Predicting Heart Aliment using Jupyter

Soorya P
AJC17MCA-I053
Department of Computer Applications
Amal Jyothi College of Engineering
Kanjirapally
sooryap@mca.ajce.in

Anit James
Assistant Professor
Department of Computer Applications
Amal Jyothi College of Engineering Kanjirapally
anitjames@amaljyothi.ac.in

***Abstract*—The heart is more important to the human body than any other circulatory organ. Its characteristic is to provide and pump blood to other organs and the brain. So, it's far very critical to have a healthy heart however researchers discovered the risk of heart failure increases each day. Many heart specialists can diagnose coronary heart sickness with their enjoy and competencies. However, some professionals missing the expertise or expertise to are expecting cardiovascular ailment inside the early stages, a small mistake can price a patient's life. consequently, it is vital to apply particular methods and algorithmic tools to estimate the occurrence of cardiac problems inside the early degrees. They compare the likelihood of heart disease in a person and look for targets to focus on. Typically, it takes years of experience and a battery of medical examinations. The primary purpose of the paintings in this study is to find the best typeset of rules for showing maximum accuracy. Decision tree and Logistic Regression are used to broaden the prediction system if you want to Analyze and expect the opportunity of a coronary heart ailment.**

Keywords—Data Mining, Classification, Prediction, Heart Ailment, Jupyter

I. INTRODUCTION

The heart is more important than any other circulatory organ to the human body. Its job is to bring blood to the brain and other organs and pump it out. As a result, having a healthy coronary heart is critical, but researchers discovered that the chance of coronary heart failure is rising every day. With their experience and capabilities, several cardiac specialists can identify coronary heart disease. However, if a few physicians lack the intelligence or understanding to diagnose cardiovascular disease in its early stages, even a minor error can have a major influence on a patient's lifestyle. Consequently, it's critically essential to use specific methods and algorithmic equipment to estimate the prevalence of cardiac issues in the early stages. They compare the likelihood of coronary heart disease in a person by examining the goals that they are aiming after. Years of training and a thorough clinical test may be necessary. The primary objective of this study project is to find the unique set of rules that would provide the highest level of accuracy. If you need to analyze and count on the chance of coronary heart illness, you can use a decision tree and Logistic Regression to produce a prediction device.

Heart disease symptoms vary greatly depending on the type of difficulty a person is experiencing. A few signs are difficult for the typical individual to recognize. Typical indications and symptoms include chest discomfort, dyspnea, and coronary heart palpitations. Angina, or angina pectoris, is a type of chest discomfort that happens when a portion of the heart no longer receives adequate oxygen. It is frequent in many types of coronary heart disease. Angina can be brought on by traumatic events or physical activity, and it usually lasts less than 10 minutes. Health issues can also develop as a result of rare forms of coronary heart disease. The symptoms of a coronary heart attack are similar to those of angina, except that they can appear at any time during rest and are more severe. The signs and symptoms of a heart attack can sometimes be mistaken for indigestion. Heartburn, stomach discomfort, and a heavy feeling in the chest are all possible symptoms. Other indications and symptoms of a heart attack include discomfort that moves throughout the body, such as from the chest to the palms, neck, back, stomach, or jaw, dizziness, lightheadedness, as well as copious sweating.

To produce predictions in this study, a comparative comparison of the two data mining classification techniques, Decision Tree and Logistic Regression, is used. The analysis is done at several levels of cross-validation and several percentages of percentage split evaluation methods respectively. In this study, the "Heart Disease" dataset from the UCI machine learning repository was used to make heart disease predictions. We must determine which of the two methods, Logistic Regression, and Decision Tree, will provide the most accurate results. And finally, we can understand using which algorithm has the best accuracy in the prediction of heart diseases.

