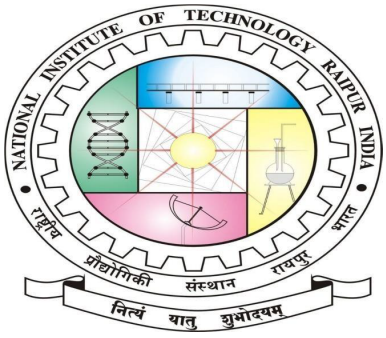
National Institute of Technology, Raipur



Major Project Report
April 2022

Web-Based Machine Learning Application for Heart Disease Prediction

Under the guidance of
Dr. Neelam Shobha Nirala

Submitted By-
Praphooll Markndey(16111018)
Chandrika Rani Tudu(18111018)
Durgesh Kumar(18111023)
Jitendra Rathore(18111028)
MD Samar Siddiqui(18111035)
Prachi Dewangan(18111041)
Sarita Kanwar(18111047)
Surjeet Singh(18111053)
8th Semester,Biomedical Engineering

Web-Based Machine Learning Application for Heart Disease Prediction

Department of Biomedical Engineering
National Institute of Technology, Raipur

Durgesh, Jitendra, Chandrika, Prachi, Sarita, Surjeet, Praphool, MD Samar

April 16, 2022

Abstract

Heart disease is one of the most common disease. Cardiovascular illnesses have been the leading cause of death worldwide in both industrialised and developing countries during the last few decades. The death rate can be reduced if heart disorders are detected early and clinicians are constantly monitored. However,it is not possible to precisely monitor patients every day in all circumstances, and a doctor's 24-hour consultation is not available because it requires more intelligence, time, and knowledge. In this project, we developed web application on the basis of our previous researched models for predicting heart disease based on a patient's various heart attributes and detecting impending heart disease using Machine Learning techniques on a data set that is publicly available on the Kaggle, with the results being evaluated using a confusion matrix and cross validation. Early detection of cardiovascular disease can aid in making lifestyle adjustments in high-risk individuals, reducing consequences and perhaps saving lives, which might be a major breakthrough in medicine. With the results, we came to the conclusion that using random forest algorithm gave us better results, with an accuracy of 93.70 % and random forest algorithm predicted better results than other algorithms.

Keywords-Cardiovascular disease; Machine learning; Random forest

1 Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke (Cardiovascular diseases (CVDs), 2021) [4]. Medical organisations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many a times, this data can be very noisy. These data sets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately. Heart disease is also referred to as a "silent killer" because it causes death without causing noticeable symptoms. Early detection of cardiac disease is critical for implementing lifestyle modifications in high-risk people and, as a result, reducing consequences. This study tries to predict future heart illness by evaluating patient data and using machine-learning algorithms to classify whether they have heart disease or not.

2 Objective

As we know highest death rate reported by heart disease in world as well as India. It's due to lots of people living in village don't aware about the heart disease and if they are, they can't even effort the diagnostic charges.Using the machine learning and low cost portable devices(such as

portable gulucometer, cholesterol,ECG,BP,etc) we are trying to solve this problem with the early prediction of heart disease so everyone can take precaution and save their life. Machine learning have ability to accurately predict that whether a person having heart disease or not and need cardiologist or not.

3 Related work

Nada Alay who is working in the data analytics and ml field, developed a ml model and build a web application to help doctors in diagnosing heart diseases. He collected data-set for the model from UCI ML which has 14 attributes. He applied 4 ML algorithm that are Support Vector Machine Random Forest, Ada Boost, Gradient Boosting, out of which he concluded that gradient boosting has highest precision that is 76, so he used gradient boosting for web application. For making web app he used Flask and create API to load the model, get user input from the HTML template, make the prediction, and return the result and also he used HTML template for the front end to allow the user to input heart disease symptoms of the patient and display if the patient has heart disease or not [1].

4 Methodology

The proposed work predicts heart disease by exploring the below mentioned eight classification algorithms and does performance analysis. The objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart disease. Fig. shows the entire process involved.

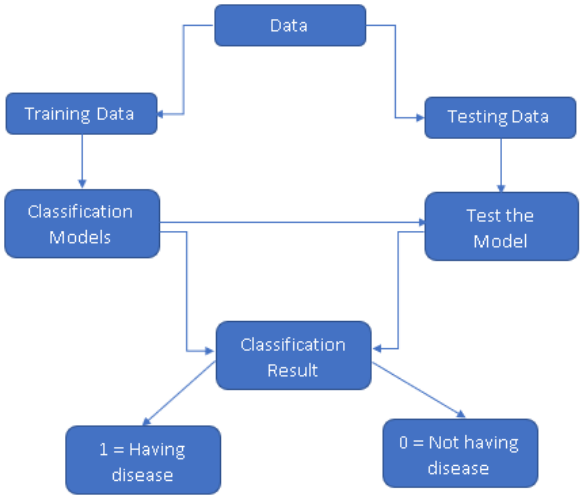


Figure 1: Generic Model Predicting Heart Disease

4.1 Data-set

We have taken this data-set from Kaggle. This data-set contains 12 attributes and 1189 instances. Below table shows you the details of attributes of data-set using in our project-

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0
5	39	1	3	120	339	0	0	170	0	0.0	1	0
6	45	0	2	130	237	0	0	170	0	0.0	1	0
7	54	1	2	110	208	0	0	142	0	0.0	1	0
8	37	1	4	140	207	0	0	130	1	1.5	2	1
9	48	0	2	120	284	0	0	120	0	0.0	1	0
10	37	0	3	130	211	0	0	142	0	0.0	1	0

Figure 2: Screenshot of Data set

Sl. no.	Attribute Description	Distinct Values of Attributes
1	age: represent the age of a person	Multiple values between 29 & 71
2	sex: describe the gender of person (0-Female, 1-Male)	0,1
3	CP: represents the severity of chest pain patient is suffering	0,1,2,3
4	Trestbps: resting blood pressure (in mm Hg on admission to the hospital)	Multiple values between 94 & 200
5	Chol: It shows the cholesterol level of the patient. (serum cholesterol in mg/dl)	Multiple values between 126 & 564
6	FBS: It represent the fasting blood sugar in the patient	0,1
7	restecg: resting electrocardiograph results	0,1,2
8	thalach: shows the max heartbeat of patient	Multiple values from 71 to 202
9	exang: used to identify if there is an exercise induced angina. If yes=1 or else no=0	0,1
10	oldpeak: describes patient's depression level. (ST depression induced by exercise relative to rest)	Multiple values from 0 to 6.2
11	slope: describes patient condition during peak exercise. It is divided into three segments (the slope of the peak exercise ST segment)	0,1,2
12	target: It is the final column of the data-set. It is class or label Column. It represents the number of classes in data set. This data set has binary classification i.e. two classes (0,1). In class "0" represent there is less possibility of heart disease whereas "1" represent high chances of heart disease. The value "0" Or "1" depends on other 13 attribute	0,1

Table 1: Heart Disease Data-set Attribute Description

4.2 Description of Nominal Attributes

• Sex: 1 = male, 0= female
• CP: chest pain type
– Value 1: typical angina
– Value 2: atypical angina
– Value 3: non-anginal pain
– Value 4: asymptomatic
• FBS: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
• restecg: resting electrocardiographic results
– Value 0: normal
– Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
– Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
• slope: the slope of the peak exercise ST segment
– Value 1: up sloping
– Value 2: flat
– Value 3: down sloping 42 rldv5: height at rest
• exang: exercise induced angina (1 = yes; 0 = no)
• target: 1 = heart disease, 0 = Normal

4.3 Data Analysis and Feature selection

We have firstly analysis the data and found some correlations between target(it shows whether a person having heart disease or not) and other attributes and based on their conclusion we do feature selection.Following graphs shows correlation between target and other attributes

Correlation of Sex and Target

As we can see in this graph of heart disease frequency for gender we can conclude on the reference of this graph that heart disease are more in male as compared of women.

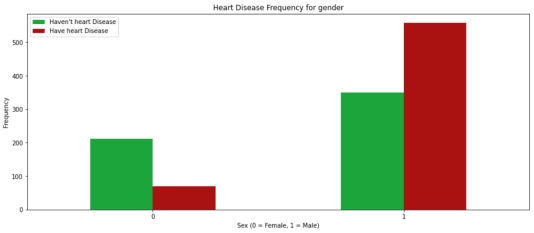


Figure 3: Heart disease frequency for gender

Correlation between Target, Maximum heart-rate and age

People age 65 and older are much more likely than younger people to suffer from heart disease. Aging can cause changes in the heart and blood vessels. The most common aging change is increased stiffness of the large arteries, called arteriosclerosis or hardening of the arteries. This causes high blood pressure, or hypertension, which becomes more common as we age.

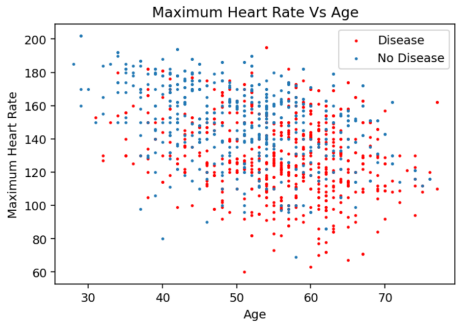


Figure 4: Maximum heart rate vs age

Correlation between Target and chest pain

In this bar plot graph, we are analyzing heart disease frequency according to chest pain type. Here x-axis represents four categorical data of chest pain (CP) type (0,1,2,3). Where,
0-Typical angina
1-Atypical angina
2-Non-anginal
3-Asymptomatic
y-axis represents frequency of target like how many patients have disease or not according to used dataset. As we can see from the graph that in chest pain type 4 asymptomatic has more risk of heart disease as compare to other chest pain types.

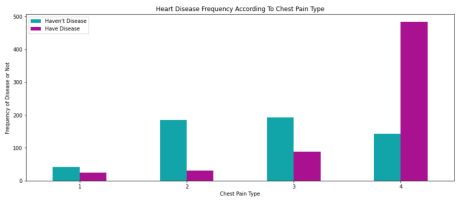


Figure 5: Heart disease frequency according to chest pain type

Correlation Between St segment and target

The plot below is showing us - low ST Depression yields people at greater risk for heart disease. While a high ST depression is considered normal healthy. The "slope" hue, refers to the peak exercise ST segment, with values: 0: upsloping , 1: flat , 2: downsloping).

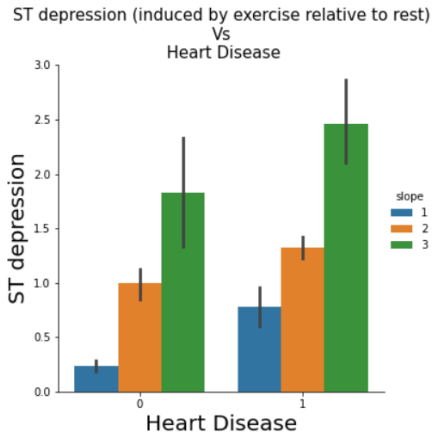


Figure 6: ST depression vs heart disease

And in conclusion of Correlation matrix we can say highest correlation shows with target are **slope**(type of St depression slope),**CP**(type of chest pain),**exang**(Exercise angina),**old peak**.

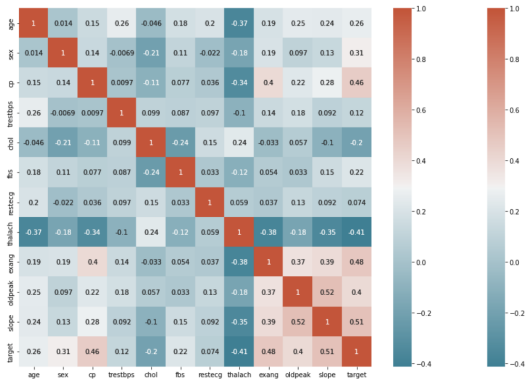


Figure 7: Correlation matrix

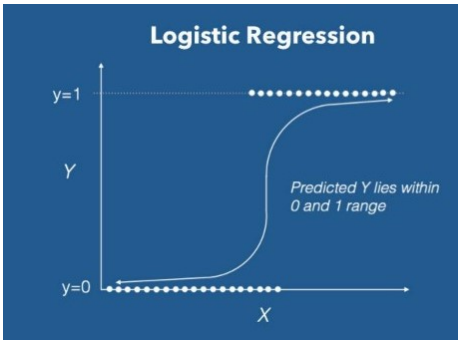


Figure 8: Logistic regression

Based on the above analysis we selected the features to predict heart disease.

4.4 Training And Testing

Data is splitted into two categories training data and testing data. We have used 80%(i.e. 950 patient's data) data for training the model and remaining 20% for testing and validation.

4.5 Algorithms and Techniques Used

Different ML algorithms, such as Logistic Regression, Naive Bayes Classifier, KNN(KNearest Neighbors), Decision Tree, Random Forest, Support Vector Machine,XG Boost and Stochastic Gradient Descent approaches [12], use the properties listed in Table 1 as input. The input data set is divided into two parts: 80% is used for training, while the remaining 20% is used for testing. A training data-set is a collection of data that is used to train a model. The testing data set is used to evaluate the trained model's performance. The performance of each method is computed and analysed using several metrics such as accuracy, precision, recall, and F-measure scores, as discussed below [6]. The many algorithms investigated in this research are given below.

4.5.1 Logistic Regression

The classification algorithm logistic regression is mostly used for binary classification problems. Instead of fitting a straight line or a curve in logistic regression, the logistic regression algorithm employs a hyper plane. the logistic function for squeezing a linear output between 0 and 1 equation. There are 13 self-contained units' logistic regression is useful for a variety of variables classification[13].

4.5.2 Naïve Bayes Classifier

Based on the Bayes Theorem, Naive Bayes is a basic but powerful categorization algorithm. It assumes predictor independence, which means that the attributes or features should not be associated with one another or related in any manner. Even if there is a dependency, all of these characteristics or attributes contribute to the probability separately, which is why it is called Naïve.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{1}$$

$$P(c|x) = P(x_1|c) * P(x_2|c) * .... * P(x_n|c) * P(c) \tag{2}$$

where,
$P(c|x)$ = Posterior Probability
$P(x|c)$ = Likelihood
$P(c)$ = Class Prior Probability
$P(x)$ = Predictor Prior Probability

4.5.3 KNN(K-Nearest Neighbors)

Hodges et al. established the K-Nearest Neighbour rule in 1951, which is a non parametric pattern categorization technique. The K-Nearest Neighbour approach is one of the most often used. The most basic yet extremely successful classification techniques. It does not make any assumptions about the data and is commonly used for When there is little or no prior knowledge, classification tasks are required concerning the data distribution This algorithm entails determining the k. the data points in the training set that are the closest to the data point for which a When the target value is unavailable, the average value of the data is assigned there's evidence for it[8].

4.5.4 Decision Tree

A supervised learning algorithm is a decision tree. This method is mostly used to solve classification difficulties. With continuous and categorical properties, it functions flawlessly. This is how the algorithm works depending on the data, splits the population into two or more related groups the most important predictors Decision First, the tree algorithm calculates the entropy of every single property The data-set is then divided into two halves with the use of maximal predictors or factors Minimum entropy or information gain These are the first two steps recursively applied to the remaining attributes[3].

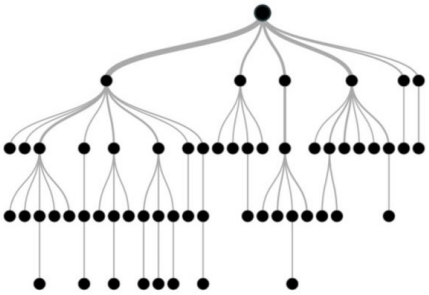


Figure 9: Decision tree

4.5.5 Random forest

Random Forest is another supervised machine learning technique that is widely used. This method can be used for both regression and classification tasks, but it performs better in the latter tasks. The Random Forest approach, as its name implies, takes into account Before producing an output, many decision trees are used. So, there you have it a collection of decision trees This method is founded on a notion that a greater number of trees will converge in the proper direction decision. It uses a vote method for classification before moving on to the next step. In regression, the mean of all the data is used to determine the class, but in classification, the mean of all the data is used to determine the class each of the decision trees' outputs It works well with huge data-sets that have a lot of dimensions.

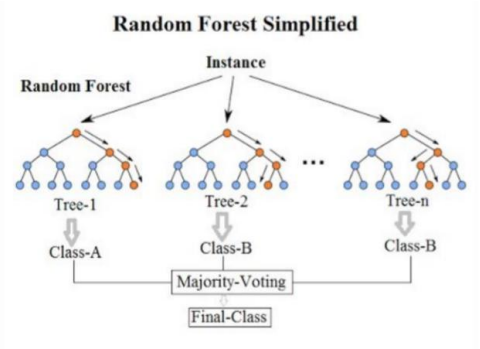


Figure 10: Random forest

4.5.6 Support Vector Machine

The Support Vector Machine (SVM) is a widely used supervised machine learning technique (with a pre-defined target variable) that may be used as both a classifier and a predictor. It finds a hyper-plane in the feature space that distinguishes between the classes for classification. The training data points are represented as points in the feature space by an SVM model, which is mapped in such a way that points belonging to different classes are separated by as wide a margin as possible. The test data points are then mapped into the same area and categorised according to where they fall on the margin [10].

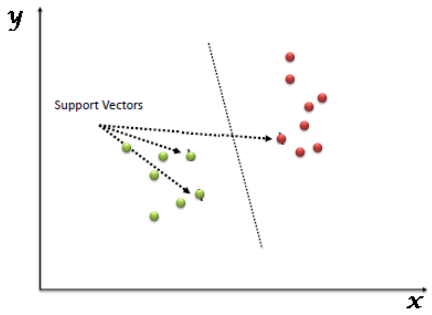


Figure 11: Support vector machine

4.5.7 XG Boost

XG Boost stands for extreme Gradient Boosting. XG Boost is a gradient boosting-based decision-tree-based ensemble Machine Learning technique. Artificial neural networks surpass all other algorithms or frameworks in prediction issues involving unstructured data (images, text, etc.). However, decision tree-based algorithms are now considered best-in-class for small-to medium structured/tabular data.

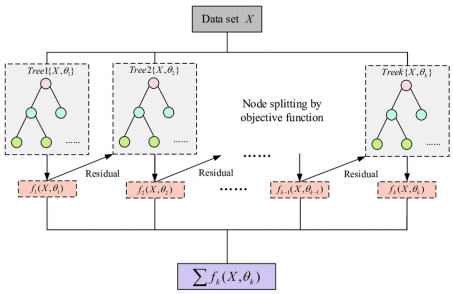


Figure 12: XG Boost

4.5.8 Stochastic Gradient Descent

Gradient Descent is a well-known optimization strategy in Machine Learning and Deep Learning, and it may be used to nearly all learning algorithms. A function's gradient is its slope. It determines how much a variable change in reaction to changes in another variable. Gradient Descent is a mathematically defined convex function whose output is the partial derivative of a collection of input parameters. The higher the slope, the greater the gradient. Gradient Descent is used iteratively to determine the best values of the parameters to find the smallest feasible value of the given cost function, starting with an initial value[12].

