Aftab Alam Khan ,2021, 9:1
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

Heart Disease Prediction Using Machine Learning

Aftab Alam Khan, Amit Rai, Alka Joshi,Gaurav Kumar, Ram Narayan
Department of CSE
Tula's Institute,Dehradun
aftabalam710@gmail.com, gauravkumar33329@gmail.com, raiamit2211@ieee.org,
alkajoshi1998@gmail.com, ramnarayan@tulas.edu.in

Abstract-It is well known that the heart is one of the most important and crucial parts of the human body. The diagnosis of heart disease through the traditional method has not been considered reliable in many aspects. In today's era, there are numbers of ML algorithms and models present to dig out meaningful information to diagnose and deciding. In this paper, various machine learning models like Support Vector Machine (SVM), Logistic regression, K NN(K Nearest Neighbour), Naïve Bayes, Decision Tree, and Random Forest is employed. The most objective of our work was to predict the patient is having any heart problem or not so after testing these models, the model with greater accuracy is taken for predicting the final result. Today the healthcare industry is information-rich however still very poor in knowledge or mostly the data are not publically available. This paper has used a Cleveland dataset containing 303 individuals and 14 attributes like age, sex, chest pain (cp), resting blood pressure (trestps), cholesterol (chol), etc. With the help of these attributes, our Random Forest Model with an accuracy of 88% provide the final result to the end-user on an interactive web-application designed using HTML, CSS, Flask framework, and JavaScript. With the help of this patients can diagnosis themselves at zero cost. This work will help the end user to get the preliminary prediction of their heart disease and will save them from severe complications.

Keywords:- Machine Learning, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, Gini Impurity, Entropy, Cost Function , Information Gain. Sigmoid Function, are Euclidian, Manhattan, and Minkowski Metric.

I. INTRODUCTION

The work done in this research paper is regarding Heart Disease Prediction with the help of Machine Learning Algorithms. Heart Diseases are generally referred to as the narrowed or blocked vessels that lead to heart attack, chest pain (angina), or stroke. In some conditions, it affects our heart muscles, valves, or rhythm. According to a report by Global Burden of Diseases in 2016, 1.7 million Indians died due to heart disease out of the world's 17.3 milliondeaths. A similar study by The Lancet found that heart-related diseases were found more among

Table 1. Dataset available on UCI Repository.

Dataset	0	1	2	3	4	Total
Cleveland	164	55	36	35	13	303
Hungarian	188	37	26	28	15	294
Switzerland	8	48	32	30	5	123
Long beach VA	51	56	41	42	10	200

International Journal of Science, Engineering and Technology

An Open Access Journal

people from rural India than those in Urban India.[1]

Our work is basically devoted to the medical practitioner and the health care servicemen who need an accurate as well as the precise result for the diagnosis of heart disease in a patient. It is the biggest cause of morbidity in the world; as a result, the healthcare industry holds a good amount of raw data in bulk. Data Mining and ML can be seen as a blessing in the field of Data Science for extracting information from the data which are used to make informed decisions and prediction. A large number of models were already proposed to the healthcare sector for Heart Disease Prediction but the race for accuracy can never be stopped. This research work is also in context to provide greater accuracy by training six classification algorithms (Logistic Regression, Naïve Bayes, K-NN, Decision Tree, Support Vector Machine, and Random Forest). By comparing all these models the model with the highest accuracy and precise result will be selected for predicting the result. The Dataset used for the training of the model is the Cleveland dataset taken from the UCI repository. There were some other Datasets like Hungarian Dataset, Switzerland Dataset, and VA Dataset on the UCI Repository [7] but those did not become too useful for our work due to various missing values. Our training will classify whether a patient is diagnosed positive or negative for heart disease or not.

II. DATASET

To perform the modelling and training of classification algorithms we decided to merge all the four datasets [7].

After merging all the info we obtained a complete of 920 data records which were processed and after proper analysis, we found that it consists of an outsized amount of missing values and that we were just able to extract 299 meaningful records. Whereas alone within the Cleveland dataset we obtained 303 records with only 6 missing values (4 in thal and 2 in ca) which we replaced with their mean values in order that we will use the Cleveland dataset in our whole model training. At last, the decimal values of the target which were 1, 2, and 3 were converted in binary 1 and 0 to binary 0.