II. METHODOLOGY

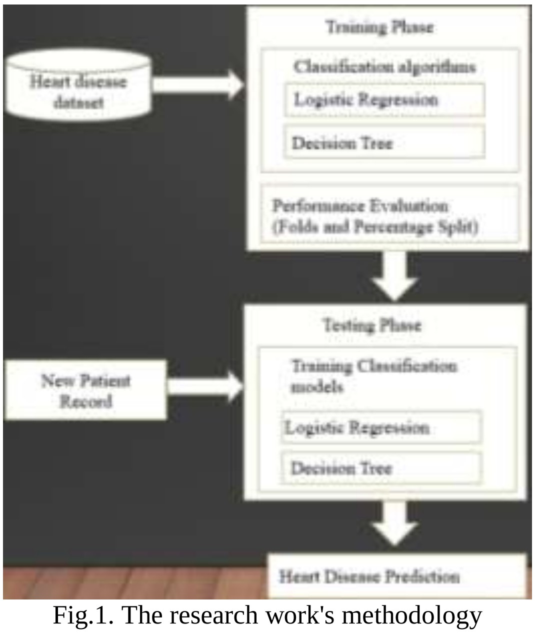


Fig.1. The research work's methodology

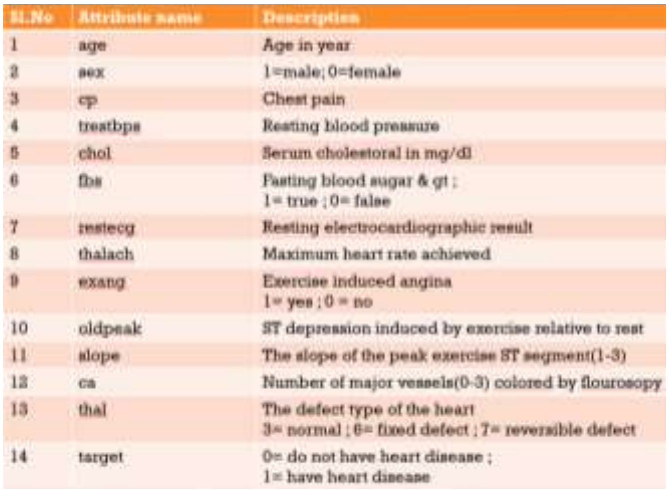
The strategy indicated in Fig. 1 for constructing the classification model needed to predict heart disease in

DOI: 10.5281/zenodo.6183068
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

patients can be used to predict heart disease. The model serves as a foundation for doing cardiac disease prediction using machine learning approaches. To make predictions, a classifier must first be trained using the records, after which a classification model must be created, which is then fed with a fresh unknown record and a prediction made. The Performance Evaluation of the two classification algorithms is part of the study approach. The training portion of this study involves using the heart disease dataset to train two classification methods, Decision Tree and Logistic Regression, and then creating a classification model.

A. Dataset Description

This study was based on the "Heart Disease" dataset from the UCI machine learning library. Despite the fact that this database has 76 features, all published studies only employ a subset of 14 of them. The Cleveland database, in particular, is the only one that machine learning researchers have used thus far. The "target" field shows whether the patient has heart illness or not. It ranges from 0 (no presence) to 4 (present). The purpose of this research is to find a way to anticipate heart illness regardless of the type. A numeric data type that ranges from 29 to 77 years old is used to represent the patient's age. Figure 2 depicts all of the properties and their values.

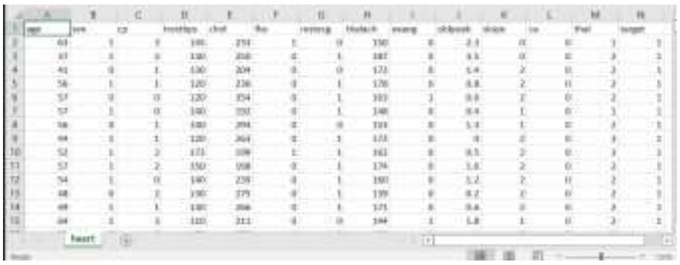


Sl.No	Attribute name	Description
1	age	Age in year
2	sex	1=male; 0=female
3	cp	Chest pain
4	treatbps	Resting blood pressure
5	chol	Serum cholestoral in mg/dl
6	fbs	Fasting blood sugar & gt ; 1= true ; 0= false
7	restecg	Resting electrocardiographic result
8	thalach	Maximum heart rate achieved
9	exang	Exercise induced angina 1= yes ; 0 = no
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	The slope of the peak exercise ST segment(1-3)
12	ca	Number of major vessels(0-3) colored by flourosopy
13	thal	The defect type of the heart 3= normal ; 6= fixed defect ; 7= reversible defect
14	target	0= do not have heart disease ; 1= have heart disease

Fig.2. Attribute and Description of the dataset used for research

B. Statistical Analysis

Kaggle.com was used to acquire the data. The dataset was then downloaded and saved in Microsoft Excel as a CSV file, which was then used for data analysis. The UCI machine learning repository's "Heart Disease" dataset.