4.6 Source code

The coding portion were carried out to prepare the data, visualize it, pre-process it, building the model and then evaluating it. The code has been written in Python programming language using Google Colab as IDE. The experiments and all the models building are done based on python libraries. The code is available in the Git repository given in following link:

Go to the url for dataset and source code: `bit.ly/Major-Project`

5 Evaluation Metrics and results

We evaluated the output of our work by using confusion matrix. We have calculated the confusion matrix for all algorithms and plotted accuracy graph for all classifiers.

5.1 Confusion Matrix

A confusion matrix, also known as an error matrix, is a table that describes how well a classification model (or "classifier") performs on a set of test data for which the true values are known. It enables the visualisation of an algorithm's performance. It provides for easy identification of class confusion, such as when one class is frequently mislabeled as the other. The number of right and incorrect predictions is summarised with count values and broken down by each class, not only the amount of errors committed, which is the key to the confusion matrix[9].

Algorithm	True positive	False positive	False negative	True negative
Logistic Regression	85	17	21	115
Naive Bayes Classifier	87	15	19	117
KNN	84	18	18	118
Decision Tree	93	9	18	118
Random Forest	94	8	7	129
SVM	86	16	16	120
XG Boost	91	11	10	126
SG Descent	82	20	20	116

Table 2: Values obtained for confusion matrix using different algorithm

5.2 Accuracy

Random forest was found to be the best algorithm with accuracy of 93.70%, followed by XG Boost and Decision Tree.

The accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Where,

True Positive (TP) = Observation is positive, and is predicted to be positive.

False Negative (FN) = Observation is positive, but is predicted negative.

True Negative (TN) = Observation is negative, and is predicted to be negative.

False Positive (FP) =Observation is negative, but is predicted positive.

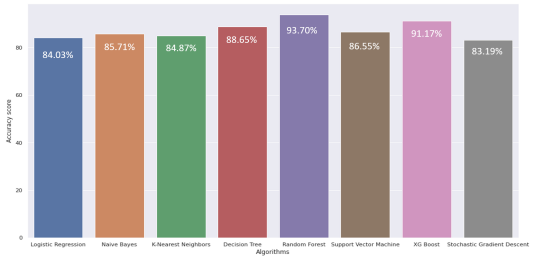


Figure 13: Accuracy graph

5.3 Recall

The ratio of the total number of correctly categorised positive examples divided by the total number of positive examples is known as recall. The class is correctly recognised if the recall is high (a small number of FN). The following formula is used to calculate recall[11]:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

5.4 Precision

We divide the total number of successfully classified positive cases by the total number of anticipated positive examples to get the precision value. A high precision shows that a positive example is, in fact, positive (a small number of FP)[7]. The precision is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

5.5 F1 Score

F1 score is defined as the harmonic mean between precision and recall. It is used as a statistical measure to rate performance. In other words, an F1-score is a mean of an individual's performance, based on two factors i.e. precision and recall [5].

$$F1 = 2 * \frac{precision * recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{6}$$

where,
TP = number of true positives
FP = number of false positives
FN = number of false negatives

Algorithm	Precision	Recall	F1-score	Accuracy(in %)
Logistic Regression	0.84	0.84	0.84	84.03
Naive Bayes Classifier	0.86	0.86	0.86	85.71
KNN	0.85	0.85	0.85	84.87
Decision Tree	0.89	0.89	0.89	88.66
Random Forest	0.94	0.94	0.94	93.70
SVM	0.87	0.87	0.87	86.55
XG Boost	0.91	0.91	0.91	91.18
SG Descent	0.83	0.83	0.83	83.19

Table 3: Analysis of Machine learning algorithm

5.6 Feature Importance

Feature importance scores play an important role in a predictive modeling project, including providing insight into the data, insight into the model, and the basis for dimensionality reduction and feature selection that can improve the efficiency and effectiveness of a predictive model on the problem[2].The feature importance describes which features are more relevant than other and assign them a score. It can help with better understanding of the attributes which are more relevant to predict whether Person heart disease or not. In this project random forest algorithm gave us highest precision, recall and accuracy and we have calculated and plotted feature importance for that, in conclusion we can say slope have highest feature importance score.

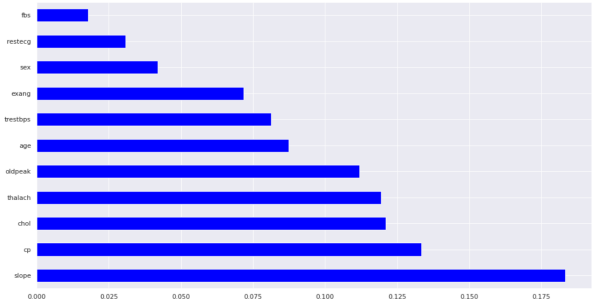


Figure 14: Feature importance

6 Improvisation

In improvisation, we have improved our dataset. Earlier 1189 patient data was used to train and test our machine learning models and maximum accuracy achieved by Random forest algorithms and it was 93.70 %. Currently we are using 3235 patients data to train and test our machine learning models and it is giving us maximum accuracy by Random forest model and it is 96.75%. By improving dataset, our model's accuracy have improved by 3%. Based upon training with dataset feature importance also changed.

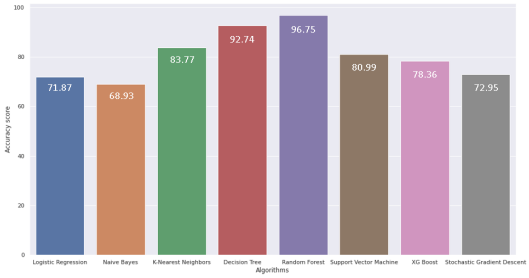


Figure 15: Improved Accuracy

Algorithm	True positive	False positive	False negative	True negative
Logistic Regression	249	97	85	216
Naive Bayes Classifier	186	53	148	260
KNN	283	54	51	259
Decision Tree	305	18	29	295
Random Forest	320	7	14	306
SVM	282	71	52	242
XG Boost	273	79	61	234
SG Descent	262	103	72	210

Table 4: Values obtained for confusion matrix using different algorithm

Algorithm	Precision	Recall	F1-score	Accuracy(in %)
Logistic Regression	0.72	0.72	0.72	71.87
Naive Bayes Classifier	0.71	0.69	0.68	68.93
KNN	0.84	0.84	0.84	83.77
Decision Tree	0.93	0.93	0.93	92.74
Random Forest	0.97	0.97	0.97	96.75
SVM	0.81	0.81	0.81	80.99
XG Boost	0.78	0.78	0.78	78.36
SG Descent	0.73	0.73	0.73	72.95

Table 5: Analysis of Machine learning algorithm

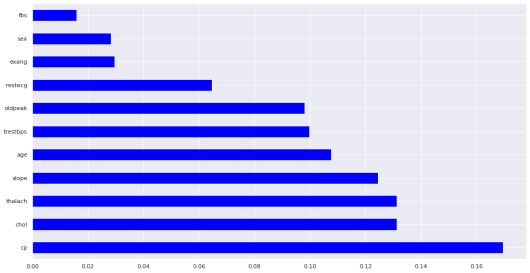


Figure 16: improved Feature Importance

7 Implementation

We have developed a web application to solve and reduce the deaths due to Heart Diseases. We have used machine learning classifiers to early prediction of heart disease. Death rate mostly comes from rural areas as if they feel chest pain they ignore it as they feel its due to gastric problem and they were not aware about, it can be due to heart problem and they don't get medications in time. We have studied about risk factor that influences heart disease. Some are age ,sex, family history, maximum heart rate achieved, Cholesterol level , sugar level chest angina and with the help of ECG (St depression) and including all these risk factor, we can predict if the patient having risk of heart disease or not. We have used heart disease data set (that include these parameters and labeled by a cardiologist if they have heart disease or not) to train

our machine learning models and they precisely predicted, if they have risk of heart disease or not . And to measure these attributes we need portable Glucometer, cholesterol meter,BP machine, portable ECG Machine, by using these we can get the values of attributes through which we can get the final result.

Portable Glucometer- The level of glucose in a person's blood is detected by a glucometer, which delivers readings.Pricking the skin — most typically the tip of the finger — and applying the blood sample to a test strip inserted in the metre is how the reading is obtained.

Cholesterol Meter- An electronic metre is included in some recent cholesterol home test kits. This metre works similarly to a blood glucose metre for diabetics. The test strips are inserted into the electronic equipment, and the amount of cholesterol is automatically measured by a small computer.

Automated BP Machine- The cuff then inflates until it is snugly wrapped around your arm, cutting off blood flow, before the valve opens to deflate it. Blood begins to flow around your artery once the cuff reaches your systolic pressure. This causes a vibration, which the metre detects and records as your systolic pressure.

Portable automatic ECG machine- The sensors act as electrodes, picking up and recording the electrical activity of your heart.ECG are a form of consumer electronics that generally include sensors. You can touch the sensors with one or two fingers, or wear the sensor on your wrist or torso.

We have developed a web application to serve rural area and home healthcare to early predication of risk of heart disease. firstly we have imported the data-set which should be .csv file, after this we have processed the data by selecting attributes and parameters.After that we have trained with different machine learning algorithms.then after we have designed an User interface using HTML, CSS and Javascript after that we devloped backend using a python's framework flask and added trained machine learning models to backend to predict the risk of heart disease of a patient.

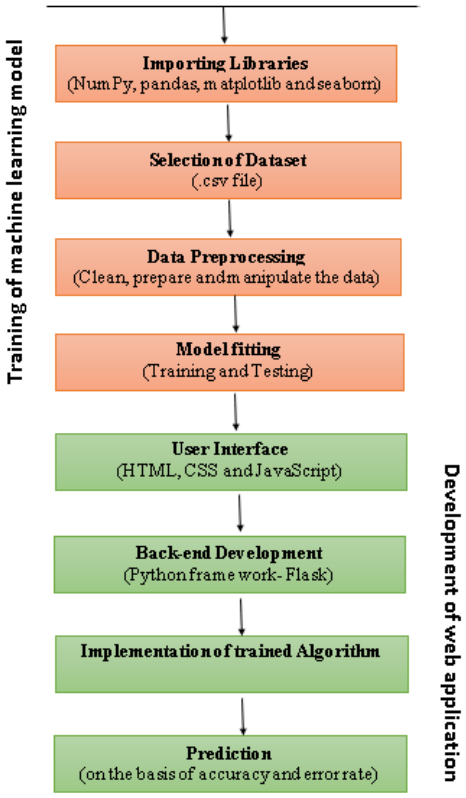


Figure 17: Flow-Chart

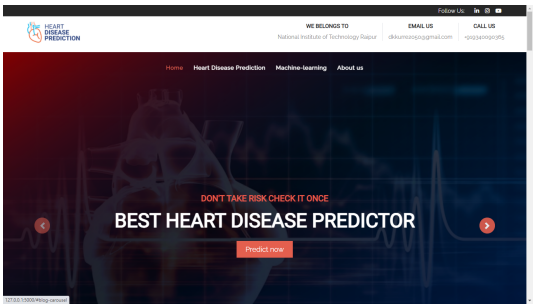


Figure 18: User Interface: Home page

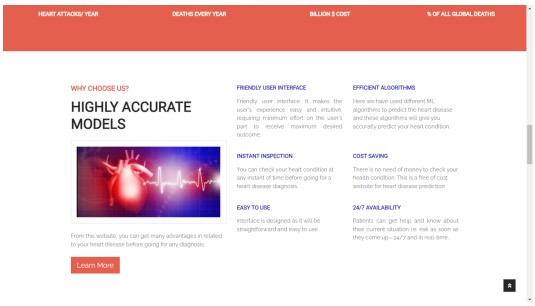


Figure 19: Features of our webapp

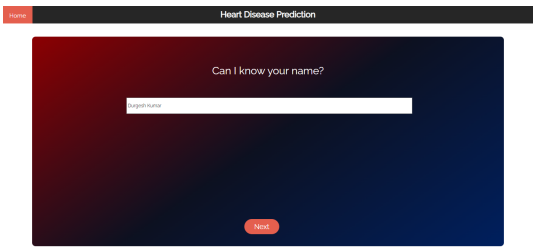


Figure 20: Enter name your Name

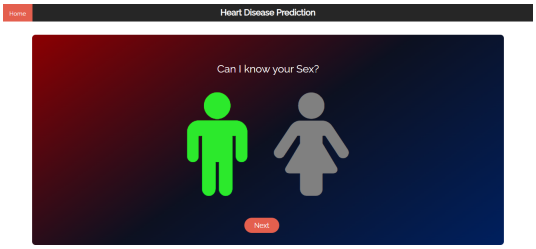


Figure 21: Select your Gender

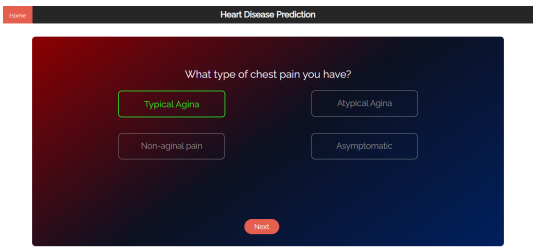


Figure 22: select your chest pain type

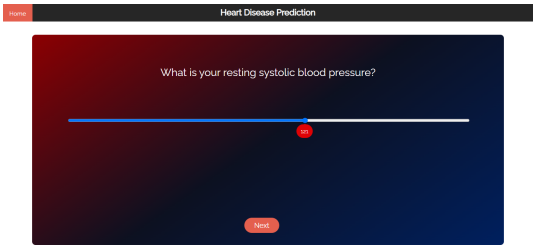


Figure 23: Select your systolic Blood Pressure

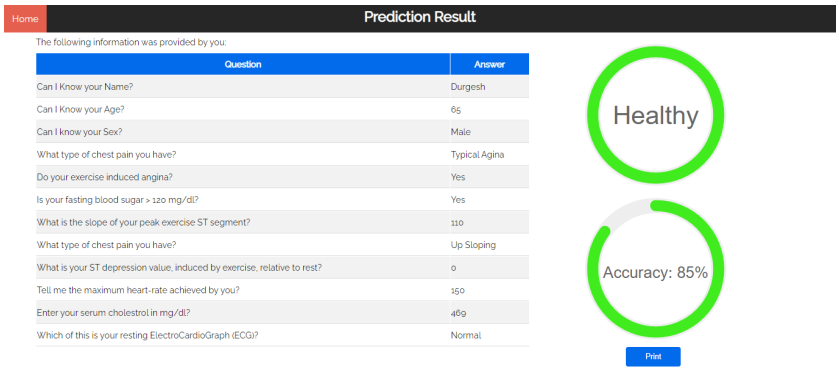


Figure 24: Result page

8 Limitation

We feel there will some limitation while implementing this project to real life (i.e. rural areas). internet coverage in rural area can be a problem as we are making web application that can not be work offline as currently smart phones do not have ability to run machine learning application as machine learning need more processing capabilities to run machine learning app individually so we are planning for web application that will be cloud based so using internet it can be run in any devices mobile, laptop, tablets. Another limitation we can feel is we need to spread awareness among villagers about heart disease for that we can take help of local govt hospital nurses or MITANIN(who spread awareness and distributes medicine in rural areas). Another limitation is currently we have less dataset(database) as it is machine learning project so accuracy is directly proportional to good dataset.

9 Conclusion

As we seen in many reports and news that heart related disease are a matter of concern for the whole world and increasing number of deaths due to heart disease it has become mandatory to develop a such a system which can predict heart disease effectively and accurately because predicting the disease before becoming infected decrease the risk of death. A lot of research is being done on this by the researchers of the world. Our previous work is based on application of machine learning algorithm in which we have used 8 algorithms on a kaggel data-set, where we had very good result