The work is carried to diagnose the Heart Disease in Binary Outcome, [6] therefore:

1. **Positive (+):** 1 (Patient is diagnosed with heart disease).
2. **Negative (-):** 0 (Patient does not have heart disease).
3. We have used 3 types of data to diagnose heart disease.
4. **Continuous (#):** This is quantitative data that will be measured.
5. **Ordinal Data:** Categorical data that has an order thereto (0, 1, 2, 3, etc.)
6. **Binary Data:** Data whose unit can combat only two possible states (0 &1).

2. We had Predictor [Y] (Positive or Negative diagnosis of Heart Disease) which is determined by 13 features[X]:

Table 2. Attributes Table (Attributes required to Diagnose Heart Disease).

Sr. No.	Attribute with description	Attribute Values
1	Age – represents age of person	(15-80)
2	sex	**1**=Male **0**=Female
3	Cp – chest pain type	**1**= typical angina **2**= atypical angina **3**= non-anginal pain **4**= asymptotic
4	trestbps – resting blood pressure	(80-200)mmhg
5	Chol – serum cholesterol in mg/dl	(120-400)mg/dl
6	fbs – fasting blood sugar > 120	**1**=true **0**=false
7	restecg – resting electrocardiography results	**Value 0** = normal **Value1** = ST-T wave Abnormality **Value 2** = Probable or definite left ventricular hypertrophy
8	thalach – maximum heart rate achieved	(80-200)bps
9	exang – exercise induced angina	**1**=yes **0**=no
10	oldpeak – ST depression induced angina	(0-6)mV
11	slope – of the peak exercise ST segment (Ordinal)	**Value 1**=up-sloping **Value 2**=flat **Value 3**= down sloping
12	ca - number of major blood vessels (0-3, Ordinal) coloured by fluoroscopy	0,1,2,3
13	thal – thalesamia	3=**normal** 6=**fixed** 7=**reversible defect**

3. Procedure used in ML algorithms:

Our research work is implemented through six classification models. The objective was to provide

International Journal of Science, Engineering and Technology

An Open Access Journal

every medical practitioner and health professional an accurate diagnosis of patient's heart disease.

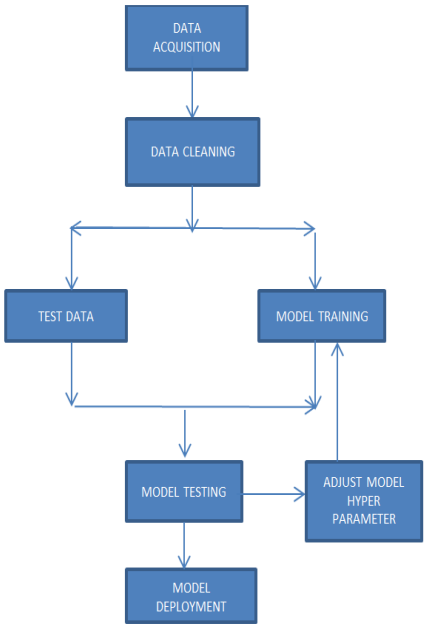


Fig 1.Machine Learning Flowchart.

III. MODEL CLASSIFICATION

1. Logistic Regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary variable, but it also can be applied for multi-class classification uses the one versus rest approach. It can only apply to a tangle where two classes are linearly separable and can be the simplest approach for numerical data features. [2][3][4][5]

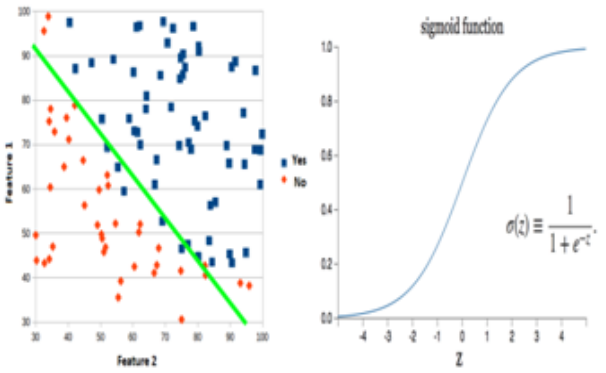


Fig 2. Logistic Regression and Sigmoid Function Plot.