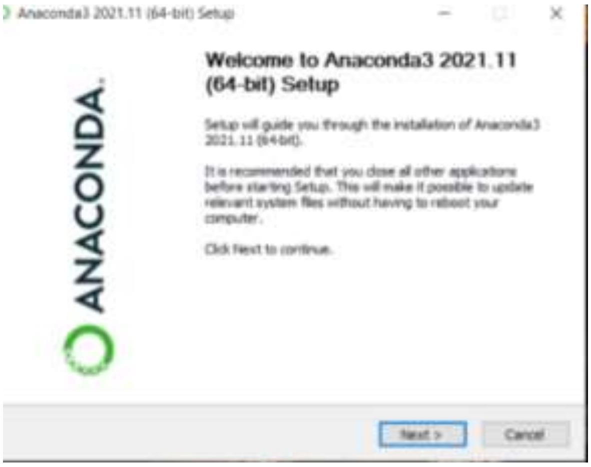
age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1

Fig.3. Dataset

C. Tool used – Jupyter

Project Jupyter develops open-source software, open standards, and interactive computing services in a variety of programming languages. Jupyter Notebook is a web platform that lets you create and share code, graphics, and text documents. It can be used for data science, statistical modeling, machine learning, and other applications.

C.1. Installing Jupyter



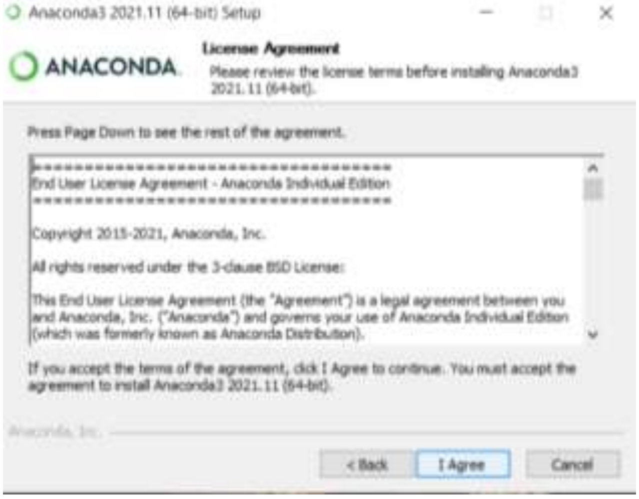
Welcome to Anaconda3 2021.11 (64-bit) Setup

Setup will guide you through the installation of Anaconda3 2021.11 (64-bit).

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.

Fig.4.1. Anaconda installation.



License Agreement

Please review the license terms before installing Anaconda3 2021.11 (64-bit).

Press Page Down to see the rest of the agreement.

===================================
End User License Agreement - Anaconda Individual Edition
===================================

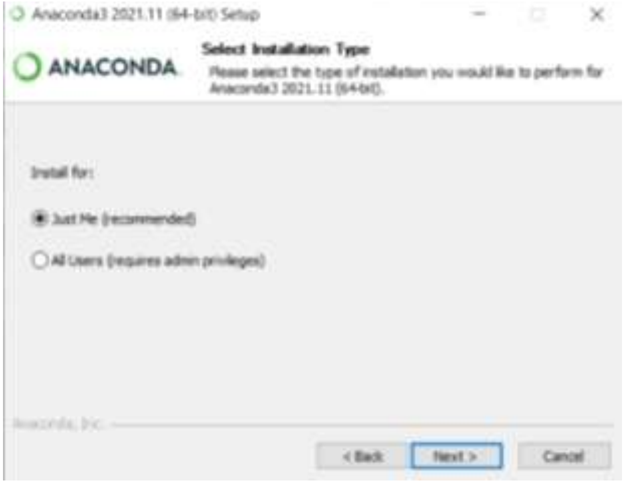
Copyright 2015-2021, Anaconda, Inc.

All rights reserved under the 3-clause BSD License:

This End User License Agreement (the "Agreement") is a legal agreement between you and Anaconda, Inc. ("Anaconda") and governs your use of Anaconda Individual Edition (which was formerly known as Anaconda Distribution).

If you accept the terms of the agreement, click I Agree to continue. You must accept the agreement to install Anaconda3 2021.11 (64-bit).

Fig.4.2. License Agreement.



Select Installation Type

Please select the type of installation you would like to perform for Anaconda3 2021.11 (64-bit).

