Capstone Project Report

Prediction of likeliness of cardiovascular disease using machine learning algorithms

Mritunjay Ashish, MT20102, IIITD

Naveen, MT20078, IIITD

Shubham Sharma,MT20315, IIITD

Sumit Patiyal, PhD Scholar, CCB, IIITD

Dr. G.P.S. Raghava, Professor & HOD CCB, IIITD

Abstract ----------

A cardiovascular disease (CVD) is a disorder that has something to do with the heart or blood vessels. Cardiovascular illnesses include heart attacks, heart failure, arrhythmia, coronary heart disease, and many others. Smoking, blood pressure, cholesterol level, diabetes, and being overweight all increase the risk of various cardiovascular illnesses. Age, gender, nutrition, and alcohol consumption can all be important risk factors for cardiovascular disease. Cardiovascular disorders are the biggest cause of death on a global scale. According to the 2019 study, almost 18 million individuals died from Cardiovascular illnesses, with heart attacks and stroke accounting for 85 percent of these deaths. In India, the projected number of CVD deaths in 1990 was 2.26 million, and It is projected that the number of fatalities would increase to 4.77 million by 2020. In the rural area, CVD grew from 1.6 percent to 7.4 percent, while in the city, it increased from 1% to 13.2 percent. In recent years, a number of researchers have worked on stroke prediction using a variety of machine learning techniques, including Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Xgboost (XGB), Naive Bayes, and others. In this paper, we describe a model that combines data preprocessing and scaling approaches to clean the dataset, different classifiers applied to this dataset, and obtained the highest accuracy of 88.52 percent and Roc AUC of 88.74 percent by applying Random forest over the test data.

Introduction ----------

The name "Cardiovascular" is derived from the combination of two terms, "Cardio" and "Vascular." The term "cardio" refers to the heart, whereas the term "vascular" refers to all veins in the body. Cardiovascular diseases (CVD) are the most well-known and deadly infections in the world, killing the majority of people each year. Illness (CVD) than any other illness. In today's fast-paced society, cardiovascular illness is on the rise. According to the WHO, 17 million people die each year from cardiovascular diseases such as heart attacks and strokes [1]. In big populations, most cardiovascular diseases (CVD) are caused by fundamental risk factors such as cigarette smoking, a poor eating routine, weight, physical lethargy, and hazardous alcohol use. Individuals with Cardiovascular disease (CVD) or high levels of Cardiovascular risk (due to the existence of at least one risk factor, such as hypertension, hyperlipidemia, diabetes, or well managed

illness) require an introduction and instruction using brief prescriptions, as seen below. In general, cardiovascular disease (CVD) is characterized by the formation of fatty deposits [5] inside the pipes (atheroscoria) as a result of blood accumulations. It has also been related to organ damage, including brain, heart, kidney, and eye impairment. Strokes and coronary events are typically triggered by stressful conditions and are caused, for the most part, by a blood clot that prevents blood from reaching the heart or brain[13]. A typical cause of cardiovascular failure and strokes is a combination of multiple risk factors, such as cigarette smoking, a poor diet, and being overweight. Heart disease is a potentially fatal ailment that should not be underestimated. According to Harvard Health Publishing [4], males are more likely than women to develop cardiac disease. Males are roughly twice as likely as women to suffer a heart attack over their lives, according to research.

Stroke is the third leading cause of death globally; hence, stroke prediction is an important metric that, if done early, can save many lives. Several research comparing the efficacy of predictive data mining approaches and other machine learning technologies to forecast various illnesses have been conducted [3]. In recent years, several experts have been working on cardiovascular illness and determining the best approach to anticipate the condition. However, more research and patient data from hospital records become available as time passes. Using various computer technologies, many open sources of access to patient data and examinations can be used to properly diagnose patients and detect this condition before it becomes fatal [2]. We all know that data mining and machine learning technologies are the greatest predictors of cardiovascular disease. In our study, we have also used a variety of machine learning techniques to increase the accuracy of our predictions [17].
The rest of the report is broken down into six sections:-

1. Introduction
2. Literature review
3. Proposed Methodology
4. Results and Analysis
5. Android mobile App for Heart Disease Prediction
6. Conclusion and Future scope

Literature Review ----------

This area contains the Analysis and Prediction of Cardiovascular Disease using ML Algorithms, the author and team tell us that CVD (Cardiovascular disease) is the most significant reason for the blocking of veins, resulting in a heart attack or stroke. This paper uses various machine learning tree-based algorithms for finding any cardiovascular disease. The project contains a heart disease dataset that is taken from the UCI repository. The dataset comprises age, sex, cp, trestbps, chol, FBS, restecg, thalach, ca, and target. Random Forest with an accuracy of 85.71 percent, Decision Tree with an accuracy of 74.28 percent, Logistic Regression with an accuracy of 74.28 percent, Support vector machine with an accuracy of 77.14 percent, and K-Nearest Neighbor with an accuracy of 68.57 percent were the machine learning algorithms used on the dataset [6].

There are Extreme Learning Machine (ELM) Techniques for Heart Disease Diagnosis. The author Salam Ismaeel and Team tell us about the different factors responsible for a heart attack in people and how to use them in extreme machine learning approaches to and early-stage warnings. The dataset used in this paper taken from Cleveland Clinic Foundation in which details from 300 patients have been collected containing many attributes such as CP, RBP, Serum cholesterol,, Restecg etc. achieved, Exercise-induced angina, Oldpeak, Slop, Ca, Thal, and Class Target the predicted attribute. ELM method consists of 3 layers of neural network with a sigmoid activation function. The system consists of an accuracy of 80% in figuring out the CV disease [7].

This work integrates to create an explainable AI model for stroke prediction, researchers used machine learning and symbolic reasoning.. The author Nicoletta Prentzas and the team tell us about the recent advancements of ML and AI in healthcare and talk about the drawbacks of algorithms like Blackbox nature and the Inability of the algorithm to explain its answer. The dataset in this paper is a stroke risk study(ACSRS) Dataset, which contains 1054 observations of both asymptomatic and stroke patients. The methodology of this paper is such that the RF Classifier used for the classification of stroke and an explainable AI model (XAI) for using it in real-time. The accuracy of the random

forest will be 78%, and answers to users were correct most of the time with good explanations [8].

This area comparing applying ways of selecting features to Improve CV Prediction, the author Saba Bashir and team describe the use of data science to predict heart disease and stroke. This work selects many features schemes and classifiers by using the UCI dataset. The research paper uses DT, LR, NB, and many more Classifiers. After doing preprocessing steps (Data Cleaning, Data Transformation, Data reduction, Binning, Feature Selection) on the dataset, we can get accuracies in the range of 82.22% for DT, 82.56% for LR, 84.17% for RF, 84.24% for Naive Bayes, and 84.85% for Logistic Regression SVM [9].

