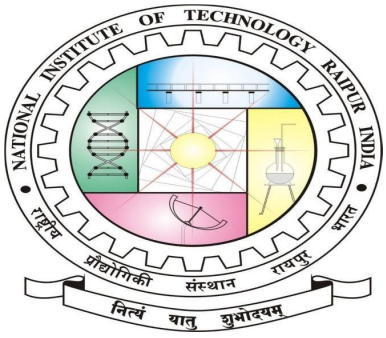
**National Institute of Technology, Raipur**




**Minor Project Report**

November 2021

***Heart disease prediction using machine learning***

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)

7th Semester,Biomedical Engineering

# Heart disease prediction using machine learning

*Department of Biomedical Engineering*
National Institute of Technology, Raipur

November 2021

**Abstract**

**This report covers a mini-project assigned to seventh-semester students as part of the minor project in our course offered by the Department of Biomedical Engineering at National Institute of Technology, Raipur. 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 and 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 Website, 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.**

***Keywords-Cardiovascular disease; Machine learning***

## 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 132 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 we are trying to solve this problem with the early prediction of heart disease. Machine learning can accurately predict that whether a person having heart disease or not and need cardiologist or not.

## 3 Data-set

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

| 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

| | 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 1: Screenshot of Dataset

### 3.1 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 Algorithms and Techniques Used

Different ML algorithms, such as Logistic Regression, Naive Bayes Classifier, KNN(K-Nearest 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.

### 4.1 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].

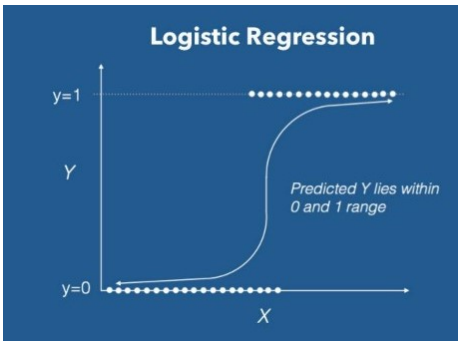



Figure 2: Logistic regression

### 4.2 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

### 4.3 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].

### 4.4 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].

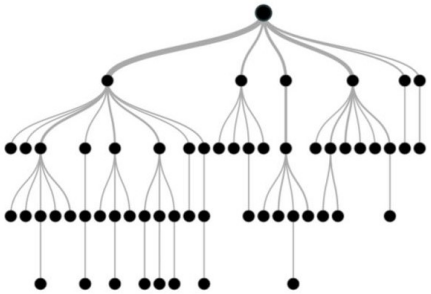


Figure 3: Decision tree

### 4.5 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.

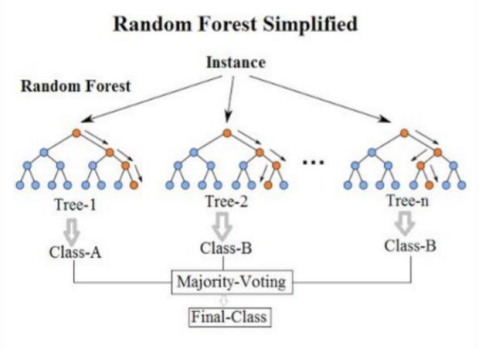



Figure 4: Random forest

### 4.6 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].

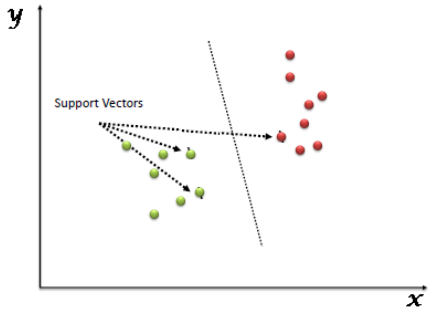



Figure 5: Support vector machine

### 4.7 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.

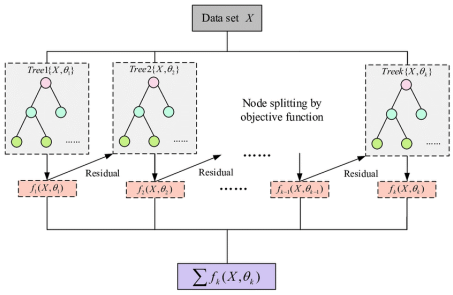



Figure 6: XG Boost

### 4.8 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].

## 5 Proposed Model

The proposed work predicts heart disease by exploring the above 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.

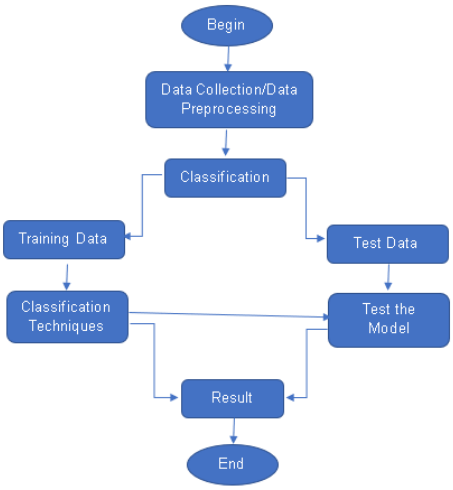



Figure 7: Generic Model Predicting Heart Disease

## 6 Feature Importance

Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. 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].

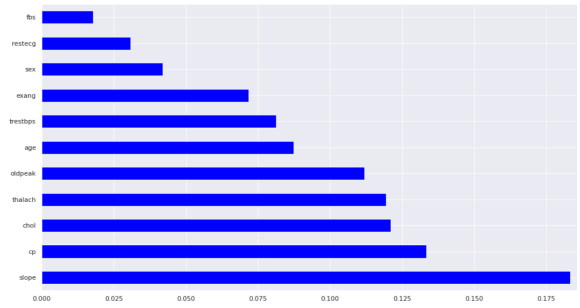



Figure 8: Feature importance

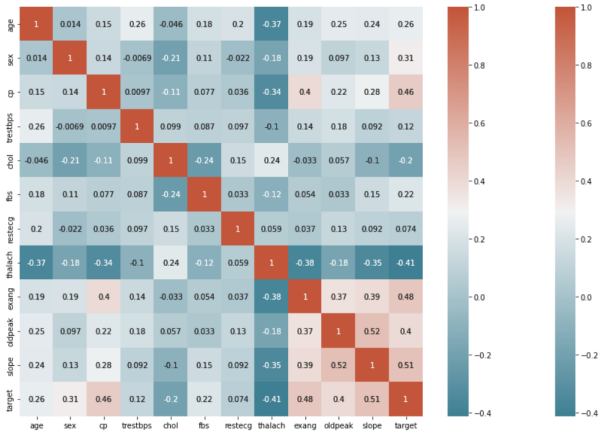



Figure 9: Correlation Matrix

# 7 Evaluation Metrics

For the evaluation of our output from our training the data the accuracy was analyzed by:

## 7.1 Correlation Matrix

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large data set and to identify and visualize patterns in the given data[1].

A correlation matrix consists of rows and columns that show the variables. Each cell in a table contains the correlation coefficient.

## 7.2 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

## 7.3 Accuracy

Random forest was found to be the best algorithm with accuracy of 93.70%, followed by XG Boost and Decision Tree. Artificial neural networks are also employed for the prediction of diseases. Supervised networks have been used for diagnosis and they can be trained using the Back Propagation Algorithm.

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.

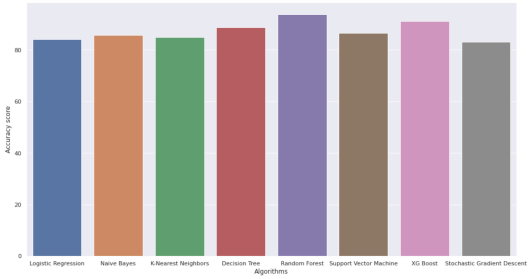



Figure 10: Accuracy graph

### 7.4 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]:

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

### 7.5 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:

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

### 7.6 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

## 8 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:

`https://github.com/Durgesh2050/Minor-Project`

## 9 Implementation

We are trying to implement this project to solve and reduce the deaths due to Heart Diseases . We are going to use machine learning models 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 desperation) and including all these risk factor, we can predict if the patient 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 model and they precisely predicted, if they have heart disease or not .And to measure these attributes we need portable Glucometer, cholesterol meter,BP machine, portable ECG Machine using these we can predict .

**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. The chemicals in the strip react with the glucose in the blood.

**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**- 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. The sensors act as electrodes, picking up and recording the electrical activity of your heart.

And we making a web application to serve rural area and home healthcare to early predication of heart disease. First we design an UI by using WEKA tool and TANAGRA tool. Then we select the data-set which should be .csv file and .arff file , after this we are processing the data by selecting attributes and parameters. Here we can select the algorithm from which we want to implement to get our result, and after all this on the basis of model we will get our result.

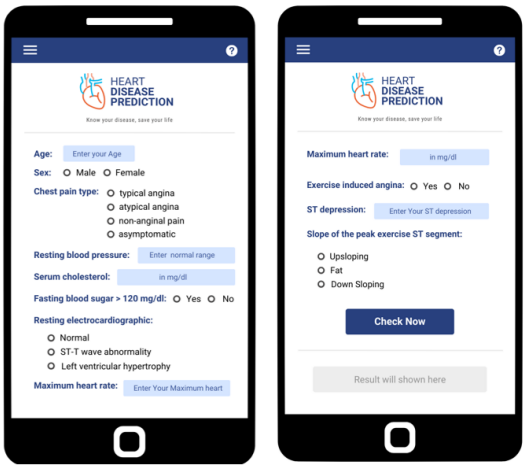



Figure 11: User Interface

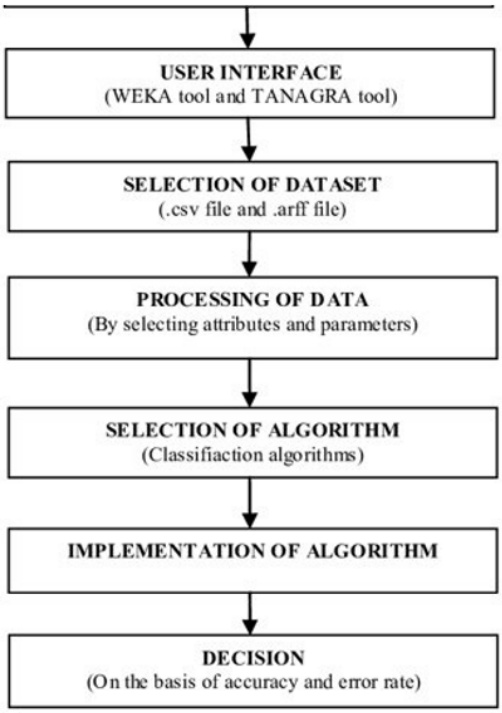



Figure 12: Flow-Chart

## 10 Conclusion/Future work

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 system to 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 paper is based on application of machine learning

algorithm which we have chosen 8 algorithm on a 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 93.70% in random forest. We can update this project in future by adding more attributes to the data set and more interactive to the users and can also be done as a Web application. we will modify the system by connecting it to the hospital's database.

# References


[1] Agustin Garcia Asuero, Ana Sayago, and AG Gonzalez. "The correlation coefficient: An overview". In: *Critical reviews in analytical chemistry* 36.1 (2006), pp. 41–59.

[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.