# CARDIAC DISEASE PREDICTION

## B.Tech. Final Year Report

*Submitted by*

**Ankit Kumar**

**1613101140**

**Under supervision of**

**Mr. Ashutosh Upadhyay**

**ASSISTANT PROFESSOR**

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

**GALGOTIAS UNIVERSITY**

**GREATER NOIDA, UTTAR PRADESH, INDIA**

# Tables of contents

# 1.ABSTRACT

The health care industry produces a huge amount of data. This data is not always made use to the full extent and is often underutilized. Using this huge amount of data, a disease can be detected, predicted or even cured. A huge threat to human kind is caused by diseases like heart disease, cancer, tumour and Alzheimer’s disease.

In today’s era deaths due to heart disease has become a major issue approximately one person dies per minute due to heart disease. This is considering both male and female category and this ratio may vary according to the region also this ratio is considered for the people of age group 25-69. This does not indicate that the people with other age group will not be affected by heart diseases. This problem may start in early age group also and predict the cause and disease is a major challenge nowadays.

In this paper, we try to concentrate on heart disease prediction. Using machine learning techniques, the heart disease can be predicted. The medical data such as Blood pressure,hypertension, diabetes, cigarette smoked per day and so on is taken as input and then these features are modelled for prediction. This model can then be used to predict future medical data.

**The algorithms like K- nearest neighbour, Naïve Bayes, and decision tree**

**are used. The accuracy of the model using each of the algorithms is calculated. Then the one with**

**a good accuracy is taken as the model for predicting the heart disease**.

Keywords— Classification, Heart Disease, Decision Tree.

# 2. INTRODUCTION

## I.OVERALL DISCRIPTION

In our day to day life, people are undergoing a routine and busy schedule which leads to stress and anxiety. In addition to this, the percentage of people who are obese and addicted to cigarette goes up drastically.  This leads to diseases like heart disease, cancer, etc. The challenge behind these diseases is its prediction. Each person has different values of pulse rate and blood pressure. But medically proven, the pulse rate must be 60 to 100 beats per minute and the blood pressure must be in the range of 120/80 to 140/90. Heart disease is one of the major cause of death in the world. [11] The number of people affected by heart disease increases irrespective of age in both men and women. But other factors like gender, diabetes, BMI also contribute to this disease. In this paper, we have tried prediction and analysis of heart disease by considering the parameters like age, gender, blood pressure, heart rate, diabetes and so on. Since numerous factors are involved in heart disease, the prediction of this disease is challenging. Some of major symptoms of heart attack are:

- Chest tightness.
- Shortness of breath.
- Nausea, Indigestion, Heartburn, or stomach pain.
- Sweating and Fatigue.
- Pressure in the upper back Pain that spreads to the arm.

The following are the type of heart disease: Heart means “cardio”. Hence all heart diseases concern to category of cardiovascular diseases. The different kinds of heart disease are:

- Coronary heart diseases.
- Angina pectoris
- Congestive heart failure.
- Cardiomyopathy
- Congenital heart diseases.

Coronary heart disease or coronary artery disease is the narrowing of the coronary arteries. The coronary arteries supply oxygen and blood to the heart. It causes a large number of people to become ill or to face death. It is one of the popular type of heart disease. High blood glucose from diabetes can damage blood vessels and nerves that control heart and blood vessels. If a person has diabetes for a longer time, there are high chances for that person to have heart disease in future. With diabetes, there are other reasons which contribute to heart disease. They are smoking which raises the risk of developing heart disease, high blood pressure makes the heart work harder to pump blood and it can strain heart and damage blood vessels, abnormal cholesterol levels also contribute to heart disease and obesity. Also, family history of heart disease can be a cause of having heart disease. But this history is not considered in this paper for prediction of heart disease. The other risk factors include age, gender, stress and unhealthy diet. Chance of having a heart disease increases when a person is getting older. Men have a greater risk of heart disease. However, women also have the same risk after menopause. Leading a stressed life can also damage the arteries and increase the chance of coronary heart disease.

So, in this paper based on the factors mentioned above we try to predict the risk of heart disease. A large amount of work has been done related to heart prediction system by using various techniques and algorithms by many authors. These techniques may be based on deep-learning, machinelearning, data mining and so on. The aim of all those papers is to achieve better accuracy and to make the system more efficient so that it can predict the chances of heart attack.

## II. PURPOSE

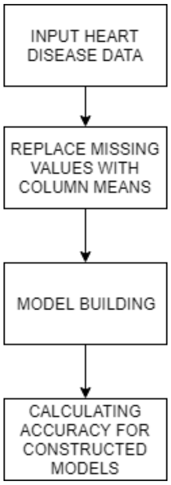
The main purpose of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set Heart disease prediction system aims to exploit Machine learning on medical data set to assit the prediction in the heart diseases.