This area consists of Adjustable Supervised Learning Method on a Cell Phone to Predict CVD from Actual Electrocardiographic Surveillance, The author Zhanpeng Jin and Team want to tell us about the use of an Electrocardiogram (ECG), which uses the skin to assess the electrical actions of the heart. This method is widely accepted in various technologies such as rate monitoring systems, smartness bands, etc. The data is then transferred to the smartphone or other device in real-time. The author finds out that most present systems lack early feedback mechanisms to users in case of a future heart stroke chance. The author uses ANN-based machine learning techniques and data coming from ECG to combine them and train the device. The dataset used here is the ECG dataset which can be gathered using an ECG device. The author proposes a HeartToGo platform that can continuously monitor and record the ECG data in real-time. It can generate signals if any abnormal ECG data is found [13].

.

Proposed Methodology ----------

The proposed approach worked on a heart disease dataset and predicted the chance of heart disease. First, we have to analyze the dataset; then, we apply preprocessing on the dataset; after that, we create dummy variables (One Hot Encoding) for categorical values, then In an 80:20 ratio, we split the data into training and testing. After splitting the dataset, we perform 5-fold cross-validation and apply various machine learning classifiers with the grid search algorithm. The proposed approach's framework is depicted in Figure 1.

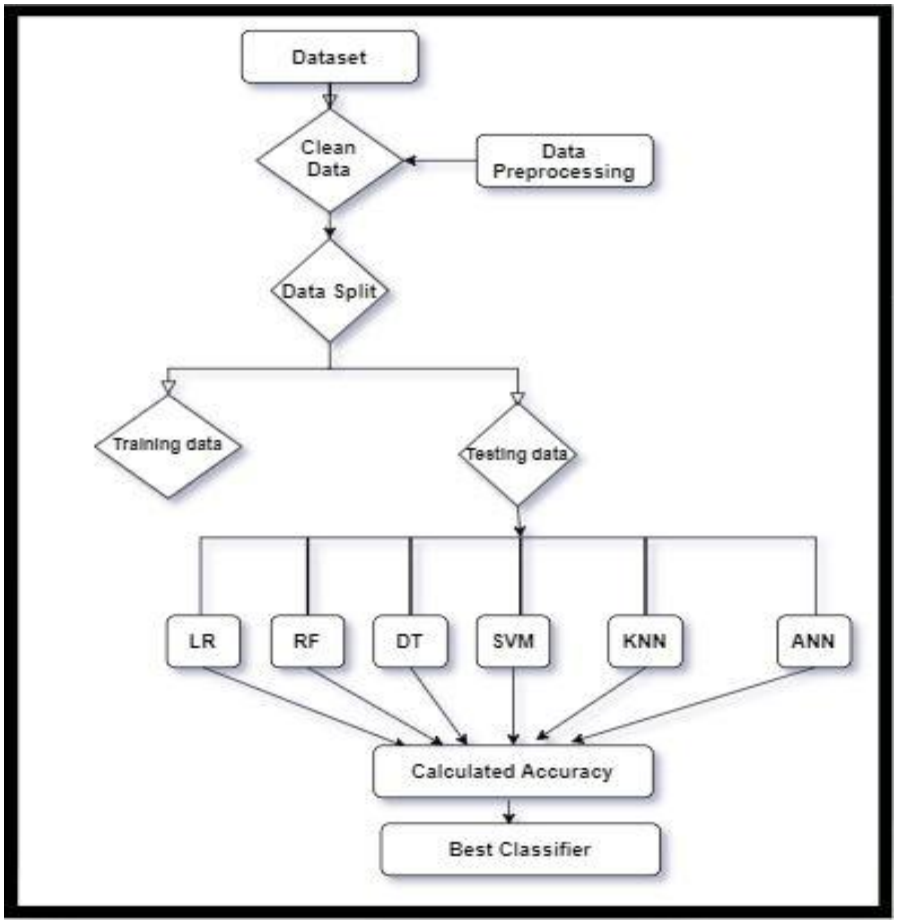


Fig-1. Framework of Proposed apporach.

(1) Dataset Description

We worked on the heart disease dataset, which was received from the UCI (University of California at Irvine) repository; this data set has 14 attributes with 303 instances, including age, gender, cp, trestbps, cho, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and target. We worked on 10 attributes. We remove less impactful attributes like exang, oldpeak, slope and thal. The target field describes heart disease in the patient. The patient's heart problem is described in the target field. This is an integer value, with 0 indicating no heart disease and 1 indicating heart disease. The following are the characteristics that were used for this study: -

(i) Age—age of the patient in years.
(ii)Sex—(1 = male; 0 = female).
(iii) Cp—displays the type of chest-pain experienced by the individual using the following format :-

0 - typical angina
1 - atypical angina
2 - non-anginal pain
3 - asymptotic

(iv) Trestbps—resting blood pressure (in mmHg).
(v) Chol—displays the serum cholesterol in mg/dl (unit).
(vi) Fbs—fasting blood sugar.If fasting blood sugar > 120 mg/dl then: 1 (true) else : 0 (false)
(vii) Restecg—resting electrocardiographic results.
(viii) Talach—maximum heart rate achieved.
(ix) Ca—number of major vessels (0–3) colored by fluoroscopy.
(x) Target (T)—no disease = 0 and disease = 1, (angiographic disease status).

Table 1 shows the value type of attributes in the dataset.

S.No.	Attributes	Value Type
1.	Age	Numerical
2.	Sex	Nominal
3.	Cp	Nominal
4.	Threstbps	Numerical
5.	Cho	Numerical
6.	Fbs	Nominal
7.	Restecg	Nominal
8.	Thalach	Numerical
9.	Ca	Nominal
10.	Target	Nominal

Fig-2 describes the distribution of target variables. In the target variable 0 refers to “No disease” and 1 refer to “disease”. The number of class 0 in target variable is 138 and number of class 1 in target variable is 165.

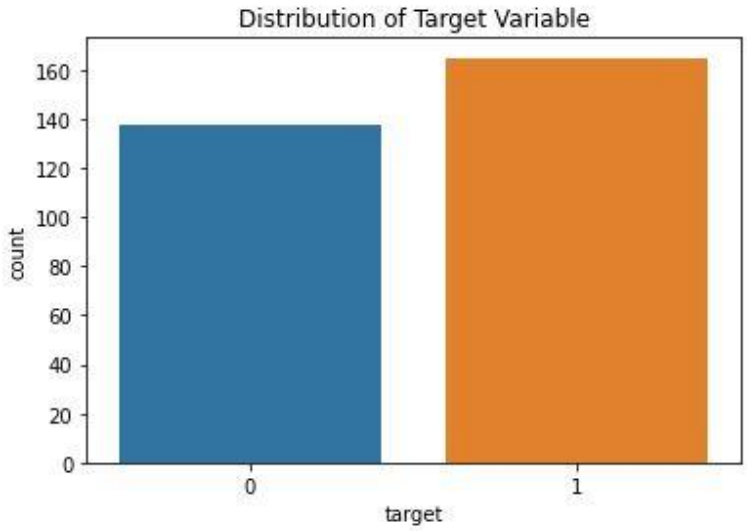


Fig-2. Class distributions, 0: no disease || 1: disease

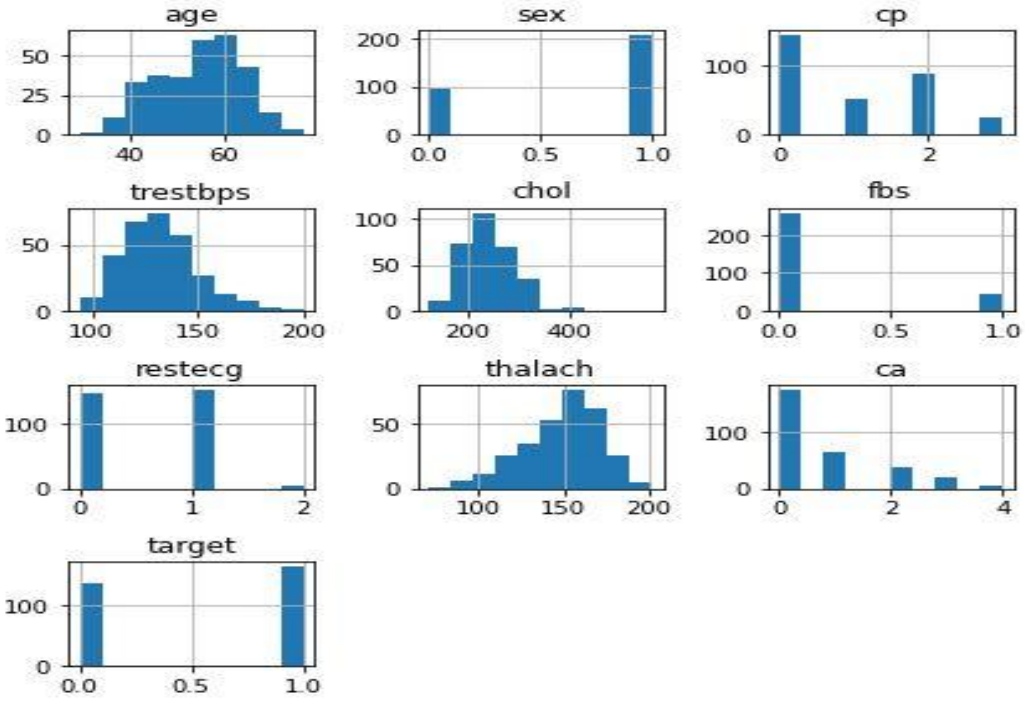


Fig-3. Histogram of attributes.

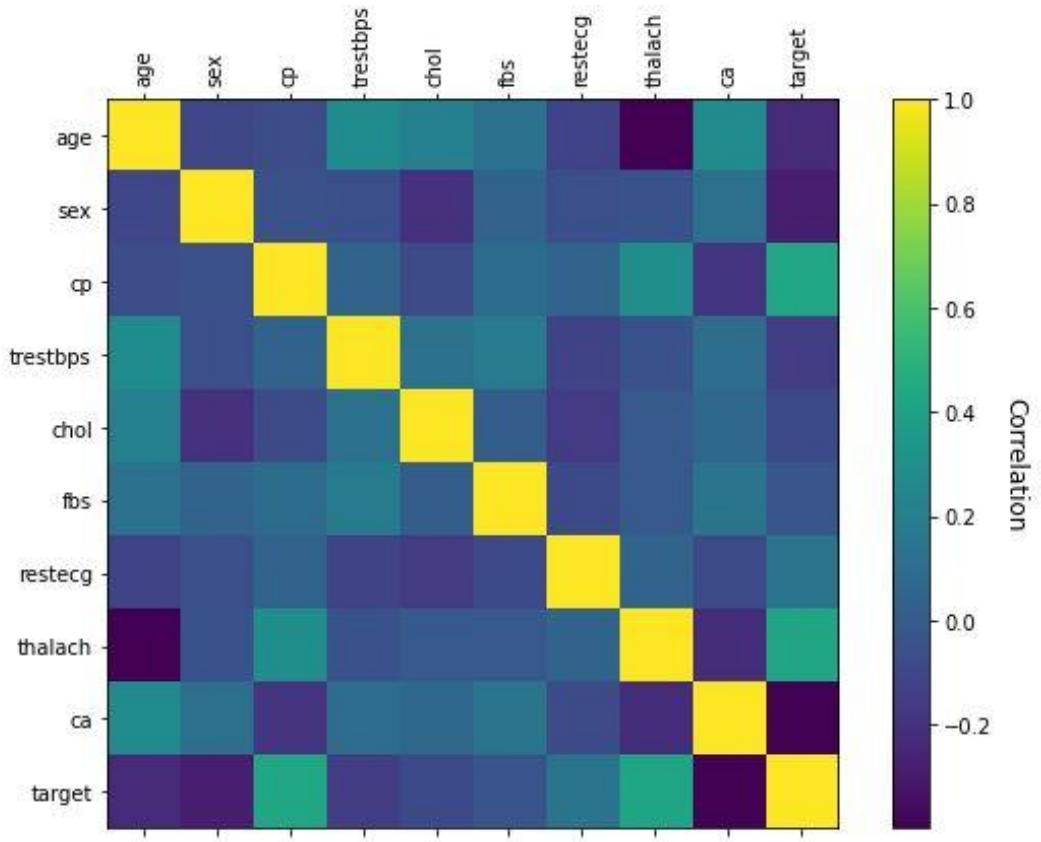
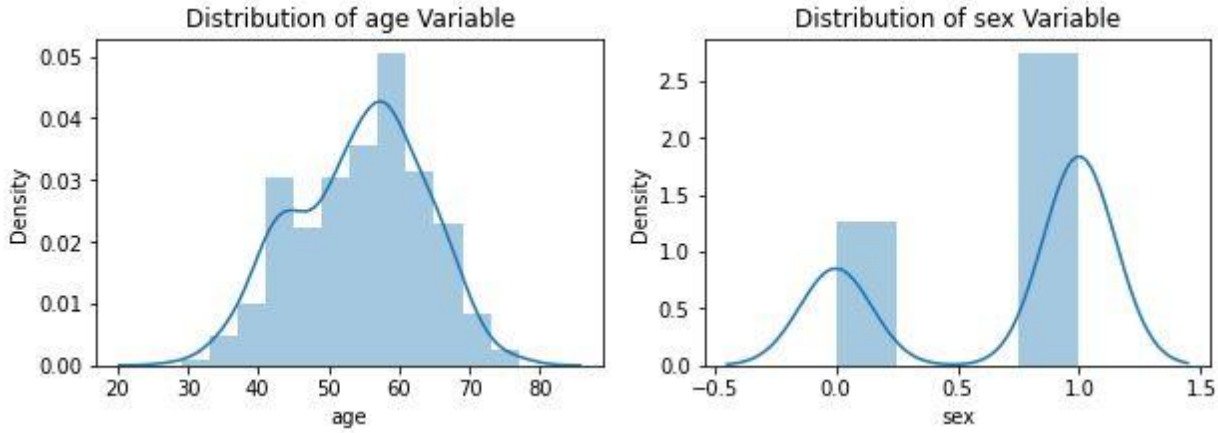
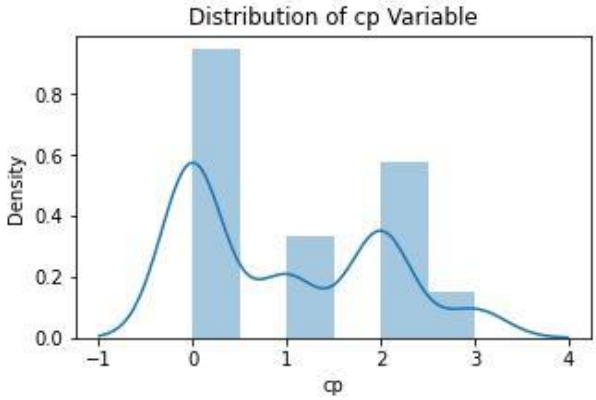
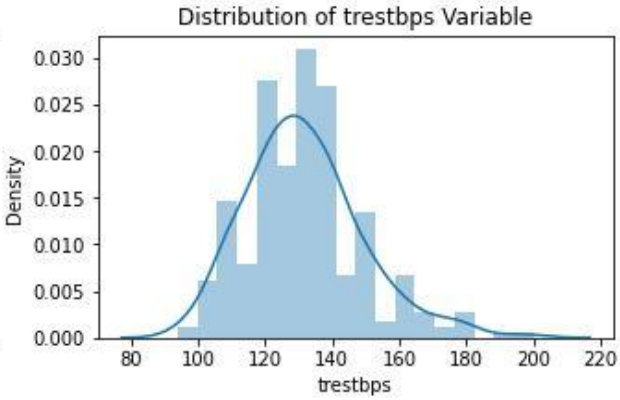


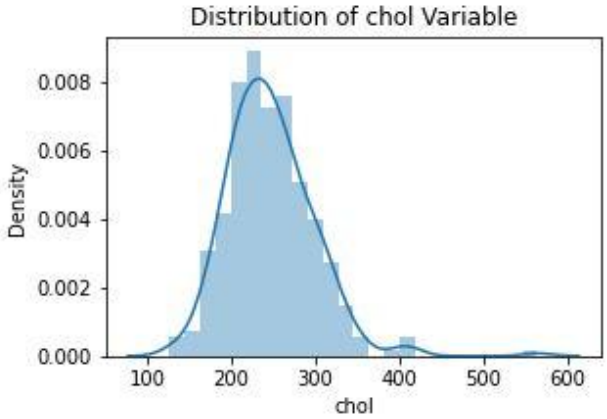
Fig-4. Correlation Matrix.

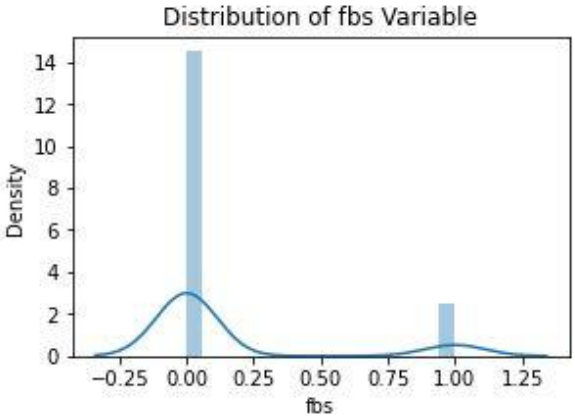
Fig-3 shows the histogram plot of features, Fig-4 shows the correlation matrix between the features and Fig-5 shows the distribution plots of all the features.

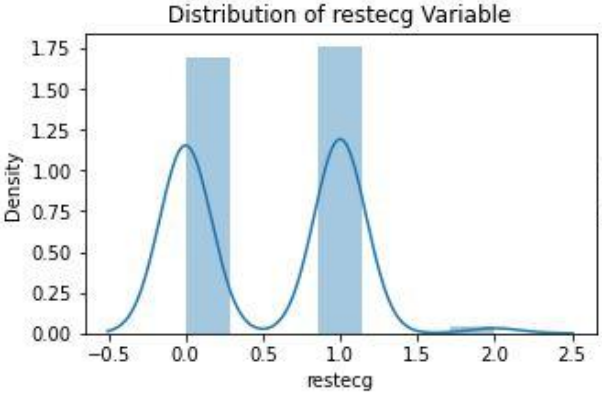


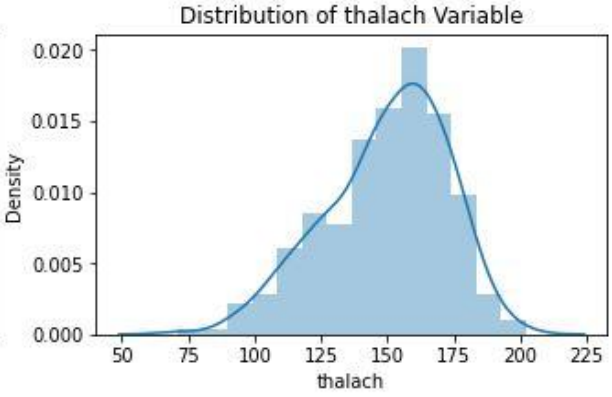

Distribution of cp Variable
Density
cp


Distribution of trestbps Variable
Density
trestbps


Distribution of chol Variable
Density
chol


Distribution of fbs Variable
Density
fbs


Distribution of restecg Variable
Density
restecg


Distribution of thalach Variable
Density
thalach

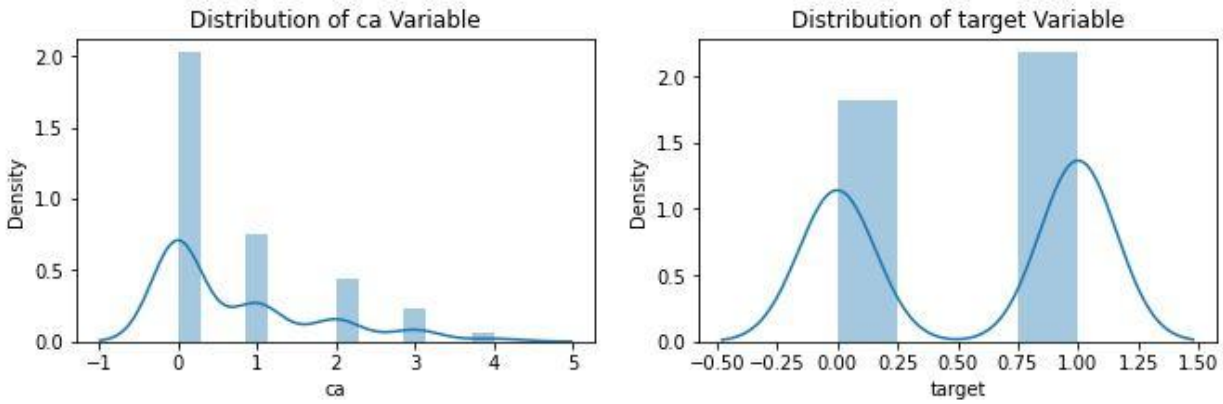


Fig-5. Distribution Plot of dataset Variables.

(2) Cross-Validation - We used 5 - fold cross-validation on the dataset. In this approach, the dataset is divided into 5 parts in which 4 parts are used in the model's training process, and the remaining part is used for the testing process.

(3) Machine Learning Classifiers - We apply different machine learning classifiers on the dataset, such as Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbor, Support Vector Machine,Artificial Neural Network [15]. The accuracy of the classifiers was calculated using the confusion matrix. A classifier that the highest accuracy bags can identify as the best classifier [15].

Logistic regression - A supervised machine learning technique used to address classification issues is logistic regression. True or false, Spam or not Spam, and other discrete values are predicted using classification issues. The sigmoid function is used to model the data in logistic regression.

Random Forest -The RF is created via supervised machine learning. Random forest is a user-friendly and adaptable algorithm. Random forest is a method that, in most cases, produces good results without the use of hyper-parameter tweaking. It creates a forest out of a collection of decision trees.

Decision tree - The DT is a supervised machine learning approach for segmenting data based on specific features. It splits the dataset into smaller and smaller subgroups as it develops the tree. There are two sorts of nodes in the

tree: decision nodes and leaf nodes. In a DT, decision nodes indicate outcomes, whereas leaf nodes are where the data is split.

Support Vector Machine(SVM) - SVM, or supervised learning, is the most widely used machine learning method. It's used to solve categorization issues. This technique tries to build a decision boundary that can divide n-dimensional space into distinct classes so that fresh data points may be rapidly assigned to the appropriate category. We forecast the incidence of stroke across a pre-defined time period in our stroke prediction issue, making it a binary classification problem that fits within the SVM framework [14].

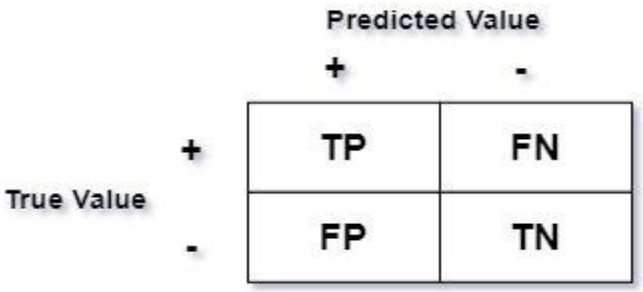
K-Nearest Neighbor(KNN) - K-NN is a supervised learning approach that may be used for both regression and classification, however classification is the most prevalent use. The K-NN technique implies that new and old data are similar, and it assigns new data to the category that is closest to the general categories. During the training phase, the K-NN approach saves the data, and when fresh data is received, it is classified into a class that is highly close to the most recent data.

Artificial Neural network - It's a hardware or software system based on the human brain and nervous system operating neurons. ANN is established on artificial neurons.The primary purpose of ANN was to solve problems in a similar manner to a human brain.

Grid Search - It is a tuning algorithm that is used for computing optimal values of hyperparameters. Grid search techniques find the optimal hyperparameters for the machine learning model. Then models are evaluated through cross-validation techniques.

Evaluation Parameters - For the assessment, we utilized the confusion matrix, accuracy, area under the receiver operating characteristic (AUC-ROC), sensitivity, specificity, f1-score, Matthews correlation coefficient (MCC) score, and Cohen kappa score. The confusion matrix is a table-like structure with true and projected values. In the below we describe confusion matrix and evaluation parameters formulas :-

Confusion Matrix -



$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$\text{Specificity} = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

$$\text{Sensitivity} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{F1-Score} = \frac{True\ Positive}{True\ Positive + 1/2(False\ Positive + False\ Negative)}$$

Result and Analysis ----------

For the sake of our research, we divided our dataset 80:20 for training and testing. We can see that our work achieves the best accuracy of 88.52% by using the random forest algorithm. We have used different ML classifiers for our work and the algorithms used are logistic regression, k-NN, SVM, DT, RF, and NN[11]. By the use of Logistic Regression, we got an accuracy of 88.52%. When we use the K- Nearest neighbor algorithm, then we got an accuracy of 88.52%. The accuracy of the SVM classifier is 86.89%. The accuracy by the use of the Decision tree algorithm is 83.51%. The accuracy by the use of neural networks is 75.41% [6].

Classifier	ACC	ROC AUC	Specificity	Sensitivity	F1-Score	MCC	Kappa
LR	88.52	88.58	90.32	87.50	88.89	0.77	0.77
KNN	88.52	88.42	87.88	90.62	89.23	0.77	0.77
SVM	86.89	87.02	90.00	84.38	87.10	0.74	0.74
DT	83.61	83.89	89.29	78.12	83.33	0.68	0.67
RF	88.52	88.74	93.10	84.38	88.52	0.77	0.77
NN	75.41	75.43	77.42	75.00	76.19	0.51	0.51

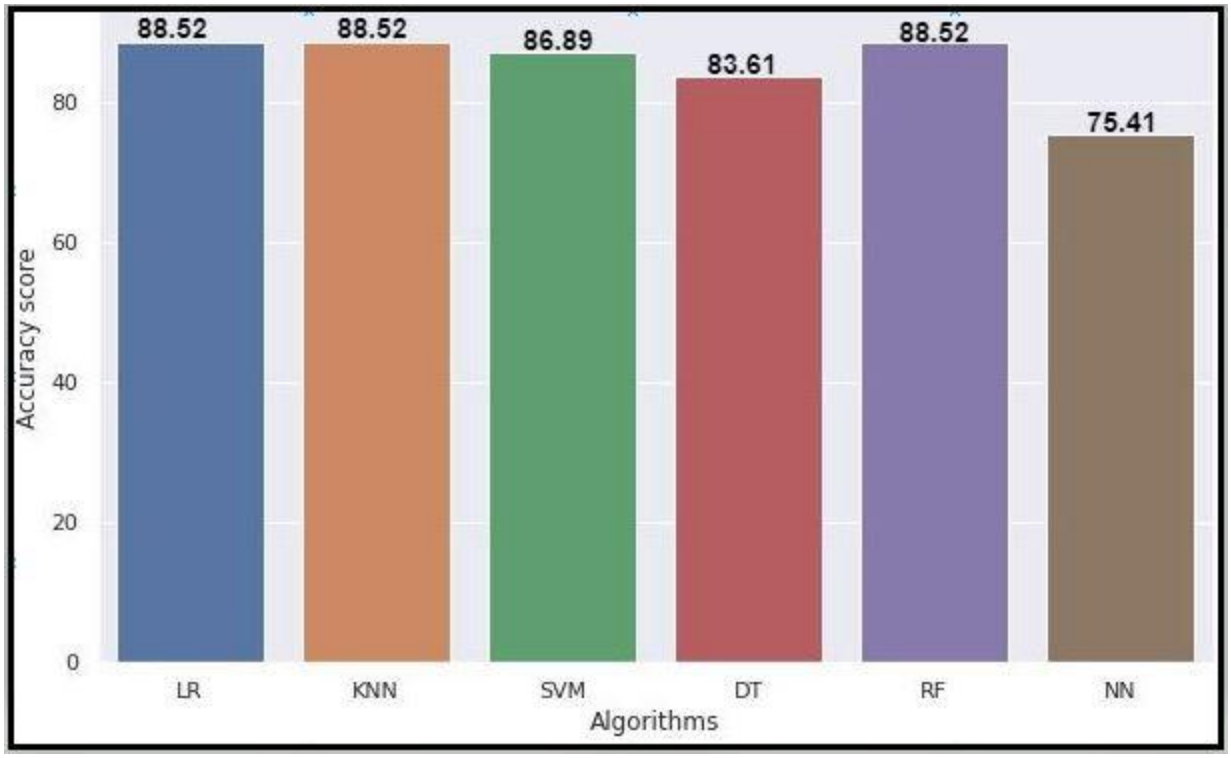


Fig-6. Accuracy achieved for classifiers on the dataset.

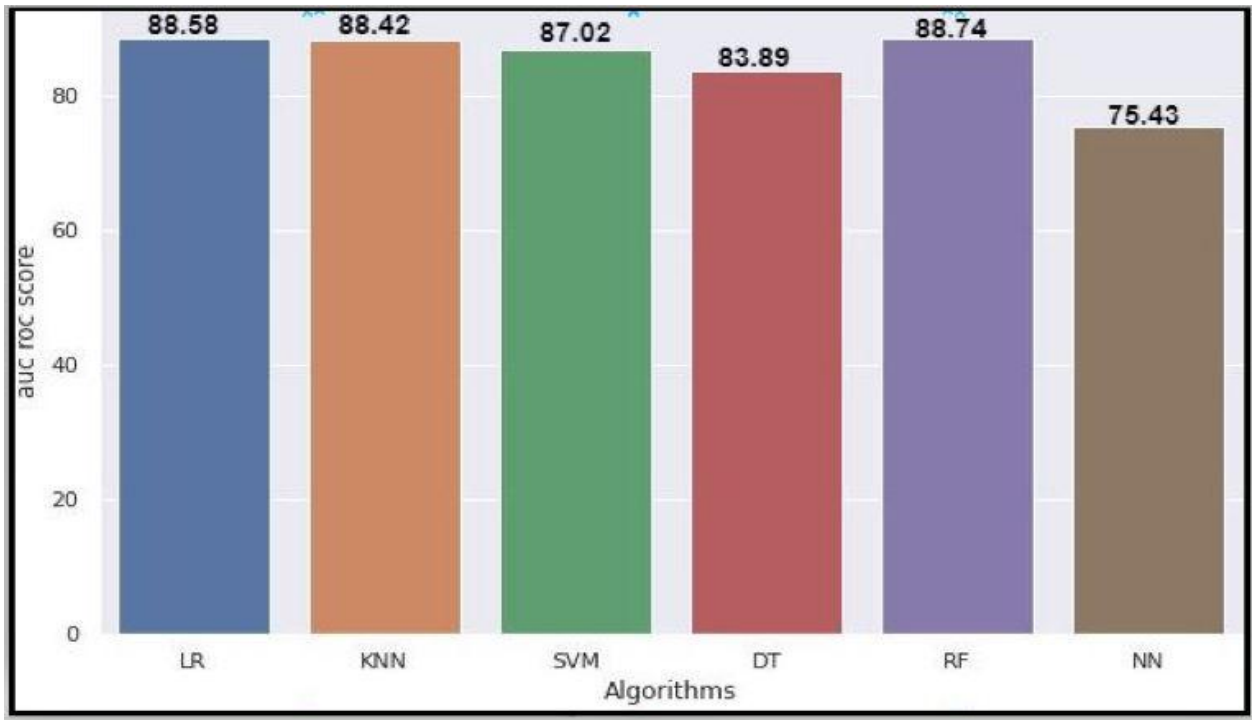


Fig-7. AUC ROC Score achieved for classifiers on the dataset..

Android Mobile App For Heart Disease Prediction ----------

We develop an android mobile app that helps predict heart disease. In this application, we create different-different fields like name, age, sex, cholesterol level, maximum heart rate, chest pain, fasting blood sugar(FBS), restecg, etc. We make this app as simple as possible so that maximum people easily use this app. We keep only three mandatory information like name, age, and sex. Our app also gives results according to only these three basic information, but if the user provides critical information like cp, fbs, etc., the app predicts better results. Our app also provides the feature-by-feature probability that affects heart disease. If the prediction result of heart disease comes above 60 then our app gives the prediction result in red color which shows danger jone and if the result comes out with less than 60 then prediction result comes in green color.

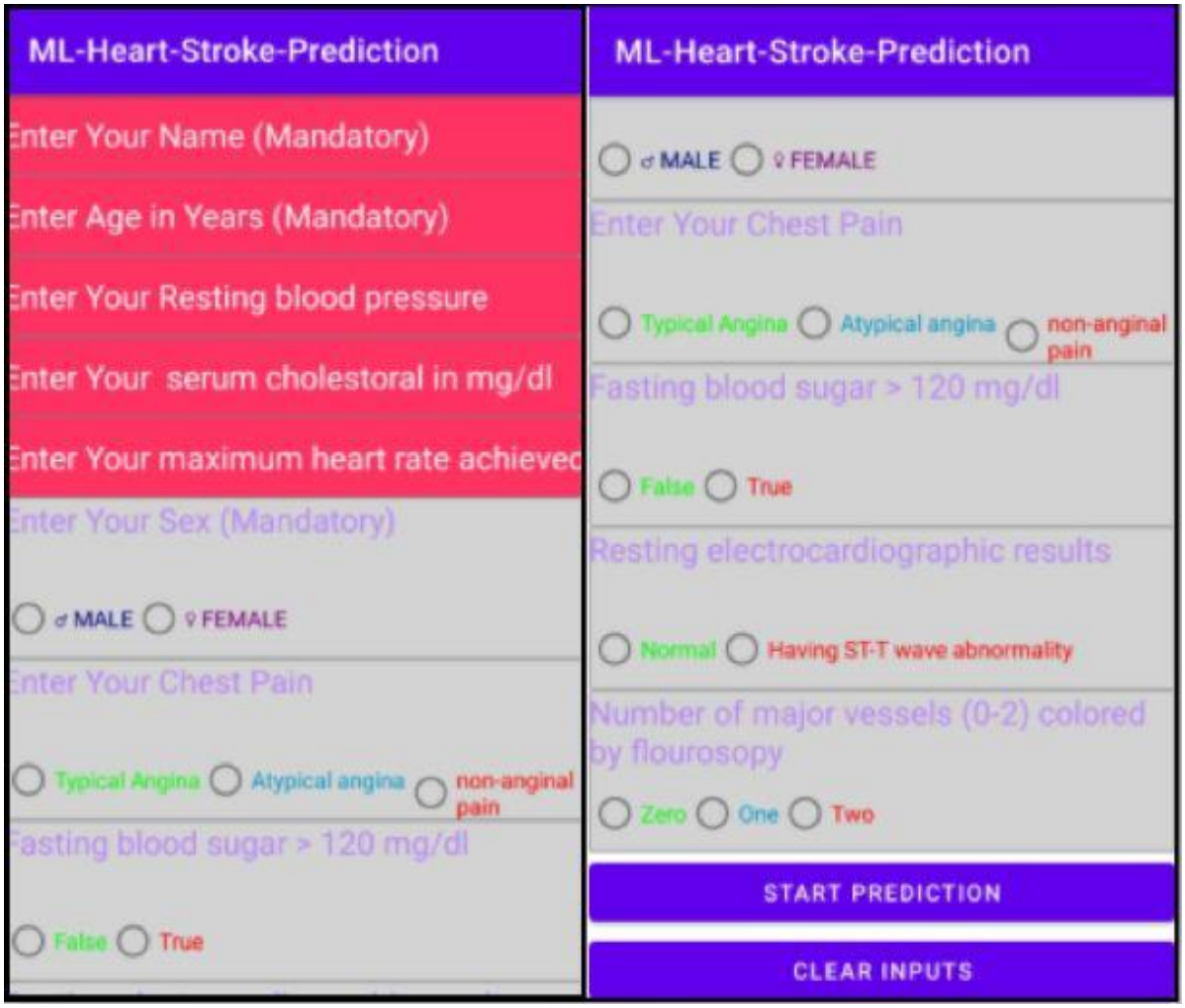
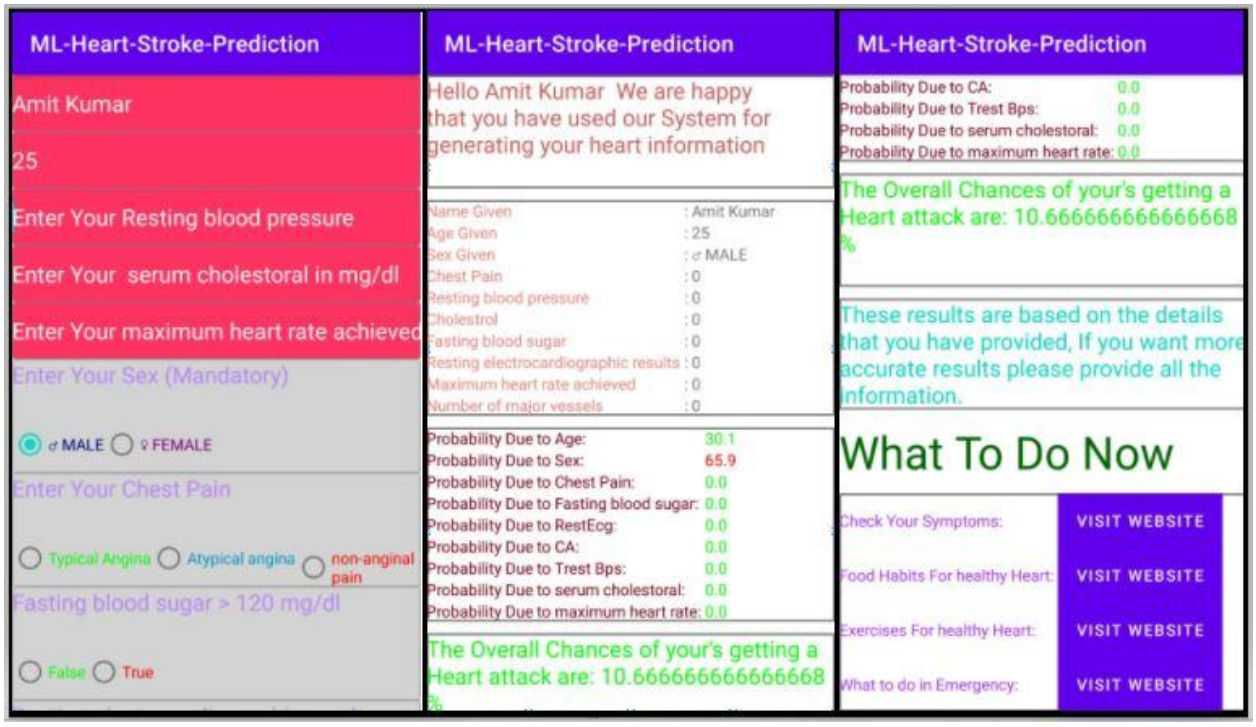
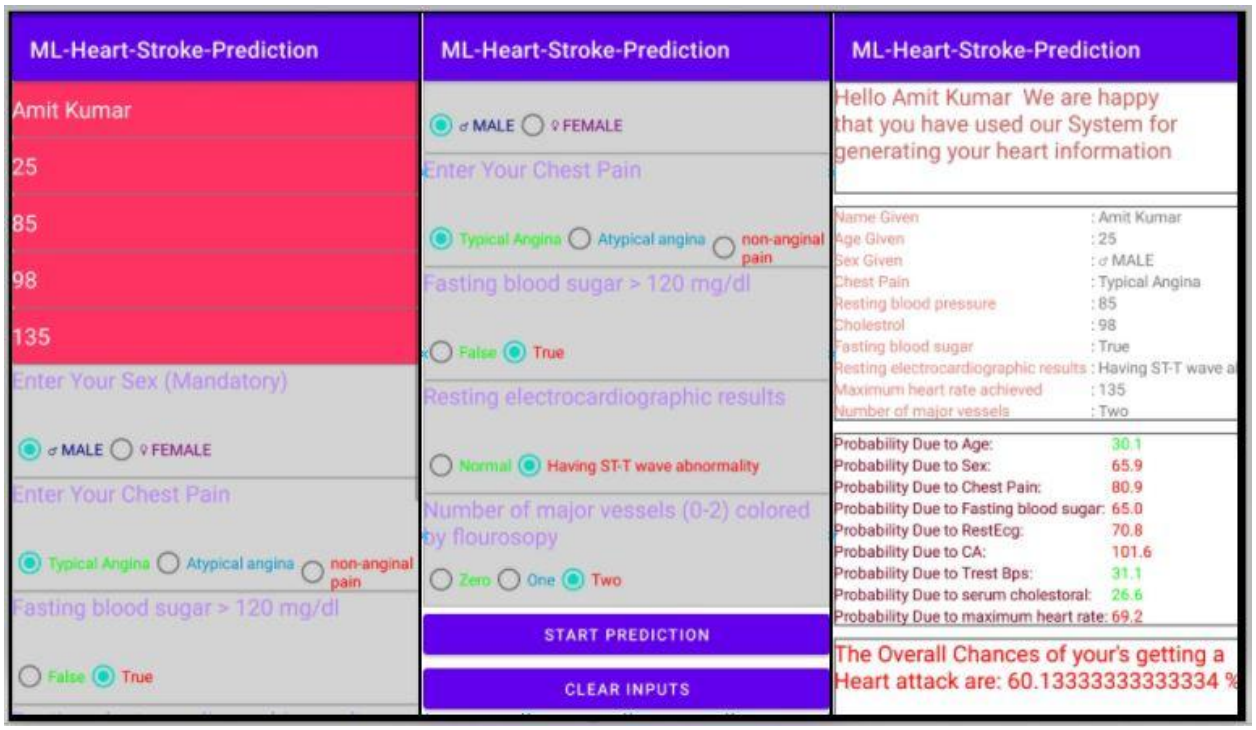


Fig-8 Full View of Mobile App.



In the above image users can only give mandatory information so our app predicts heart disease based on these three information. But in the below image the user can give his maximum information and predict the result.



Conclusion and Future Scope ----------

In this study, we applied the idea of machine learning to create a smartphone app for the prediction of cardiovascular illness. We attained an accuracy of 88.52 percent and an auc roc score of 88.74 percent using a random forest classifier [16]. Our method can assist people in predicting cardiac disease at an early stage. In recent years, it has become critical to forecast or diagnose illness at an early stage. So, in the future, if we combine deep learning techniques with machine learning techniques [18], we anticipate that the findings will be more accurate, which will benefit both patients and physicians [10].

Acknowledgement ----------

We, Mritunjay Ashish(MT20102), Naveen(MT20078), and Shubham Sharma(20315), are thankful to Mr. Sumeet Patiyal and Dr. G.P.S.Raghava for their guidance and support throughout the project timeline.

References -

1) https://www.ijeat.org/wp-content/uploads/papers/v9i3/B3986129219.pdf

2) https://www.hindawi.com/journals/cin/2021/8387680/

3) https://dl.acm.org/doi/pdf/10.1145/1835804.1835830

4) https://www.health.harvard.edu/

5) Poirier, Paul, et al. "Obesity and cardiovascular disease:pathophysiology, evaluation, and effect of weight loss: an update of the 1997 American Heart Association Scientific Statement on Obesity and Heart Disease from the Obesity Committee of the Council on Nutrition, Physical Activity, and Metabolism." Circulation 113.6 (2006): 898-918.

6) N. K. Kumar, G. S. Sindhu, D. K. Prashanthi and A. S. Sulthana, "Analysis and Prediction of Cardiovascular Disease using Machine Learning Classifiers," 2020

6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020,pp. 15-21, doi: 10.1109/ICACCS48705.2020.9074183.

7) S. Ismaeel, A. Miri and D. Chourishi, "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis," 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), Ottawa, ON, Canada, 2015, pp. 1-3, doi: 10.1109/IHTC.2015.7238043.

8) N. Prentzas, A. Nicolaides, E. Kyriacou, A. Kakas and C. Pattichis, "Integrating Machine Learning with Symbolic Reasoning to Build an Explainable AI Model for Stroke Prediction," 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), Athens, Greece, 2019, pp. 817-821, doi: 10.1109/BIBE.2019.00152.

9) S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 2019, pp.619-623, doi: 10.1109/IBCAST.2019.8667106.

10) D. Tiwari, M. Ashish, N. Gangwar, A. Sharma, S. Patel and S. Bhardwaj, "Potato Leaf Diseases Detection Using Deep Learning," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 461-466, doi: 10.1109/ICICCS48265.2020.9121067.

11) https://www.ijraset.com/fileserve.php?FID=27198

12) M. Bhatia and D. Motwani, "Use of Ensemble Learning for Prediction of Heart Disease," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), 2020, pp. 1016-1023, doi: 10.1109/ICOEI48184.2020.9142964.

13) M. Gjoreski, M. Simjanoska, A. Gradišek, A. Peterlin, M. Gams and G. Poglajen, "Chronic Heart Failure Detection from Heart Sounds Using a Stack of Machine-Learning Classifiers," 2017 International Conference on Intelligent

Environments (IE), 2017, pp. 14-19, doi: 10.1109/IE.2017.19.

14) R. S. Jeena and S. Kumar, "Stroke prediction using SVM," 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), Kumaracoil, India, 2016, pp. 600-602, doi: 10.1109/ICCICCT.2016.7988020.

15) S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.

16) T. Obasi and M. Omair Shafiq, "Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 2393-2402, doi: 10.1109/BigData47090.2019.9005488.

17) M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), Bangkok, Thailand, 2017, pp. 158-161, doi: 10.1109/IEMECON.2017.8079581.

18) C. Hung, W. Chen, P. Lai, C. Lin and C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea (South), 2017, pp. 3110-3113, doi: 10.1109/EMBC.2017.8037515.