

# Real-Time Cardiovascular Disease Prediction using Machine Learning

<sup>1</sup>Aashna Shroff, <sup>2</sup>Anushka Tare, <sup>3</sup>Bhavna Arora.

<sup>1,2</sup>Student, <sup>3</sup>Assistant Professor, Atharva College of Engineering, Mumbai, India.

<sup>1</sup>shroff.aashna22@gmail.com, <sup>2</sup>anu.tare22@gmail.com, <sup>3</sup>bhavnaarora@atharvacoe.ac.in

**Abstract :** Cardiovascular Diseases are the number one source of explanation for death globally, more people die annually from CVDs than from the other cause. Roughly estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the death rate. Coming to CVD, the silver lining is that heart attacks are highly preventable and simple lifestyle modifications. It is, however, difficult to identify high risk patients because of the multifactorial nature of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, et cetera. This is where machine learning and data mining comes to the rescue. Our aim is to develop a machine learning model capable of identifying if a person has a cardiovascular disease or not. We can do this by extracting the features of a dataset and training a machine learning model with the highest accuracy. We trained four machine learning algorithms and compared their accuracy: Naive Bayes, Logistic Regression, Decision Tree and Random Forest to identify people with high risk of getting cardiovascular disease. For real-time prediction, we then deployed the trained machine learning model which had the best accuracy to the internet using Flask. The user can enter their data into the website after either logging in or registering themselves and get the result of the prediction in real-time.

**Keywords** — Algorithms, Accuracy, Cardiovascular Diseases, Machine Learning, Random Forest, Real-time Analysis, Predictions.

## I. INTRODUCTION

Machine learning algorithms are used in a wide variety of applications, for example email filtering, computer vision, where it is difficult or unattainable to develop conventional algorithms to perform the needed tasks. Prognosis is a branch of medicine that specializes in predicting the future health of patients and Machine Learning is considered to be a powerful tool for it. In this project, we aimed to develop the machine learning model which has the highest accuracy and precision in predicting if a patient is at a risk of having heart diseases. Thus, we trained four classification algorithms namely Logistic Regression, Naive Bayes, Decision Tree and Random Forest to find out which one has the highest accuracy as these algorithms are considered to be very accurate and easy to use.

## II. NEED & MOTIVATION

Raised blood pressure, glucose, and lipids as well as overweight and obesity may be demonstrated in individuals with high risk of CVD. Identifying those at highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths. The early diagnosis of heart disease is needed today since it plays a vital role in making prior decisions related to lifestyle changes and thus taking utmost precautions at the right time. According to study there are around 80% of preventable cases of heart disease if diagnosed early.

Early diagnosis of CVD diseases would contribute immensely to study the disease and analyse the state of the

patient. This would result in taking necessary steps to cure the patient in time.

Thus, this drives us to develop an application that could help reduce the fatality rate by providing early diagnosis results so that concerned people can take precautionary measures well in advance and also receive medical recommendations.

## III. BASIC CONCEPT

The main aim of our project is to build an application system that carries out real-time analysis by accepting real-time data & predicting the risk of a patient getting cardiovascular disease in the future by analyzing the required vital data of patients. We require a trained machine learning model with the best accuracy to do the necessary prediction. In order to gain the best accuracy model, we compared the accuracy of four trained machine learning models and then selected the model with the highest accuracy for further prediction. The algorithms we used to compare are Logistic Regression, Naive Bayes, Decision Tree and Random Forest algorithms.

Following this, the next step was to deploy this machine learning model and make it accessible for the users so they can avail the real-time prediction feature. In order to achieve this, we developed a website which allows the user to enter their data and obtain real-time prediction result. Security of the system is maintained by features like 'Login' and 'Register'. After entering the data the system will respond by highlighting whether or not the patient is at risk

#### IV. REPORT ON PRESENT INVESTIGATION

As far as the market is concerned, the use of Machine Learning techniques in the field of Medicine and Cardiology to be precise is quite low. As a result the rate of heart diseases are increasing day by day with a low recovery percentage. Present system still works on curing the disease and not focusing much on taking or promoting precautionary measures for this chronic problem. The System fails in analysing the data and predicting patterns out of the data efficiently. Also, the system puts more load and stress over human forces of the concerned field which could prove fatal over time.