## III.MOTIVATION AND SCOPE

Here the scope of the project is that integration of clinical decision support with computer based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

# 3.PROPOSED SYSTEM

In this paper, comparison of various machine learning methods is done for predicting the 10 year risk of coronary heart disease of the patients from their medical data. The following is the flowchart for proposed methodology:


INPUT HEART DISEASE DATA
REPLACE MISSING VALUES WITH COLUMN MEANS
MODEL BUILDING
CALCULATING ACCURACY FOR CONSTRUCTED MODELS

The heart disease data set is taken as input. It is then pre-processed by replacing non-available values with column means.

Four different methods were used in this paper. The different methods used are depicted in the output is the accuracy metrics of the machine learning models. The model can then be used in prediction

**K-Nearest Neighbours (KNN)**

KNN is a non-parametric machine learning algorithm. The KNN algorithm is a supervised learning method. This means that all the data is labelled and the algorithm learns to predict the output from the input data. It performs well even if the training data is large and contains noisy values.

The data is divided into training and test sets. The train set is used for model building and training.A k- value is decided which is often the square root of the number of observations. Now the test data is predicted on the model built. There are different distance measures. For continuous

variables, Euclidean distance, Manhattan distance and Minkowski distance measures can be used.However, the commonly used measure is Euclidean distance. The formula for Euclidean distance is as follows:

$$d = \sqrt{\sum^{k} (x_i - y_i)^2}$$

i =1

### Naive Bayes algorithm (NB)

This is a classification algorithm which is used when the dimensionality of the input is very high. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It is based on Bayes theorem. The Bayes theorem is as follows:

P(Y/X) = P(X/Y) P(X)

This calculates the probability of Y given X where X is the prior event and Y is the dependence event. It needs less training data. It can be used for binary classification problems and is very simple.

### Decision trees

Decision trees is one of the ways to display an algorithm. It is a classic machine learning algorithm. In cardiac disease, there are several factors such as cigarette, BP, Hypertension, age etc. The challenge of the decision tree lies in the selection of the root node. This factor used in root node must clearly classify the data. We make use of age as the root node.

The decision tree is easy to interpret. They are non-parametric and they implicitly do feature selection.

### Tools

Hardware used:-

| | |
|---|---|
| Monitor | : 15” color monitor |
| RAM | : 4GB |
| Hard Disk Space | : 500GB |

| | | |
|---|---|---|
| Processor | : | intel core i3 |

Software used:-

| | |
|---|---|
| Operating System | : Windows 10 |
| IDLE | : Python 3.x, PyCharm |
| Front end | : Anakonda , Jupiter notebook |
| Back end | : Python |
| Database | : EXCEL SHEET |

## TECHNOLOGY

**Python** is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms,including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library. Python was conceived in the late 1980s as a successor to the ABC language. Python 2.0, released 2000, introduced features like list comprehensions and a garbage collection system capable of collecting reference cycles. Python 3.0, released 2008, was a major revision of the language that is not completely backward-compatible, and much Python 2 code does not run unmodified on Python 3. Due to concern about the amount of code written for Python 2, support for Python 2.7 (the last release in the 2.x series) was extended to 2020. Language developer Guido van Rossum shouldered sole responsibility for the project until July 2018 but now shares his leadership as a member of a five-person steering council.

Python interpreters are available for many operating systems. A global community of programmers develops and maintains CPython, an open source reference implementation. A non-profit organization, the Python Software Foundation, manages and directs resources for Python and CPython development.

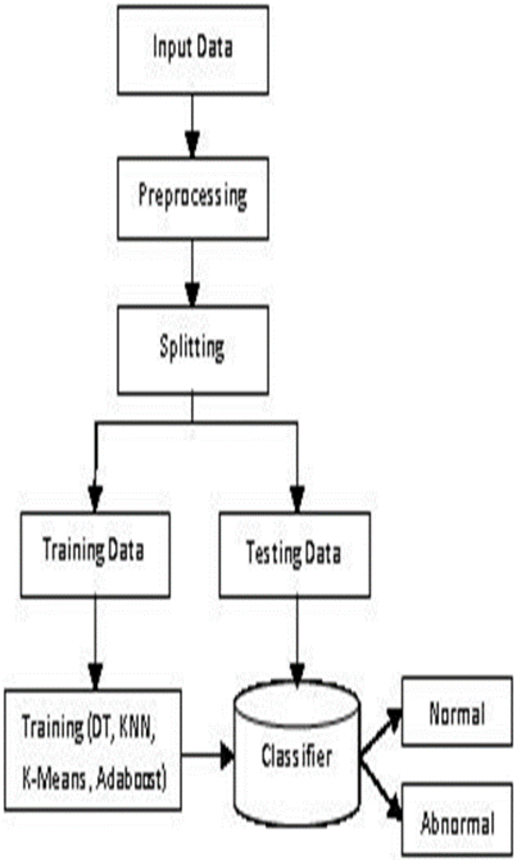
# 4.Existing System

Cardiac disease can be managed effectively with a combination of lifestyle changes, medicine and in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive.

