. The most applications may be a net application which accepts the user's various parameters as input and calculates the result. The predictive accuracy is displayed alongside the result.

V.METHODOLOGY

As per the information and knowledge we've gathered, we have a tendency to found that these following tasks should be allotted so as to induce correct predictions. The tasks that we have a tendency to are about to do are givey by

Data Preprocessing:

The dataset we've a bent to obtained isn't absolutely correct and error free. Hence, we have a tendency to are getting to initial do the following operations on it.

Data Cleaning:

Na values inside the dataset is that the most important issue for US as a result of it will prune the predictive performance deeply so, lets go have a tendency to are getting to fill the fields with most frequent price.

Feature Scaling:

Due to the vary in Data quality wide, on other ml algorithms, deducted functionality won't work correctly whereas not feature scaling. as associate degree Exemplary proof,most classifiers measure gap interval between a pair Over levels via geometrician range.Even when one does all told that's it, choices incorporates a wide region vary of the ideals,of gap are dominated by this specific feature. Therefore, the vary of all choices have to be compelled to be scaled therefore Every role certainly contributes to the roughly Similar to the last word mobile.Now lets go to have a tendency to are getting to scale the various fields therefore on induce them nearer As to the principles. The key code example is the age of time merely a pair of Werte, i.e., 0,1 and steroid also has high ratings like 100.And So,therefore on induce them nearer to each various we have a tendency to are getting to have to be compelled to be halving them back.

Factorization:

During this verse, we have the tendency to appointed a assuming Towards the beliefs which the algorithmic rule did not necessarily overwhelm them.As an example, assignment assuming to zero and one within the Category Era so0the algorithmic rule might not think about one such as bigger hardly zero therein segment.

Random forest algorithm:

It is a supervised learning formula that is employed for each classification still as regression. However , it's mainly used for classification issues. As we know, a forest is formed of trees and additional trees suggests that a more strong forest. Similarly, random forest formula creates call trees on knowledge samples then gets the prediction from every of them and at last selects the simplest answer by suggesting that of voting. It's an ensemble methodology that is better than one call tree as a result of it reduces the over-fitting by averaging the result. When exploitation the Random Forest formula to resolve regression issues, the mean square error (MSE) is employed to understand however the information branches from every node.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2$$

Where *N* is the number of data points,
fi is the value returned by the model and
yi is the actual value for data point *i*.

Fig. 1.Mean squared error formula

This formula calculates the gap of every node from the anticipated actual price, serving to come to a decision that branch is that the higher decision for your forest. Here, Yi is that the price of the information purpose you're testing at a precise node and fi is the value came back by the choice tree. Once performing Random Forests supported classification knowledge, you must apprehend that you just are typically exploiting the Gini index, or the formula accustomed to decide however nodes on a decision tree branch. This formula uses the category and chance to see the Gini of every branch on a node, deciding that of the branches is additional doubtless to occur.

$$Gini = 1 - \sum_{i=1}^{C}(p_i)^2$$

Fig. 2.Gini index

Here, pi represents the frequency of the category you're perceptive within the dataset and c represents the amount of categories. Entropy uses the chance of a precise outcome so as to create a choice on however the node ought to branch. Unlike the Gini index, it's additional mathematical intensive because of the power operate utilized in hard it.

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Fig. 3.Entropy formula

K-nearest neighbors (KNN) algorithm:

KNN could be a non-parametric machine learning rule. The KNN rule is a supervised learning technique. this implies that all the information is labeled and therefore the rule learns to predict the output from the input file. It performs well albeit the training data is large and contains noisy values. The data is subdivided into test and training sets. This is used for the construction and training of models. A k- value which is typically the origin of the quantity of observations is calculated. Currently the check data is expected on the model designed. There are completely different distance measures. Euclidean distance, Manhattan distance, and Minkowski distance measurements are used for the continuous variables. However, the normally used measure is Euclidean distance. The formula for Euclidean distance is as follows

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Fig. 4.Euclidean distance formula

It ought to even be noted that each one 3 distance measures are solely valid for continuous variables. within the instance of categorical variables the Hamming distance should be used. It additionally brings up the difficulty of standardization of the numerical variables between zero and one, there's a mix of numerical and categorical variables within the dataset.

