**A REPORT**

*On AI Mini Project*

**Heart Disease Prediction**

Made By

Saquib Mir (TE A 47)

Pratik Mhatre (TE A 46)

Pranav Lature (TE A 42)

*A report submitted in partial fulfillment of the requirements for the Award of Degree of*

BACHELOR OF ENGINEERING

*In*

COMPUTER ENGINEERING

Under Supervision of

Asst. Prof. D.M Kanade , KKWIEER



**K. K. WAGH INSTITUTE OF ENGINEERING EDUCATION AND RESEARCH, NASHIK, MAHARASHTRA.**

**DEPARTMENT OF COMPUTER ENGINEERING**

**Chapter – 1**

**DETAILS OF THE PROJECT**

**1.1 Problem Statement**

Heart disease is the leading cause of death among all other diseases , even cancers .It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors.

Heart 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.

**1.2 Objectives / Motivation Scope**

The main objective 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 data mining techniques on medical data set to assist in the prediction of the heart diseases.

* Provides new approach to concealed patterns in the data.
* Helps avoid human biasness.
* To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.
* Reduce the cost of medical tests

**Chapter – 2**

**METHODOLOGICAL DETAILS**

**2.1 Dataset**

The dataset provides the patients information. It includes over 302 records and 14 attributes. Variables Each attributes a potential risk factor. There are both demographic, behavioral and medical risk factors.

Provides new approach to concealed patterns in the data.

•Helps avoid human biasness.

•To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.

•Reduce the cost of medical tests.

Provides new approach to concealed patterns in the data.

•Helps avoid human biasness.

•To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.

•Reduce the cost of medical tests.

Provides new approach to concealed patterns in the data.

•Helps avoid human biasness.

•To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.

•Reduce the cost of medical tests.

Provides new approach to concealed patterns in the data.

•Helps avoid human biasness.

•To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.

•Reduce the cost of medical tests.

**Attributes:**

* Age : age of the patient
* Sex : sex of the patient (1 = male; 0 = female)
* Cp : chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 0 = asymptomatic)
* Trtbps : resting blood pressure (in mm Hg)
* Chol : serum cholesterol (in mg/dl)
* Fbs : fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
* Restecg : resting electrocardiographic results (0 = normal; 1 = ST-T wave abnormality; 2 = left ventricular hypertrophy)
* Thalachh : maximum heart rate achieved
* Exng : exercise-induced angina (1 = yes; 0 = no)
* Oldpeak : ST depression induced by exercise relative to rest
* Slp : slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
* Caa : number of major vessels (0-3) colored by fluoroscopy
* Thall : thallium stress test result (3 = normal; 6 = fixed defect; 7 = reversible defect)
* Output : presence of heart disease (1 = yes; 0 = no)

**2.2 Data Pre-processing**

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pre-processing task.

