Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

A SURVEY: HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

Noor Mohd[1], Juhi Sharma[1], Deepak Upadhyay[2]

[1]Department of Computer Science & Engineering Graphic Era Deemed to be University, Dehradun, Uttarakhand, India
[2]Department of Computer Science & Engineering, Graphic Era Hill University, Dehradun, Uttarakhand India, 248002

ABSTRACT

This Heart attacks and the deaths caused by heart diseases are increasing at every part of the world. Heart troubles are caused by people's lives being under more stress. One of the most important organs in our body is probably the heart. It assists in controlling blood circulation and flow to all of the body's organs. Today, this disease is responsible for the greatest number of fatalities. Generally a person suffering from any kind of heart problem shows few signs and symptoms from which we can easily judge it and thus can help the individual on time. The symptoms are like high intensity chest pain, fast rate of heartbeat, and pain with some uneasiness in breathing are recorded. This information is dissected on customary premise. In this survey, an outline of the coronary illness furthermore, its present techniques is right off the bat presented. Besides, a top to bottom examination of the most important AI methods accessible on the writing for coronary illness forecast is momentarily expounded. Various machine learning algorithms that performs well in every detection mechanism are Naive Bayes, Decision Tree, Support Vector Machine, Artificial Neural Network, Random Forest, K-Nearest Neighbours and many more. The performance of every machine learning model depends on how accurately it makes predictions. Therefore, one of the key components of evaluating the model's efficiency is accuracy. In our research, Random Forest provides the best accuracy of 91% in predicting heart disease.

Keywords: Parasitic Heart disease, artificial neural network, Machine Learning, Naïve Bayes', SVM, Classification Technique, Decision Tree, CVD, Accuracy.

INTRODUCTION

Among all lethal infection, coronary episodes sicknesses are considered as the most predominant. Clinical professionals lead diverse studies on heart sicknesses and assemble data of heart patients, their manifestations and sickness movement. Progressively are accounted for about patients with normal sicknesses who have common side effects. In this era everyone is busy in making their life better and all are working hard to achieve their goals in studies, career, jobs, relationships and every part of their life. This increasing sense of being the best and above all in all phases of life have made every one's life a hectic journey. Knowingly or unknowingly we all deal with such stress in our daily life and it leads to stress on our brain, heart, and mind, in fact in all parts of our body. This increasing stress leads to the heart diseases and that is the reason that today most of the individuals

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

are suffering from some or the other heart problem. Ignoring home made fresh food and opting junk food and fast food is also one of the reason of this heart disease rate increase. People are ignoring healthy life style and make a fight over the idea of the food whenever incapacitated the go for their own medication due to all of these little remissness it's anything but a huge threat that is the coronary disease.

The term 'coronary illness' incorporates the assorted sicknesses that influence heart. The quantity of individuals experiencing coronary illness is on the ascent (wellbeing subjects, 2010). Data mining has been used in a grouping of usages like publicizing, customer relationship the board, planning, and prescription assessment, ace conjecture, web mining and flexible enlisting. Of late, data mining has been applied viably in clinical consideration blackmail and perceiving abuse cases.

Information examination ends up being significant in the clinical field. It's anything but a huge base to essential decisions. It helps with making a complete report suggestion. Potentially the primary businesses of data examination is that it helps in getting human inclination far from clinical end with the help of fitting quantifiable treatment. By utilization of information digging for exploratory investigation in light of nontrivial data in enormous volumes of information. The medical services businesses gather gigantic measures of information that contain some secret data, which is helpful for settling on viable choices for giving proper outcomes and making powerful choices on information, a few information mining procedures are utilized to better the experience and end that have been given. The system will operate on the data collected using data mining to give a customer arranged approach to manage the latest and secret models for data. The data which is executed can be used by the clinical consideration experts to further develop nature of organization and to diminish the level of negative medicine sway.

Data burrowing holds fantastic potential for the clinical consideration industry to engage prosperity systems to intentionally utilize data and examination to perceive weaknesses and best practices that further develop mind and decrease costs. According to (Wurz and Takala, 2006) the opportunities to additionally foster consideration what\'s more, decline costs all the while could apply to as much as 30% of as a rule clinical consideration spending. The productive utilization of data mining in extraordinarily obvious fields like e-business, displaying and retail has provoked its application in different endeavors and regions. Among these areas simply finding is medical care. The medical care climate is still „information rich“ yet „knowledge poor“. There is an abundance of information accessible inside the medical services frameworks. Be that as it may, there is an absence of viable investigation apparatuses to find stowed away connections and patterns in the information for African classes. Coronary illness is the sort of infection which can cause the demise. Every year an excessive number of people groups are kicking the bucket because of heart illness. Due to the weakening of the heart muscle, coronary disease might occur. The heart's failure to pump blood is a similar way to describe cardiovascular collapse. Also known as coronary corridor sickness, this condition affects the heart (CAD). Deficient blood flow to courses can lead to computer-aided design. Utilizing symptoms such a fast heartbeat, chest pain, high blood pressure, heart failure, and others, one might identify coronary disease.

