**-**

**Project Report**

**HEART DISEASE PREDICTION**

Made by

SIDDARTH MOHAN

Abstract  
  
**The prediction of cardiac disease helps practitioners make more accurate decisions regarding patients' health. Therefore, the use of machine learning (ML) is a solution to reduce and understand the symptoms related to heart disease. In this project we have used different machine learning models to predict heart disease out of which naïve byes model gave a accuracy of 82%. K-Neighbours model gave a accuracy of 61% and decision tree which gave a accuracy of 88%.**

Table of Contents

|  |  |  |
| --- | --- | --- |
| 1. | Introduction………………………………………… | 1 |
| 2.  3. | Motivation……………………………………………  Problem Statement………………………………1 | 1 |

4. Objectives…………………………………………1

5. Tools Used………………………………………2

6. Dataset Description………………………………2

7. Data Preprocessing…………………………………3

8. Data mining tasks Perform………………………3

9. Algorithm……………………………………………..4

10. Output…………………………………………………5

11. Visualization Screenshots….......................6

12. Conclusion……………………………………10

13. References in IEEE format………………………10

1.Introduction

Heart attacks and strokes are usually acute events and are mainly caused by a blockage that prevents blood from flowing to the heart or brain. The most common reason for this is a build-up of fatty deposits on the inner walls of the blood vessels that supply the heart or brain. The cause of heart attacks and strokes are usually the presence of a combination of risk factors, such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol, hypertension, diabetes and hyperlipidaemia.

The most important behavioural risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioural risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These “intermediate risks factors” can be measured in primary care facilities and indicate an increased risk of developing a heart attack, stroke, heart failure and other complications.

2.MOTIVATION

Machine Learning is used across many spheres around the world.

The healthcare industry is no exception. Machine Learning can play an essential role in predicting presence/absence of Locomotor disorders Heart diseases and more.

Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis.

3.PROBLEM STATEMENT

Heart disease is the leading cause of death for both men and women.

Given clinical parameters about a patient, can we predict whether or not they have heart disease using machine learning techniques?

4.OBJECTIVES

1.Predict heart disease with different machine learning algorithms.

2.To understand the Dataset thoroughly.

3.To perform various Data Preprocessing and Data mining Tasks.

5.TOOLS USED

1. KAGGLE (For dataset)
2. Google Colab

6.Dataset Description

The dataset consists of 303 entries.There are 14 columns in the dataset, which are described below.

**1.Age:** displays the age of the individual.

**2.Sex:** displays the gender of the individual:0 = male 1 = female

**3.Chest-pain(cp) type:** displays the type of chest-pain experienced by the individual using the following format:

1 = typical angina 2 = atypical angina 3 = non — anginal pain 4 = asymptotic

**4.Resting Blood Pressure(trestbps):** displays the resting blood pressure value of an individual in mmHg (unit)

**5.Cholesterol(chol):** displays cholesterol in mg/dl (unit)

**6.Fasting Blood Sugar(fbs):** compares the fasting blood sugar value of an individual with 120mg/dl.

If fasting blood sugar > 120mg/dl then : 1 (true) else : 0 (false)

**7.Resting ECG(restecg):** displays resting electrocardiographic results

0 = normal

1 = having ST-T wave abnormality

2 = left ventricular hypertrophy

**8.Max heart rate achieved(thalach) :** displays the max heart rate achieved by an individual.

**9.Exercise induced angina(exang) :**1 = yes 0 = no

**10.oldpeak:**ST depression induced by exercise relative to rest:

**11.slopePeak** exercise ST segment :

1 = upsloping

2 = flat

3 = downsloping

**12.Number of major vessels (0–3) colored by flourosopy(ca)** : displays the value as integer or float.

**13.Thal :** displays the thalassemia :

3 = normal

6 = fixed defect

7 = reversible defect

**14.Target:**0 = absent 1=present

7.Data Preprocessing

Data Duplication

Data deduplication is a technique for eliminating duplicate copies of repeating data. A related and somewhat synonymous term is single-instance (data) storage. This technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent.

Dummy variables:

Adding dummy variables for columns with categorical values.

Standard scalar:

Scaling the values to the normalized values for the column.

Feature Selection:

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features

8.Data mining tasks Performed

Kneighbours:-

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry.

Decision Tree:-

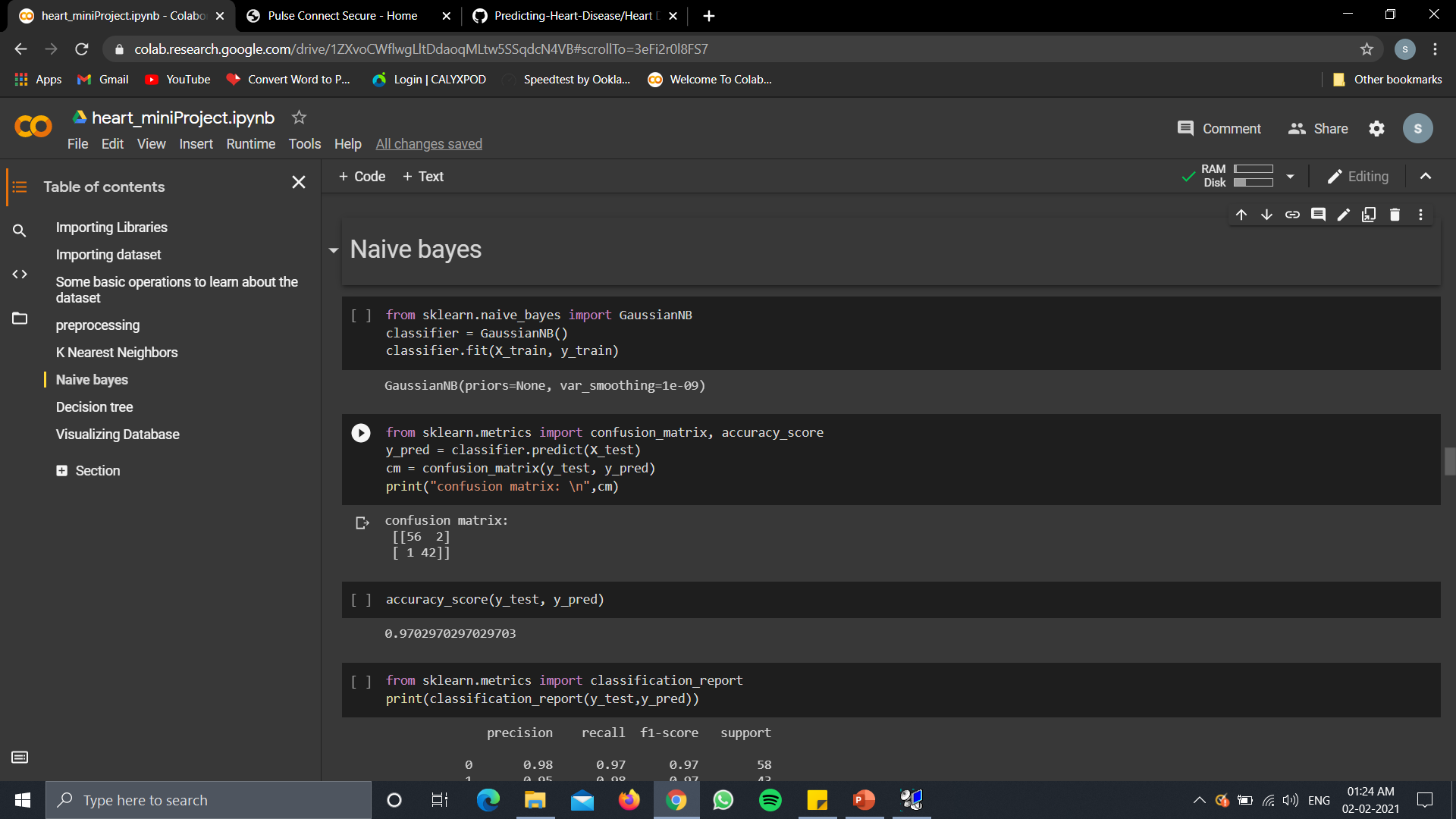
A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.