The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

The healthcare environment is still „information rich‟ but „knowledge poor‟ Their is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in the data for African.

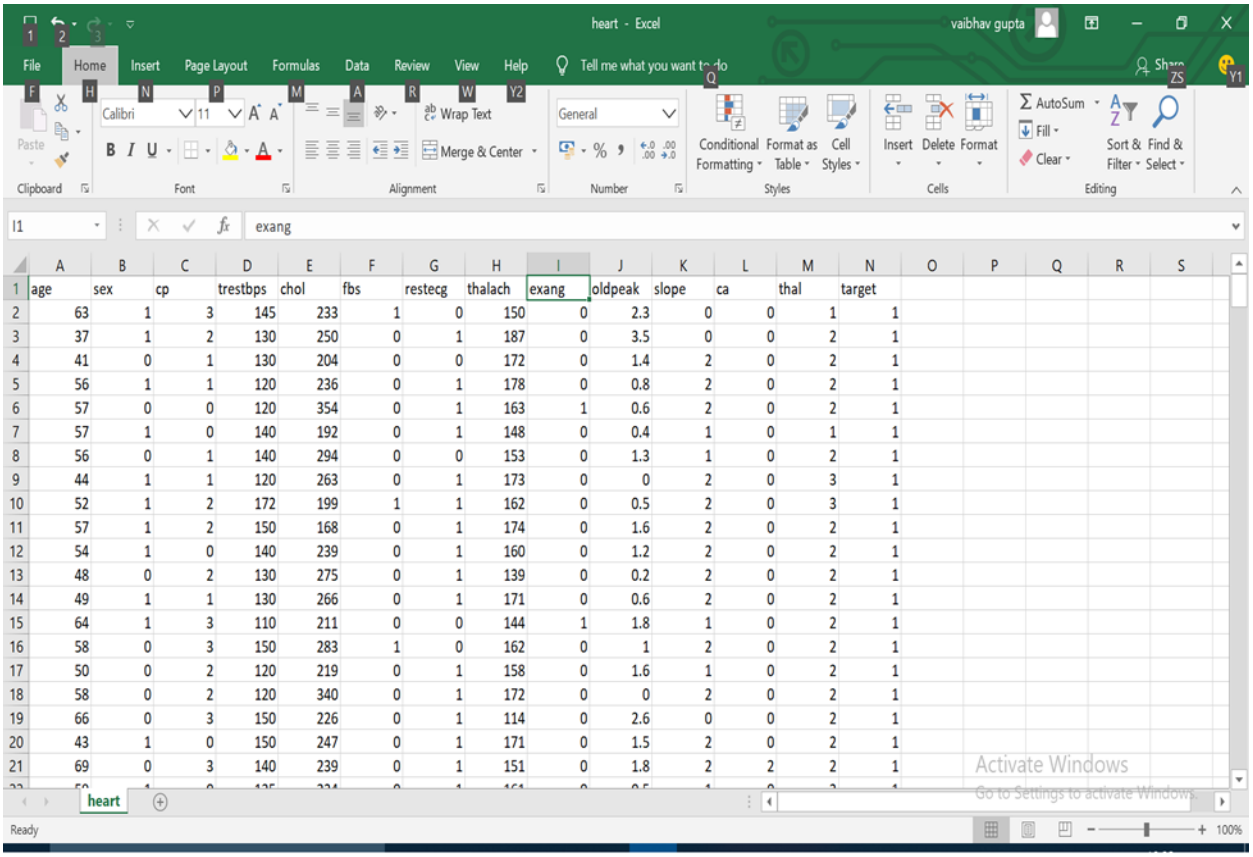
# 5.IMPLEMENTATION



**Data source:**

The dataset used is Framingham taken from Kaggle. There were various attributes as follows:

– age, sex, cholesterol level, BP etc.