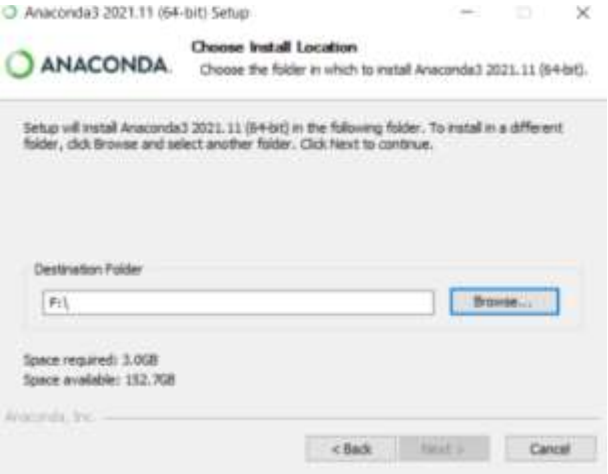
Install for:

● Just Me (recommended)

○ All Users (requires admin privileges)

Fig.4.3. Select installation type

DOI: 10.5281/zenodo.6183068
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam



Choose Install Location

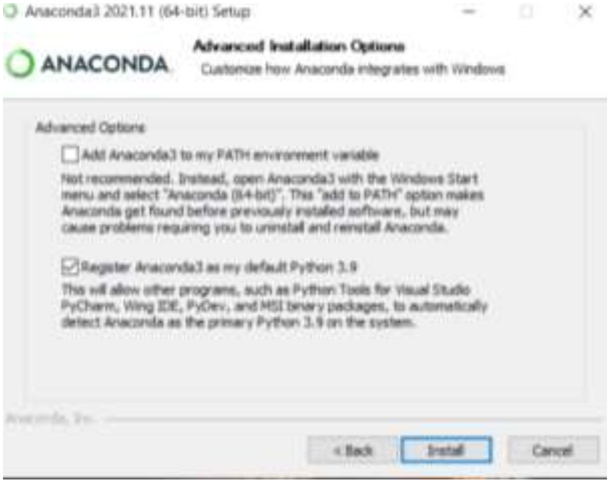
Choose the folder in which to install Anaconda3 2021.11 (64-bit).

Setup will install Anaconda3 2021.11 (64-bit) in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder: F:\

Space required: 3.0GB
Space available: 152.7GB

Fig.4.4. Select a place for installation.



Advanced Installation Options

Customize how Anaconda integrates with Windows

Advanced Options

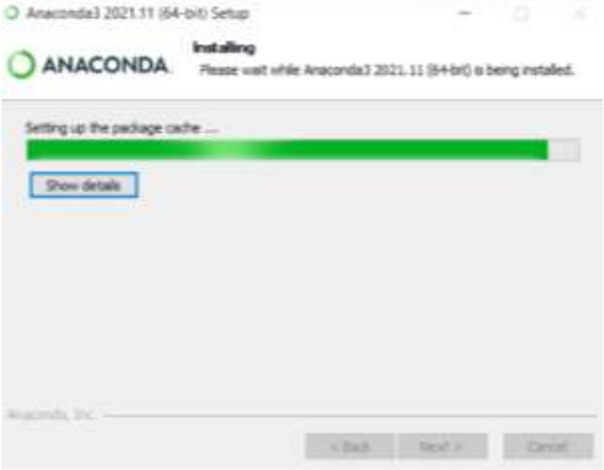
□ Add Anaconda3 to my PATH environment variable

Not recommended. Instead, open Anaconda3 with the Windows Start menu and select "Anaconda (64-bit)". This "add to PATH" option makes Anaconda get found before previously installed software, but may cause problems requiring you to uninstall and reinstall Anaconda.

☑ Register Anaconda3 as my default Python 3.9

This will allow other programs, such as Python Tools for Visual Studio PyCharm, Wing IDE, PyDev, and MSI binary packages, to automatically detect Anaconda as the primary Python 3.9 on the system.

Fig.4.5. Options for advanced installation

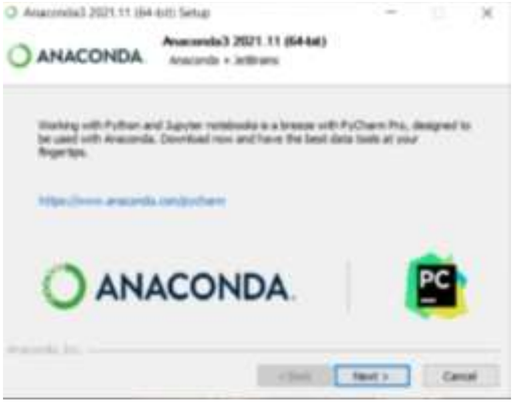


Installing

Please wait while Anaconda3 2021.11 (64-bit) is being installed.

