A REPORT PROJECT ON

Heart Diseases Prediction using Machine Learning

Submitted by:

5th Semester Students of

SILIGURI INSTITUTE OF TECHNOLOGY

- **Manasi Dey.(TL)**
- **Sweata Ghosh.**
- **Shairee Roy.**
- **Moumita Dey.**
- **Swarnali Chakraborty.**
- **Payel Sinha.**

Under the supervision of

Mr. ARPAN SAMANTA

Sikharthy Infotech Pvt. Ltd

Department of CSE

Date: 15.09.22

I hereby forward the documentation prepared under my supervision by **Shairee Roy** entitled **Siliguri Institute Of Technology** be accepted as fulfilment of the requirement for the Degree of **Bachelor of Computer Science Engineering(CSE)** from **Siliguri Institute Of Technology** affiliated to **Maulana Abul Kalam Azad University of Technology (MAKAUT).**

Department of CSE

Shairee Roy

Project Guide
Sikharthy Infotech Pvt. Ltd.

Arpan Samanta

ABSTRACT

In recent times, HeartDisease prediction is one of the most complicated tasks in medical field. In the modern era,approximately one person dies per minute due to heart disease. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. This paper makes use of heart disease dataset available.In UCI machine learning repository. The proposed work predicts the chances of HeartDisease and classifies patient's risk level by implementing different data mining techniques such as naïve Bayes, Decision Tree, Logistic Regression and Randomforest. Thus, this paper presents a comparative study by analysing the performance of different machine Algorithms. The trial results verify that Random Forest algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.

INTRODUCTION

The work proposed in this paper focus mainly on various data mining practices that are employed in heart disease prediction. Human heart is the principal part of the human body. Basically, it regulates blood flow throughout our body. Any irregularity to heart can cause distress in other parts of Body. Any sort of disturbance to normal functioning of the heart can be classified as a Heartdisease. In today's contemporary world, heart disease is one of the primary reasons for occurrence of most deaths. Heart disease may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hypertension. According to

The World Health Organization more than 10 million die due to Heartdiseases every single year around the world. A healthy lifestyle and earliest detection are only ways to prevent the heart related diseases. The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis [1]. Even if heart diseases are found as the prime source of death in the world in recent years, they are also the ones that can be controlled

and managed effectively. The whole accuracy management of a disease lies on the proper time of detection.of that disease. The proposed work makes an attempt to detect these heart diseases at early stage to avoid disastrous consequences.Records of large set of medical data created by medical experts are available for analysing and extracting valuable knowledge from it. Data mining techniques are the means of extracting valuable and hidden information from the large amount of data available. Mostly the medical database consists of discrete information. Hence, decision making using discrete data becomes complex and tough task. Machine Learning (ML) which is subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis,detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect heart disease as early stage. This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and thereby analyse the given data.

After analysis of data ML techniques help in heart disease prediction and early diagnosis. This paper presents performance analysis of various ML techniques for predicting heart disease at early stage.

PROBLEM STATEMENT

Heart Disease prediction using machine learning.

MOTIVATION

Being extremely interested in everything having a relation with Machine Learning, the independent project was a great occasion to give me the time to learn and confirm my interest for this field. The fact that we can make estimations, predictions and give the ability for machines to learn by themselves is both powerful and limitless in terms of application possibilities. We can use Machine Learning in Finance, Medicine, and almost everywhere. That's why I decided to conduct my project around Machine Learning.

OVERVIEW OF TECHNICAL AREA

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser,with Zero configuration required, free access to GPUs and Easy Sharing.

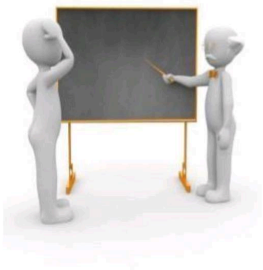
The Jupyter Notebook application allows you to create and edit documents that display the input and output of a Python or R language script. Once saved, you can share these files with others. Python and R language are included by default, but with customization, Notebook can run several other kernel environments.

Python is used for machine learning because A great choice of libraries is one of the main reasons Python is the most popular programming language used for AI. A library is a module or a group of modules published by different sources like PyPi which include a pre-written piece of code that allows users to reach some functionality or perform different actions. Python libraries provide base level items so developers don't have to code them from the very beginning every time. ML requires continuous data processing, and Python's libraries let you access, handle and transform data.

BASIC CONCEPTS OF MACHINE LEARNING

At a high-level, machine learning is simply the study of teaching a computer program or algorithm how to progressively improve upon a set task that it is given. On the research-side of things, machine learning can be viewed through the lens of theoretical and mathematical modeling of how this process works. However, more practically it is the study of how to build applications that exhibit this iterative improvement.

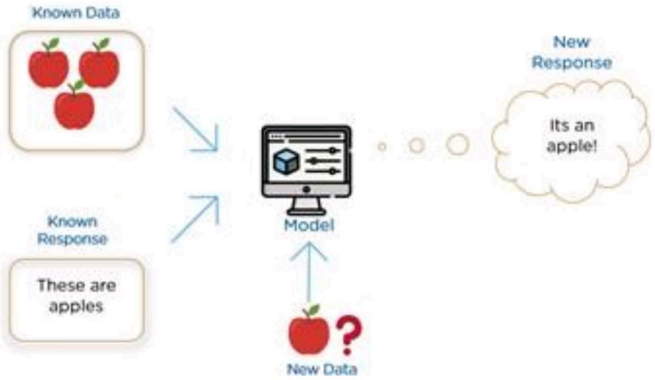
SUPERVISED LEARNING | UNSUPERVISED LEARNING | REINFORCEMENT LEARNING



SUPERVISED LEARNING

Supervised learning is the most popular paradigm for machine learning. It is the easiest to understand and the simplest to implement. It is very similar to teaching a child with the use of flash cards.

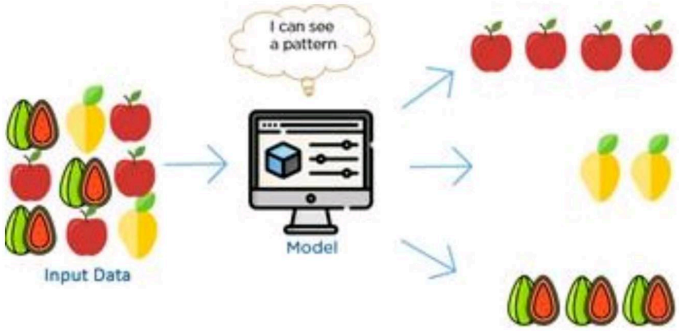
Given data in the form of examples with labels, we can feed a learning algorithm these example label pairs one by one, allowing the algorithm to predict the label for each example, and giving it feedback as to whether it predicted the right answer or not. Over time, the algorithm will learn to approximate the exact nature of the relationship between examples and their labels. When fully trained, the supervised learning algorithm will be able to observe a new, never-before-seen example and predict a good label for it.



UNSUPERVISED LEARNING

Unsupervised learning is very much the opposite of supervised learning. It features no labels. Instead, our algorithm would be fed a lot of data and given the tools to understand the properties of the data. From there, it can learn to group, cluster, and/or organize the data in a way such that a human (or other intelligent algorithm) can come in and make sense of the newly organized data.

What makes unsupervised learning such an interesting area is that an overwhelming majority of data in this world is unlabeled. Having intelligent algorithms that can take our terabytes and terabytes of unlabeled data and make sense of it is a huge source of potential profit for many industries. That alone could help boost productivity in a number of fields.



Reinforcement Learning

Reinforcement learning is fairly different when compared to supervised and unsupervised learning. Where we can easily see the relationship between supervised and unsupervised (the presence or absence of labels), the relationship to reinforcement learning is a bit murkier.

Some people try to tie reinforcement learning closer to the two

by describing it as a type of learning that relies on a time-dependent sequence of labels, however, my opinion is that that simply makes things more confusing.

I prefer to look at reinforcement learning as learning from mistakes. Place a reinforcement learning algorithm into any environment and it will make a lot of mistakes in the beginning. So long as we provide some sort of signal to the algorithm that associates good behaviors with a positive signal and bad behaviors with a negative one, we can reinforce our algorithm to prefer good behaviors over bad ones. Over time, our learning algorithm learns to make less mistakes than it used to.

REGRESSION ANALYSIS IN MACHINE LEARNING

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as temperature, age, salary, price, etc.

DependentVariable: The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called target variable.

IndependentVariable: The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variables, also called as a predictor.

Why do we use Regression Analysis?

As mentioned above, Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future predictions such as weather conditions, sales prediction, marketing trends, etc., for such a case we need some technology which can make predictions more accurately. so for such a case we need Regression analysis which is a statistical method and used in machine learning and data science.