I1 exang

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
| 2 | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.3 | 0 | 0 | 1 | 1 |
| 3 | 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 | 0 | 3.5 | 0 | 0 | 2 | 1 |
| 4 | 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 | 0 | 1.4 | 2 | 0 | 2 | 1 |
| 5 | 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 | 0 | 0.8 | 2 | 0 | 2 | 1 |
| 6 | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 | 0 | 2 | 1 |
| 7 | 57 | 1 | 0 | 140 | 192 | 0 | 1 | 148 | 0 | 0.4 | 1 | 0 | 1 | 1 |
| 8 | 56 | 0 | 1 | 140 | 294 | 0 | 0 | 153 | 0 | 1.3 | 1 | 0 | 2 | 1 |
| 9 | 44 | 1 | 1 | 120 | 263 | 0 | 1 | 173 | 0 | 0 | 2 | 0 | 3 | 1 |
| 10 | 52 | 1 | 2 | 172 | 199 | 1 | 1 | 162 | 0 | 0.5 | 2 | 0 | 3 | 1 |
| 11 | 57 | 1 | 2 | 150 | 168 | 0 | 1 | 174 | 0 | 1.6 | 2 | 0 | 2 | 1 |
| 12 | 54 | 1 | 0 | 140 | 239 | 0 | 1 | 160 | 0 | 1.2 | 2 | 0 | 2 | 1 |
| 13 | 48 | 0 | 2 | 130 | 275 | 0 | 1 | 139 | 0 | 0.2 | 2 | 0 | 2 | 1 |
| 14 | 49 | 1 | 1 | 130 | 266 | 0 | 1 | 171 | 0 | 0.6 | 2 | 0 | 2 | 1 |
| 15 | 64 | 1 | 3 | 110 | 211 | 0 | 0 | 144 | 1 | 1.8 | 1 | 0 | 2 | 1 |
| 16 | 58 | 0 | 3 | 150 | 283 | 1 | 0 | 162 | 0 | 1 | 2 | 0 | 2 | 1 |
| 17 | 50 | 0 | 2 | 120 | 219 | 0 | 1 | 158 | 0 | 1.6 | 1 | 0 | 2 | 1 |
| 18 | 58 | 0 | 2 | 120 | 340 | 0 | 1 | 172 | 0 | 0 | 2 | 0 | 2 | 1 |
| 19 | 66 | 0 | 3 | 150 | 226 | 0 | 1 | 114 | 0 | 2.6 | 0 | 0 | 2 | 1 |
| 20 | 43 | 1 | 0 | 150 | 247 | 0 | 1 | 171 | 0 | 1.5 | 2 | 0 | 2 | 1 |
| 21 | 69 | 0 | 3 | 140 | 239 | 0 | 1 | 151 | 0 | 1.8 | 2 | 2 | 2 | 1 |

heart

# 6.RESULT

The machine learning models is evaluated using the AUC-ROC metric. This can be used to understand the model performance.

The ROC curve is the Receiver Operating Characteristic curve. The AUC is the area under the ROC curve. If the AUC scoreis high, the model performance is high and vice versa. The ROC curve of the machine learning algorithms. The comparison of AUC score of the various algorithms is as follows:

| Algorithm | AUC score |
|---|---|
| NB | 0.68 |
| KNN | 0.56 |
| Decision tree | 0.53 |

The accuracy of the algorithms is calculated. The accuracy results are tabulated as follows:

| Method | Accuracy |
|---|---|
| KNN | 67.21% |
| NB | 85.25% |
| Decision tree | 81.97% |

**The accuracy of NAÏVE BAYES algorithm is good when compared to other algorithms.**

## Algorithm for Naïve Bayes:

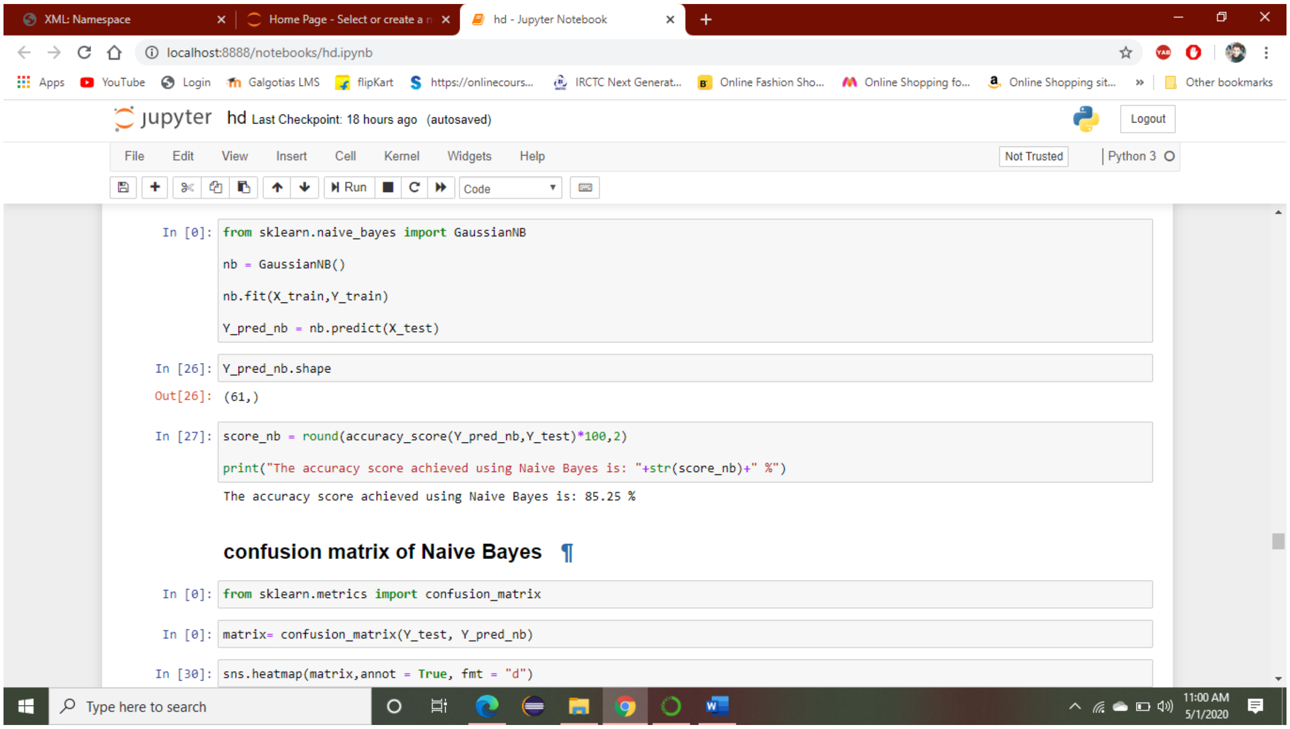
Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A/B) = [P(B/A)\ P(A)]/P(B)$$