This Cost function can easily get impacted by outliers and give a wrong best fit line that will miss classifying our data points. To decrease the effect of outliers we wrap up our Cost into the Sigmoid function which squashes the values between 0 and 1.

1.1 Optimizer / Cost function used for Logistic Regression:

$$\text{Max} \sum_{i=1}^{n} y_i w_i^T x_i$$

This Cost function can easily get impacted by outliers and give a wrong best fit line that will miss classifying our data points.

To decrease the effect of outliers we wrap up our Cost into the Sigmoid function which squashes the values between 0 and 1.

1.2 Sigmoid Function:

$$f(z) = \frac{1}{1+e^{-z}}$$

2. K-Nearest Neighbour:

K-NN is a classification algorithm that operates on a very simple principle which at first stores all the data, then during prediction it calculates the distance from query point(x) to all points in the data and then sorts those data point in ascending order from query point(x) and predict the majority label of the 'K' closest point. For measuring distance between query point and training data point we use distance metrics that are Euclidian, Manhattan, and Minkowski metric.

The K value must be odd if it is even then there may be a situation that we may get an equal number of values for both categories and our model may get confused.

3. Naïve Bayes:

Naïve Bayes is a type of supervised learning algorithm which comes under the Bayesian Classification. It uses conditional probability for doing its predictive analysis.

$$P\left(\frac{class}{data}\right) = \frac{P\left(\frac{data}{class}\right) * P(class)}{P(data)}$$

This algorithm can be easily applied on categorical data after calculating conditional probability of feature with respect to target class but, for numerical data we have to calculate mean standard deviation with respect to target and then use Gaussian distribution to achieve desired probability.

International Journal of Science, Engineering and Technology

An Open Access Journal

4. Support Vector Machine:

SVM is a supervised machine learning algorithm that supports statistical learning theory used for solving both classifications and regression problems. The target of the Support Vector machine algorithm is to seek out a hyperplane in an 'n' dimensional space (No. of features) that distinctly classify the information.

It gives the simplest prediction even when the training data sample is restricted. To separate the 2 classes of knowledge points many possible hyperplanes might be chosen. Our objective is to seek out a plane that features a maximum margin (the maximum distance between both the classes).

When given a training data set SVM separates the info into different categories employing a hyperplane. New examples are then predicted supported by which side of margin they fall. [2][3][4][5]

4.1 Loss Function for SVM:

The loss function for maximizing the margin between the data points is:

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

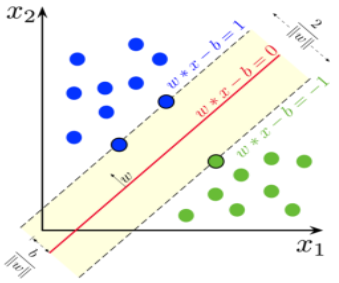


Fig 3. SVM for binary classification. [8]

5. Decision Tree:

Decision Tree is a simple predictive statistical procedure given by J.S. Ross. It infers its decision-making supported by its features provided and therefore in the final model we obtain a recursive nested if-else. The Decision tree is usually used for linearly inseparable data and decision boundaries are always parallel to the feature axis. It is often used for both regressions also as a classification task. For classification, three matrices are used which are:

5.1 Entropy:

$$H(S) = -\sum_{i=1}^{n} log_2(P_i)$$

5.2 Gini Impurity:

$$G_i = 1 - \sum_{k=1}^{k} (P_{i,k})^2$$

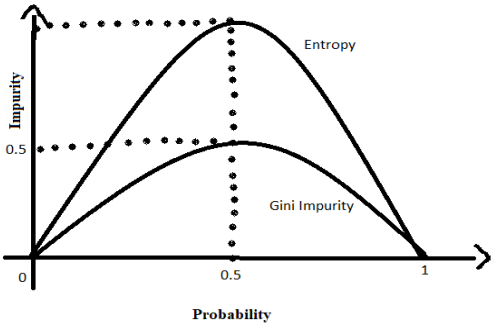


Fig 4. Entropy vs. Gini Impurity.

5.3 Information Gain:

$$\boldsymbol{Gain(S, A)} = \boldsymbol{H(S)} - \sum \frac{|\boldsymbol{S_v}|}{|\boldsymbol{S}|} \boldsymbol{H(S_v)}$$

Range 0 to 1

5.4 Mean Squared Error:

For the regression task, we use Mean Squared Error (MSE) for splitting features.

$$\boldsymbol{MSE(t)} = \frac{\mathbf{1}}{\boldsymbol{N_t}} \sum_{\boldsymbol{i \in D_t}} \left(\boldsymbol{y^{(i)}} - \boldsymbol{\hat{y}_t}\right)^2$$

The decision tree is a weak learner, there's a high chance of overfitting during the training of the model. To beat the prospect of overfitting we use Tree-Pruning by adjusting the value of ccp-α (Cost Complexity Parameter), which decrease the amount of maximum depth.

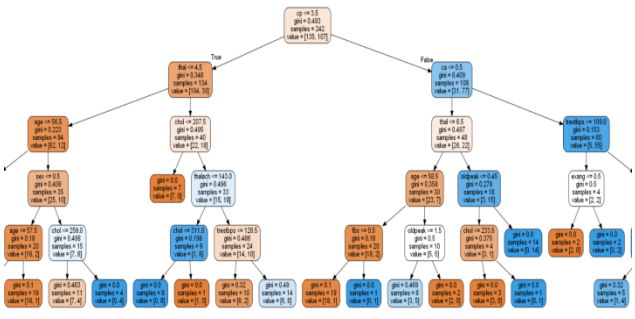


Fig 5. Decision tree Visualization.

6.Random Forest:

While using the Decision Tree we have low bias and high variance, to overcome this problem we used Random forest. Random forest is an ensemble

International Journal of Science, Engineering and Technology

An Open Access Journal

technique that consists of bagging and boosting. This algorithm is an extension over bagging in which we also do feature sampling with replacement in addition to row sampling with replacement.

In this algorithm, decision trees are used as base learners. Whenever we provide test queries in the Random forest model it gets processed by all the base learners and then the major voted outcome is predicted.[3][5]

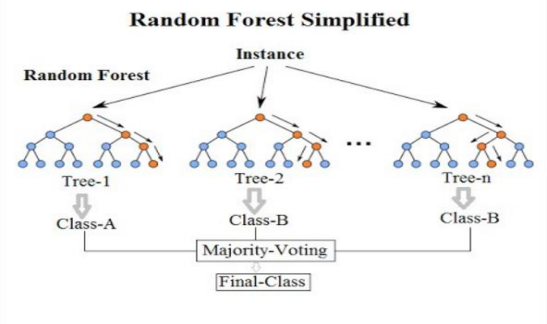


Fig 6. Random forest. [8]

It can handle higher dimensionality data variable and missing values as well as it maintains accuracy for missing data.

IV. MODEL COMPARISON AND ANALYSIS

After completing Exploratory Data Analysis and understanding the Pair Plots we get to know that:
Our data points are not linearly separable i.e., the target class is overlapping. So we decided not to choose Logistic Regression as well SVM, since SVM and Logistic Regression do not work well with linearly inseparable data.

We have 8 categorical and 5 numerical features therefore we have to reject K-NN since K-NN is not feasible to work with categorical features.

We had also seen that Naïve Bayes assumes all predictor (feature) are independent, which rarely happens in real-life scenarios, taking example from our dataset we know that cholesterol, blood pressure, and many other features are dependent on the age of a patient, similarly, it happens with other features also so, Naïve Bayes algorithm treats all the feature as an independent entity which leads us to reject this model for the prediction of heart disease in a patient.

Decision Tree works very well with the categorical features as well as linearly inseparable data but it provides low bias and high variance. High variance can lead to overfitting.

By comparing all these models' characteristics and accuracy we reached the conclusion to use the Random forest model which reduced high variance and provided greater accuracy (88%) to predict the result.

Table 3. Classification model accuracy table.

Sr. No	Model	Accuracy
1	Logistic Regression	85%
2	K-NN	83%
3	Naïve Bayes	83%
4	Support Vector Machine	85%
5	Decision Tree	77%
6	Random Forest	88%

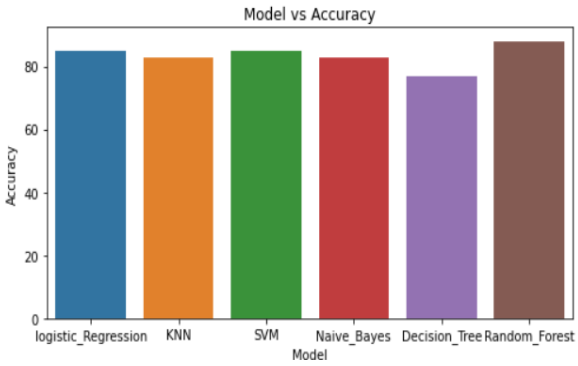


Fig 7. Classification Model Accuracy plot.

V. RESULT

Through confusion matrix we had tried to describe the performance of a classification model (or"classifier") on a set of test data for which the truevalues are known. Logistic Regression, K-NN,Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest.

International Journal of Science, Engineering and Technology

An Open Access Journal

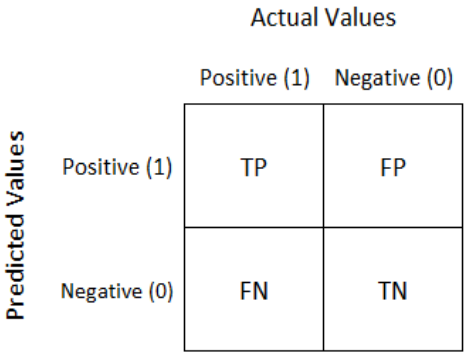


Fig 8. Confusion Matrix.

Table 4. Confusion matrix of Classification Models.

Logistic Regression	[[25 4] [5 27]]
K-NN	[[26 3] [4 28]]
Naïve Bayes	[[26 3] [7 25]]
Support Vector Machine	[[27 2] [5 27]]
Decision Tree	[[23 6] [8 24]]
Random Forest	[[27 2] [4 28]]

After training and testing, we used the Random Forest model and deployed it in a web application, in which we requested the 13 attributes from the user in a web form under some basic validation (with defined max and min values and prescribed). This input gets redirected to the Random Forest model and after complete processing, it predicts whether the user has heart disease or not. The result is shown on the result page of the web application.

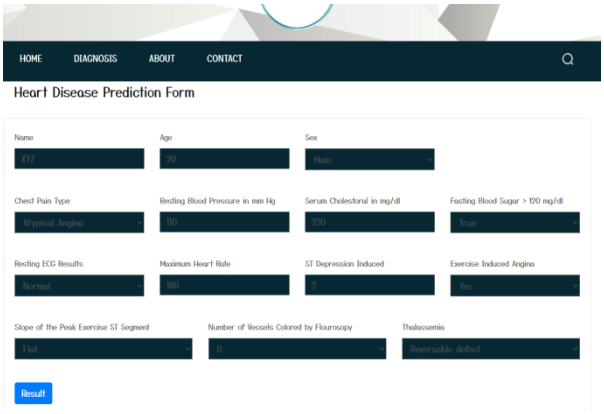
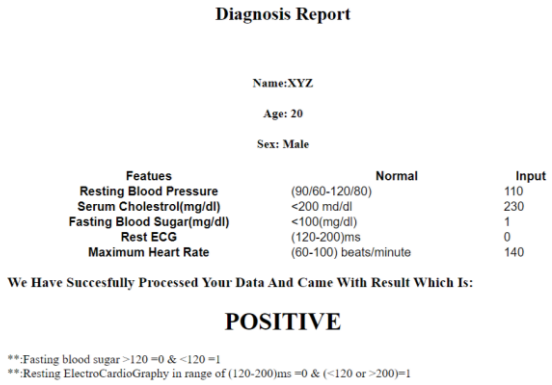


Fig 9. Diagnosis Web Page.