There are many types of heart sicknesses and side effects, for example

1) Coronary illness in blood vessels: chest torment, windedness, torment in neck throat.

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

2) Heart sickness brought about by unusual pulses: moderate heartbeat, uneasiness, chest torment and so on. Diagnosing a disease with the most efficient and effective way is a very important function of any health-care unit. If a disease detection is done at early it becomes easier to save life of a patient. Machine learning methods are very helpful in early detection of the disease. Because of the instant results given by the machine learning models the important time of patient can be saved and treatment can be started on time before it is too late. Heart disease is one such dangerous situation where a patient's life is in danger because it is very difficult to diagnose this disease. In this paper we aim to perform the survey on earlier work done in this domain also we are working to present a model to detect heart disease at early stages.

1.2 Overview of Machine Learning Algorithms

Machine Learning is a complete package of algorithms that enables a system or application to learn efficiently and provide the results on certain other similar data based on the leaning it got from previous data. This process of continuous learning and prediction makes the system more efficient for future data analysis work and for future predictions. While working with machine learning algorithms we need to feed training data to the algorithm using which the machine learning model is trained to do task, and get results on the same way with some other data of similar type. Thus it allows us to develop intelligent systems for our business.

The four main categories of machine learning algorithms are reinforcement learning, semi-supervised learning, unsupervised learning, and supervised learning.

Supervised Learning means that the algorithm will be given some historic labelled data from which it can learn and then the algorithm works with some new data and predict the output. These algorithms works by matching the pattern to provide the output with maximum accuracy. Thus, we can say that supervised machine learning algorithms works by learning from historic data to generate suitable equations for the label assigned to a data. These algorithms keep on improving the performance by learning from the difference between its calculated result and exact value or correct value. These algorithms uses regression and classification. These algorithms are applicable on areas like Images segmentation, Emotion detection, spam detection, fraud detection, and speech recognition.

On the other hand the unsupervised ML algorithms does not depend upon learning from the old historic data, in fact these algorithms works with the completely unlabelled data. These algorithms discovers the underlying patterns and relation between the data without any information about it in labels. These algorithms works on the basis of clustering and association. Unsupervised ML algorithms are useful in areas like Analysis of clusters and finding association rules between clusters and participants of cluster. Anomaly detection is also an area where these unsupervised machine learning algorithms performs efficiently. Examples of unsupervised machine learning algorithms includes k-means algorithm, DBSCAN algorithm, Principal component analysis, association rules. Search engines, network analysis and building recommendation systems for ecommerce websites are a few areas where these algorithms are useful.

Whereas reinforcement learning enables the system to learn in a virtual environment with simulations using trial and error method. Systems developed using these algorithms does not

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

initially have the labelled data for learning but it learns while dealing with data in the virtual simulated environment. This algorithm learns from its own results in a virtual environment. In contrast to this the semi-supervised learning have to deal with both labelled and unlabelled data. The different ML algorithms are used in medical imaging also in [21, 22].

According to the nature of the problem the system build using reinforcement learning learns to organize the data as per the structure of the problem and also make predictions for the unlabelled data.

1.3 Machine Learning Techniques used in different papers

There are various machine learning techniques that are used in papers that are used in this survey. These techniques have particular results and to test the performance of various classifiers using the dataset related to heart disease. We used various ml algorithms Logistic Regression (LR), Naïve Bayes (NB), decision tree (DT), random forest (RF). We addressed various Machine learning classifiers used for this study.

1.3.1 **Naïve Bayes:** Naive Bayes' classifier is a supervised approach to learning, and is very useful in applications. It's named "naive," as it's based mostly on trying to simplify premise that the values of the attributes are linearly independent. The classifier Naïve Bayes (NB) is simpler than other classifiers and has a fast detection speed. Naïve Bayes algorithm is a type of classification algorithm that works on the basis of the Bayes' theorem. It assumes that all the presence of all the features in a class are independent of each other. The model developed with Naïve Bayes' algorithm are easy to implement and also are capable of working with large datasets. The Naïve Bayes algorithms Bernoulli Naive Bayes, Gaussian Naive Bayes and Multinomial Naïve Bayes are available in three types.
Bayes' theorem helps in predicting the posterior probability, the basic equation of naïve bayes' is given as,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

1.3.2 **Logistic Regression:** This machine learning algorithm works by establishing a relationship between the variables and the available classes. Logistic regression machine learning algorithm comes under the category of classification algorithms, it works by predicting the class of a variable or instance data. Thus, finding a feature class relationship is the core functionality in logistic regression machine learning algorithms. For linear regressions the predicted value (y) of class can be 0 or 1 for a feature. The probability of a feature belonging to a class can be somewhere between 0 and 1. Logistic function of this algorithm can be defined as,

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Where, right side are called odds and the left side we have logit or log-odds function.
Odds here is the probability ratio of success to failure.