The major limitation seen in the present investigation is that the models are not trained to fetch the real time data and give the desired results which would benefit the doctors to start the treatment quite early and prevent any further complications.

#### V. AIM & OBJECTIVES

##### 5.1 Aim

Our aim is to develop a machine learning model capable of identifying if a person has a cardiovascular disease or not. We can do this by extracting the features of a dataset and training a machine learning model with the highest accuracy. We trained four machine learning algorithms and compared their accuracy: Naive Bayes, Logistic Regression, Decision Tree and Random Forest to identify people with high risk of getting cardiovascular disease.

##### 5.2 Objectives

1. To determine significant risk factors based on medical dataset which could contribute to heart disease.
2. To develop machine learning models to predict future possibility of heart disease by implementing Naive Bayes, Logistic Regression, Decision Tree and Random Forest.
3. To compare all the models and determine the accuracy of the models.
4. To take real time data and give the required results using the most efficient model trained, among the 4 algorithms to give best results.

#### VI. IMPLEMENTED SYSTEM

We are proposing that we need a machine learning model which gives the best accuracy and precision for the prediction. So we compare the accuracy of four prominent machine learning classifier algorithms: Naive Bayes, Decision Tree, Logistic Regression and Random Forest Algorithm. Additionally, to solve the problem of real time data analysis we built a website where users can directly feed their data and get the prediction. The data the user has to feed includes age, gender, height, weight, systolic and diastolic blood pressure, cholesterol, glucose, smoker or not, if they

consume alcohol or not and active or not. After the user has submitted the necessary data, the machine learning model predicts and shows the results, telling whether the user is at a high risk of getting cardiovascular disease or not. For authentication of the user, we have integrated a login page to the website and also added a registration page for new users.

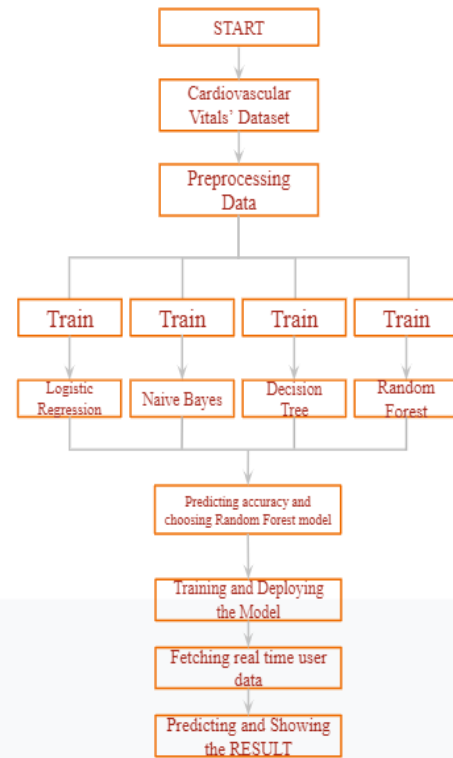


Diagram 1: BLOCK DIAGRAM

#### VII. DATASET & CONNECTIVITY

To train our model we used a dataset called “Cardiovascular Disease Dataset”, in our dataset we have parameters like { id, age, gender, height, weight, systolic BP, diastolic BP, cholesterol, glucose, smoke, alcohol, active, cardio} where ‘cardio’ describes if the patient has cardiovascular diseases or not. All the other parameters are necessary for determining if the patient has cardiovascular disease or not except for the ‘id’ parameter. This dataset is essential in building the machine learning model, and for the connectivity of this project we need a stable internet connection with python and flask pre-installed.

For the smooth running of our system, we linked the website to a database using phpMyAdmin. We used SQL to query from the database and maintained the database for the login and registration of the users into the website so they can access the real-time prediction.

