JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Machine Learning Based Cardiovascular Disease Prediction

Nikita Waghmode[1] , Sanidhya Keche[2] , Ganesh Shillare[3] , Poonam Badadhe[4] , Chandrashekhar Patil[5]

Department of E&TC, SKNCOE, SPPU, Pune

[1]nikwaghmode@gmail.com, [2]sanidhyakeche2000@gmail.com, [3]ganeshshillare@gmail.com,
[4]poonam.badadhe_skncoe@sinhad.edu, [5] chandrashekhar.patil_skncoe@sinhgad.edu

***Abstract*- Cardiovascular disease is increasing worldwide, even among young people. Cardiovascular disease prediction is a complex task that requires detailed prior knowledge. In the proposed model, heart disease prediction is made from data collected from Kaggle. In this project, we present a comparison of learning systems such as decision tree, Support Vector Machines and Random forest. We use different algorithms by using different classifiers from which we can conclude with best accuracy.**

Keywords: **Machine Learning(ML), Decision Tree(DT),SVM, Random Forest(RF), Kaggle.**

I. INTRODUCTION

The healthcare industry has vast use of Machine Learning applications. Machine Learning can play an essential role in predicting presence/absence of disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important intuitions to doctors who can then change their diagnosis and dealing methodsper patient basis. In this project, it proposes the comparative analysis of classifiers like Decision Tree, SVM and Random Forest. Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future Heart Disease by analysing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. Machine Learning techniques can be a boon in this regard. By collecting the data from various sources, classifying them under suitable headings &finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

II. LITERATURE REVIEW

Heart disease prediction can be perform using the algorithms- support vector machine, Random Forest and Decision Tree. The objective is to effectively study whether the patient has any heart disease. The health professional enters the input values from the patient's health report. The data is then fed into the machine learning model which provides the probability of having the heart disease.

[1]According to paper(1)the model has been trained on a dataset with attributes like gender, age, resting blood pressure, cholesterol, fasting blood sugar, etc. It is a web-based machine learning application where the user inputs his medical details based on these attributes to predict his heart disease. The algorithm calculates the probability of having a heart disease and the result is displayed on the web page itself.

[2] According to paper (2) raw dataset contains unbalanced data of class distribution. BY applying 3 sampling techniques, accuracy and recall rates increased drastically. SVM gives best accuracy.

[3]Ruby Hasan, proposed in "COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR HEART DISEASE PREDICTION" Machine learning based solutions are widely used in healthcare sector for analysing patients' data, predicting diseases and suggesting possible treatments. With a number of machine learning techniques available today, it is important to identify the most efficient and accurate technique especially in critical domains like healthcare. In decision tree algorithm the tree is built using certain conditions which gives True or False decisions. Decision tree for a tree like structure having root node, leaves and

branches base on the decision made in each of tree. Decision tree also help in the understating the importance of the attributes in the dataset. Dataset splits in training and testing by using some methods.

[4] M.Snehith Raja, Nageswara Rao Sirisala, M .Anurag, Ch. Prachetan Reddy in paper "MACHINE LEARNING BASED HEART DISEASE PREDICTION SYSTEM", Predicting the occurrence of disease at early stages is a major challenge nowadays Random Forest algorithm is an efficient algorithm which is an ensemble learning method for regression and classification techniques. The algorithm constructs N of Decision trees and outputs the class that is the average of all decision trees output. So accuracy of prediction at early stages is achieved effectively.

[5] Pranav Motarwar , Ankita Duraphe, G Suganya , M Premalatha, in paper "COGNITIVE APPROCH FOR HEART DISEASE PREDICTION USING MACHINE LEARNING", says about prediction of patterns to prevent and control diseases is a challenging and a prominent requirement in medical domain. In this paper, they propose a machine learning framework to predict the possibility of having heart disease using various algorithms. Data visualization was used on the dataset to visualize correlation or dependency between any of the featured attributes in the following dataset. Feature selection was used to select the best attributes in the dataset. This process presents the highest data quality for performing classification algorithms. Further enhancement techniques are used to increase the basic accuracy of each algorithm. The dataset was trained with 80% data, 242 instances. The rest 20% data, 61 instances are predicted."

[6] According to paper (6), Health Care Field having enormous data, for processing the data that must use any advanced techniques which will be helpful to provide the effective results and making effective decisions on data and getting the appropriate results. This shows that whether the patient has the heart disease by considering the parameters such as age, gender, chol, exang, restecg, thal, slope, oldpeak, trestbps, fps, cp and thalach. That experiment is performed by training the dataset containing of 304 records with 13 different parameters.

PROPOSED SYSTEM

The data is pre-processed into the required format. The data is then divided into two parts training and testing data. This system is implemented using the following classifiers.

1. Decision Tree
2. Support Vector Machine
3. Random Forest

I. SYSTEM ARCHITECTURE

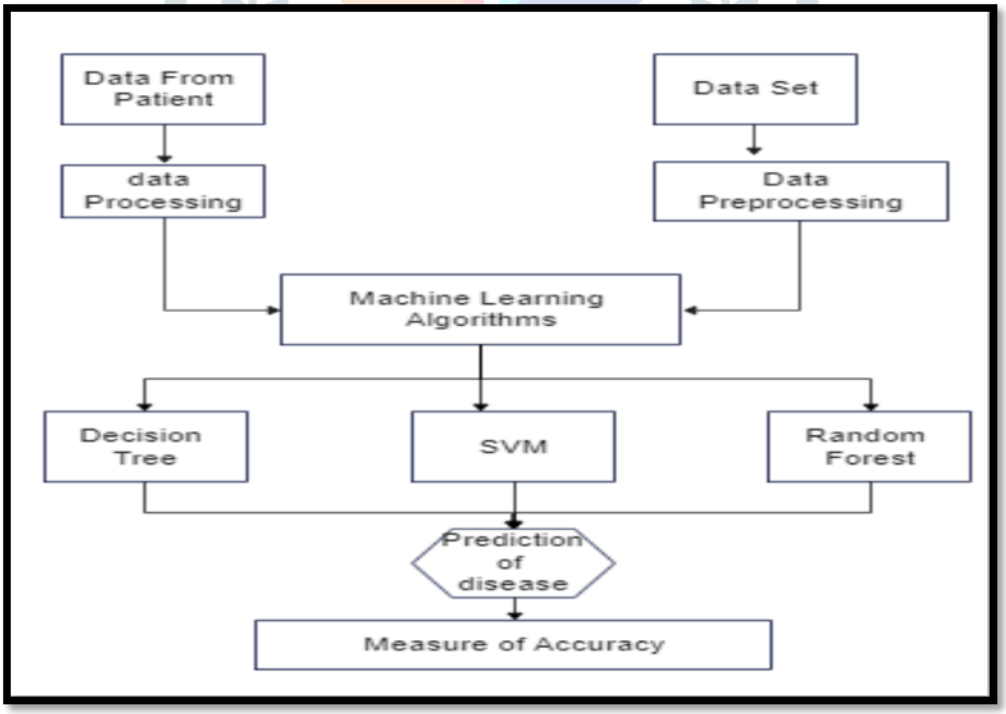


Fig 1: Block Diagram of Heart Disease Prediction Mode

BLOCK DIAGRAM DESCRIPTION

Dataset collection is collecting data which contains patient details. Attributes selection process selects the useful attributes for the prediction of heart disease. After identifying the available data resources, they are further selected, cleaned, made into the desired form.Different classification techniques as stated will be applied on preprocessed data to predict the accuracy of heart disease.

Accuracy measure compares the accuracy of different classifiers

I. IMPLEMENTATION

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre-processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1.) Collection of Dataset
2.) Selection of attributes
3.) Data Pre-Processing
4.) Balancing of Data
5.)Disease Prediction

Collection of dataset

Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 70% of training data is used and 30% of data is used for testing. The dataset consists of 13 attributes; all attributes are used for the system.

Selection of attributes

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. Various attributes of the patient like age, sex, chest pain type, blood pressure, cholesterol, exercise, depression, ECG, blood sugar, thallium etc. are selected for the prediction.

Pre-processing of Data

In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Pre-processing of data is required for improving the accuracy of the model.

Balancing of Data

Imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling

(a) Under Sampling: In under Sampling, dataset balance is done by the reduction of the size of the ample class.
(b) Over Sampling: In over Sampling, dataset balance is done by increasing the size of the scarce samples.

Prediction of Disease

Various machine learning algorithms like SVM, Decision Tree, Random forest Tree are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.

II. METHODS FOR PREDICTION

Figure 1 shows the process of predicting heart diseases using machine learning algorithms. The process of data collection, pre-processing, classification and prediction for predicting of results, whether the person has presence or absence of heart disease is detected. Numerous classification techniques, such as SVM, Decision Tree, and Random Forest are studied in this paper to determine the best suited machine learning algorithm for accuracy.

SVM Algorithm –

SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems (primarily – classification).

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed a Support Vector Machine.

In the SVM algorithm, to maximize the margin between the data points and the hyperplane. The loss function that helps maximize the margin is hinge loss.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

$$c(x, y, f(x)) = (1 - y * f(x))_+ \quad \ldots\ldots\ldots\ldots$$

Hinge loss function (function on left can be represented as a function on the right) **{I}**

The cost is 0 if the predicted value and the actual value are of the same sign. If they are not, then calculate the loss value. We also add a regularization parameter the cost function. The objective of the regularization parameter is to balance the margin maximization and loss. After adding the regularization parameter, the cost functions looks as below.