Setting up the package cache ...

Fig.4.6. Installing



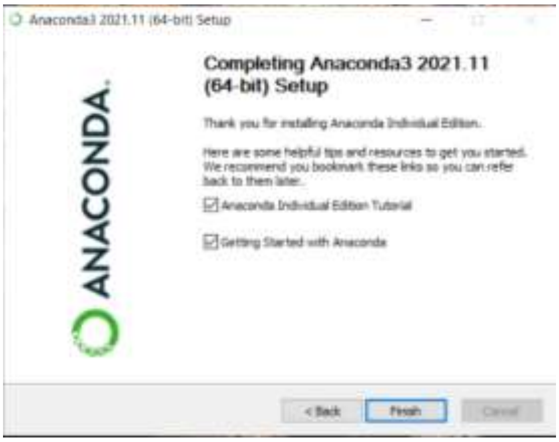
Anaconda3 2021.11 (64-bit)

Anaconda + JetBrains

Working with Python and Jupyter notebooks is a breeze with PyCharm Pro, designed to be used with Anaconda. Download now and have the best data tools at your fingertips.

https://www.anaconda.com/pycharm

Fig.4.7. Anaconda3



Completing Anaconda3 2021.11 (64-bit) Setup

Thank you for installing Anaconda Individual Edition.

Here are some helpful tips and resources to get you started. We recommend you bookmark these links so you can refer back to them later.

☑ Anaconda Individual Edition Tutorial

☑ Getting Started with Anaconda

Fig.4.8. Completing Anaconda3 2021.11 setup

III. Performance Metrics

This section explains the metrics that were utilized in the research.

A. *Precision*

Precision refers to the proportion of important occurrences found between the retrieved instances.

Precision = TP/(TP+FP)

B. *Recall*

The recall is the small percentage of appropriate examples that have been retrieved out of the total amount of relevant examples. The recall Eq. is given.

TP/ (TP + FN) = Recall

C. *F Score*

The real positive rate (recall) and precision are weighted averaged.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

IV. Overview of Data Mining Tools

Data mining has a wide range of applications, ranging from marketing and advertising for goods, functions, and products to artificial intelligence research, biological sciences, crime investigations, and high-level government intelligence.

DOI: 10.5281/zenodo.6183068

Data mining technologies have been developed over decades due to their extensive use and the complexity involved in developing information mining applications. Every tool has its own set of benefits and drawbacks. A series of data mining tools have been developed by a research community and data analysis hobbyists, and they are available for free under one of the open-source licenses now in use. A tool developed using an open-source development paradigm is the outcome of a community effort, not necessarily backed by a single entity, but rather the result of contributions from a global and informal development team. This technique of development enables the incorporation of a variety of experiences. For obtaining data from databases, data boring offers a number of excavation tactics. Data mining tools predict future trends and behaviors, allowing firms to make proactive, informed decisions. The invention and use of data mining algorithms demand the use of incredibly powerful computing tools. As the number of tools available expands, selecting the most suited tool gets more challenging.

V. Logistic regression

Logistic Regression is a categorization problem-solving Machine Learning method. It is a probability-based predictive analytic method.

A Logistic Regression model is similar to a Linear Regression model, but instead of using a linear function, it uses a more sophisticated cost function known as the 'Sigmoid function' or the 'logistic function.'

According to the logistic regression hypothesis, the cost function should be restricted to a range of 0 to 1. As a result, linear functions fail to describe it since they can have a value larger than 1 or less than 0, which logistic regression's hypothesis states is not conceivable.

$$0 \leq h_\theta(x) \leq 1$$

Logistic regression hypothesis expectation

VI. Decision tree

Data is continually segregated based on a given parameter in Decision Trees, a kind of Supervised Machine Learning. The tree can be explained using two entities: decision nodes and leaves. The leaves symbolize the outcomes or consequences of the decisions made. At decision nodes, the data is separated.

VII. Experimental Resuls

In this research project, the analysis and identification of the best classification method are completed, and the findings are presented.