#### VIII. FEASIBILITY

##### 8.1 TECHNICAL FEASIBILITY :

Technical Feasibility, includes the development of the machine learning model. This model is developed using Jupyter Notebook and is deployed to the internet using Flask.

It requires a stable internet connection and pre-installed libraries like numpy, pandas, sklearn etc.

**8.2 Economic Feasibility :**

Economic feasibility consists of an economical and logistical outlook for this model. By using this model we can further avoid the expenses of required pathological services. This model can be deployed to use in facilities like hospitals, military, research centers for greater analysis of a patient's medical background.

**8.3 Legal Feasibility :**

The product fulfills all the legal requirements as it is only a prediction model and predicting if a patient is at high risk or at a low risk of getting cardiovascular disease, also the necessary data taken from the user to make prediction is not stored by our system.

**IX. METHODOLOGY**

- To train our model we used a dataset called "Cardiovascular Disease Dataset" which is available for free on Kaggle website. Below is the screenshot of our dataset, in our dataset we have parameters like {id, age, gender, height, weight, systolic BP, diastolic BP, cholesterol, glucose, smoke, alcohol, active, cardio} where 'cardio' describes if the patient has cardiovascular diseases or not. All the other parameters are necessary for determining if the patient has cardiovascular disease or not except for the 'id' parameter. This dataset is essential in building the machine learning model, and for the connectivity of this project we need a stable internet connection with python and flask pre-installed.
- Our Cardiovascular dataset comprises of the following attributes:

CARDIOVASCULAR DISEASE DATASET			
Sr NO.	Attributes	Description	Type
1.	Age	Patient's age	Numeric
2.	Gender	Gender of patient: male =1, female=2;	Nominal
3.	Height	Height in cm	Numeric
4.	Weight	Weight in kg	Numeric
5.	Cholesterol	Normal =1, Above normal =2, Well above normal =3;	Numeric
6.	Systolic Blood Pressure	Blood pressure in mmHg	Numeric
7.	Diastolic Blood Pressure	Blood pressure in mmHg	Numeric
8.	Glucose	Normal =1, Above normal =2, Well above normal =3;	Numeric
9.	Smoker or not	Yes or no ( 1 or 0 )	Binary
10.	Alcohol consumption	Yes or no ( 1 or 0 )	Binary
11.	Active or not	Yes or no ( 1 or 0 )	Binary

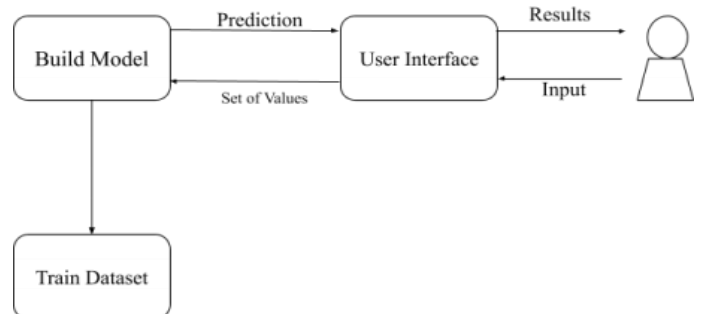
**Table 1: CARDIOVASCULAR DISEASE DATASET**

- We load the necessary libraries and preprocess the data.
- We conduct exploratory data analysis to understand the data properly.
- Then we split the data into two parts: train and test.
- We load the machine learning models with the dataset and train the model.
- Further, we test the machine learning model and check its accuracy.
- We compare the accuracies of all the machine learning models and we learn that the random forest algorithm has the highest accuracy hence it is the best algorithm.
- Further, now for real time-data analysis we used the random forest machine learning algorithm to make the prediction.
- We build a website where users can enter their data and get real-time predictions.
- For security, we built a login page and for new users we built a registration page, after logging in the users can easily access the machine learning model.
- They have to enter data related to their health including their age, gender, height, weight, systolic and diastolic blood pressure, cholesterol, glucose, smoker or not, if they consume alcohol or not and are active or not.
- The machine learning model makes the prediction using this data, and the results of the prediction are seen on the directed page.