where A and B are events and P(B) ? 0.

- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as **evidence**.
- P(A) is the **priori** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance (here, it is event B).
- P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.



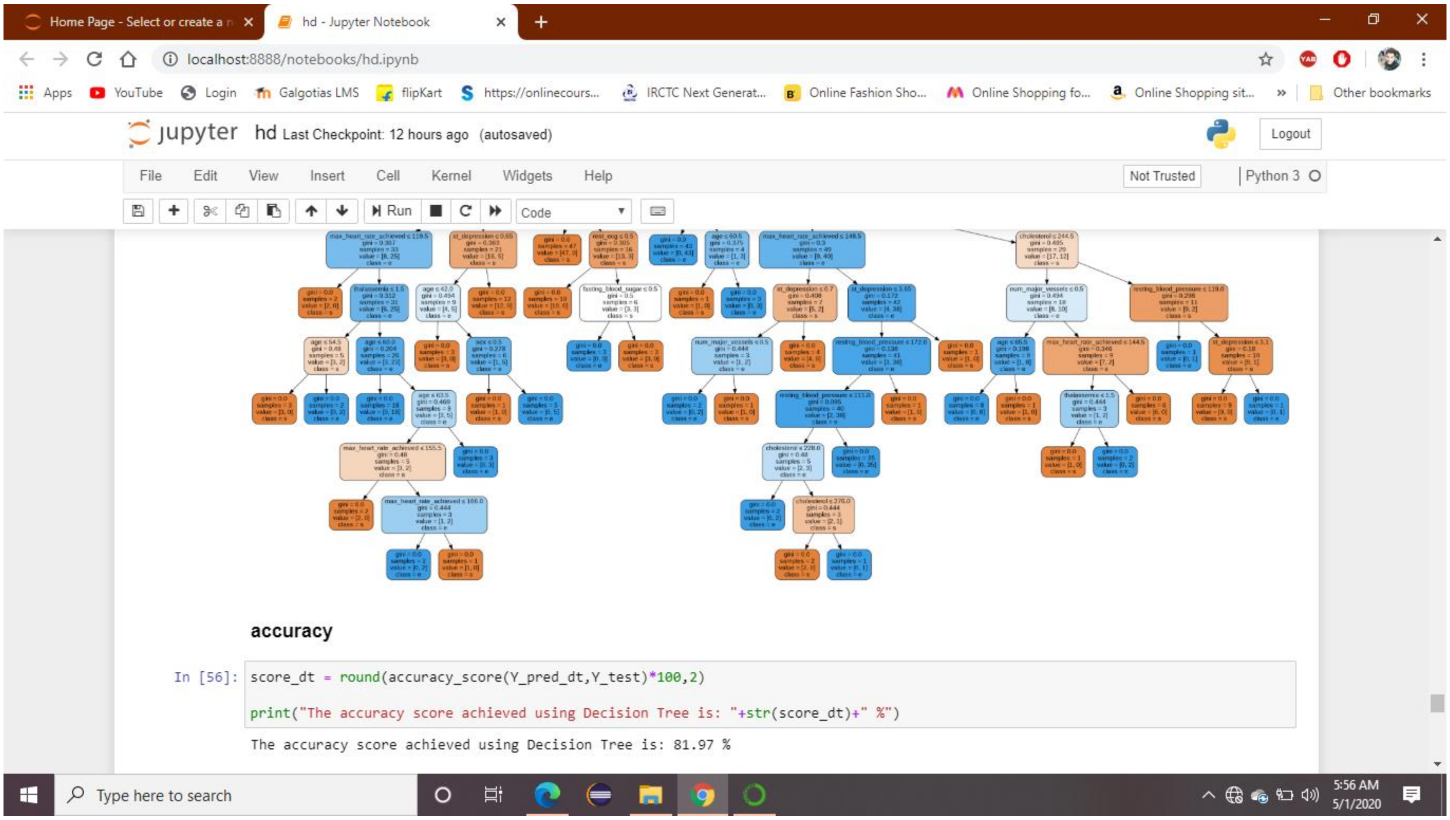
```
In [0]: from sklearn.naive_bayes import GaussianNB

        nb = GaussianNB()

        nb.fit(X_train,Y_train)

        Y_pred_nb = nb.predict(X_test)

In [26]: Y_pred_nb.shape

Out[26]: (61,)

In [27]: score_nb = round(accuracy_score(Y_pred_nb,Y_test)*100,2)

         print("The accuracy score achieved using Naive Bayes is: "+str(score_nb)+" %")

         The accuracy score achieved using Naive Bayes is: 85.25 %
```