- If Anaconda is not already installed, download and install it.
- When the installation is complete. Anaconda Navigator should now be open.

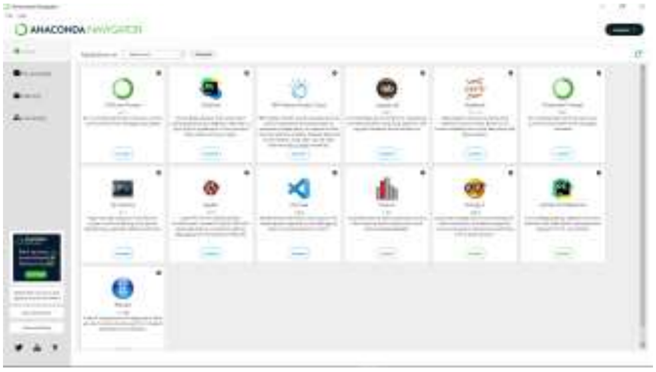
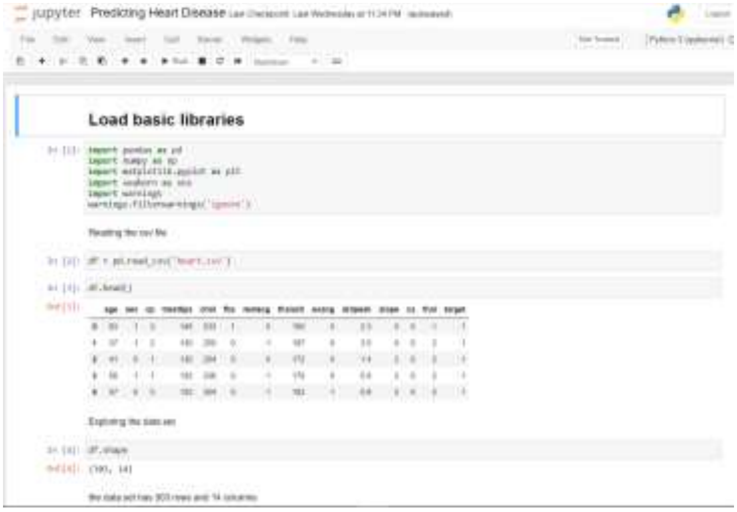


Fig.5.1

- In Jupyter Notebook, click the Launch button. In the browser, a local host will be immediately opened.

Fig.5.2

- Download the dataset. The downloaded dataset should be in the CSV extension.
- Upload the downloaded dataset to Jupyter localhost by selecting the file and clicking the Upload button in the right corner. The file will then be successfully uploaded.
- Then click on the “New” dropdown button and select “Python 3”. Then it redirects to Jupyter notebook space.
- We can write codes in that notebook space to achieve the desired result.



Load basic libraries

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        import warnings
        warnings.filterwarnings('ignore')
```

Reading the csv file

```
In [2]: df = pd.read_csv('heart.csv')

In [3]: df.head()
Out[3]:
   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  ca  thal  target
0   63    1   3       145   233    1        0      150      0      2.3      0   0     1       1
1   37    1   2       130   250    0        1      187      0      3.5      0   0     2       1
2   41    0   1       130   204    0        0      172      0      1.4      2   0     2       1
3   56    1   1       120   236    0        1      178      0      0.8      2   0     2       1
4   57    0   0       120   354    0        1      163      1      0.6      2   0     2       1
```

Exploring the data set