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+ \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

Loss function for SVM **{II}**

Now that we have the loss function, we take partial derivatives with respect to the weights to find the gradients. Using the gradients, we can update our weights.

$$\frac{\delta}{\delta w_k} \lambda \parallel w \parallel^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} (1 - y_i \langle x_i, w \rangle)_+ = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases} \quad \ldots\ldots\ldots\ldots$$

Gradients **{III}**

When there is no misclassification, i.e. our model correctly predicts the class of our data point, we only have to update the gradient from the regularization parameter.

$$w = w - \alpha \cdot (2\lambda w) \quad \ldots\ldots\ldots\ldots\ldots\ldots$$

Gradient Update — No misclassification **{IV}**

When there is a misclassification, i.e. our model make a mistake on the prediction of the class of our data point, we include the loss along with the regularization parameter to perform gradient update.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w) \quad \ldots\ldots$$

Gradient Update — Misclassification **{V}**

Decision Tree Algorithm-

Decision tree is a non-parametric supervised learning algorithm. Used to build automated predictive models. It is a tree-like structure that is used as a decision-support tool in both classification and regression problems. For both categorical and continuous input and output variables. A decision tree is a classifier in the form of a tree structure with two types of nodes:

Decision node: Specifies a choice or test of some attribute, with one branch for each outcome

Leaf node: Indicates a decision or classification

The decision tree model is generated by putting the best attribute of the dataset as root of the tree or attribute selection it uses two methods-

1. Information Gain:

Information Gain= Entropy(S)[(Weighted Avg) *Entropy(each feature ……..{VI}

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(S) = - $P(yes)P(yes) - P(no)P(no)$ **……. {VII}**

Where, S= Total number of sample

P (yes) = probability of yes

P (no) = probability of no

2. Gini Index

Gini Index= 1- ∑jPj2……{VIII}

There are several subsets in the training package. Each subset contains information for the same attribute. The decision tree is associated with decision rules and conditions along the path form and if else rule to allow the algorithm to decide on the basis of the situation, where the outcome is the contents of the leaf node. As the algorithm has more number of layers it becomes more complex. So over fitting problem will occur frequently when building a decision tree model. The use of a decision tree aids in the visualization of logic. It generates all possible decision outcomes. The decision tree algorithm's main goal is to prioritize the attribute that can provide the highest level of accuracy.

Random Forest Algorithm –

It is the most used supervised machine learning algorithm for classification and regression. It uses ensemble learning method in which predictions are based on the combined results of various individual models. Finally voting is used to find the class of the predicted value.

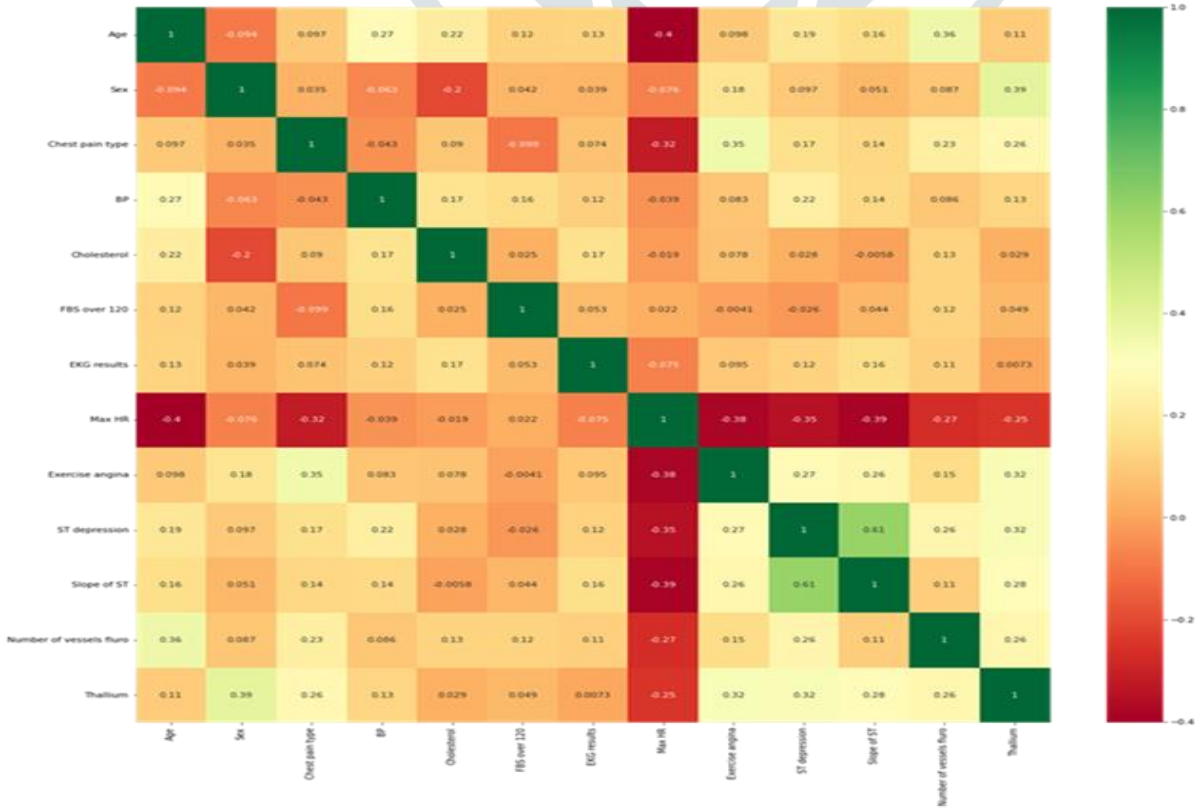
It contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Greater the number of trees gives the higher accuracy in Random Forest Algorithm.

Table1: Comparison of SVM, Decision Tree and Random Forest Algorithm

Technique	Primary Problem	Predictor	Raw Implementation	Time taken	Accuracy	Overfitting Issues	Complexity
SVM	Binary	Numeric	Easy	Less	Low	No	Simple
Decision Tree	Multiclass or binary	Numeric Categorical	Difficult	Moderate	Medium	Possible	Simple
Random forest	Multiclass or binary	Numeric Categorical	Difficult	More	Highest	Reduces	More Complex

Co-Relation Map:



III. RESULT ANALYSIS

The three distinct classifier techniques: Decision tree, SVM and Random Forest are employed to build the model. Feature selection is performed to assist us in selecting the features that will improve prediction. The random forest algorithm generates a decision tree with multiple levels. Accuracy is one of the measurements used to evaluate classification models. Informally, accuracy refers to our model's percentage of correct predictions as shown in Table-

$$\textbf{Accuracy} = \frac{(TP + TN)}{Total} \text{..... \{IX\}}$$

Table 2: comparative analysis of accuracy in percentage

Algorithms	Accuracy 1(216,54)	Accuracy2(175,95)
SVM	77.78%	83.16%
Decision Tree	85.19%	72.63%
Random Forest	85.19%	86.32%

Discussion: In Table2: Column 1 gives accuracy of dataset1{Taken From KAGGLE},training data 216 and testing data 54 attributes are used. Here we get minimum accuracy of SVM and Decision Tree and Random Forest gives same accuracy. In Column2 accuracy of dataset2 {Taken From GITHUB}, training data 175 and testing data 95 attributes are used. Here we get maximum accuracy of Random Forest, so we can conclude that Random Forest can be used for better result.

IV. CONCLUSION

Heart diseases are predicted with system that would early prediction the heart diseases effectively so that immediate medical attention can be given. In this regard, ML can provide good approaches for accelerating the diagnosing process. Furthermore, data mining technologies, particularly tree-based algorithms, are used in this study to provide accurate predictions for Heart disease patients. This study could be utilised as a model for developing a healthcare system for Heart disease patients in the future. The results for algorithms based on accuracy are like SVM is 77.78%, Decision Tree 85.19% and Random Forest 85.19%.

V. REFERENCES

[1]TanishaRakshit ,"Comparative Analysis and Implementation of Heart Stroke Prediction using Various Machine Learning Techniques", *International Journal of Engineering Research & Technology (IJERT);* ISSN: 2278-0181; Published by, www.ijert.org; NTASU - 2021 Conference Proceedings .

[2]PoojaAnbuselvan ,"Heart Disease Prediction using Machine Learning Techniques", *International Journal of Engineering Research & Technology (IJERT)*; http://www.ijert.org ISSN: 2278-0181; Vol. 9 Issue 11, November-2020

[3]Ruby Hasan," Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction"Published in: *ITM Web of Conferences* 40, 03007 (2021) ;https://doi.org/10.1051/itmconf/20214003007.

[4]Pranav Motarwar, AnkitaDuraphe, G Suganya , M Premalatha," Cognitive Approach for Heart Disease Prediction using Machine Learning" Published in 2020 *International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE*

[5]Shaik Farzana, DuggineniVeeraiah," Dynamic Heart Disease Prediction using MultiMachine Learning techniques" Published in *UNIVERSITY OF WESTERN ONTARIO. Downloaded* on May 26, 2021 at 10:48:55 UTC from IEEE Xplore.

[6]MangeshLimbitote, "A survey on Prediction Techniques of Heart Disease using Machine Learning*", International Journal of Engineering Research & Technology (IJERT);* http://www.ijert.org ISSN: 2278-0181; Vol. 9 Issue 06, June-2020.