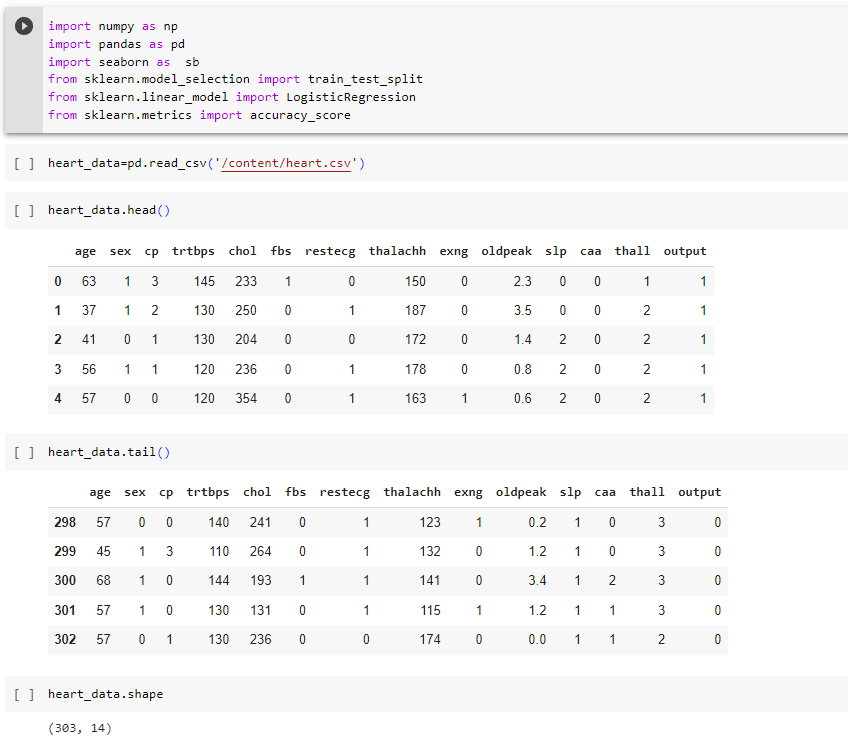


Fig.1. Data Pre-processing

Here, we imported required libraries for the project, then we fetched an .csv(comma separated values) file which contains our dataset then we displayed first 5 record and last 5 using head function and Tail Function & Shape Funtion to get the size of datset .

* **Balancing of Data**

Imbalanced datasets can be balanced in two ways. They are Under Sampling

and Over Sampling

(a) Under Sampling :

In Under Sampling, dataset balance is done by the reduction of the size of the ample

class. This process is considered when the amount of data is adequate.

(b) Over Sampling :

In Over Sampling, dataset balance is done by increasing the size of the scarce samples.

This process is considered when the amount of data is inadequate.

**2.3 Exploratory Data Analysis**

* **Missing variables**

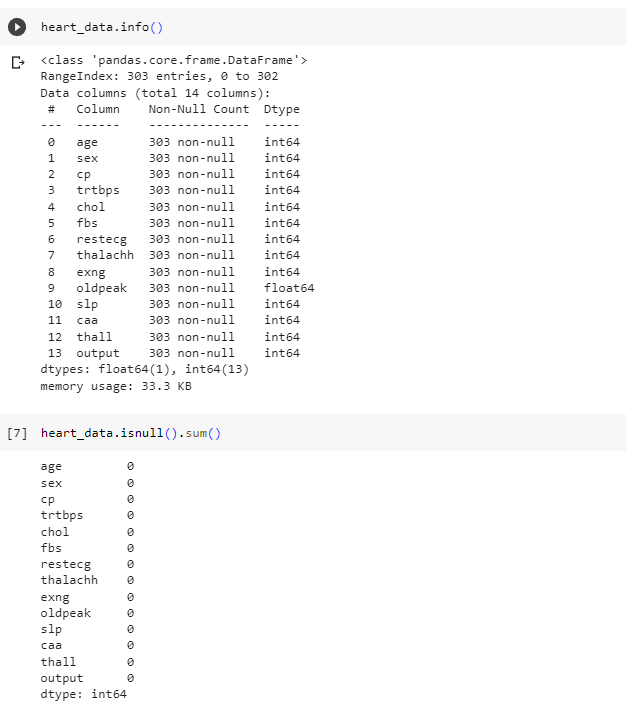


Fig.2. Missing Variables

The output of the code shows the number of null values in each column of heart data as a series with the column names as the index. Since all the values in the output are 0, it indicates that there are no missing or null values in any of the columns in heart data. Therefore, the dataset is complete, and we can proceed with data analysis without the need for imputing or dropping missing values.

* **Data Distribution**

Data distribution is a function that specifies all possible values for a variable and also quantifies the relative frequency (probability of how often they occur). Distributions are considered any population that has a scattering of data. It’s important to determine the kind of distribution that population has so we can apply the correct statistical methods when analyzing it.