```
In [4]: df.shape
Out[4]: (303, 14)
```

the data set has 303 rows and 14 columns

Fig.5.3

- Here in the above picture, the basic libraries are imported.
- The dataset is named “heart.csv”.
- df = PD.read_csv(‘heart.csv’)

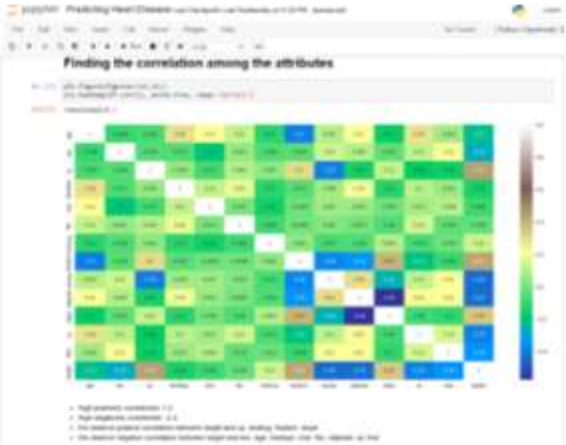
  code is used to read the dataset.
- df.head()

  is used to explore the data set.
- df.shape

  is used to know the data sets rows and columns

Fig.5.4

- To get the description of the given dataset.



Finding the correlation among the attributes

Fig.5.5

- Finding the correlation among the attributes.

Fig.5.6

- To get the Cross Table.

Cross Tables

Fig.5.7

Fig.5.8

- For scaling the data.

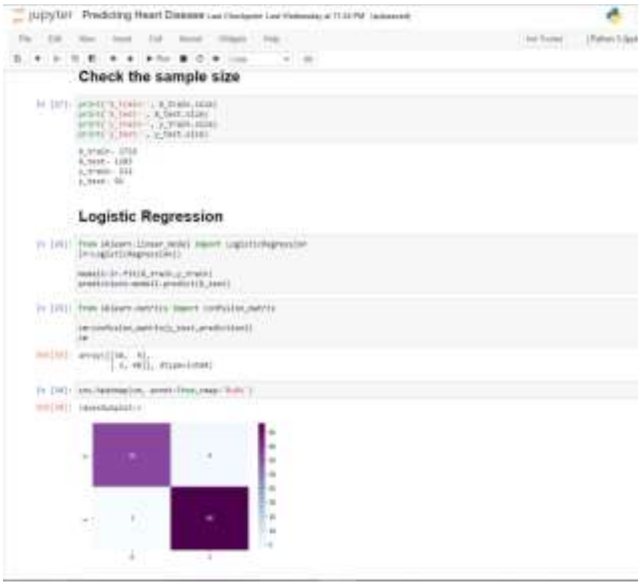
Preparing the data for model

Scaling the data

Fig.5.9

- For checking the sample size.
- Using the Logistic Regression algorithm to predict the heart disease from the given data set.

DOI: 10.5281/zenodo.6183068
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam



jupyter Predicting Heart Disease

Check the sample size

Logistic Regression

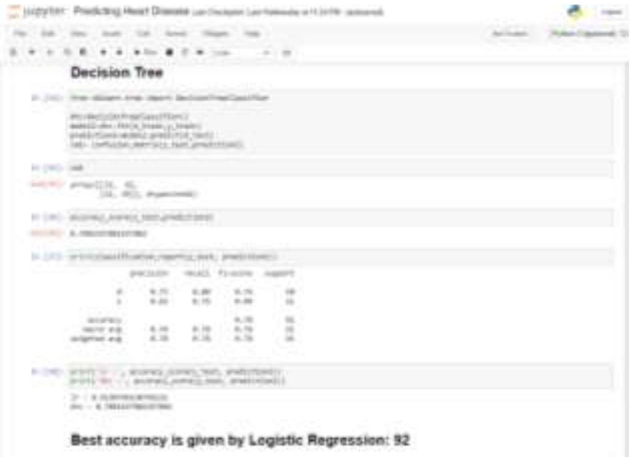
Fig.5.10

- Using the algorithm to find the accuracy.

jupyter Predicting Heart Disease

Fig.5.11

- Using the Decision Tree algorithm to predict heart disease from the given data set.
- Using the algorithm to find the accuracy.



jupyter Predicting Heart Disease

Decision Tree

Best accuracy is given by Logistic Regression: 92

Fig.5.12

- When comparing both the algorithm. Logistic Regression has more accuracy than Decision Tree.

VIII. Conclusion

Finally, the Logistic Regression classification technique holds up well in the study of medical data classification, particularly for heart illnesses, in terms of performance and accuracy.

References

1. H. Benjamin Fredrick David and S. Antony Belcy, "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES", Department of Computer Science and Engineering, Manonmaniam Sundaranar University, India, OCTOBER 2018.
2. Sarangam Kodati & Dr. R. Vivekanandam, "Analysis of Heart Disease using in Data Mining Tools Orange and Weka", Sri Satya Sai University, the Year 2018.
3. Beant Kaur & Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques", Dept of Computer Engineering, Punjabi University, Patiala, India.
4. Radhanath Patra & Bonomali Khuntia, "Predictive Analysis of Rapid Spread of Heart Disease with Data Mining", Predictive Analysis of Rapid Spread of Heart Disease with Data Mining.

DOI: 10.5281/zenodo.6183068
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam