



# VIT

**Vellore Institute of Technology**  
(Deemed to be University under section 3 of UGC Act, 1956)

## **School of Computer Science and Engineering**

### **J Component report**

**Programme : B.Tech(CSE with specialization in AI & ML)**

**Course Title : Machine Learning Essentials**

**Course Code : CSE1015**

**Slot : A1**

**Title: Prediction of Heart Disease at an Early Phase**

**Team Members: Gautam Arora | 20BA11053**

**Narayan Subramanian | 20BA11207**

**Sarvesh Chandak | 20BA11221**

**Faculty: Dr. R. Rajalakshmi**

**Sign:**

**Date:**

**29.04.2022**

# **CSE1015 – Machine Learning Essentials**

## **J Component Report**

### **A project report titled**

## **Prediction of Heart Disease at an Early Phase**

*By*

Reg. No: 20BAI1053

Name: Gautam Arora

Reg. No: 20BAI1207

Name: Narayan Subramanian

Reg. No: 20BAI1221

Name: Sarvesh Chandak

BACHELOR OF TECHNOLOGY  
IN  
COMPUTER SCIENCE AND ENGINEERING

*Submitted to*

**Dr. R. Rajalakshmi**

**School of Computer Science and Engineering**



**VIT<sup>®</sup>**

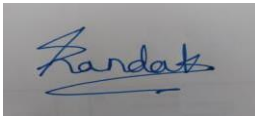
**Vellore Institute of Technology**

(Deemed to be University under section 3 of UGC Act, 1956)

*April 2022*

## **DECLARATION BY THE CANDIDATE**

We hereby declare that the report titled “**Prediction of Heart Disease at an Early Phase**” submitted by me to VIT Chennai is a record of bona-fide work undertaken by us under the supervision of **Dr. R. Rajalakshmi, Associate Professor, SCOPE, Vellore Institute of Technology, Chennai.**



Sarvesh Chandak



Narayan Subramanian



Gautam Arora

---

## **ACKNOWLEDGEMENT**

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. R. Rajalakshmi**, School of Computer Science and Engineering for her consistent encouragement and valuable guidance offered to us throughout the course of the project work.

We are extremely grateful to **Dr. R. Ganesan, Dean**, School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, for extending the facilities of the school towards our project and for his unstinting support.

We express our thanks to our **Head of the Department** for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the courses.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

## **BONAFIDE CERTIFICATE**

Certified that this project report entitled “**Prediction of Heart Disease at an Early Phase**” is a bona-fide work of **Gautam Arora (20BAI1053)**, **Narayan Subramanian (20BAI1207)**, **Sarvesh Chandak (20BAI1221)** carried out the “J”-Project work under my supervision and guidance for CSE1015 – Machine Learning Essentials.

**Dr. R. Rajalakshmi**

SCOPE

## **TABLE OF CONTENTS**

<b>Ch. No</b>	<b>Chapter</b>	<b>Page Number</b>
1	Abstract	6
2	Introduction	7-8
3	Problem Statement	9
4	Proposed Methodology	10-13
4	Literature Survey	14-18
5	Dataset Description	19-23
6	Results	24-28
7	Appendix (Implementation)	29-34
8	Conclusion and Future Work	35
9	Reference	35-36

## **ABSTRACT**

Machine Learning is used across many ranges around the world. The healthcare industry is no exclusion. Machine Learning can play an essential role in predicting presence/absence of locomotors disorders, heart diseases and more. Such information, if predicted well in advance, can provide important intuitions to doctors who can then adapt their diagnosis and dealing per patient basis. We work on predicting possible heart diseases in people using Machine Learning algorithms. In this project we perform the comparative analysis of classifiers like decision tree, Naïve Bayes, Logistic Regression, SVM and Random Forest and we propose an ensemble classifier which perform hybrid classification by taking strong and weak classifiers since it can have multiple number of samples for training and validating the data so we perform the analysis of existing classifier and proposed classifier.

Keywords: SVM; Naive Bayes; Decision Tree; Random Forest; Logistic Regression; Adaboost; XG-boost; python programming; confusion matrix; correlation matrix.



## **INTRODUCTION**

Cardiovascular diseases (CVDs) are a group of disorders involving the heart and blood vessels. They are the number one cause of death globally, taking an estimated of 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Among these, 38% people were under the age of 70 years. Stroke, heart failure, arrhythmia and myocardial infarction are some of the most common cardiovascular diseases with high mortality rates around the world.

Early and effective detection of heart diseases is critical in the treatment and management of cardiovascular diseases, wherein machine learning can be a powerful tool in detecting a potential heart disease diagnosis. Even if heart diseases are found because the prime source of death within the world in recent years, they are also those which will be controlled and managed effectively. The whole accuracy in management of a disease lies on the right time of detection of that disease. But they are not detected in the early stages due to the impractical costs of the tests available. Thus, a fast, real-time and reliable system that predicts the chances of a patient having heart disease in an optimized manner is required.

The heart attack occurs when arteries which supply oxygenated blood to heart does not function due to completely blocked or narrowed. Various types of heart diseases are:

1. Coronary heart disease
2. Cardiomyopathy
3. Cardiovascular disease
4. Ischaemic heart disease
5. Heart failure
6. Hypersensitive heart disease
7. Inflammatory heart disease
8. Valvular hear disease

Common risk factors of heart disease include:

1. High blood pressure
2. Abnormal blood lipids
3. Use of tobacco
4. Obesity
5. Physical inactivity

6. Diabetes
7. Age
8. Gender
9. Family genetics, etc

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future heart disease by analysing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

## **Problem Statement**

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either it is expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyse the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data

## **Proposed Methodology**

### **Existing System**

Heart disease is even being highlighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being done to predict the possibility of this deadly disease in prior. So that various tools & techniques are regularly being experimented with to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can conclude. This technique can be very well adapted to do the prediction of heart disease. As the well-known quote says “Prevention is better than cure”, early prediction & its control can be helpful to prevent & decrease the death rates due to heart disease.

### **Proposed System**

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre-processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

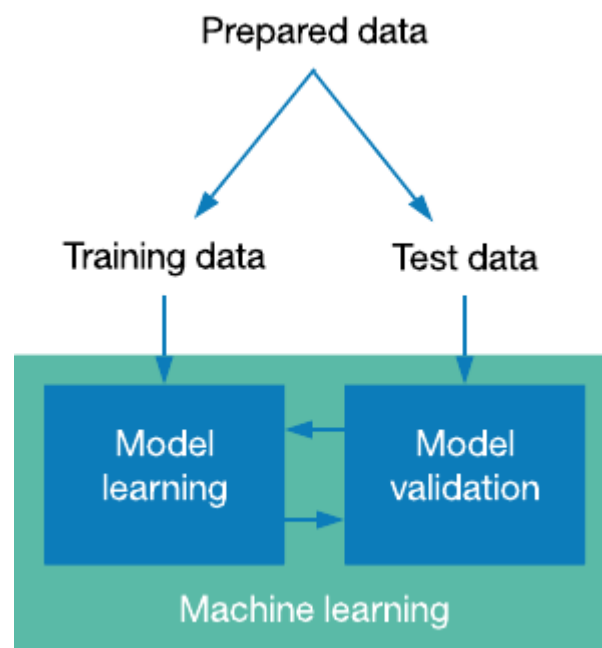
- 1) Collection of Dataset
- 2) Selection of attributes

3) Data Pre-Processing

4) Disease Prediction

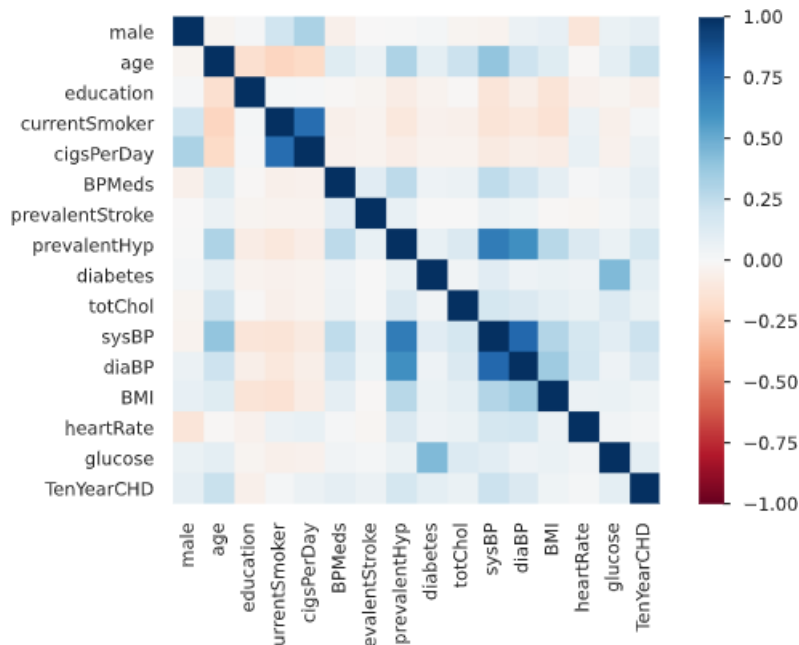
## Collection Of Dataset

Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 70% of training data is used and 30% of data is used for testing.



## Selection Of Attributes

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model.



## Pre-Processing of Data

Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Pre-processing of data is required for improving the accuracy of the model.



## **Literature Survey**

### **1. Intelligent Heart Disease Prediction System using Data Mining Techniques**

The healthcare industry collects huge amounts of healthcare records which, regrettably, are not “mined” to discover hidden statistics for powerful selection making. Discovery of hidden styles and relationships frequently goes unexploited. Advanced records mining strategies can assist treatment this case. This study has advanced a prototype Intelligent Heart Disease Prediction System (IHDPS) using facts mining strategies, particularly, Decision Trees, Naïve Bayes and Neural Network. Results show that every approach has its particular power in figuring out the targets of the defined mining goals. Using scientific profiles together with age, sex, blood pressure and blood sugar it could expect the likelihood of patients getting a heart disease. It permits widespread information, e.g., Styles, relationships between clinical factors associated with coronary heart disorder, to be installed.

### **2. Securing Data Transmission from adversaries in Wsn using efficient key management techniques**

The sensor nodes that make up the wireless sensor networks have inherent resource limitations. These nodes are generally deployed in diverse, unattended locations for capturing real time data which they convert to digital format and transmit to CPUs or base stations for further analysis. Such transmissions take place in wireless form, easily susceptible to different types of attacks by adversaries. Adversaries attempt to disclose the data illegally or just disrupt the data transmission, affecting its integrity. Establishing security in such areas is not an easy task. Literatures throwing light upon the preventive measures have collectively put forward that there is no single escape solution that could be commonly applied to all type of attacks. Key management plays a prominent part in the upkeep of security in the wireless domain. The key management techniques focus upon the capability of nodes and the security demand imposed by any specific application. Key management can be implemented through public key management, pairwise key management, group based key management and dynamic key management techniques.



### **3. Improvement to data classification in heart diseases using Hybrid Optimization Techniques:**

The clinical area is growing at a speedy pace with new diseases are cropping up every day with need for invention of appropriate route of remedy. The coronary heart is a muscular organ that is the dimensions of clenched human fist and is liable for blood movement. Although, heart/cardiac sickness is the call given to sicknesses affecting the heart in preferred, there are numerous sicknesses which come beneath this name along with coronary artery sicknesses (cas), cardiomyopathy, cardiovascular sickness (cvs) and so on relying upon the circulation of blood throughout the body. To assist clinicians within the prognosis of heart ailment, the coronary heart disorder statistics prediction has been so designed to research scientific records with scientific knowledge. There can be an enhancement within the exceptional of clinical diagnostic selections for heart sickness.

### **4. Towards Deep Learning Models resistant to Adversarial Attacks**

Recent work has demonstrated that deep neural networks are vulnerable to adversarial examples—inputs that are almost indistinguishable from natural data and yet classified incorrectly by the network. In fact, some of the latest findings suggest that the existence of adversarial attacks may be an inherent weakness of deep learning models. To address this problem, we study the adversarial robustness of neural networks through the lens of robust optimization. This approach provides us with a broad and unifying view on much of the prior work on this topic. Its principled nature also enables us to identify methods for both training and attacking neural networks that are reliable and, in a certain sense, universal. In particular, they specify a concrete security guarantee that would protect against any adversary. These methods let us train networks with significantly improved resistance to a wide range of adversarial attacks. They also suggest the notion of security against a first-order adversary as a natural and broad security guarantee. We believe that robustness against such well-defined classes of adversaries is an important stepping stone towards fully resistant deep learning models.

### **5. Computer aided diagnostic model for heart disease prediction using machine learning techniques**

The advanced methods in computing and communication technologies have enabled the healthcare sector to gather and save regular patient records which assists to make medical decisions. The saved medical information can be investigated to make the needed medical decisions that might be forecast, analysis, image examination, and line of treatment. Presently, various ML techniques have been commonly employed to classify and predict diseases. In this research paper, they made a review to study the existing ML models for Heart disease prediction. Besides, a review of CDSS takes place along with the survey of Outlier Detection (OD) based heart disease prediction models. A detailed comparative analysis is also made to identify the characteristics of the reviewed prediction models.

## **6. An Empirical Study and Analysis of Heart Disease Prediction using Machine Learning Techniques**

The significant contribution of this dissertation is divided into two parts. First, an effective approach to earlier detection and classification of heart disease is described. Next, a Fourier transform based medical recommendation model is presented for the earlier diagnosis of heart diseases. The algorithms such as naive bayes, smoothing Laplace transform models are used for effective health data classification processes. The outcomes are analyzed regarding factors such as accuracy, sensitivity, precision, and specificity. It has been found from the analysis that the proposed system provides comparatively better accuracy and prediction measures than the existing techniques

## **7. An Optimized Feature Selection Based on Genetic Approach and Support Vector Machine for Heart Disease**

This Work presents a coronary illness expectation model. Among the new innovation Shrewd Gadget empowered medical services assumes a fundamental part. The clinical sensors utilized in medical care give immense volume of clinical information in constant way. The speed of information age in medical care is high, so the volume of information is likewise high. The capacity of anticipating the commonness of diabetes and hypertension in the Indian ladies utilizing the particular limits files are inspected. The limit for midsection conditions is  $\geq 35\%$  edges, 12% have more than WC edge in hypertension and 13% in instance of diabetic patients. The weight record cut point limit esteem is 25.02 kg/m<sup>2</sup> furthermore 34% individuals have more than BMI cut-point if there should arise an occurrence of hypertension and 25% have more than BMI cut point if there should be an occurrence of diabetes. If there should arise an

occurrence of midriff to stature proportion cut-off as it were 1% have missed the end in hypertension and 0% if there should arise an occurrence of diabetes, from these three boundaries have a similar capacity in foreseeing the hypertension and diabetes in Indian ladies. The wellbeing observing framework is proposed utilizing choice tree to view as most critical variable that causes the coronary illness. The Increase proportion-based choice tree is tried with various choice tree calculations while applying casting a ballot choice system to see as more precise and powerful technique. In the wake of applying casting a ballot component, the precision of various calculations increments. By utilizing information gain choice tree calculation, it further upgrades the precision in diagnosing the coronary illness. The choice tree can be utilized in the forecast of coronary illness. Coronary illness determination is viewed as a difficult issue which can offer a modernized gauge about the degree of coronary illness with the goal that beneficial activity can be made simple. Along these lines, coronary illness determination has expected enormous consideration around the world among the medical services climate. Enhancement calculations assumed a huge part in coronary illness determination with great effectiveness. The Hereditary Calculation (GA) for choosing the more critical elements to get coronary illness. The exploratory aftereffects of the GA-SVM are contrasted and the different existing element determination calculations, for example, Alleviation, CFS, Sifted subset, Information gain, Consistency subset, Chi-squared, one trait based, separated quality, Gain proportion, and Hereditary calculation.

## **8. Secure and Robust Machine Learning for Healthcare**

Recent years have witnessed widespread adoption of machine learning (ML)/deep learning (DL) techniques due to their superior performance for a variety of healthcare applications ranging from the prediction of cardiac arrest from one-dimensional heart signals to computer-aided diagnosis (CADx) using multi-dimensional medical images. Notwithstanding the impressive performance of ML/DL, there are still lingering doubts regarding the robustness of ML/DL in healthcare settings (which is traditionally considered quite challenging due to the myriad security and privacy issues involved), especially in light of recent results that have shown that ML/DL are vulnerable to adversarial attacks. In this paper, we present an overview of various application areas in healthcare that leverage such techniques from security and privacy point of view and present associated challenges. In addition, we present potential methods to ensure secure and privacy-preserving ML for healthcare applications.

## **9. Accurate and adversarially robust classification of medical images and ECG time-series with gradient-free trained sign activation neural networks**

Adversarial attacks in medical AI imaging systems can lead to misdiagnosis and insurance fraud as recently highlighted by Finlayson et. al. in science 2019. They can also be carried out on widely used ECG time-series data as shown in Han et. al. in Nature Medicine 2020. At the heart of adversarial attacks are imperceptible distortions that are visually and statistically undetectable but cause the machine learning model to misclassify data. Recent empirical studies have shown that a gradient-free trained sign activation neural network ensemble model requires a larger distortion than state of the art models. We apply them on medical data in this study as a potential solution to detect and deter adversarial attacks. We show on chest X-ray and histopathology images, and on two ECG datasets that this model requires a greater distortion to be fooled than full-precision, binary, and convolutional neural networks, and random forests. We show that adversaries targeting the gradient free sign networks are visually distinguishable from the original data and thus likely to be detected by human inspection. Since the sign network distortions are higher we expect an automated method could be developed to detect and deter attacks in advance. Our work here is a significant step towards safe and secure medical machine learning.

# Dataset Description

## **Demographic:**

- Sex: male or female(Nominal)
- Age: Age of the patient

## **Behavioural:**

- Current Smoker: whether or not the patient is a current smoker (Nominal)
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

## **Medical( history)**

- BP Meds: whether or not the patient was on blood pressure medication (Nominal)
- Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
- Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
- Diabetes: whether or not the patient had diabetes (Nominal)

## **Medical(current)**

- Tot Chol: total cholesterol level (Continuous)
- Sys BP: systolic blood pressure (Continuous)
- Dia BP: diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
- Glucose: glucose level (Continuous)

Predict variable (desired target)

- 10 year risk of coronary heart disease CHD (binary: "1", means "Yes", "0" means "No")

## Overview

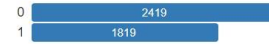
Overview	Alerts 25	Reproduction
Dataset statistics		Variable types
Number of variables	16	Categorical 8
Number of observations	4238	Numeric 8
Missing cells	0	
Missing cells (%)	0.0%	
Duplicate rows	0	
Duplicate rows (%)	0.0%	
Total size in memory	529.9 KiB	
Average record size in memory	128.0 B	

## Variables

male

Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	33.2 KiB

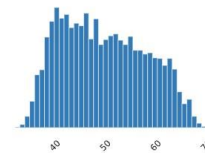


Toggle details

age

Real number (ᐃᓐ)

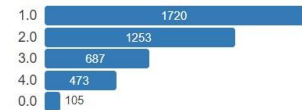
Distinct	39	Minimum	32
Distinct (%)	0.9%	Maximum	70
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	49.58494573	Memory size	33.2 KiB



education

Categorical

Distinct	5
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	33.2 KiB



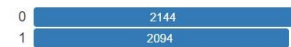
Toggle details

currentSmoker

Categorical

HIGH\_CORRELATION  
HIGH\_CORRELATION  
HIGH\_CORRELATION  
HIGH\_CORRELATION

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	33.2 KiB



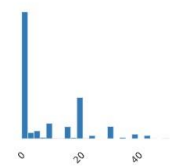
Toggle details

cigsPerDay

Real number (ᐃᓐ)