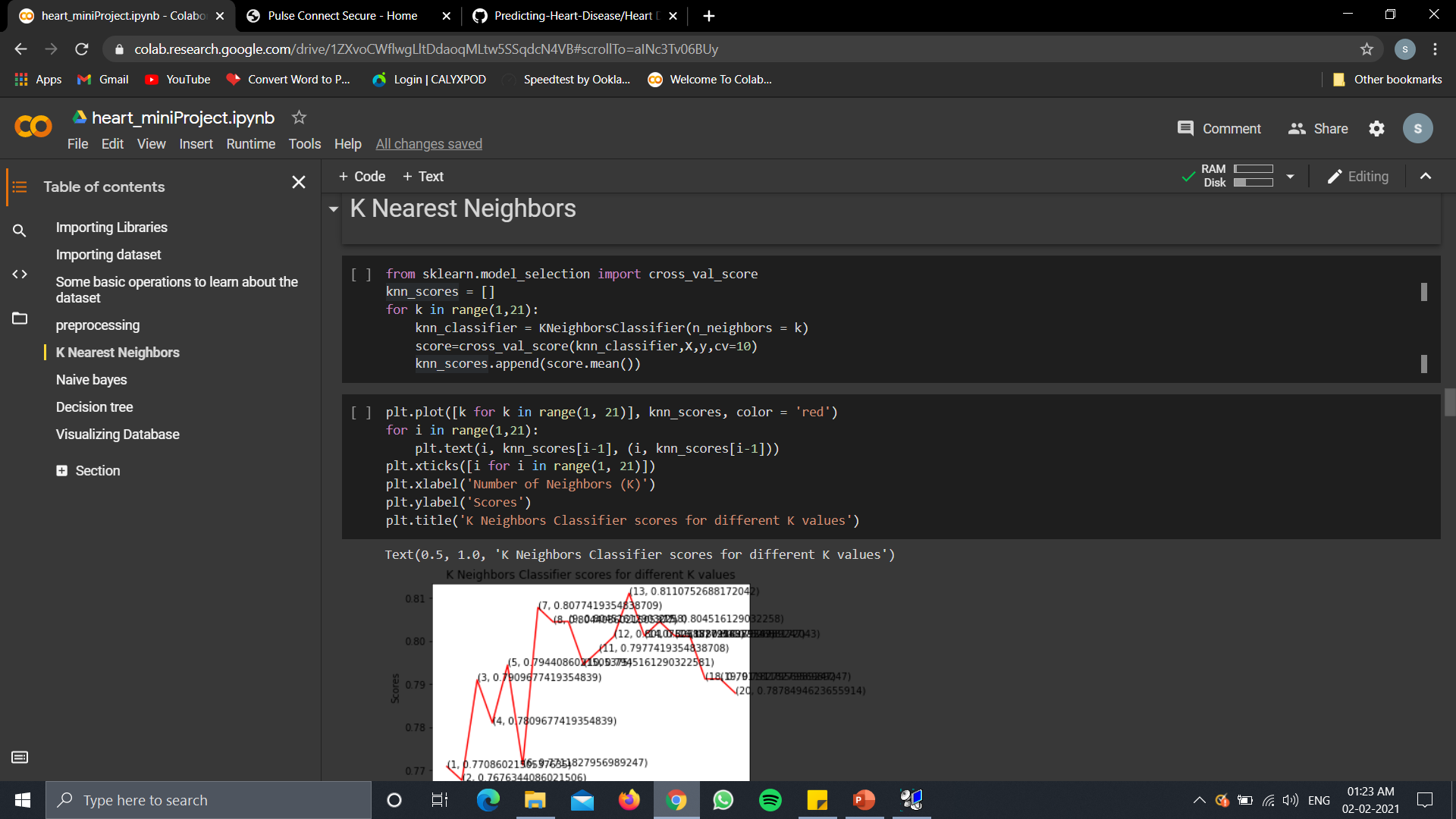
Naive Baye's:

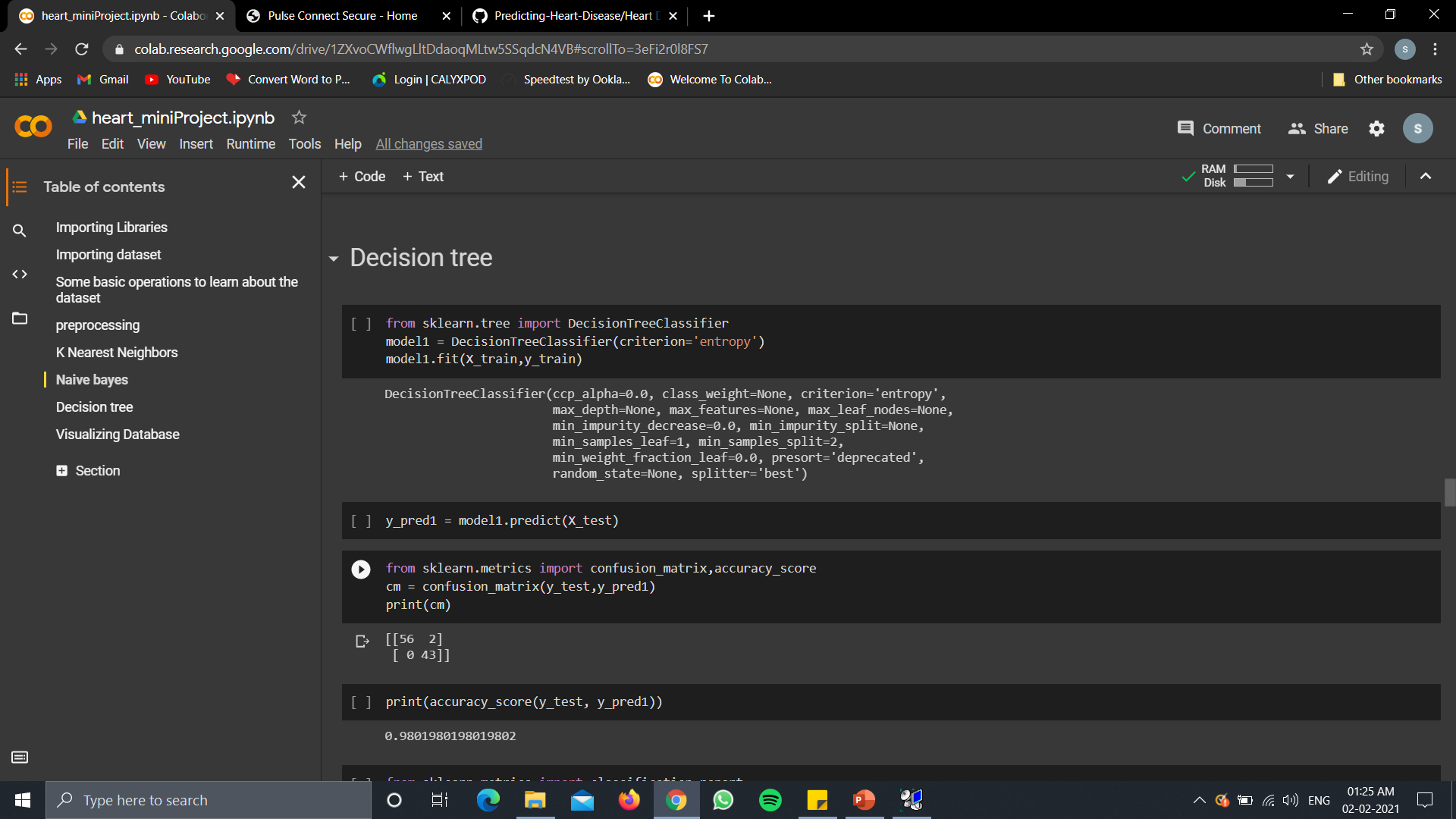
Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

It is mainly used in text classification that includes a high-dimensional training dataset.Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

10.Output







11.Visualization

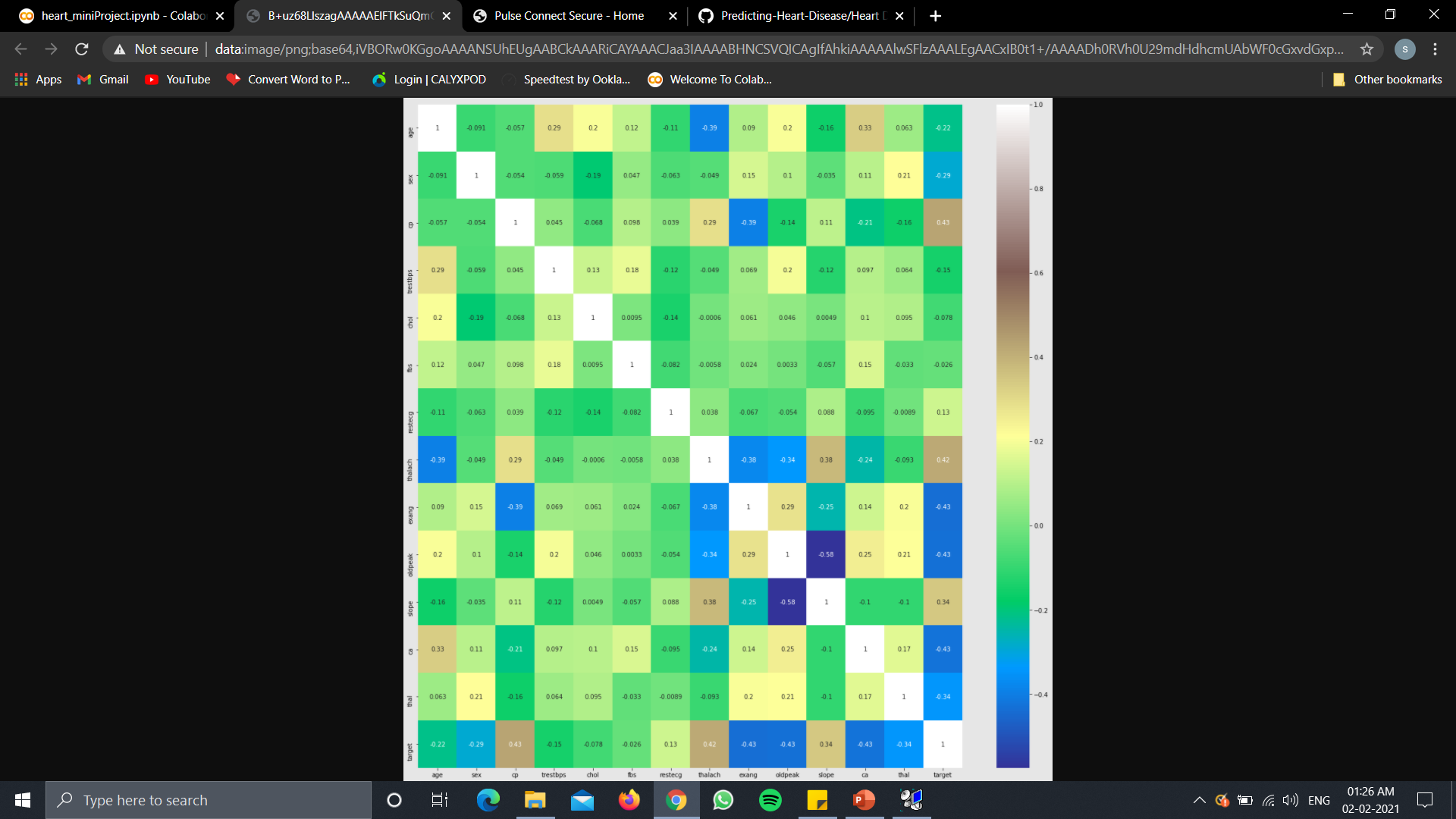


Fig1:Heatmap

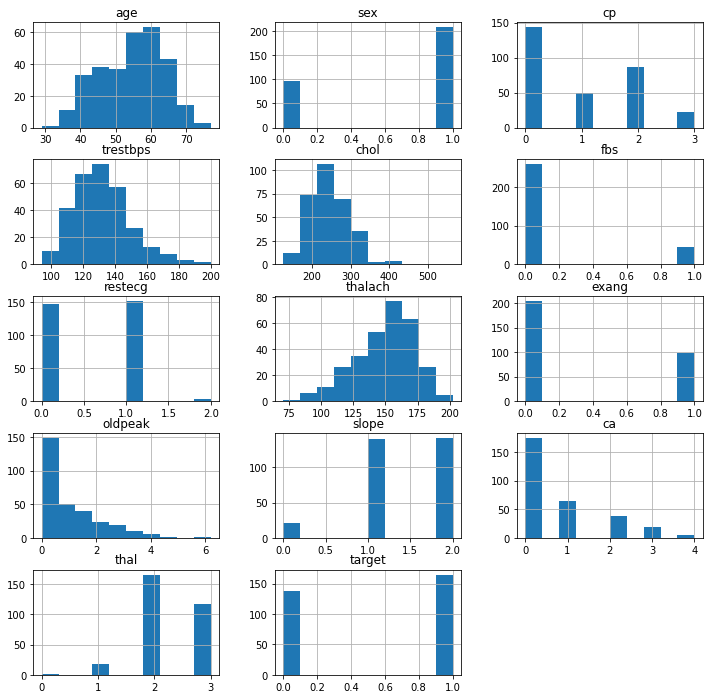


Fig2:Histogram

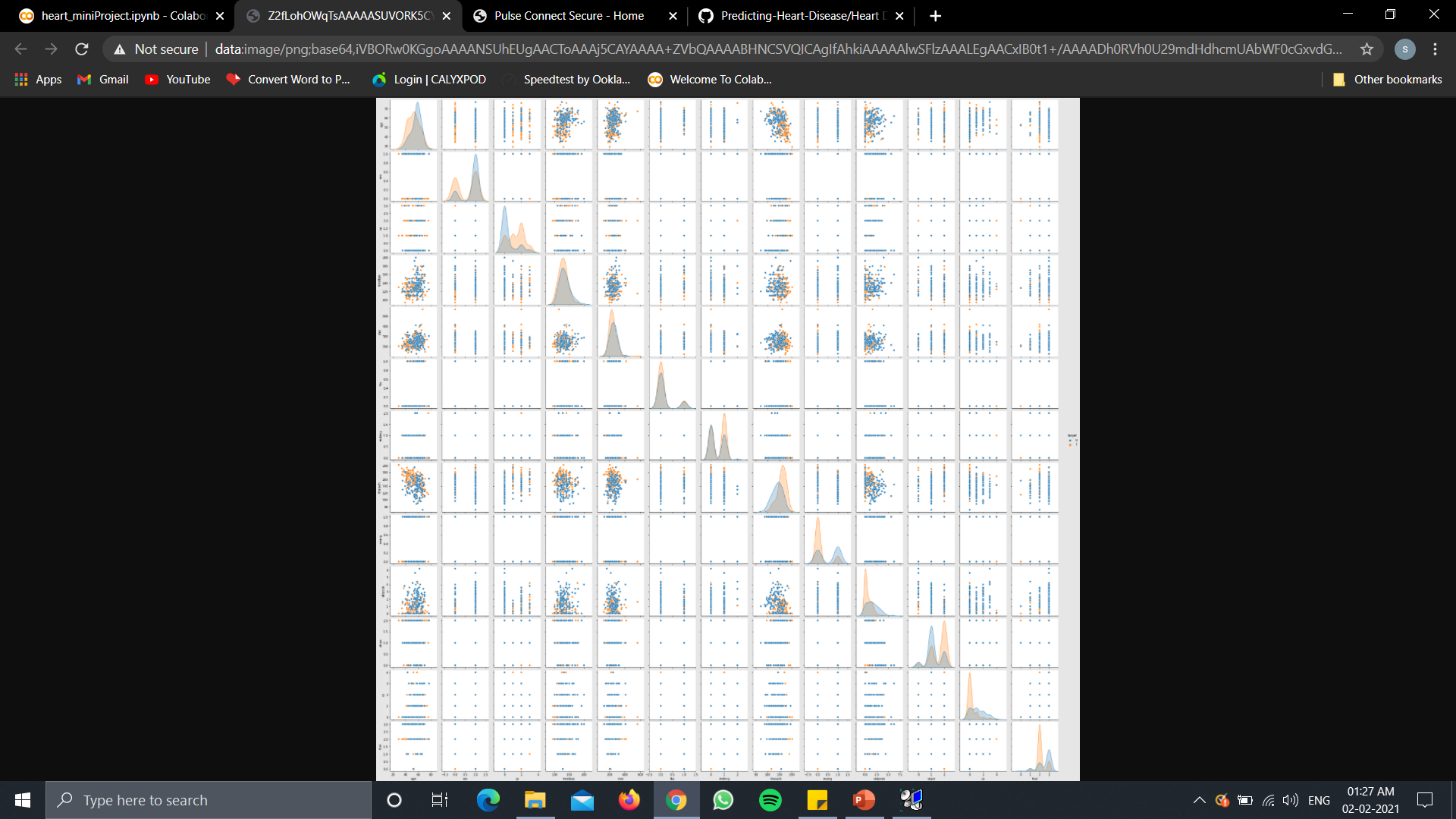


Fig3:Pairplot

12. CHECKING FOR DIFFERENT FEATURES OF EVERY ALGORITHM TO DECIDE WHICH ONE IS BEST:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **K-NEAREST NEIGHBORS** | **NAÏVE BAYES** | **DECISION TREE** |
| ACCURACY | 0.61 | 0.82 | 0.88 |
| PRECISION | 0.63 | 0.81 | 0.95 |
| RECALL | 0.51 | 0.84 | 0.78 |
| F2 SCORE | 0.57 | 0.83 | 0.86 |
| TIME(TRAINING) | 0.0065 | 0.0024 | 0.0058 |
| TIME (TESTING) | 0.0097 | 0.0019 | 0.0015 |
| SPACE | 135168 | 27854 | 250848 |

12.Conclusion

* Thus, the team has successfully implemented a machine learning model using KNN, naive byes and decision tree for predicting heart disease, by checking for accuracy, precision, recall, f2Score, training time, testing time, space occupied.
* We have come to a conclusion that decision tree is giving us the best result with respect to accuracy but it takes more space and also requires more training time.
* But considering our problem statement and the size of our dataset we can say that we can compromise on space and training time as long as we are getting good accuracy from the model.
* As it is very important to correctly predict whether the patient has a heart disease.

13.References in IEEE format

1. <https://www.kaggle.com/ronitf/heart-disease-uci>
2. <https://pdfs.semanticscholar.org/36ee/7596ab9b01e8850bf853ad8e3326318bdde7.pdf>
3. Bhatla, Nidhi, and Kiran Jyoti. "An analysis of heart disease prediction using different data mining techniques." *International Journal of Engineering* 1.8 (2012): 1-4.
4. Thomas, J., and R. Theresa Princy. "Human heart disease prediction system using data mining techniques." *2016 international conference on circuit, power and computing technologies (ICCPCT)*. IEEE, 2016.