Hamming Distance

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

Fig. 5.Hamming distance formula

Logistic Regression:

Logistic Regression offers the prospect of a conclusion that can only have 2 values. The prediction relies on the use of numerical and categorical predictors like 1 or many. A supply regression creates a logistical curve that is constrained to zero-to - one values. Logistic regression appreciates a simple regression, but the curve does victimize the log of the target variable's "odds," rather than the potential. In addition, the predictors in each cluster would not necessarily be distributed or have equal variance. At logistic regression intervals, constant (b0) pushes the curve left and right and hence the slope (b1) determines the curve gradient. The supply regression of y on x is written in terms of related degree odds magnitude relationship by a straight transformation. Finally, we must write the equation in terms of log-odds (logit) that could be a linear representation of the predictors, taking the natural log of each side. The constant (b1) is that for a 1 unit change in x, the amount the logit (log-odds) changes to. Regression of supply manages any number of numerical and categorical variables.

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}$$

Fig. 6.Probability of any event

Logistic regression uses most probability Estimation (MLE) to find the logistic regression coefficients. When the first performance is calculable, this method is performed continuously till LL (Log Likelihood) doesn't vary considerably.

$$\beta^1 = \beta^0 + [X^T W X]^{-1}.X^T(y - \mu)$$

β is a vector of the logistic regression coefficients.

W is a square matrix of order N with elements $n_i \pi_i (1 - \pi_i)$ on the diagonal and zeros everywhere else.

μ is a vector of length N with elements $\mu_i = n_i \pi_i$.

Fig. 7.Logistic regression formula

VI. TECHNOLOGY USED

Anaconda

Anaconda distribution could be a free and open source distribution of the R and Python programming languages for computing such as data science, machine learning applications, large-scale data processing, predictive analytics, etc. that aims to change package management and preparation. Anaconda has many packages further as conda packages and virtual atmosphere. It conjointly includes a user interface referred to as Anaconda Navigator. It is graphically different to the statement interface.

Anaconda Navigator

Anaconda Navigator can be a graphical desktop interface included in Anaconda Navigator Distribution that allows users to launch applications and control conda packages, environment, and networks. Navigator can rummage around for packages on the Anaconda Navigator Cloud or in a very native Anaconda Navigator Repository, install them in an environment, run and upgrade the packages. It is on the Windows, macOS, and UNIX market.The applications are JupyterLab,Jupyter Notebook,Spyder,Orange,RStudio,Visual Studio Code.

Spyder

Spyder (Scientific Python Development Environment) is an integrated software environment (IDE) with open source cross-platform for scientific programming in the Python language. Spyder also incorporates NumPy, SciPy, Matplotlib, Pandas, IPython, SymPy and Cython as an alternative open source program with a range of excellent packages within the scientific Python stack. It is cross-platform on the market via Anaconda.

Python

Python is a general programming language, which is high-level. The language constructs and object-oriented methodology are aimed at helping programmers write simple, logical code that comes in for small and large scale. This embraces many programming paradigms, as well as programming that is procedural, object-oriented and sensible.

NetBeans IDE 8.0.1

NetBeans could be an environment of free, open source Integrated Development for developers of computer code. Professional desktop, business, cloud, and mobile applications with the Java language, C / C++, and even dynamic languages such as PHP, JavaScript, Groovy, and Ruby are customary. Apache NetBeans runs on a number of platforms, including Windows, Linux, waterproof OS X, and Solaris. NetBeans is also spoken as a standard parts platform which is used to develop Java desktop applications. The NetBeans Team released an update to NetBeans 8.0, with improvements to the HTML5, JavaScript, and CSS3 options. NetBeans 8.0 was released at an equal time as Java 8, a temporary order that made it clear that Oracle was making an effort to own approved NetBeans instead of the usual Java IDE. It supports JDK eight with tools and editor enhancements for profile, lambda, and stream operation; Java ME Embedded 8 support; and hence the ability to deploy, run, rectify, or profile Java SE applications on an embedded device, such as Raspberry PI, directly from NetBeans IDE, in conjunction with many enhancements to Java Editor. Together with PrimeFaces, it extended support for wizard and Java electrical engineering; introduced new tools for HTML5, specifically for AngularJS; and brought enhancements to support for PHP and C / C++.

VII. CONCLUSION

In the medical field, machine learning algorithms are used extensively to detect diseases and diagnose the heart patient based on the data set and the attributes provided. Researchers are applying the work completely with different algorithms to assist health care professionals within the identification of heart condition. In the proposed work Random Forest algorithm, K Nearest Neighbour algorithm and Logistic regression algorithm is helpful to order the data set because they provide accurate results, with these results heart diseases among people is predicted. The accuracy of the ensemble model with logistic regression is 95.06% and without logistic regression is 98.77%. Therefore in this project, the algorithm Random Forest and K Nearest Neighbour algorithm is used to provide better accuracy.The results thus obtained shows 98.77% accuracy with minimum time. Thus heart diseases prediction system successfully diagnoses the medical data and predicts the heart diseases and intimates the patient to improve their health.

VIII. EXPERIMENTAL RESULT

ALGORITHM	**ACCURACY**
Random Forest	**100 %**
K-nearest Neighbour	**91. 36 %**
Logistic Regression	**87. 65 %**
Ensemble model with Logistic Regression	**95. 06 %**

Ensemble model without Logistic Regression	**98. 77 %**

REFERENCES

1. Youness Khourdifi, Mohamed Bahaj, **“**Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization”
2. Vikas Chaurasia, Saurabh Pal, “Early Prediction of Heart Diseases Using Data Mining Techniques”.
3. Sonam Nikhar , A.M. Karandikar ,“Prediction of Heart Disease using Machine Learning Algorithms”.
4. Gagandeep Kaur, Anshu Sharma, Anurag Sharma ,“Heart Disease Prediction using KNN classification approach”.
5. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni ,“Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”
6. Youness Khourdifi, Mohamed Bahaj ,“The Hybrid Machine Learning Model Based on Random Forest Optimized by PSO and ACO for Predicting Heart Disease”.
7. Beant Kaur, Williamjeet Singh ,“Review on Heart Disease Prediction System using Data Mining Techniques”
8. Amita Malav, Kalyani Kadam, Pooja Kamat ,“Prediction of Heart Disease using K-means and Artificial Neural network as Hybrid approach to improve accuracy”
9. Kathleen H. Miao, Julia H. Miao, and George J. Miao ,“Diagnosing Coronary Heart Disease Using Ensemble Machine Learning”
10. G. Parthiban, S.K.Srivatsa ,“Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients”

AUTHORS PROFILE

N.Saranya,B.Tech,MBA,M.E.
Assistant Professor, Department of Computer Science and Engineering,Sri Shakthi institute of Engineering and Technology, Coimbatore.
saranya.ramya@gmail.com

P. Kaviyarasu UG Student,Department of Computer Science and Engineering,Sri Shakthi institute of Engineering and Technology, Coimbatore.
kavi422815@gmail.com

A. Keerthana UG Student, Department of Computer Science and Engineering, Sri Shakthi institute of Engineering and Technology, Coimbatore.
keerthanaarivoli98@gmail.com

C. Oveya UG Student, Department of Computer Science and Engineering, Sri Shakthi institute of Engineering and Technology, Coimbatore.
oveyac@gmail.com