**Chapter – 3**

**RESULTS**

**3.1 Librarys**

**- Numpy**

**- Pandas**

**- Matplotlib**

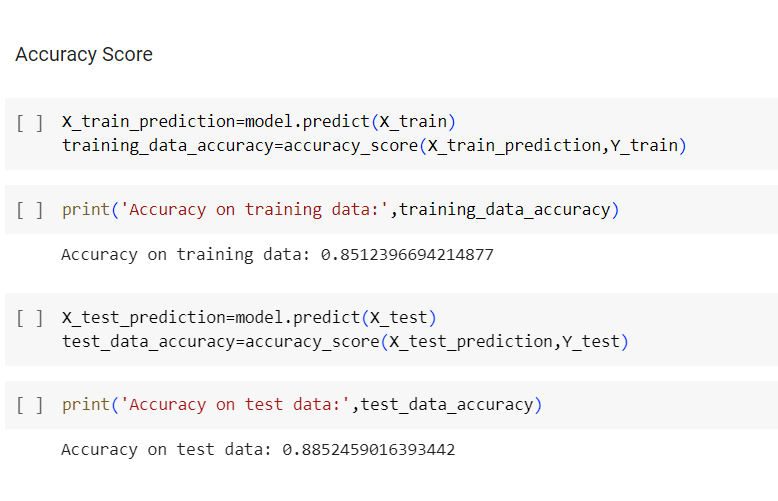
**- Seaborn**

**3.2 Algorithms**

**3.2.1 Logistic Regression**

Logistic regression is one of the most popular Machine Learning algorithms,which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables .

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.



**3.2.2 Random Forest**

Random Forest is a supervised learning algorithm. It is an extension of machine learning classifiers which include the bagging to improve the performance of Decision Tree. It combines tree predictors, and trees are dependent on a random vector which is independently sampled. The distribution of all trees are the same. Random Forests splits nodes using the best among of a predictor subset that are randomly chosen from the node itself, instead of splitting nodes based on the variables. The time complexity of the worst case of learning with Random Forests is O(M(dnlogn)) , where M is the number of growing trees, n is the number of instances, and d is the data dimension.

**Assumptions:**

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

● There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

● The predictions from each tree must have very low correlations.

**Algorithm Steps:**

It works in four steps:

● Select random samples from a given dataset.

● Construct a Decision Tree for each sample and get a prediction result from each Decision Tree.

● Perform a vote for each predicted result.

● Select the prediction result with the most votes as the final prediction.

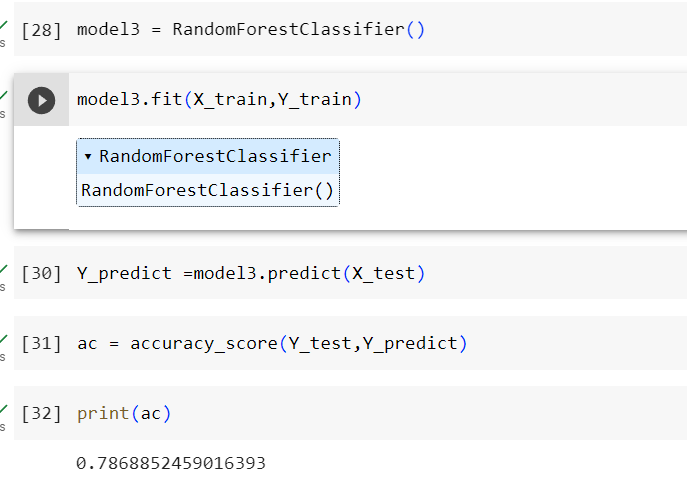


Fig 5. Random Forest

**3.2.3 Gaussian Naive Bayes algorithm**

The Gaussian Naive Bayes algorithm is a probabilistic classification algorithm that is commonly used in machine learning. It is based on Bayes' theorem, which is a fundamental concept in probability theory.