HIGH\_CORRELATION  
HIGH\_CORRELATION  
HIGH\_CORRELATION  
HIGH\_CORRELATION  
ZEROS

Distinct	33	Minimum	0
Distinct (%)	0.8%	Maximum	70
Missing	0	Zeros	2173
Missing (%)	0.0%	Zeros (%)	51.3%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	8.941481831	Memory size	33.2 KiB



Toggle details

BPMeds

Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	33.2 KiB



Toggle details

prevalentStroke  
Categorical

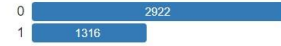
Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	33.2 KiB



Toggle details

prevalentHyp  
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	33.2 KiB



Toggle details

diabetes  
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	33.2 KiB



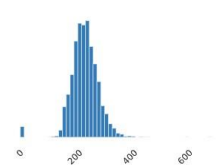
Toggle details

totChol  
Real number (ℝ<sub>≥0</sub>)

ZEROS

Distinct	249
Distinct (%)	5.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	233.92874

Minimum	0
Maximum	696
Zeros	50
Zeros (%)	1.2%
Negative	0
Negative (%)	0.0%
Memory size	33.2 KiB



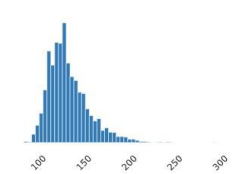
Toggle details

sysBP  
Real number (ℝ<sub>≥0</sub>)

HIGH\_CORRELATION  
HIGH\_CORRELATION  
HIGH\_CORRELATION  
HIGH\_CORRELATION

Distinct	234
Distinct (%)	5.5%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	132.3524068

Minimum	83.5
Maximum	295
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	33.2 KiB



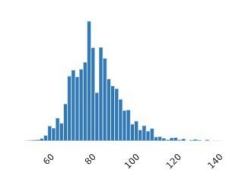
Toggle details

diaBP  
Real number (ℝ<sub>≥0</sub>)

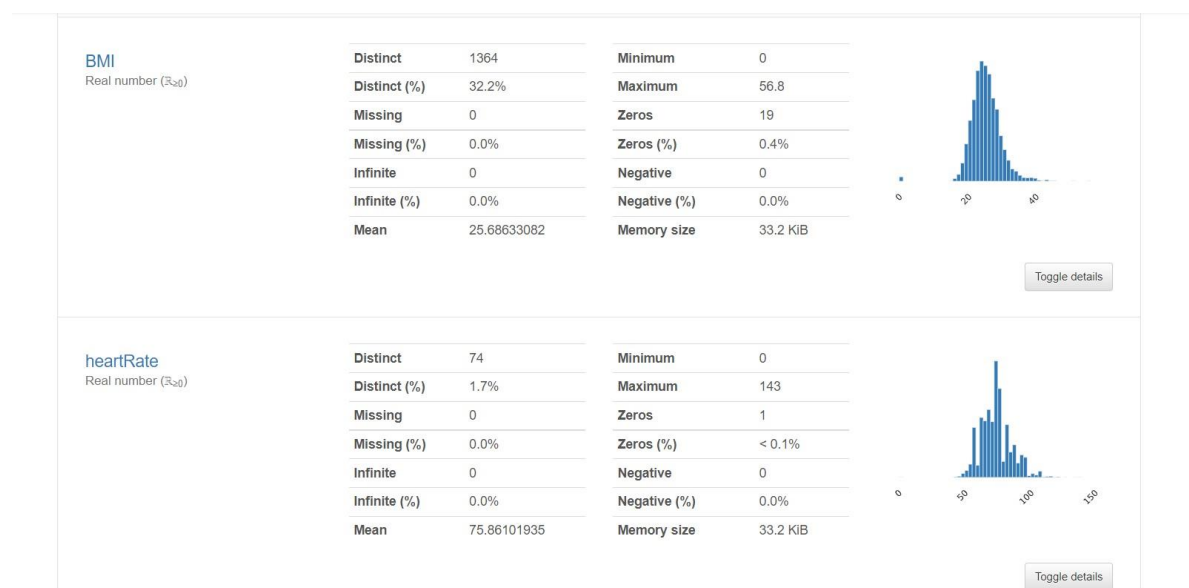
HIGH\_CORRELATION  
HIGH\_CORRELATION  
HIGH\_CORRELATION  
HIGH\_CORRELATION

Distinct	146
Distinct (%)	3.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	82.8934639

Minimum	48
Maximum	142.5
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	33.2 KiB