1.3.3 **Random Forest (RF):-**Random forest is a supervised learning method which is used for regression and classification. A random subset of the total set of labelled samples is trained for each

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

tree. In the classification method, most approved class among all the tree in the model shows the classifier's result. Random forest algorithm is quite efficient and it is one of the accurate learning algorithms. Random forest algorithm achieves a (96%) detection rate. Random forest algorithm provides a higher detection rate as compare to other classifiers. Random forest algorithm works with the labelled data and gains the capability of predicted the result for the new data. Random forest algorithm is applicable on classification problems and regression problems. While working with the regression problems we use Mean-square-error that tells about the difference between the predicted and actual node value and helps in choosing the right branch for efficient performance of our random forest model.

$$MSE = \frac{1}{N}\sum_{k=1}^{n}\binom{n}{k} \quad (fi - yi)^2$$

Here,
N is the number of data points,
fi is the value predicted by the model,
yi is the actual value for the data point i.

1.3.4 **Decision Tree:-**Decision tree is a supervised learning approach used for classification, regression and branching method. Three underlying elements decision nodes, branch and leaf nodes are of the tree. Decision node defines a check over a specific attribute. To this attribute, each branch represents one of its possible values. Ultimately, node represents the category the object belongs to. Various algorithms for the decision tree exist. This classifier is used for performing classification and creates a predictive model.

1.3.5 **Long Short Term Memory:** LSTM is a special kind of sequential recurrent neural network that can remember the information it have gone through before and this makes it special. Neural networks works on a structure like neurons in our brain they pass information and predict results. Earlier introduced RNNs had problems of vanishing gradient but now with use of LSTM we are able to handle this problem. Neural networks works in similar fashion as our brain does. RNNs had the problem of vanishing gradient that makes that incapable of remembering information for long time. Now, LSTM have overcome this issue and are able to remember long term dependencies. Architecture of the LSTM network is also similar to the RNN. LSTM network have three parts and each of these parts performs specific functions. First part of the LSTM network in figure[4444444] is responsible for deciding whether the coming information from previous timestamp is important to remember or it can neglect it from remembering because of unimportant information. Second part is responsible for adding and updating new information from the coming data, this helps the model in learning. The third part of this LSTM network is responsible for sending the updated information to the next timestamp. These three parts are called as forget gate, input gate, and output gates respectively.

Similar to RNN the LSTM network also have the concept of hidden state. If t is the current time stamp them h(t-1) will refer to the hidden state of the previous time stamp whereas h(t) refers to the hidden state of current timestamp. Additionally LSTM also have the concept of cell state that is also called as long term memory c(t-1) refers to the cell state of previous time stamp whereas c(t) refers to the cell state of the current time stamp. One of the important and foremost step in an LSTM

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

network is to decide if the information coming is important to be remembered for long term or it is not-important and can be forget. Here comes the responsibility of the forget gate that is the first gate in LSTM network and performs this step.

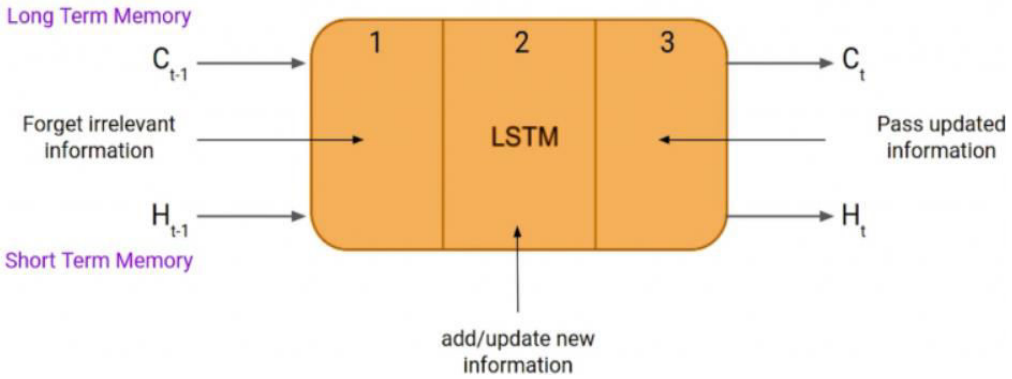

Figure. 1: Long Short Term Memory Network

Equation giving details about function of the forget gate is given below,

$$Forget\ gate, f_t = \sigma(X_t * U_f + H_{t-1} * W_f)$$

Where,

X_t is the input for current time stamp,
U_f is the weight associated input,
H_{t-1} is the hidden state of previous time stamp,
W_f is the matrix of weights associated with the hidden state.