**X. DESIGN DETAILS**

**1. Sequence Diagram:**

This diagram represents the flow of our project, first we have our user interface where the user enters all the necessary data needed for prediction which includes their age, gender, height, weight etc and all these values are passed as arguments to the trained machine learning model which uses

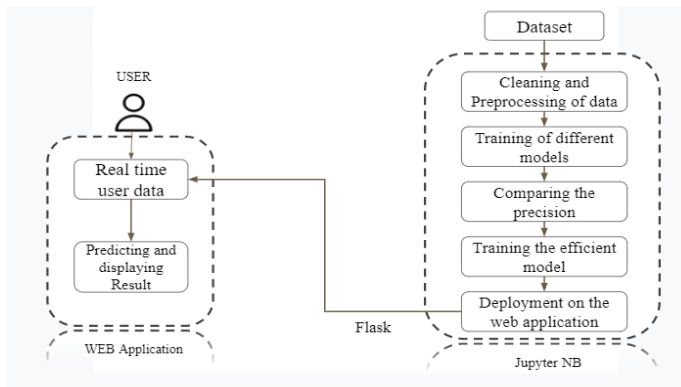


this data to predict whether the user is at like of getting cardiovascular disease or not and then this prediction is shown to the user.

**2. Architecture Diagram:**

This diagram represents the entire architecture of our project which includes all the hardware and software environments, and the middleware connecting them. For the purpose of building our machine learning model and also comparing the

accuracy of our models we used Jupyter Notebook and in order to deploy our application to the Internet we used Flask.