Diagnosis Report

Name:XYZ

Age: 20

Sex: Male

Featues	Normal	Input
Resting Blood Pressure	(90/60-120/80)	110
Serum Cholestrol(mg/dl)	<200 md/dl	230
Fasting Blood Sugar(mg/dl)	<100(mg/dl)	1
Rest ECG	(120-200)ms	0
Maximum Heart Rate	(60-100) beats/minute	140

We Have Succesfully Processed Your Data And Came With Result Which Is:

POSITIVE

**:Fasting blood sugar >120 =0 & <120 =1
**:Resting ElectroCardioGraphy in range of (120-200)ms =0 & (<120 or >200)=1

Fig 10. Result Web Page.

VI. CONCLUSION

As the numbers of heart disease patients are increasing gradually, the necessity of a system to diagnose heart disease with better accuracy is also increasing. We have studied 6 machine learning models which were trained and tested individually with the help of Cleveland Dataset Out of these 6 models we found that Random Forest was most effective having an accuracy of 88% whereas Logistic Regression, Naive Bayes, Support Vector Machine, Decision tree and K-NN model have 85%, 83%, 85%, 77% and 83% respectively. Between Random Forest and K-NN algorithm, we have selected Random Forest as we were dealing with more categorical data.

Finally, using the Random Forest model we have created an interactive web-application using HTML, CSS, JavaScript, and Flask framework which has a simple GUI and user-friendly interface.[5] Based on the input entered by the user, our web application can predict whether a user has heart disease or not. The diagnosis will be at zero cost which will help the

International Journal of Science, Engineering and Technology

An Open Access Journal

end-user to get the preliminary prediction of their heart disease and will save them from severe complications.

VII. FUTURE SCOPE

With reference of the knowledge gained through this research work, one can build a better and accurate model using lesser amount of features after gaining more data in this field.With the help of arduino and sensors we can design a cost effective IOT device which can measure the required attributes for diagnosis at home, that can predict whether a person is having risk of heart disease or not, and if the person is at risk of being heart disease this IOT device with the help of a smartphone can suggest nearby cardiologists and can provide a proper controlling diet plan according to the symptoms.

REFERENCES

[1] Report on medical certification of cause of death, census India by Dr Vivek Baliga.

[2] Hardi Rathod, Pratik Kumar Singh , Nikita Tikone, Suresh Babu et. al. "Health Classification and Prediction Using Machine Learning", International Research Journal of Engineering and Technology(IRJET), Volume 07 Issue 03 March 2020.

[3] Apurb Rajdhan, Milan Sai, Avi Agarwal, Dundigalla Ravi , Dr. Poornam Ghuli, et. al. "Heart Disease Prediction using Machine Learning" International Research Journal of Engineering and Technology(IRJET), Volume 09 Issue 04 April 20DOI:http://dx.doi.org/10.17577/IJERTV9IS0404

[4] V.V.Ramalingam , Ayantan Dandapath , M Karthik Raja et.al."Heart Disease Prediction Using Machine Learning Techniques"International Research Journal of Engineering and Technology(IRJET) Volume 7 ,2018.

[5] Rishabh Magar, Rohan Memane, Suraj Raut, Prof. V.S. Rupnar, et. al. "Heart Disease Prediction Using Machine Learning", International Research Journal of Engineering and Technology(IRJET) Volume 7 ,Issue 6,2020.

[6] Jarar Zaidi, "Project: Predicting Heart Disease with Classification Machine Learning algorithms", https://github.com/jzaidi143/Project-Predicting-Heart-Disease-with-Classification-Machine-Learning-Algorithms

[7] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[8] https://en.wikipedia.org/wiki/Supportvector_machine