Here, if value of ft =0 then the network will forget that information and if it is 1 it will remember the information

related work

AI is a type of computerized reasoning that empowers the machine to make and execute calculations that can gain from past encounters. We utilized an assortment of classifiers from regulated and unaided figuring out how to anticipate and decide the dataset's exactness. Cardiovascular illnesses allude to different problems that can influence the heart and circulatory framework. Coronary illness has been normal for quite a while is as yet perhaps the most serious sicknesses today. As per WHO, cardiovascular illnesses (CVDs) are the number 1 reason for death universally. Nine out of ten CVD demise are because of cardiovascular failures and strokes, and 33% of these demise happen rashly in individuals under 70 years of age [1]. Our examination can distinguish individuals who are bound to be determined to have coronary illness dependent on their clinical history. It predicts dependent on factors like sugar level, circulatory strain, chest torment, cholesterol, and so forth. Along these lines, individuals will actually want to know about themselves in advance and avoid potential risk. To check the exactness and investigate various models, we utilized both KNN and strategic relapse models. We utilized a dataset of individuals with and without coronary illness to foresee coronary illness. We utilized 14 distinct traits of a patient to foresee in the event that they are defenceless to coronary illness. The more productive of these calculations is KNN which gives us an exactness of

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

75.409%. We utilized numerous graphical portrayals to introduce the consequences of the undertaking's expectations and such.

This undertaking was roused by a lot of work on the identification of CVDs utilizing Machine Learning calculations. ML calculations have been utilized to make a few proficient heart infection expectations. Utilizing the past and current machine learning and profound learning models, the model joining IHDPS had the option to anticipate the probability of an individual getting coronary illness pretty precisely [2]. It zeroed in on fundamental credits like age, sex, pulse, and glucose. Be that as it may, new models which utilize profound learning and neural organization are more productive, exact and dependable. Another neural organization model had the order force of 77% to effectively group the presence of Coronary Heart Disease (CHD) and 81.8% to precisely order the shortfall of CHD cases on testing information, which is 85.70% of the all out of their dataset [3]. While the review esteems got from other AI strategies, like SVM and arbitrary timberland, are practically identical to that of our proposed CNN model, our model predicts the negative cases with higher exactness. They express that other AI models, like SVM and arbitrary woods, produce review esteems equivalent to their proposed Convolutional Neural Network (CNN) model. The CNN model, then again, predicts the negative cases with more noteworthy exactness. We gathered the dataset from Kaggle to prepare the model [4].

The information was assembled from different occasions. The data set is taken from the UCI repository [5] in many studies related to heart disease prediction. It at first had 76 characteristics in any case, a subset of 14 ascribes was utilized. The dataset incorporates a assortment of people and their accounts of coronary illness, also as other ailments. The dataset comprises of the clinical history of 303 unique patients of various traits spread across. This dataset gives well insights concerning the patient's clinical attributes, like age, chest torment types, blood pressure, sugar level, angina, etc, which empowers us in deciding if the patient has been determined to have coronary illness.

There is number of works has been done identified with illness forecast frameworks utilizing diverse AI calculations in clinical Centres.

S. k. Mohan et al, [6] proposed a hybrid model using ML techniques working on the precision in the assumption for cardiovascular disease. The assumption model is made with different mixes of features and a couple known game plan methodologies. We produce a further developed display level with an exactness level of 88.7%.Similarly they instructed about Different information mining approaches and assumption strategies,

S. Nikhar et al. [7] also wrote a paper on the similar topic where they explained how different machine learning classifiers can be used to work on dataset related to this disease. They have used various classifiers like Naïve Bayes, Decision Tree to perform the testing and predictions to see which classifier gives the best result in terms of accuracy.

A.Gavhane et al. [8] proposed a framework for prediction of heart disease that is based on neural networks. By using several layers of the neural networks they perform the task of prediction on the

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

dataset created using multi-facet perceptron model. Feed forward network of neural networks is used here to develop an efficient model.

A. Kishore et al, [9] developed a framework using RNN artificial neural network from the set of deep learning algorithms to make an intelligent model to predict the chances of a patient to be affected with heart disease. The developed model performs very well in the prediction task together with discussing many structures and modules related to this study

L. Rao et al, [10] also uses ml techniques to build a model for heart disease prediction. Machine Learning Techniques are used on medical data related to heart disease.

S. Krishnan et al., [11] proposed a prediction and classification model that worked on classification model to detect disease rate in male patients. Different machine learning tools and methods coronary illness is anticipated in this System.

A. Golande et al,[12] proposed a prediction model developed using machine learning techniques that also incorporates in it some data mining techniques. This model is basically developed to help medical practitioners in their work. Normally used system used are choice tree, k-nearest and Naïve Bayes. Other one of a kind portrayal based techniques used are pressing computation, Part thickness, continuous immaterial smoothing out.