### XI. TESTING

1. We use confusion matrix to find the values for recall, accuracy and precision and it is as follows:

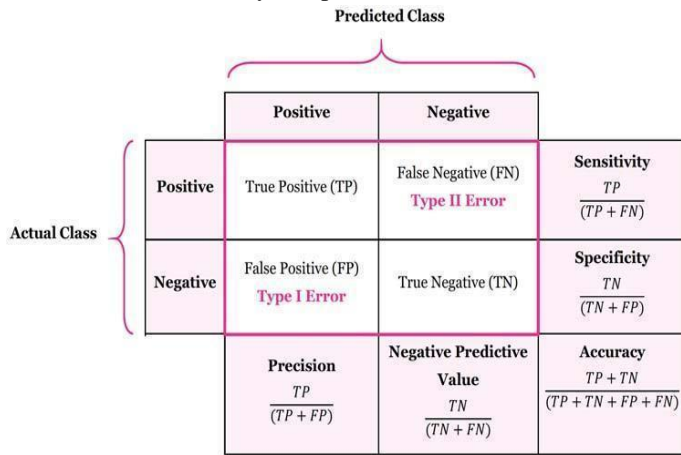


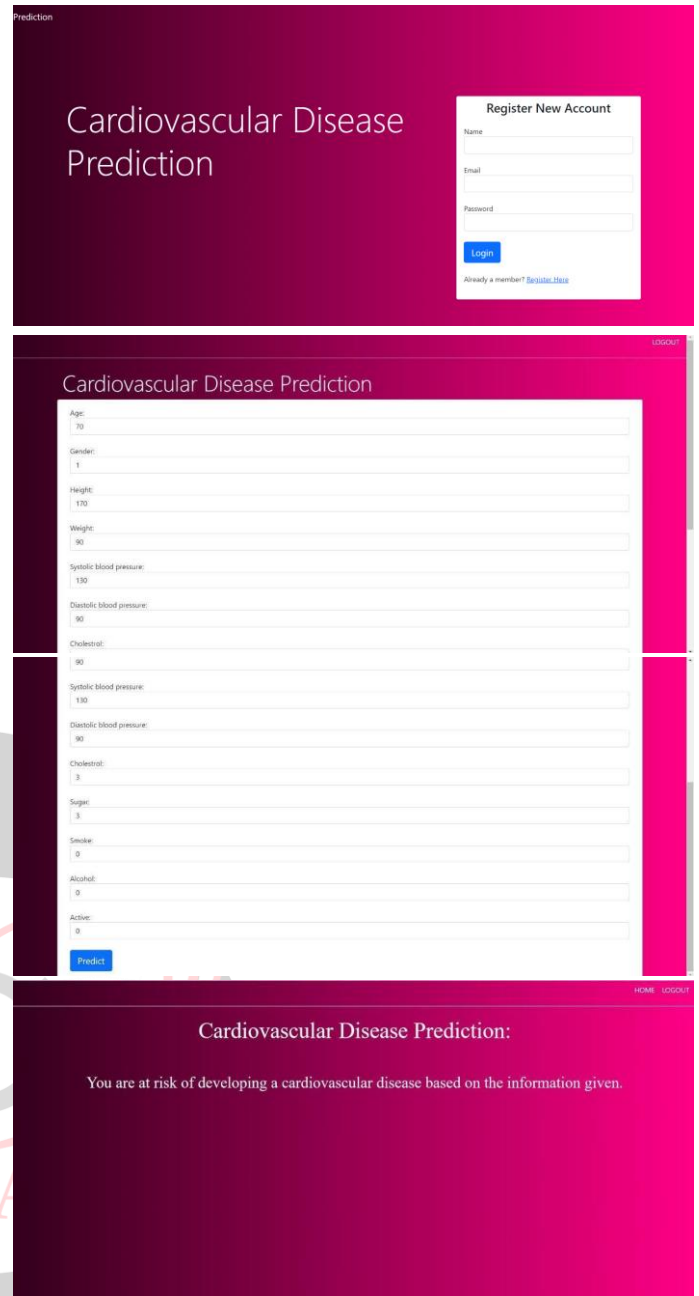
Diagram 2: CONFUSION MATRIX VALUES

2. We tested the accuracy, precision and recall values of all the machine learning algorithms and the table below is the comparison of all these values:

Algorithm	Precision	Recall	Accuracy
Random Forest	0.724	0.630	0.719
Decision Tree	0.633	0.705	0.633
Naive Bayes	0.722	0.220	0.568
Logistic Regression	0.799	0.453	0.670

Diagram 3: ACCURACY OF ALL ALGORITHMS

3. Further, we tested the accuracy of the Random Forest model performing real time data analysis: When entered data of a person who has been already detected with cardiovascular disease:



Screenshots: OUR IMPLEMENTED WEBSITE

### XII. RESULT ANALYSIS

We calculated the accuracies of four machine learning algorithms namely : Naive Bayes, Logistic Regression, Decision Trees and Random Forest algorithm trained by our dataset and found out that the Random Forest algorithm had the best accuracy which is 0.719. Thus, for the real time data analysis, we used this Random Forest model for real-time prediction of the data filled by the user. Then we deployed the Random Forest machine learning model on the internet using Flask, whereas on the frontend the user will see a website on which it has to 'login' or 'register' in order to access the prediction model, then they can fill in all the necessary data needed for prediction. Subsequently, after filling the data, the machine learning model predicts if a patient is at risk of having a cardiovascular disease or not.



```
Out[25]:
```

	Accuracy Score
Random forest	0.719463
KNN	0.677882
Logistic Regression	0.670634
Decision tree	0.635843
Naive bayes	0.568475

```
In [27]: plt.figure(figsize=(5,5))
sns.barplot(x=scores_frame.index,y=scores_frame["Accuracy Score"])
plt.xticks(rotation=45) # Rotation of Country names..
```

```
Out[27]: (array([0, 1, 2, 3, 4]),
[Text(0, 0, 'Random forest'),
Text(1, 0, 'KNN'),
Text(2, 0, 'Logistic Regression'),
Text(3, 0, 'Decision tree'),
Text(4, 0, 'Naive bayes')])
```