**confusion matrix of Naive Bayes ¶**

```
In [0]: from sklearn.metrics import confusion_matrix

In [0]: matrix= confusion_matrix(Y_test, Y_pred_nb)

In [30]: sns.heatmap(matrix,annot = True, fmt = "d")
```

Code snippet of the Naïve Bayes algo we have imported the GaussianNB file

# 7.OUTPUT

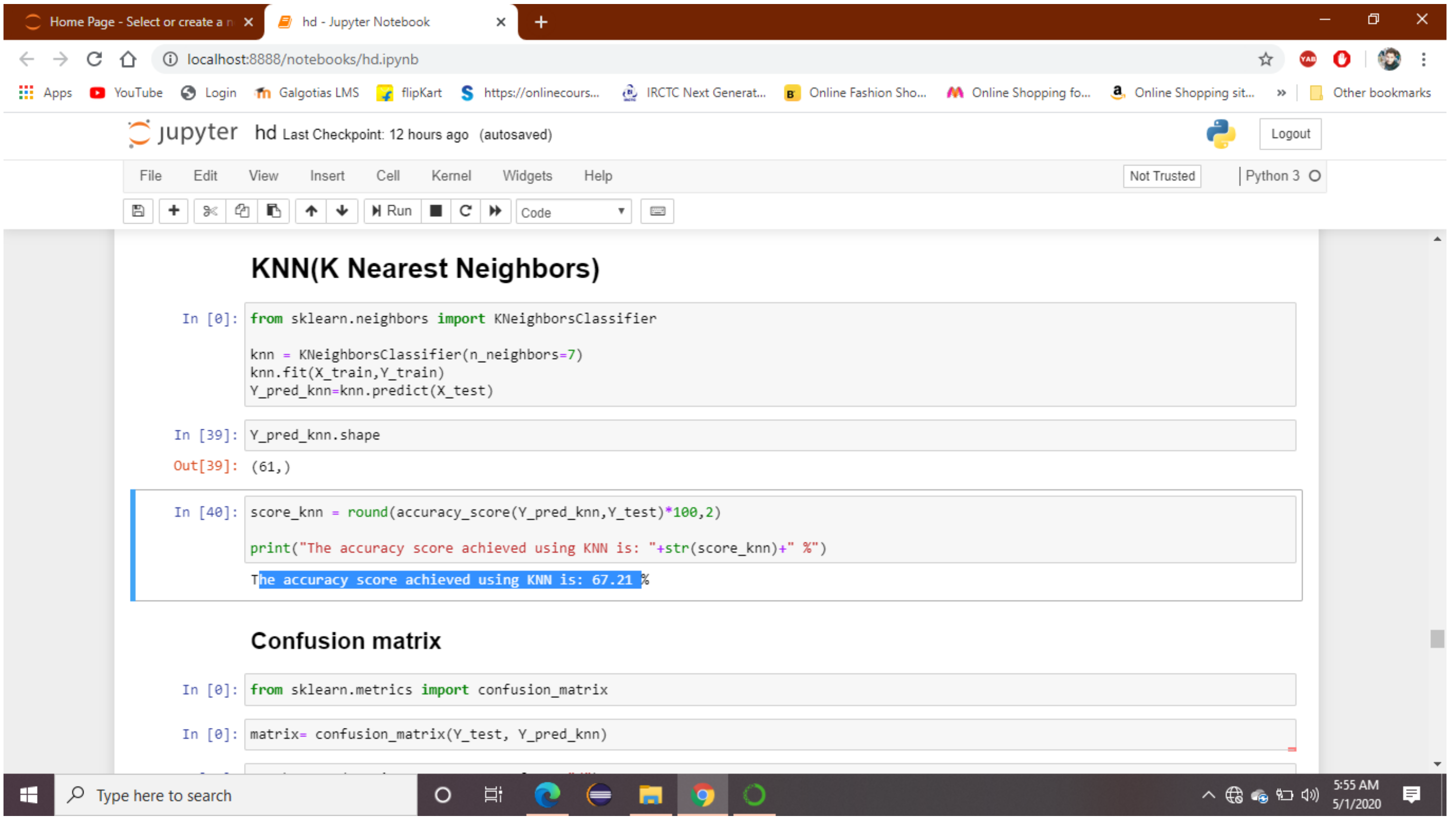


accuracy

```
In [56]: score_dt = round(accuracy_score(Y_pred_dt,Y_test)*100,2)

         print("The accuracy score achieved using Decision Tree is: "+str(score_dt)+" %")
```

```
The accuracy score achieved using Decision Tree is: 81.97 %
```