V.V. Ramalingam et Al,[13] developed method for Heart infection expectation utilizing AI methods in which AI calculations and methods used with various clinical and medical data to computerize the functioning of the model developed for the checking of a big amount of data related to heart patients . This media=cal data is complex and not easy to understand for a new person and thus it is better to allow use of these AI enabled models to work in predictions with complex dataset.

Various researchers, lately, have been using a couple AI calculations and strategies that are used with different clinical and medical data to make automatic disease detection possible. As time goes on, several analysts have been using a few AI techniques to assist the medical care sector and the professionals in the investigation of heart-related ailments.

METHODOLOGY

We have worked on different classifiers to see the performance of each classifier on the publically available dataset for heart disease study. The dataset is available at UCI repository [5]. It contains 76 features for a detail study of all the related issue in study of heart attacks, heart failure and any other heart disease but for our study we have selected only 11 features to see the accuracy of our machine learning classifiers. Selected features are explained in details here.

Dataset Details:

This dataset consists of 11 features and a target variable. It has 6 nominal variables and 5 numeric variables. That are age, sex, chest pain, bp, cholestrol, fasting bp, ecg, heartrate etc.,

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

Data Cleaning & Pre-processing:
This is a very important step before starting any work with a dataset. Here we have renamed a few columns as per our better understanding. Further we have changed the names of the extracted features to the categorical feature names. So, there are total 1189 records and 11 features with 1 target variable.

```
In [7]: # converting features to categorical features

dt['chest_pain_type'][dt['chest_pain_type'] == 1] = 'typical angina'
dt['chest_pain_type'][dt['chest_pain_type'] == 2] = 'atypical angina'
dt['chest_pain_type'][dt['chest_pain_type'] == 3] = 'non-anginal pain'
dt['chest_pain_type'][dt['chest_pain_type'] == 4] = 'asymptomatic'


dt['rest_ecg'][dt['rest_ecg'] == 0] = 'normal'
dt['rest_ecg'][dt['rest_ecg'] == 1] = 'ST-T wave abnormality'
dt['rest_ecg'][dt['rest_ecg'] == 2] = 'left ventricular hypertrophy'


dt['st_slope'][dt['st_slope'] == 1] = 'upsloping'
dt['st_slope'][dt['st_slope'] == 2] = 'flat'
dt['st_slope'][dt['st_slope'] == 3] = 'downsloping'

dt["sex"] = dt.sex.apply(lambda  x:'male' if x==1 else 'female')

In [8]: dt['chest_pain_type'].value_counts()

Out[8]: asymptomatic        625
        non-anginal pain    283
        atypical angina     216
        typical angina       66
        Name: chest_pain_type, dtype: int64
```

Figure 1. Converting features to categorical features.

Exploratory data analysis:
In this step we see the details of all the categorical features created in last step. This gives us a summary of all the numerical values and the categorical values related to a feature.

Out[15]:

	age	resting_blood_pressure	cholesterol	fasting_blood_sugar	max_heart_rate_achieved	exercise_induced_angina	st_depression	target
count	1189.000000	1189.000000	1189.000000	1189.000000	1189.000000	1189.000000	1189.000000	1189.000000
mean	53.708158	132.138772	210.376787	0.212784	139.739277	0.387721	0.923549	0.528175
std	9.352961	18.369251	101.462185	0.409448	25.527386	0.487435	1.086464	0.499416
min	28.000000	0.000000	0.000000	0.000000	60.000000	0.000000	-2.600000	0.000000
25%	47.000000	120.000000	188.000000	0.000000	121.000000	0.000000	0.000000	0.000000
50%	54.000000	130.000000	229.000000	0.000000	141.000000	0.000000	0.600000	1.000000
75%	60.000000	140.000000	270.000000	0.000000	160.000000	1.000000	1.600000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	1.000000	6.200000	1.000000

Figure 2. Values of all parameters for exploratory data analysis

Distribution of Heart disease (target variable)
In this step of exploratory data analysis we created the graph to distinguish between the patients with heart diseases and the normal patients. As our data is balanced we have got a fairly close number of patients in both the categories. In our dataset we have 629 heart disease patients and 561 normal patients.