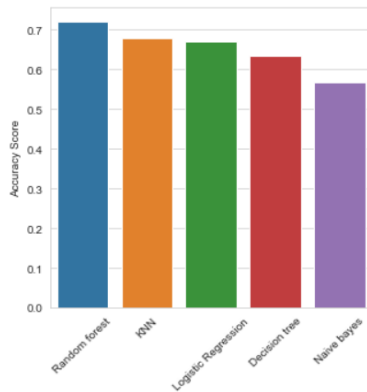


Diagram 4: COMPARISON OF ALGORITHMS

### XIII. CONCLUSION

The early prediction of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. In this project we compared various algorithms to calculate the algorithm with the best efficiency then we deployed this machine learning model to a website to help make real-time predictions. We can thus predict and avert some adverse effects of cardiovascular disease using this model. Further for its enhancement, we can train on models and predict the types of cardiovascular diseases providing recommendations to the users, and also use more enhanced and accurate models. The system we have created is reliable and secure, and makes real-time predictions based on the information specified by the user on the website.

### XIV. APPLICATIONS

- Various Private & Governmental Medical centres.
- Health related Non Governmental Organisations.
- Private Medical Practitioners.

### XV. FUTURE SCOPE

The future prospects of the following system includes giving precautionary measures to the minimal risk patients and enabling the Doctor to treat high risk patients accordingly over the required period of time. Also, adding more data to the model and making it much more efficient than it is already.

### XVI. ACKNOWLEDGEMENT

We are highly indebted to Atharva College of Engineering for their guidance and constant supervision as well as for providing necessary information regarding the project and also for their support in completing the project black book. We feel obliged that our Principal Dr. Shrikant Kallurkar has given us such a platform to showcase our technical prowess and knowledge.

We would like to express our gratitude towards our parents and the members of Atharva College of Engineering, especially Prof. Suvarna Pansambal (HOD) as the project coordinator for their kind cooperation and encouragement which helped us in completion of this project synopsis.

Our special thanks and appreciation also go to our colleagues and also to the people who have willingly helped us out with their abilities in this project.

### XVII. REFERENCES

- [1] Apurb Rajdhan , Avi Agarwal , Milan Sai , Dundigalla Ravi, Dr. Poonam Ghuli, 2020, "Heart Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT) Volume 09, Issue 04 (April 2020).
- [2] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.
- [3] Ramalingam, V V & Dandapath, Ayantan & Raja, M. (2018). Heart disease prediction using machine learning techniques: A survey. International Journal of Engineering & Technology. 7. 684. 10.14419/ijet.v7i2.8.10557.
- [4] Dinesh, Kumar G, K. Arumugaraj, K. Santhosh and V. Mareeswari. "Prediction of Cardiovascular Disease Using Machine Learning Algorithms." 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (2018): 1-7. doi: 10.1109/ICCTCT.2018.8550857.
- [5] N. Louridi, M. Amar and B. E. Ouahidi, "Identification of Cardiovascular Diseases Using Machine Learning," 2019 7th Mediterranean Congress of Telecommunications (CMT), 2019, pp. 1-6, doi: 10.1109/CMT.2019.8931411
- [6] Siva, Galla & Bindhika, Sai & Meghana, Munaga & Reddy, Manchuri & Dharmadurai, Rajalakshmi. (2020). Heart Disease Prediction Using Machine Learning Techniques. 2395-0056.
- [7] V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842
- [8] Garg, Apurv & Sharma, Bhartendu & Khan, Rizwan. (2021). Heart disease prediction using machine learning techniques. IOP Conference Series: Materials Science and Engineering. 1022. 012046. 10.1088/1757-899X/1022/1/012046.
- [9] "Heart Disease Prediction using Machine Learning", International Journal of Emerging Technologies and Innovative Research (www.jetir.org | UGC and issn Approved), ISSN:2349-5162, Vol.7, Issue 6, page no. pp 2081-2085, June-2020
- [10] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in *IEEE Access*, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.