Output for Decision tree



## KNN(K Nearest Neighbors)

```
In [0]: from sklearn.neighbors import KNeighborsClassifier

        knn = KNeighborsClassifier(n_neighbors=7)
        knn.fit(X_train,Y_train)
        Y_pred_knn=knn.predict(X_test)
```

```
In [39]: Y_pred_knn.shape

Out[39]: (61,)
```

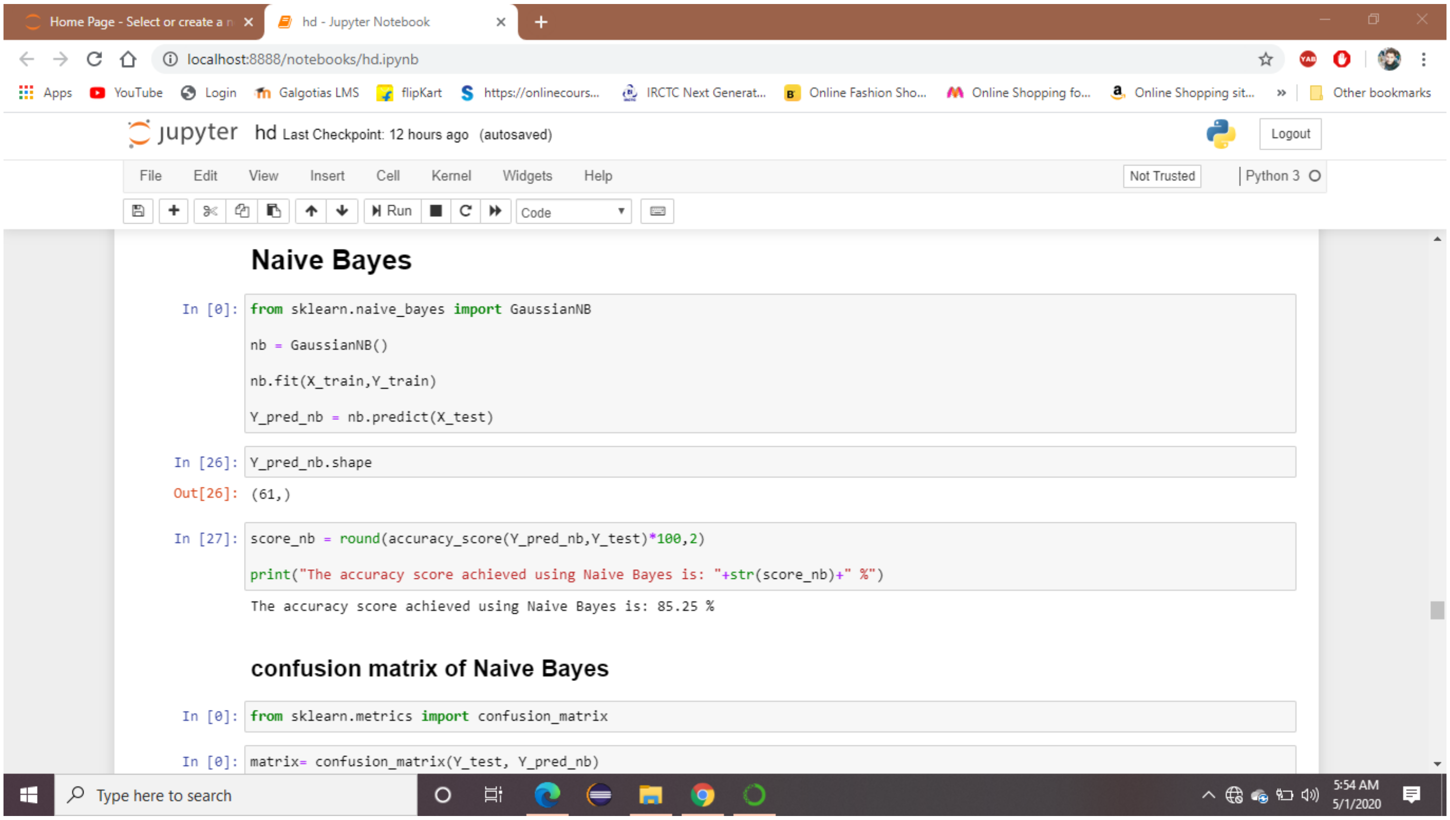
```
In [40]: score_knn = round(accuracy_score(Y_pred_knn,Y_test)*100,2)

         print("The accuracy score achieved using KNN is: "+str(score_knn)+" %")

         The accuracy score achieved using KNN is: 67.21 %
```

### Confusion matrix

```
In [0]: from sklearn.metrics import confusion_matrix
```

```
In [0]: matrix= confusion_matrix(Y_test, Y_pred_knn)
```

Output for KNN



### Naive Bayes

```
In [0]: from sklearn.naive_bayes import GaussianNB

        nb = GaussianNB()

        nb.fit(X_train,Y_train)

        Y_pred_nb = nb.predict(X_test)

In [26]: Y_pred_nb.shape
Out[26]: (61,)

In [27]: score_nb = round(accuracy_score(Y_pred_nb,Y_test)*100,2)

         print("The accuracy score achieved using Naive Bayes is: "+str(score_nb)+" %")

         The accuracy score achieved using Naive Bayes is: 85.25 %
```

### confusion matrix of Naive Bayes

```
In [0]: from sklearn.metrics import confusion_matrix

In [0]: matrix= confusion_matrix(Y_test, Y_pred_nb)
```

**Output for Naïve Bayes**

**In this project, We have used Machine Learning to predict whether a person is suffering from a heart disease or not. After importing the data, we have analysed it using plots.**

**Then, generated categorical features and scaled other features. Then applied Machine Learning algorithms: K Nearest Neighbors Classifier, Naive Bayes, Decision Tree. In the end, Naïve Bayes achieved the highest score of 85.25%.**

# 7.CONCLUSION

This paper discusses the various machine learning algorithms such as support vector machine, Naïve Bayes, decision tree and k- nearest neighbour which were applied to the data set. It utilizes the data such as blood pressure, cholesterol, diabetes and then tries to predict the possible coronary heart disease patient in next 10 years-Family history of heart disease can also be a reason for developing a heart disease as mentioned earlier. So, this data of the patient can also be included for further increasing the accuracy of the model.

This work will be useful in identifying the possible patients who may suffer from heart disease in the next 10 years. This may help in taking preventive measures and hence try to avoid the possibility of heart disease for the patient. So when a patient is predicted as positive for heart disease, then the medical data for the patient can be closely analysed by the doctors. An example would be - suppose the patient has diabetes which may be the cause for heart disease in future and then the patient can be given treatment to have diabetes in control which in turn may prevent the heart disease.

The heart disease prediction can be done using other machine learning algorithms. Logistic regression can also perform well in case of binary classification problems such as heart disease prediction. Random forests can perform well than decision trees. Also, the ensemble methods and artificial neural networks can be applied to the data set. The results can be compared and improvised.

# **REFRENCES**

[1] Monika Gandhi, Shailendra Narayanan Singh Predictions in heart disease using techniques of data mining (2015)

[2] J Thomas, R Theresa Princy Human heart disease prediction system using data mining techniques (2016)

[3] Sana Bharti, Shailendra Narayan Singh, Amity university, Noida, India Analytical study of heart disease prediction comparing with different algorithms (May 2015)

[4] Purushottam, Kanak Saxena, Richa Sharma Efficient heart disease prediction system using Decision tree (2015)

[5] Sellappan Palaniyappan, Rafiah Awang Intelligent heart disease prediction using data mining techniques (August 2008)

[6] Himanshu Sharma,M A Rizvi Prediction of Heart Disease using Machine Learning Algorithms: A Survey (August 2017)

[7] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review (2017)

[8] V.Krishnaiah, G.Narsimha, N.Subhash Chandra Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review (February 2016)

[9] Ramandeep Kaur, 2Er. Prabhsharn Kaur A Review - Heart Disease Forecasting Pattern using Various Data Mining Techniques (June 2016)

[10] J.Vijayashree and N.Ch. SrimanNarayanaIyengar Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review (2016)

[11] Benjamin EJ et.al Heart Disease and Stroke Statistics 2018 At-a-Glance (2018)

[12] Abhay Kishore, Ajay Kumar, Karan Singh, Maninder Punia, Yogita Hambir Department of Computer Engineering, Army Institute of Technology, Pune, Maharashtra Professor, Department of Computer Engineering, Army Institute of Technology, Pune, Maharashtra Heart Attack Prediction Using Deep Learning (2018)

[13] M.Nikhil Kumar, K.V.S Koushik, K.Deepak Department of CSE, VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India Prediction Heart Diseases using Data mining and machine learning algorithms and tools.(2018)

[14] Amandeep Kaur, Jyoti Arora Dept of CSC Desh Bhagat University, Punjab, India heart disease prediction using data mining techniques: a survey (2018)