IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 2, January 2024

Heart Disease Prediction using ML Techniques

Snehakumari Vijaykumar Prasad
Master of Computer Applications
Institute of Distance and Open Learning, Mumbai, Maharashtra, India

Abstract: *Artery disease is considered as one of the notable principles of life end nowadays. Heart patients are growing rapidly because of deficient health awareness and bad consumption lifestyles. Therefore, it's important to have a system that can effectively acknowledge the frequency of heart issues in 1000s of selected instantly ML depicted to be fruitful in reinforcement in resolve and prognosis from the vast quantities of data generated by the medical professional. ML disease prediction is an approach that predicts diseases based on information supplied by users. Itdetects the heart related issues of the tolerant or the user post on the details or the symptoms enter into the system and gives results based on that information. This HDP Using ML is wholly run with the assist of ML algorithms and PPL using the Cleveland dataset that's available in the UCI repository. This study focuses on recognizing ML classifiers with the extreme fairness for symptomatic purposes. Various supervised ML algorithms were requested & differentiate for execution and reliability in heart disease problems. Every single quality was ranked onthe foremost result to those offering high HDP. This research paper finds that operate on heart disease dataset gather from Kaggle six- algorithms like NB, RF, LG, KNN,SVM, DT algorithms can be used for heart disease prediction. Here, we find that a quite simple supervised algorithm can be utilize to build HDP with very high efficiency and extremely good future utility*

Keywords: HDP Heart Disease Prediction, Data Mining,Cleveland dataset, UCI repository, Naive Bayes, SVM, Machine Learning, RF, Big Data, Logistic Regression, K- NN, Decision Tree

I. INTRODUCTION

The heart is an internal organ that purifies and suppliesthe blood, to all the nerves of the human body with its vitalizing oxygen & nutrients. If the injecting work of the heart converts unfit, necessary organs like the kidneys & head hurt, if the heart halts operating completely, death happens within few seconds. Cardiovascular problem has been considered as one ofthe intricate and dangerous person illness in the whole world. Life of every living being is totally relies on the effective functioning of the heart. Diagnosis and treatment of cardiac disease are very complicated, mainly in progressing countries, due to the irregular availability of predictive tools and additional resources which influence the actual forecast & therapy of illness in heart patients. This labels heart issues a more affect to be concern with. But it is difficult to verify coronary artery disease because of major supportive venture such as abnormal pulse rate, sugar, high BP and hypercholesterolemia many other factors. The invasive– based method to examine coronary artery issues are grounded on the study of the patient's health in past, physical check-up report, and evaluation of concerning symptoms by medical specialist. Frequently there is a hold- up in the diagnosis due to failure. Expected to such hindrances, scientists have gone round approaching modernmethods like DM and ML for predicting the disease [1].

DM plays a significant bit part in constructing intelligent models for medicinal systems to recognize coronary artery disease using the present details of patients,which count in risk factors correlated with the disorder. Health professionals may anticipate help for the recognition. Different software device and many algorithms have been put forward by analyzer for progressive medicalDSS. The data mining techniques includes 13 medical attributes such as Gender, blood pressure, cholesterol like to predict the likelihood of patient getting a heart disease [2].

ML assists workstation to study and act appropriately. It assists the device to study the system model and detect the data & also has the capability to determine multifacetedmathematics on BD. The ML build HDP systems will be specified and will turn down the risk. The ethics of ML algorithms is appreciated well in the medical industry which has a huge amount of dataset. It assists medical specialist to forecast the condition and leads to improvising the therapy. In the supervised algorithms, the model is learning from thelabeled dataset. Keeps input dataset & its outcomes. Datasetare sort

DOI: 10.48175/IJARSCT-15234

and divided into learning & testing data. The learning data teach system while the testing data gathering as latest dataset to become the reliability of the prototypical.The data obtains through replicas & their productivity. MLanalytical models such as DT, SVM, LG, RF, KNN & NB is deployed to forecast whether a user is having heart issuesor not [3].

II. LITERATURE SURVEY

In [4] Shan Xu describes coronary artery disease prognosis which is primary diagnosis & it is beneficial for the patient's therapy. The system concern on giving additional accuracy with assist of SVM & NB classifiers. Itcontains 4 parts such as dataset interface having health center pre- processing data and raw data for dataset combination & feature variety used to recover individual beneficial features for receiving extra accuracy & executionthe organization of those features.

In [5] R. Perumal established a model for the forecast of cardiovascular disease using the Cleveland data which contains 303 data cases concluded attribute correction and attribute decrease by applying PCA, where they recognized& used 7 major mechanisms to teach the machine learning classifiers. They gather that SVM & LR gives nearly close to precise evaluate like 87% & 85%, correspondingly associated to that of K-nearest neighbor(k-NN) having 69%.

In [6] Sahaya Arthy study the present tasks on HDP, where DM is utilized. The DM method are commonlymaking use in coronary artery disease prediction. In addition, they express regarding directory utilized, for e.g.,the cardiovascular illness dataset from the UCI repository keeps, tools required, for e.g., R, Weka, Apache Mahout, Information dissolve, Quick Excavator, Clatter & Bottom soon. Also, they accomplish that the work of one classifier report in improved accuracy in estimate. But the utilization of crossing of 2 or other technique can grow & expand the coronary artery disease prediction with better accuracy.

In [7] N. K. Kumar learned 5 ML classifiers, namely SVM, DT, KNN, LG, and RF using UCI data having 303 reported & ten features to predict coronary artery illness. In comparison with group of classifiers, the RF attained the topmost precision as eighty five.sevenety one percent with 0.8675 ROC AUC.

In [8] A. Gupta make use of the Pearson's r from the dataset to restore the misplaced utility based on the most label and extract 28 features & learned DT, RF, LG, KNN, SVM, NB method using the factor examination of different dataset techniques; the report grounded on a RF W carry outthe maximum validity as 93.44%.

III. METHODOLOGY OF SYSTEM

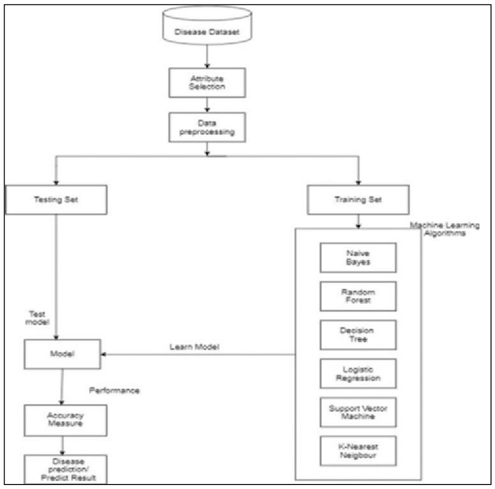


Fig 1: Proposed System

Data Collection

First step for predication system is data collection and determining about the training and testing dataset. In this project we have used 73% training dataset and 37% datasetused as testing dataset the system. Dataset has imported from UCI repository which is well verified by number of researchers and authority of the UCI. This dataset contains 303 rows and

DOI: 10.48175/IJARSCT-15234

among them 164 were negatives, 139 were positives. The dataset is stored in CSV (Comma Separated Value) file format where each row represents a single value.

Attribute Selection

Attribute of dataset are characteristics of dataset whichare utilized for system and for heart many attributes are likeheart bit rate of person, gender of the person, age of the person and many more for predication system.

Preprocessing

Preprocessing needed for achieving prestigious result from the machine learning algorithms. In this phase first weneed to gather the training dataset which is in CSV file format. For that we need to read the file data with the help of Pandas library for converting it to list array. The input message which is given by the user should get append to list array, because the machine cannot understand the file format data. On completion of converting process, it will remove null values if they are in training and testing data. For example, Random Forest (RF) algorithm does not carrynull values dataset and for this we have to handle null values from original raw data. For our project we have to transform some categorized value by dummy value means in the formof "0"and "1".

Standardization

In this, we need to apply standardization on given training dataset for scaling of each column data for getting best accuracy model. To perform scaling, we use Standard Scaler python library class which contains fit transform function and it takes input as dataset columns data which have higher values.

Data Balancing

Data balancing is essential for accurate result because by data balancing graph we can see that both the target classes are equal. Fig.3 represents the target classes where “1” represents with heart diseases patient and “0” representsno heart diseases patients.

Data Source

Considering this model, we have handed-down a dataset gained from the CHD dataset at UCI Repository. This data is including the medical report of 303 separate patients of specific age set. This data carries 14 health attributes of 303 patients that help us detect if the patient isat threat of obtaining a heart issue or not and it assists us tolist patients that are at threat of having heart issues and thosewho are not at danger.

The explanation of the dataset is given as follows:

Sr. No	Attribute	Representative Icon	Details
1.	Age	Age	Patients age in years.
2.	Sex	Sex	0=female; 1=male
3.	Chest Pain	Cp	4 types of chest pain (1-typical angina; 2- atypical angina; 3-non-anginal pain; 4 asymptomatic)
4.	Rest Blood Pressure	Trestbps	describes the Resting systolic blood pressure (in mm Hg on admission to the hospital)
5.	Serum Cholesterol	Chol	Serum cholesterol inmg/dl
6.	FastingBlood Sugar	Fbs	Fasting blood sugar 120mg/dl (0=false; 1=true)
7.	Rest Electrocardiograph	Restecg	0=normal; 1-having ST- T wave abnormality; 2- left ventricular hypertrophy
8.	Max Heart Rate	Thalach	describes the Maximum heart rate achieved
9.	Exercise-inducedangina	Exang	Exercise-induced angina(0=no; 1=yes)
10.	ST Depression	Oldpeak	Describes the ST depression induced by exercise relative to rest

11.	Slope	Slope	slope of the peakexercise ST segment(1=upsloping; 2=fat; 3-downsloping)
12.	No. of Vessels	Ca	describes the number of blood vessels (0-3) colored by fluoroscopy
13.	Thalassemia	Thal	Defect types; 3-normal; 6-fixed defect; 7-reversible defect
14.	Num (target class attribute)	Class	diagnosis of heart disease status (0-nil risk; 1-low risk; 2-potential risk; 3= high risk; 4=very high risk)

Algorithms Used

Naive Bayes Algorithm:

This derived from the Theorem of Bayes. The self- determination between the aspect of the data is the primarybelief and nearly all essential in creating a classification. It is fast and easy to forecast and carry topmost when the operation of independence carries out

Bayes Theorem on Mathematical Representation:

$$P(A|B) = (P(BA)\ P(A)) / P(B)$$

P(A|B) - defines the posterior probability of class (A, target)given predictor (B, attributes).
P(A) - defines the prior probability of class.
P(BA) - defines the likelihood which is the probability ofthe predictor given class.
P(B) - defines the prior probability of predictor.

Logistic Regression Algorithm:

It is a type of supervised algorithm for regression & classification. Instead of fitting a straight line or hyper plane, this method uses the logistic techniques to enfold the result of a linear calculation between zero & one. There is 14 independent variable that makes LR fine for classification. In this study, we obtain an accuracy of 84% by utilize this technique.

Random Forest Algorithm:

It is a type of algorithm of supervised learning. It is utilize for regression and classification, it's easy and straightforward to execute. A forest is contained of trees. This class generates DT on casually choose data samples, gets predictions from all tree, & chooses the most productive results by means of electing. The RF consists ofseveral DT. It generates a forest of trees. In the RF classifier, a greater number of trees will advance accuracy. RF classifier is also known as meta-estimator because it suitable a variety of DT on various sub-set of data and workan average to rise the analytical correctness of the prototypical & manage over-fitting. The size of the sub-setis generally alike to the size of the first (initial) input set butthe sets are drawn with renewal.

Decision Tree Algorithm:

The DT is algorithm that works on numerical as well ascategorical dataset. It is created structures like tree. It is easyand broadly utilized to hold health datasets. It's simple to execute & analyze the dataset in the tree-shaped form of graph. This classifier comes under supervised algorithm. Itis used to resolve classification & regression issues. This technique can be utilized for problem having persistent as well as categorical process & target review. It's the more operative ML techniques used for telling the trees in the graphical method.

The DT model build study based on the following 3 nodes.

- **Root node:** It is the main node, based on all other node'sfunctions.
- **Interior node:** It handles various attributes.
- **Leaf node:** It represents the result of each test.

This method separates the dataset into 2 or additionalsimilar sets grounded on the furthermost main measure.

K-Nearest Neighbor Algorithm:

This is supervised technique classification. It's a non- parametric classification method. Euclidean Distance method can be used to calculate the distance between an attribute & its neighbor. A collection of labeled points is used to spot additional point. Data are grouped grounded on similarity between them & then by using K-NN classifier the absent values of data are fulfilled. The absent numbers are fill up, several methods of prediction put on to the dataset. This is feasible for achieve improved accuracy by applying numerous blends of the techniques or classifiers. K- NN classifier is easy to withstand in the absence of making a system for further expectations. It is flexible & used for regression, classification & search. It is used whenthere is huge amount of data & uncertain decision boundaries amongst classes. Although KNN is an easy MLclassifier, its accuracy is influenced by irrelevant & noisy features.

Support Vector Machine Algorithm:

This algorithm used classification techniques of supervised for regression & classification algorithms as SVR & SVC. This technique splits data spot by means of ahyperplane with majority in huge amount. Numerous kernels are present on which the hyperplane could be clear.Here, we primarily focused on 4 kernels i.e., Sigmoid, poly,Linear and RBF. The Support Vector Machine (SVM) classifier uses smaller store because they utilize a subset of learning spot in the conclusion stage. It classifies 2 classes with the assist of a hyperplane which has the huge majorityto split the data into classes. The majority among the 2 classes depicts the lengthiest interval amid the closest data spot of those classes called as SV.

Implementation

This paper used 6 ML algorithms, using the estimation approach to predict the possibility of heart defeat in a patient admitted in the health center. Hence, as explain above the data source extract from Kaggle has patients' details, which have been gathered from various section. The data has a list of fourteen attributes, which are accumulatively used for diagnosing heart issues in a patient. For observation implementation, the Google Colab tool was used in this research.

Google Colab, is an open-source platform that accept everyone to perform and implement random Python code through the internet & is very well suitable to ML, learning & data examine. It delivers a wide range of pre- programmed operators for numerous functions associated toML, DM, statistical, and many more.

The present algorithm uses six ML models SVM, RF, LG, NB, DT, KNN. The following steps include in the proposed methodology:

First step is referred to as the collection of the data.

In the Second stage it extracts significant values.

Third is the pre-processing stage where we can explore the data. Data pre-processing deals with the missing values, cleaning of data, and normalization depending on the algorithms used. Then the dataset is split into two separate sets:

Train Dataset: It is used to fit the machine learning model.

Test Dataset: It is used to evaluate the fit ML model.

After pre-processing of data, the classifier is used to classify the pre-processed data.

Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics.

Few libraries we imported for the venture are as follows:

NumPy - To work with arrays

Pandas - To work with the CSV, record an information framework

Matplotlib - To draw diagrams using PY-plot, mark the parameters using SRC params and match them to cm. rainbow

Warnings - Neglecting all warnings appearing in themagazine due to past/future devaluation of a component **Train Test Split** - Splitting the dataset into training andtesting information

Standard Scaler - To scale every single highlight, with the goal of best adjusting the machine-learning model to the dataset.

In this model, a successful EHDPS has been established using several techniques. This framework uses 14 health specification like chest cramp, age, BP, sex, cholesterol, fasting sugar, etc. for detection.

IV. RESULT AND DISCUSSION

The objective of this project is to recognize outlook the user has heart issues or not. This study was completed on supervised type of ML algorithms by applying LG, RF, SVM, KNN, DT, NB on the UCI repository. Data was placeand break into learning set & testing set. Dataset recoding is done & supervised algorithms are utilized to obtain accuracy score. This part displays the outcome of those classification models done using PP. The decision is made and noted for both learning datasets and test data sets. A comparison of the accuracy outcome of HDP in the proposed system is given in Table 1.

Algorithms	Accuracy
Logistic Regression	84%
Nave Bayes Classifier	80%
K Nearest Neighbors Classifier	87%
Decision Tree Classifier	79%
Support Vector Classifier	83%
Random Forest Classifier	84%

Table 1: Comparison of Accuracy Values

The table above display that **K-NN Classifier show the top accuracy as 87% in differentiation with the more ML classifiers** utilize in the research study. Since KNN techniques is establish on quality comparison & is well- known classification algorithms at this time in the application clarify due to its clarity and reliability. KNN is an easy method that keeps all the approachable type & classifies latest type grounded on a homogeneity amount.

Correlation Matrix:

Fig 2. shows the Correlation Matrix of attributes.Here, few attributes are extremely correlated & few are not.

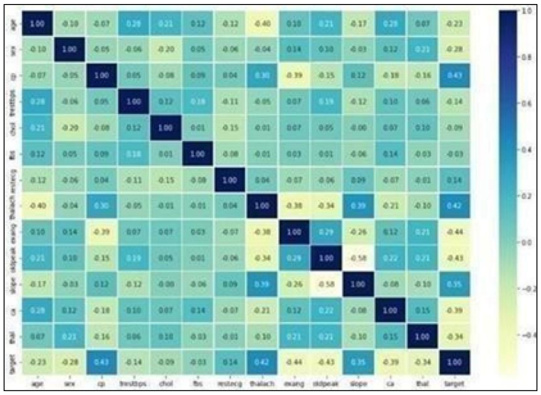


Fig 2: Correlation Matrix

Bar Graph for Target classes with variousattributes:

It's essential that the set of data we're utilizing must be pre-processed & cleaned. The rate of both target classes is shown in below graph.

DOI: 10.48175/IJARSCT-15234

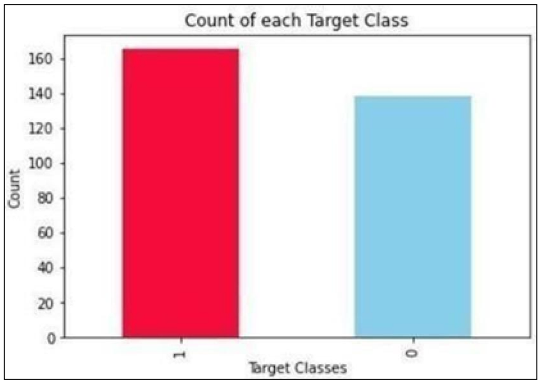


Fig 3: Count of each Target

The graph above displays the diffusion of Target classes vs rate class, used to forecast the total amount of heart disease whether the patient has cardiac disease or not(1 represents having heart disease, 0 represents no heart disease).

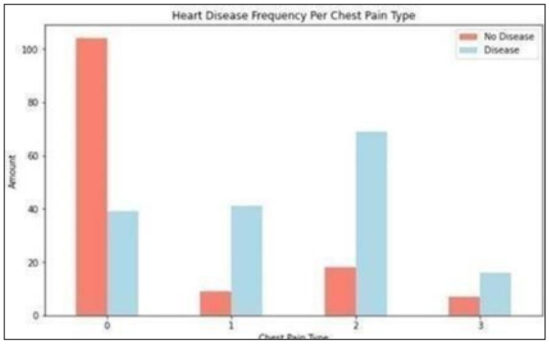


Fig 4: Heart Disease Frequency Per Chest Pain Type

The above graph shows the distribution of heartdisease based on chest pain type.

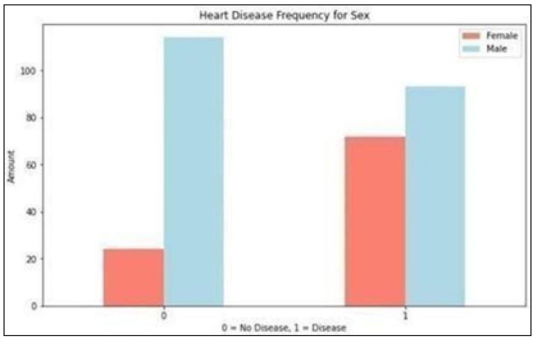
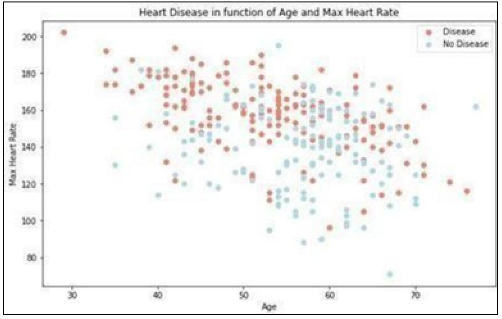


Fig 5: Frequency for Sex in Heart Disease

The above graph shows the distribution of heart disease based on sex where red represents ‘Female’ and blue represents ‘Male’ for each target values 0(No disease)and 1(Disease).

DOI: 10.48175/IJARSCT-15234



The above figure shows the distribution of heart diseasein function of Max Heart Rate & Age where red dot represents ‘Disease’ and blue dot represents ‘No Disease’.

V. CONCLUSION

Throughout this research study, we have tried to evaluate the different ML techniques & predict if someonein specific, given particular separate independent attributesand symptoms, will receive a coronary artery disease or not. The essential thought- out approach of our detail was to lookat the precision and inspect the cause after the modificationof different methods. We have utilized the Cleveland data for heart issues that carry 1025 cases & used the part split to divide the dataset into learning and testing datasets. We have evaluated fourteen attributes & execute six different methods to inspect the accuracy. By the terminate of the Execution part, we have determined that the KNN has display the topmost accuracy amount in our data which is 87% & DT take part in the Minimum with an accuracy amount of 85%. Apparently for more case and different data, one more technique may function in a Well aspect no matter how our situation, we have determined this result. Also, on the off probability that we expand the amount of Learning dataset, perhaps we can obtain the better accurateoutcome but it takes extra time to approach and the device will be slow down than straight away as it is also perplexingand will be Operate more dataset. In this approach, observethis possible stuff we hold this option, which work well forus.

REFERENCES & CITATION

[1]. Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y

[2]. Rairikar, V. Kulkarni, V. Sabale, H. Kale and A. Lamgunde, "Heart disease prediction using data mining techniques," 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, India, 2017, pp. 1-8, doi: 10.1109/I2C2.2017.8321771.

[3]. V. Ramalingam, V., Dandapath, A., & Karthik Raja,M. (2018). Heart disease prediction using machine learning techniques : a survey. International Journal of Engineerin g & Technology, 7(2.8), 684.

[4]. S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan and T. Zhu, “Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework,” 2017 IEEE 2nd International Conference on Big Data Analysis, 2017.

[5]. Perumal, R. Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques. Int. J. Adv. Sci. Technol. 2020, 29, 4225–4234

[6]. A. Sahaya Arthy, G. Murugeshwari A survey on heart disease prediction using data mining techniques(April 2018)

[7]. Kumar, N.K.; Sindhu, G.; Prashanthi, D.; Sulthana, A. Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers. In Proceedings of the 2020 6th.

[8]. Gupta, A.; Kumar, R.; Arora, H.S.; Raman, B. MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. IEEE Access 2019, 8, 14659–14674

[9]. S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp.1-5,doi: 10.1109/ICIICT1.2019.8741465.

[10]. A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.