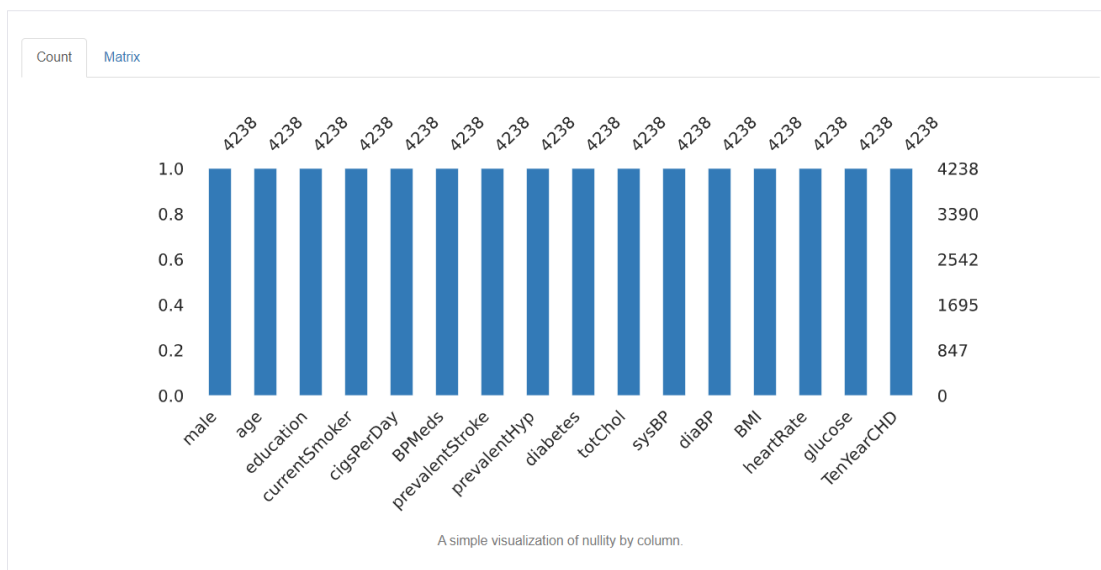


Toggle details





## Missing values



## Results

- All attributes selected after the elimination process show P-values lower than 5% and thereby suggesting significant role in the heart disease prediction.
- Men seem to be more susceptible to heart disease than women. Increase in age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease.
- Total cholesterol shows no significant change in the odds of CHD. This could be due to the presence of 'good cholesterol (HDL) in the total cholesterol reading. Glucose too causes a very negligible change in odds (0.2%)
- The model predicted with 0.88 accuracy. The model is more specific than sensitive. Overall model could be improved with more data.

### 1. Logistic Regression: -

```
confusion matrix
[[708  2]
 [130  8]]
```

Accuracy of Logistic Regression: 84.43396226415094

	precision	recall	f1-score	support
0	0.84	1.00	0.91	710
1	0.80	0.06	0.11	138
accuracy			0.84	848
macro avg	0.82	0.53	0.51	848
weighted avg	0.84	0.84	0.78	848

## 2. Naive Bayes: -

```
↳ confusion matrix  
[[663  47]  
 [106  32]]
```

Accuracy of Naive Bayes model: 81.95754716981132

	precision	recall	f1-score	support
0	0.86	0.93	0.90	710
1	0.41	0.23	0.29	138
accuracy			0.82	848
macro avg	0.63	0.58	0.60	848
weighted avg	0.79	0.82	0.80	848

## 3. Random Forest Classifier: -

```
↳ confusion matrix  
[[710   0]  
 [135   3]]
```

Accuracy of Random Forest: 84.08018867924528

	precision	recall	f1-score	support
0	0.84	1.00	0.91	710
1	1.00	0.02	0.04	138
accuracy			0.84	848
macro avg	0.92	0.51	0.48	848
weighted avg	0.87	0.84	0.77	848

## 4. K-Neighbours Classifier: -

```
↳ confusion matrix  
[[709   1]  
 [136   2]]
```

Accuracy of K-NeighborsClassifier: 83.84433962264151

	precision	recall	f1-score	support
0	0.84	1.00	0.91	710
1	0.67	0.01	0.03	138
accuracy			0.84	848
macro avg	0.75	0.51	0.47	848
weighted avg	0.81	0.84	0.77	848

### 5. Decision Tree Classifier: -

```
confussion matrix
[[696  14]
 [132   6]]
```

Accuracy of DecisionTreeClassifier: 82.78301886792453

	precision	recall	f1-score	support
0	0.84	0.98	0.91	710
1	0.30	0.04	0.08	138
accuracy			0.83	848
macro avg	0.57	0.51	0.49	848
weighted avg	0.75	0.83	0.77	848

### 6. Support Vector Classifier: -

```
confussion matrix
[[705   5]
 [129   9]]
```

Accuracy of Support Vector Classifier: 84.19811320754717

	precision	recall	f1-score	support
0	0.85	0.99	0.91	710
1	0.64	0.07	0.12	138
accuracy			0.84	848
macro avg	0.74	0.53	0.52	848
weighted avg	0.81	0.84	0.78	848

## **7. Multi-Layer Perceptron: -**

```

              precision    recall  f1-score   support

     0         0.85         0.99         0.91         710
     1         0.61         0.08         0.14         138

 accuracy          0.84         848
 macro avg         0.73         0.53         0.53         848
 weighted avg      0.81         0.84         0.79         848

confussion matrix

0.8419811320754716
```

### **Accuracy comparison of algorithms**

After performing the machine learning approach for training and testing we find that accuracy of the Logistic Regression is better compared to other algorithms. Accuracy is calculated with the support of the confusion matrix of each algorithm, here the number count of TP, TN, FP, FN is given and using the equation of accuracy, value has been calculated and it is concluded that extreme gradient boosting is best with 84.5% accuracy and the comparison is shown below

Algorithm	Accuracy
Logistic Regression	84.43
Naïve Bayes	81.95
Random Forest Classifier	84.08
KNN	83.84
Decision Tree	82.78
Support Vector Classifier	84.19
Multi-Layer Perceptron	84.19

## Accuracy Comparison with Other's ML Model

We have also made a comparison of the accuracy of algorithms used in our model with the algorithms of our reference model. Out of which we implemented most of the algorithm and some new one's also and we have improved the accuracy of all the algorithm's we have used in comparison to the reference model and the comparison is shown below

Algorithm	Accuracy of our Model	Accuracy of Different Model
Logistic Regression	84.43	79.12
Naïve Bayes	81.95	76.92
Random Forest	84.08	79.12
KNN	83.84	---
Decision Tree	82.78	75.82
Support Vector Classifier	84.19	80.21
Multilayer Perceptron	84.19	---
Ada Boost	----	73.6
XG Boost	----	81.3

## For Reference

```
Accuracy of svm: 0.8021978021978022
Accuracy of naive bayes: 0.7692307692307693
Accuracy of logistic regression: 0.7912087912087912
Accuracy of decision tree: 0.7582417582417582
Accuracy of random forest: 0.7912087912087912
```

```
Majority Voting accuracy score: 0.7912087912087912
Weighted Average accuracy score: 0.8131868131868132
Bagging_accuracy score: 0.8021978021978022
Ada_boost_accuracy score: 0.7362637362637363
Gradient_boosting_accuracy score: 0.8131868131868132
```

## **APPENDIX**

### ➤ **System Configuration**

#### **Hardware Requirements**

**Processor:** Any Update Processor

**Ram:** Min 4GB

**Hard Disk:** Min 100GB

#### **Software Requirements**

Operating System: Windows family

Technology: Python3.7

IDE: Google Colab

### ➤ **Implementation / Code**

```
# Importing modules and packages
import six
import sys
sys.modules['sklearn.externals.six'] = six
import pandas as pd
import numpy as np
import xgboost as xgb
from pandas_profiling import ProfileReport
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from collections import Counter
import pandas_profiling as pp

!pip install 'neptune-contrib[monitoring]>=0.24.9'
```

## # Preprocessing

```
from sklearn.preprocessing import StandardScaler
# data splitting
from sklearn.model_selection import train_test_split
# data modeling
from sklearn.metrics import
confusion_matrix, accuracy_score, roc_curve, classification_report
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from xgboost import XGBClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.metrics import make_scorer, accuracy_score
from sklearn.model_selection import GridSearchCV
from sklearn.neural_network import MLPClassifier
#ensembling
from mlxtend.classifier import StackingCVClassifier
from neptunecontrib.monitoring.xgboost_monitor import neptune_callback
```

```
! pip install https://github.com/pandas-profiling/pandas-
profiling/archive/master.zip
```

## # Exploratory Data Analysis

```
from google.colab import files
uploaded = files.upload()

df = pd.read_csv("framingham.csv")
df.head()

df = df.fillna(0)
df.info()

pp.ProfileReport(df)

y = df["TenYearCHD"]
X = df.drop("TenYearCHD", axis=1)
```



```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,  
random_state = 0)
```

```
print(y_test.unique())  
Counter(y_train)
```

```
scaler = StandardScaler()  
X_train = scaler.fit_transform(X_train)  
X_test = scaler.transform(X_test)
```

## # Logistic Regression

```
m1 = 'Logistic Regression'  
lr = LogisticRegression()  
model = lr.fit(X_train, y_train)  
lr_predict = lr.predict(X_test)  
lr_conf_matrix = confusion_matrix(y_test, lr_predict)  
lr_acc_score = accuracy_score(y_test, lr_predict)  
print("confusion matrix")  
print(lr_conf_matrix)  
print("\n")  
print("Accuracy of Logistic Regression:",lr_acc_score*100,'\n')  
print(classification_report(y_test,lr_predict))
```

## # Naïve Bayes

```
m2 = 'Naive Bayes'  
nb = GaussianNB()  
nb.fit(X_train,y_train)  
nbpred = nb.predict(X_test)  
nb_conf_matrix = confusion_matrix(y_test, nbpred)  
nb_acc_score = accuracy_score(y_test, nbpred)  
print("confusion matrix")  
print(nb_conf_matrix)  
print("\n")  
print("Accuracy of Naïve Bayes model:",nb_acc_score*100,'\n')  
print(classification_report(y_test,nbpred))
```

## # Random Forest Classifier

```
m3 = 'Random Forest Classifier'  
rf = RandomForestClassifier(n_estimators=20, random_state=12,max_depth=5)  
rf.fit(X_train,y_train)
```

```

rf_predicted = rf.predict(X_test)
rf_conf_matrix = confusion_matrix(y_test, rf_predicted)
rf_acc_score = accuracy_score(y_test, rf_predicted)
print("confussion matrix")
print(rf_conf_matrix)
print("\n")
print("Accuracy of Random Forest:",rf_acc_score*100,'\n')
print(classification_report(y_test,rf_predicted))

```

## # KNN (K Neighbours Classifier)

```

m5 = 'K-NeighborsClassifier'
knn = KNeighborsClassifier(n_neighbors=10)
knn.fit(X_train, y_train)
knn_predicted = knn.predict(X_test)
knn_conf_matrix = confusion_matrix(y_test, knn_predicted)
knn_acc_score = accuracy_score(y_test, knn_predicted)
print("confussion matrix")
print(knn_conf_matrix)
print("\n")
print("Accuracy of K-NeighborsClassifier:",knn_acc_score*100,'\n')
print(classification_report(y_test,knn_predicted))

```

## # Decision Tree Classifier

```

m6 = 'DecisionTreeClassifier'
dt = DecisionTreeClassifier(criterion = 'entropy',random_state=0,max_depth =
6)
dt.fit(X_train, y_train)
dt_predicted = dt.predict(X_test)
dt_conf_matrix = confusion_matrix(y_test, dt_predicted)
dt_acc_score = accuracy_score(y_test, dt_predicted)
print("confussion matrix")
print(dt_conf_matrix)
print("\n")
print("Accuracy of DecisionTreeClassifier:",dt_acc_score*100,'\n')
print(classification_report(y_test,dt_predicted))

```

## # Support Vector Classifier

```

m7 = 'Support Vector Classifier' svc
= SVC(kernel='rbf', C=2)
svc.fit(X_train, y_train)
svc_predicted = svc.predict(X_test)
svc_conf_matrix = confusion_matrix(y_test, svc_predicted)

```

```

svc_acc_score = accuracy_score(y_test, svc_predicted)
print("confusion matrix")
print(svc_conf_matrix)
print("\n")
print("Accuracy of Support Vector Classifier:",svc_acc_score*100,'\n')
print(classification_report(y_test,svc_predicted))

# ANN Multilayer Perceptron Classifier
ann_clf = MLPClassifier()

#Parameters
parameters = {'solver': ['lbfgs'],
              'alpha':[1e-4],
              'hidden_layer_sizes':(9,14,14,2),    # 9 input, 14-14 neuron in 2
layers,1 output layer
              'random_state': [1]}

# Type of scoring to compare parameter combos
acc_scorer = make_scorer(accuracy_score)

# Run grid search
grid_obj = GridSearchCV(ann_clf, parameters, scoring=acc_scorer)
grid_obj = grid_obj.fit(X_train, y_train)

# Pick the best combination of parameters
ann_clf = grid_obj.best_estimator_

# Fit the best algorithm to the data
ann_clf.fit(X_train, y_train)

MLPClassifier(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9,
              beta_2=0.999, early_stopping=False, epsilon=1e-08,
              hidden_layer_sizes=14, learning_rate='constant',
              learning_rate_init=0.001, max_iter=200, momentum=0.9,
              nesterovs_momentum=True, power_t=0.5, random_state=1, shuffle=True,
              solver='lbfgs', tol=0.0001, validation_fraction=0.1, verbose=False,
              warm_start=False)

y_pred_ann = ann_clf.predict(X_test)

from sklearn.metrics import confusion_matrix
cm_ann = confusion_matrix(y_test, y_pred_ann)
cm_ann

```

```
ann_result = accuracy_score(y_test,y_pred_ann)
ann_result
```

```
recall_ann = cm_ann[0][0]/(cm_ann[0][0] + cm_ann[0][1])
precision_ann = cm_ann[0][0]/(cm_ann[0][0]+cm_ann[1][1])
recall_ann,precision_ann
```

## **CONCLUSION AND FUTURE WORK**

Heart diseases are a major killer in India and throughout the world, application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease 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. The number of people facing heart diseases is on a raise each year. This prompts for its early diagnosis and treatment. The utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. In this paper, the seven different machine learning algorithms used to measure the performance are Logistic Regression, SVM, Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, and MLP applied on the dataset.

All the seven machine learning methods accuracies are compared based on which one prediction model is generated. Hence, the aim is to use various evaluation metrics like confusion matrix, accuracy, precision, recall, and f1-score which predicts the disease efficiently. Comparing all seven the extreme gradient boosting classifier gives the highest accuracy of 84.5%.

## **References: -**

Logistic regression To predict heart disease | Kaggle

<https://www.programmingempire.com>

GeeksforGeeks | A computer science portal for geeks

Intelligent Heart Disease Prediction System Using Data Mining Techniques

Securing Data Transmission from adversaries in Wsn using efficient key management techniques

Improvement to data classification in heart diseases using Hybrid Optimization Techniques

Towards Deep Learning Models resistant to Adversarial Attacks

Computer aided diagnostic model for heart disease prediction using machine learning techniques

An Empirical Study and Analysis of Heart Disease Prediction using Machine Learning Techniques

An Optimized Feature Selection Based on Genetic Approach and Support Vector Machine for Heart Disease