The formula for Bayes' theorem is as follows:

**P(A|B) = P(B|A) \* P(A) / P(B)**

where:

* P(A|B) is the probability of event A given that event B has occurred
* P(B|A) is the probability of event B given that event A has occurred
* P(A) is the prior probability of event A
* P(B) is the prior probability of event B

**Algorithm Steps:**

1. **Data Preparation**: The first step of the algorithm is data preparation. A labeled dataset is required for training the algorithm. Each instance in the dataset consists of a set of features and a corresponding class label.
2. **Calculating Class Priors**: The algorithm starts by calculating the prior probability of each class label in the training dataset. This is done by counting the number of instances belonging to each class and dividing it by the total number of instances.
3. **Estimating Feature Distributions**: The next step is to estimate the parameters of a Gaussian distribution (mean and variance) for each feature in the dataset. The algorithm calculates the mean and variance of each feature separately for instances belonging to each class.
4. **Predicting the Class Label**: Given a new instance with a set of features, the algorithm calculates the conditional probability of the instance belonging to each class label using Bayes' theorem. It multiplies the prior probability of the class with the product of the conditional probabilities of the features given that class. The class with the highest conditional probability is selected as the predicted class label.
5. **Model Evaluation**: After predicting the class labels for the test instances, the algorithm evaluates its performance by comparing the predicted labels with the true labels. Common evaluation metrics include accuracy, precision, recall, and F1-score.

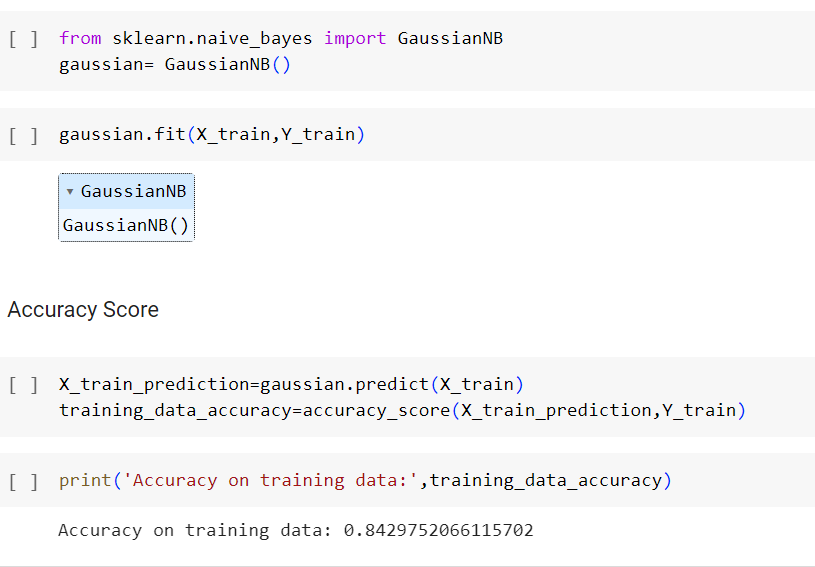


Fig 6.Gaussian Naive Bayes

**3.3 Model Comparison**

|  |  |
| --- | --- |
| **Model Name** | **Accuracy** |
| Logistic regression | 0.8512 |
| Random Forest | 0.7868 |
| Gaussian Naive Bayes | 0.8429 |

Table 1. Comparison Graph

**3.4 Output**

After trying and testing 3 different algorithms, the best accuracy is achieved by Logistic Regression , folloewd Gaussian Naive Bayes ,followed by Random Forest. Hence the best algorithm is Logistic Regression.

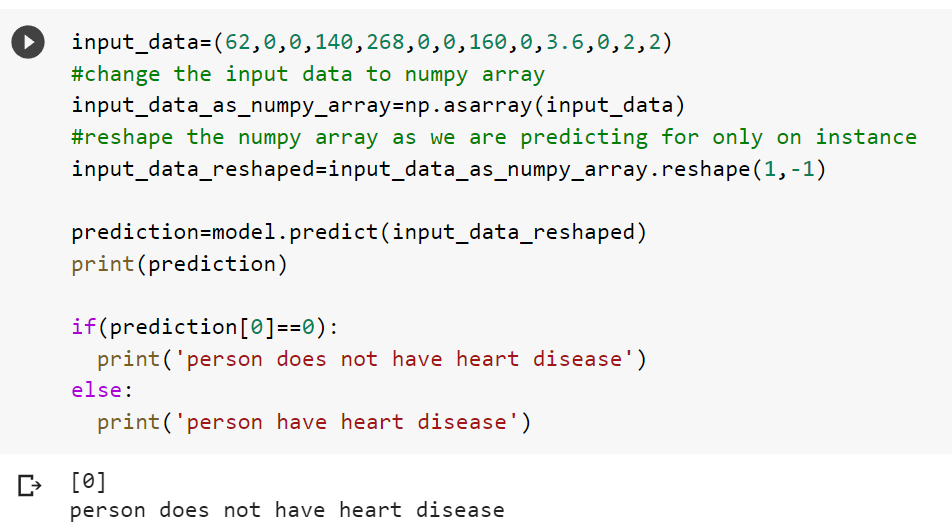


Fig.7. Output.1

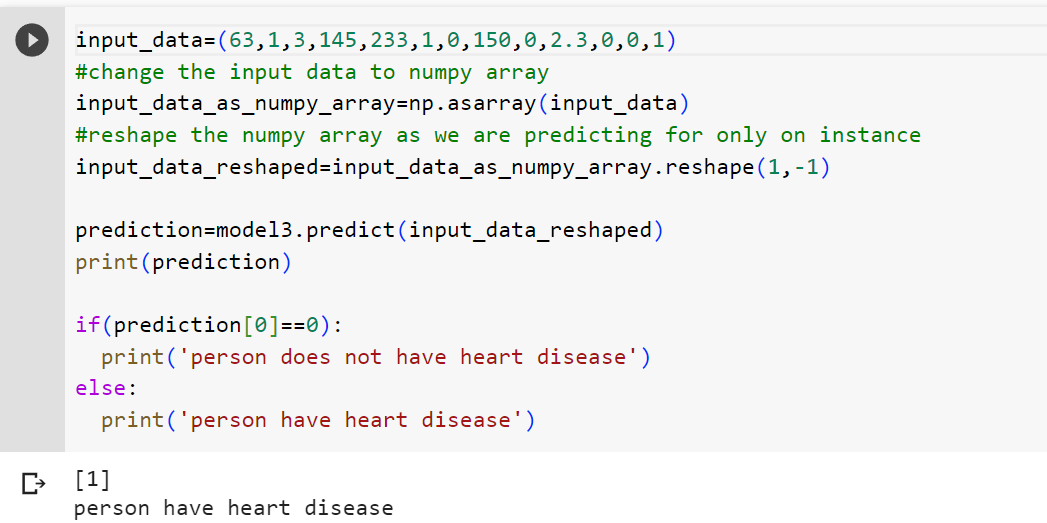


Fig.8.Output.2

**Chapter – 4**

**CONCLUSION AND FUTURE**

**Conclusion**

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 three different machine learning algorithms used to measure the performance are Logistic Regression , Random Forest, Gaussian Naive Bayes applied on the dataset.

The expected attributes leading to heart disease in patients are available in the dataset which contains 7 features and 14 important features that are useful to evaluate the system are selected among them. If all the features taken into the consideration then the efficiency of the system the author gets is less. To increase efficiency, attribute selection is done. In this n features have to be selected for evaluating the model which gives more accuracy. The correlation of some features in the dataset is almost equal and so they are removed. If all the attributes present in the dataset are taken into account then the efficiency decreases considerably.