and system based on machine learning algorithm and techniques have very accurate in predicting the heart related disease but still there is lot scope of research to be done on how to handle high dimensional data and over-fitting.We got highest accuracy of 96.75% in random forest classifier. Now we have improved our model by training model with more attributes. We have developed a web application which have very interactive user interface that can accurately predict heart disease and in back-end we have used these machine learning classifiers.

References

[1] Nada Alay. *The Lifecycle to Build a Web Application for Prediction from Scratch.* analytics vidhya, 2020.

[2] Jason Brownlee. *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python.* Machine Learning Mastery, 2020.

[3] Anis Cherfi, Kaouther Nouira, and Ahmed Ferchichi. "Very fast C4. 5 decision tree algorithm". In: *Applied Artificial Intelligence* 32.2 (2018), pp. 119–137.

[4] Gagan D Flora and Manasa K Nayak. "A brief review of cardiovascular diseases, associated risk factors and current treatment regimes". In: *Current pharmaceutical design* 25.38 (2019), pp. 4063–4084.

[5] Margherita Grandini, Enrico Bagli, and Giorgio Visani. "Metrics for multi-class classification: an overview". In: *arXiv preprint arXiv:2008.05756* (2020).

[6] SMM Hasan et al. "Comparative analysis of classification approaches for heart disease prediction". In: *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2).* IEEE. 2018, pp. 1–4.

[7] Maciej A Mazurowski et al. "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance". In: *Neural networks* 21.2-3 (2008), pp. 427–436.

[8] Hamid Parvin, Hosein Alizadeh, and Behrouz Minaei-Bidgoli. "MKNN: Modified k-nearest neighbor". In: *Proceedings of the world congress on engineering and computer science.* Vol. 1. Citeseer. 2008.

[9] V Mohan Patro and Manas Ranjan Patra. "Augmenting weighted average with confusion matrix to enhance classification accuracy". In: *Transactions on Machine Learning and Artificial Intelligence* 2.4 (2014), pp. 77–91.

[10] VV Ramalingam, Ayantan Dandapath, and M Karthik Raja. "Heart disease prediction using machine learning techniques: a survey". In: *International Journal of Engineering & Technology* 7.2.8 (2018), pp. 684–687.

[11] Takaya Saito and Marc Rehmsmeier. "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PloS one* 10.3 (2015), e0118432.

[12] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[13] Patrick Adolf Telnoni, Reza Budiawan, and Mutia Qana'a. "Comparison of machine learning classification method on text-based case in twitter". In: *2019 International Conference on ICT for Smart Society (ICISS).* Vol. 7. IEEE. 2019, pp. 1–5.