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

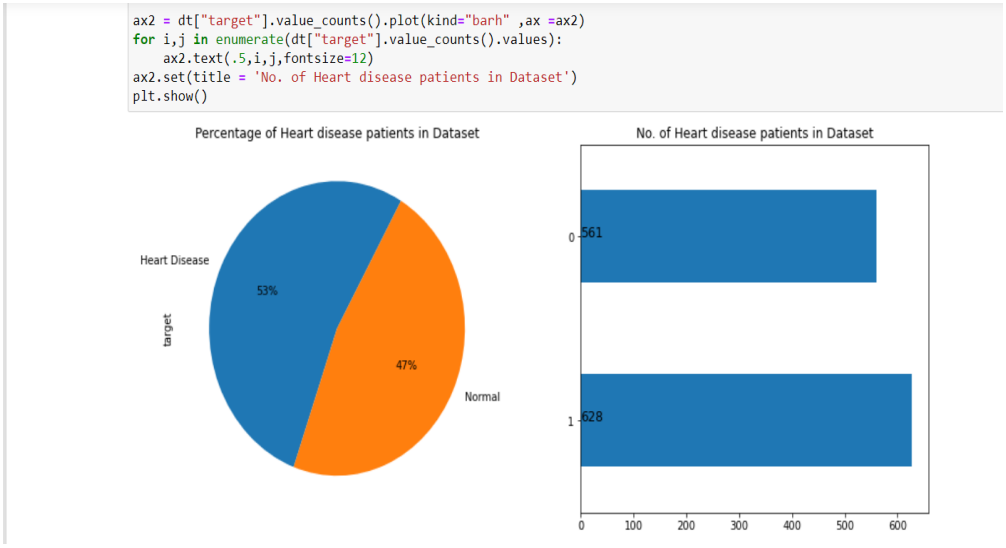

Figure 3. Division between heart patients and normal person.

Gender And Age wise distribution of the patients in the dataset.

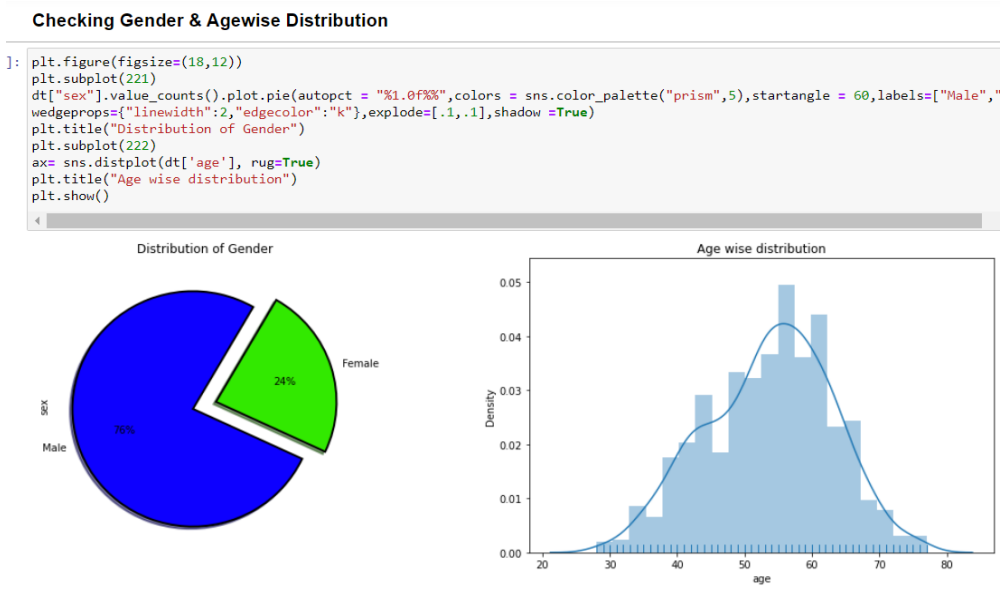

Figure 4. Demographic Distribution of data.

With the visualization of the above graph we can see that dataset males affected with this disease is higher.

Distribution of the dataset:

In this step we further divided our dataset in normal patients and heart patients' dataset on the basis of age and gender.

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

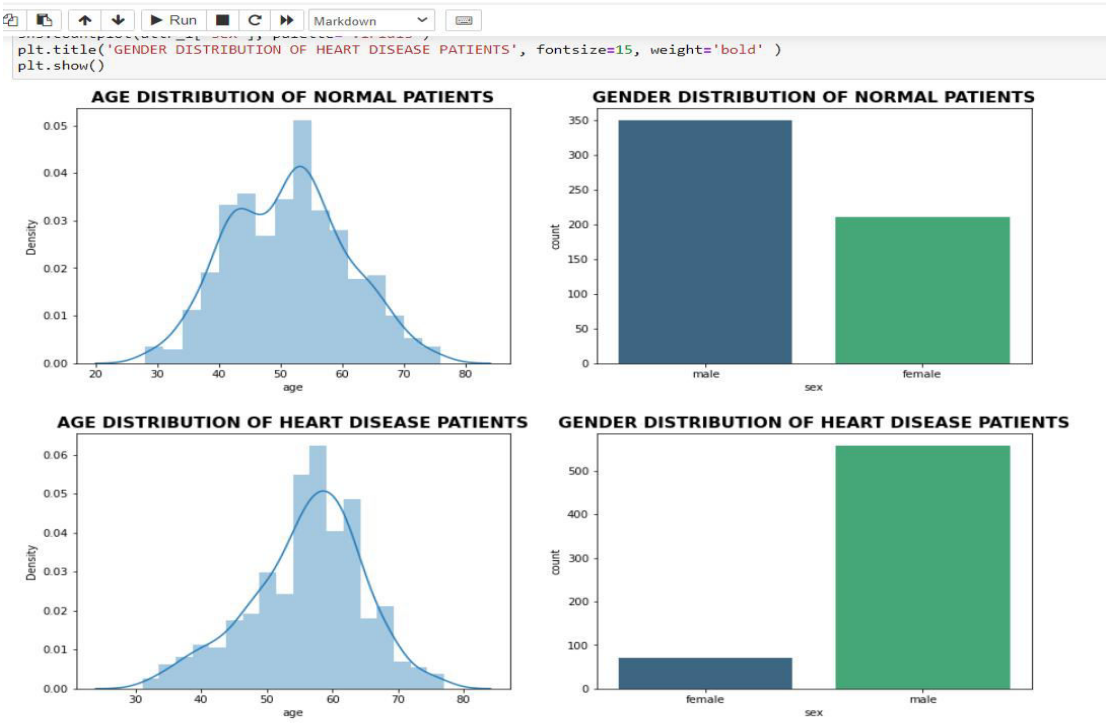

Figure 5. Distribution on basis of demographics

Similarly we have explored the dataset on the basis of chest pain in normal and heart patients, st-slope in both types of patients, rest ecg of normal patients and heart patients.

The numerical features are also analysed using the graphs to show the results and values related to blood pressure cholesterol level related to the factor of age and our graph shows that as the age of the patient increases there are more chances of him getting affected with the heart diseases because of increase in blood pressure and cholesterol level.

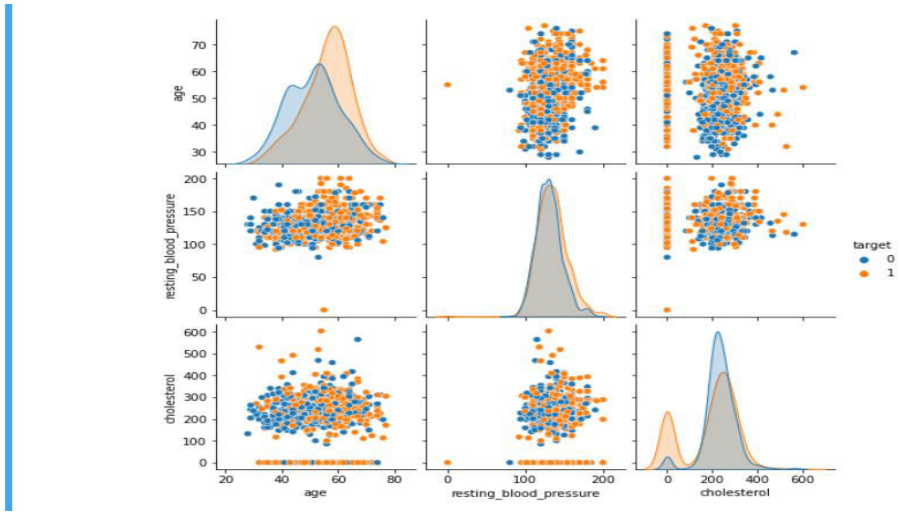

Figure 6. Graph to show BP, Cholesterol and age values.

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

Training and testing data:
For application of machine learning models we have split the dataset in 80:20 ratio for training and testing respectively.

7. Train Test Split

```
In [40]: X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y, test_size=0.2,shuffle=True, random_state=5)

In [41]: ## checking distribution of traget variable in train test split
         print('Distribution of traget variable in training set')
         print(y_train.value_counts())

         print('Distribution of traget variable in test set')
         print(y_test.value_counts())

         Distribution of traget variable in training set
         1    491
         0    446
         Name: target, dtype: int64
         Distribution of traget variable in test set
         1    123
         0    112
         Name: target, dtype: int64

In [42]: print('------------Training Set------------------')
         print(X_train.shape)
         print(y_train.shape)

         print('------------Test Set------------------')
         print(X_test.shape)
         print(y_test.shape)

         ------------Training Set------------------
         (937, 15)
         (937,)
         ------------Test Set------------------
         (235, 15)
         (235,)
```

Figure 7. Splitting of dataset.

Machine Learning Models used
We have used different machine learning models to see the performance on the dataset explored for heart disease prediction. These classifiers includes SVM, Random Forest, Multi perceptron model, Adaboost, Decision Tree.
As per the accuracy Random forest performs the best with accuracy of 91%.

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Random Forest	0.906383	0.874074	0.959350	0.848214	0.914729	0.903782	3.233475	0.815777
1	MLP	0.842553	0.811594	0.910569	0.767857	0.858238	0.839213	5.438109	0.688238
2	EXtra tree classifier	0.906383	0.885496	0.943089	0.866071	0.913386	0.904580	3.233468	0.813662
3	SVC	0.825532	0.801471	0.886179	0.758929	0.841699	0.822554	6.026006	0.652539
4	SGD	0.812766	0.776224	0.902439	0.714286	0.834586	0.808362	6.466944	0.631090
5	Adaboost	0.834043	0.813433	0.886179	0.776786	0.848249	0.831482	5.732052	0.668866
6	CART	0.838298	0.840000	0.853659	0.821429	0.846774	0.837544	5.585062	0.675725
7	GBM	0.851064	0.833333	0.894309	0.803571	0.862745	0.848940	5.144148	0.702485

Figure 8. Overall accuracy of all the models used here.

conclusion and future work
In this work we have analysed performance of different classifiers on heart disease dataset and we got the result that random forest machine learning algorithm resulted in higher performance than any

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

other machine learning model. The most important features are Max heart Rate achieved, Cholestrol, st_depression, Age. For the prediction of cardiac sickness, we have compiled several AI calculations. By analysing their highlights, we ran after finding the optimal calculation using several machine learning models that were discussed. In many situations, each calculation has produced a unique result. It is being further studied that just peripheral exactness is achieved for the predictive model of coronary disease, and going forward, more perplexing models are anticipated to increase the exactness of foreseeing the early onset of coronary illness. In the future, we will provide an approach for very accurate early diagnosis and prediction of coronary sickness with the least amount of expenditure and complexity.

References

1. World Health Organization. (n.d.). *Cardiovascular diseases*. World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases/.
2. [2] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS International Conference on Computer Systems and Applications*, 2008.
3. [3] A. Dutta, T. Batabyal, M. Basu, and S. T. Acton, "An efficient convolutional neural network for coronary heart disease prediction Expert Systems with Applications." 2020.
4. [4] ronit, "Heart Disease UCI." DATASET .
5. [5] "UCI machine learning repository: Heart disease data set," *Uci.edu*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease/. [Accessed: 14-Jul-2021].
6. [6] S. Mohan, C. Thirumalai, and S. P. Bingulac, "Gautam Srivastava ―Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques‖, Digital Object Identifier 10. 1109/ACCESS," 2019, vol. 7, p. 2019.
7. [7] A. M. Sonam Nikhar, I. S. Jacobs, and C. P. Bean, "Karandikar" Prediction of Heart Disease Using Machine Learning Algorithms" International Journal of Advanced Engineering, Management and Science (IJAEMS) Infogain Publication,[Vol-2, Issue-6," *Magnetism*, vol. III, pp. 271–350, 2016.
8. [8] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018.
9. [9] Abhay Kishore, Ajay Kumar, Karan Singh, Maninder Punia, Yogita Hambir, "Heart Attack Prediction Using Deep Learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 4, 2018.
10. [10] A.Lakshmanarao, Y.Swathi, P.Sri Sai Sundareswar, "Machine Learning Techniques For Heart Disease Prediction," *International Journal Of Scientific & Technology Research*, vol. 8, no. 11, 2019.
11. [11] "Heart disease prediction using supervised machine learning algorithms," *Regular Issue*, vol. 9, no. 4, pp. 148–151, 2020.
12. [12] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.
13. [13] V. V. Ramalingam, A. Dandapath, and M. Karthik Raja, "Heart disease prediction using machine learning techniques : a survey," *Int. j. eng. technol.*, vol. 7, no. 2.8, p. 684, 2018.

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/148

14. [14] V. Manikantan and S. Latha, "Predicting the analysis of heart disease symptoms using medicinal data mining methods"," *International Journal of Advanced Computer Theory and Engineering*, vol. 2, pp. 46–5, 2013.
15. [15] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease"," *Telematics Inform*, vol. 36, pp. 82–93, 2019.
16. [16] M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain, "Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis"," *Phys. A,Stat. Mech. Appl*, vol. 482, pp. 796–807,2017.
17. [17] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLoS One*, vol. 12, no. 4, p. e0174944, 2017.
18. [18] N. Al-milli, "Backpropagation neural network for prediction of heart disease," *''J. Theor. Appl. Inf. Technol*, vol. 56, no. 1, pp. 131–135, 2013.
19. [19] A. S. Abdullah and R. R. Rajalaxmi, "'A data mining model for predicting the coronary heart disease using random forest classifier,'" 2012, pp. 22–25.
20. "Cardiovascular diseases," *Am. Heart J.*, vol. 61, no. 4, p. 573, 1961.
21. Tiwari, P., Upadhyay, D., Pant, B., & Mohd, N. (2022). Multiclass Classification of Disease Using CNN and SVM of Medical Imaging. In *International Conference on Advances in Computing and Data Sciences* (pp. 88-99). Springer, Cham.
22. Tiwari, P., Pant, B., Elarabawy, M. M., Abd-Elnaby, M., Mohd, N., Dhiman, G., & Sharma, S. (2022). CNN Based Multiclass Brain Tumor Detection Using Medical Imaging. *Computational Intelligence and Neuroscience*, *2022*.