Below are some other reasons for using Regression analysis:

- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It helps to predict real/continuous values.
- By performing the regression, we can confidently determine the most important factor, the least important factor, and how each factor is affecting the other factors.Linear Regression

- Linear regression is a statistical regression method which is used for predictive analysis.
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.
- Linear regression shows the linear relationship between the independent variable (X- axis) and the dependent variable (Y-

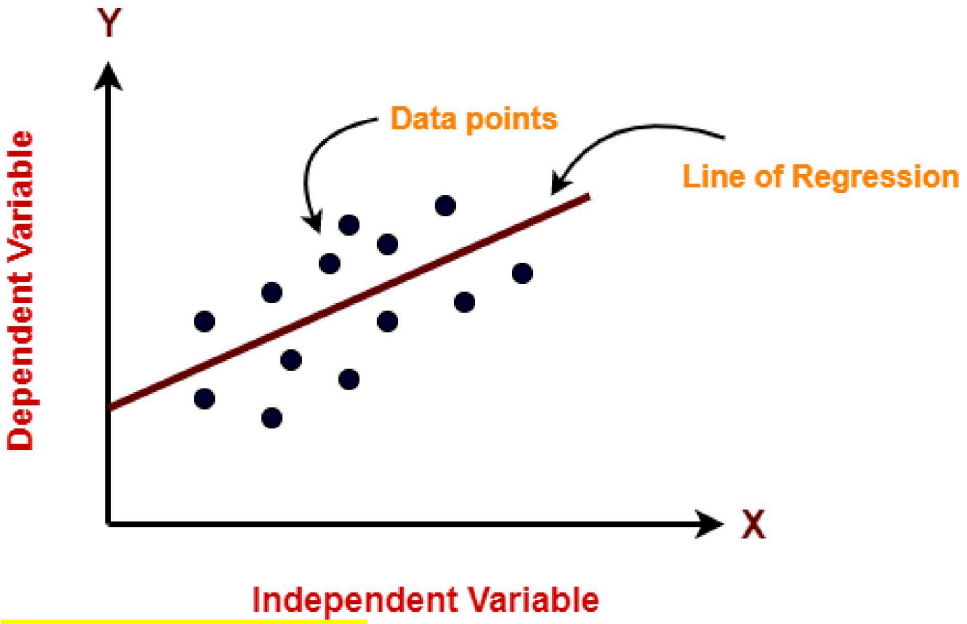
axis), hence called linear regression.

- If there is only one input variable (x), then such linear regression is called simple linear regression. And if there is more than one input variable, then such linear regression is called multiple linear regression.
- The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of the year of experience.

Below is the mathematical equation for Linear regression:

Y= aX+b

Here, Y = dependent variables (target variables), X= Independent variables (predictor variables), a and b are the linear coefficients.



Logistic Regression

- Logistic regression is another supervised learning algorithm which is used to solve the classification problems. In classification problems, we have dependent variables in a binary or discrete format such as 0 or 1.

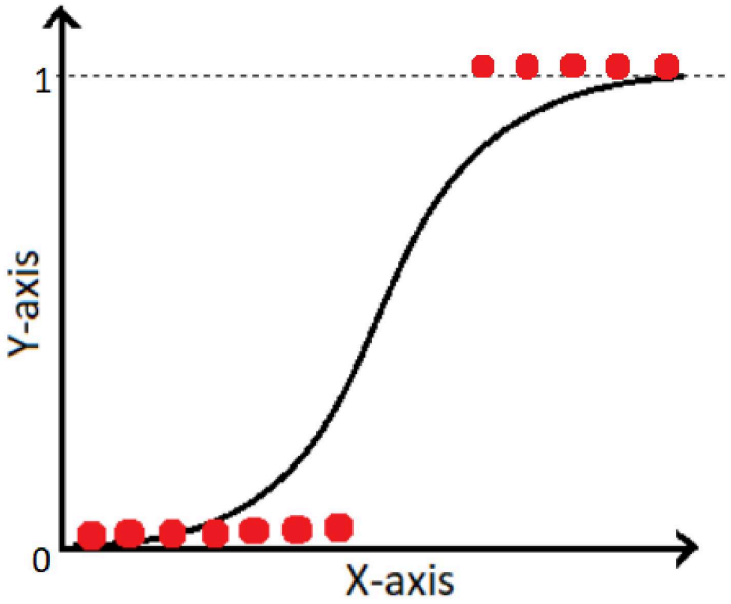
- Logistic regression algorithm works with the categorical variable such as 0 or 1, Yes or No, True or False, Spam or not spam, etc.
- It is a predictive analysis algorithm which works on the concept of probability.
- Logistic regression is a type of regression, but it is different from the linear regression algorithm in the term how they are used.
- Logistic regression uses sigmoid function or logistic function which is a complex cost function. This sigmoid function is used to model the data in logistic regression.

The function can be represented as:

f(x)=1/(1+e^-x)

f(x)= Output between the 0 and 1 value, x= input to the function and e= base of natural logarithm.

When we provide the input values (data) to the function, it gives the S-curve as follows:



PROPOSED MODEL

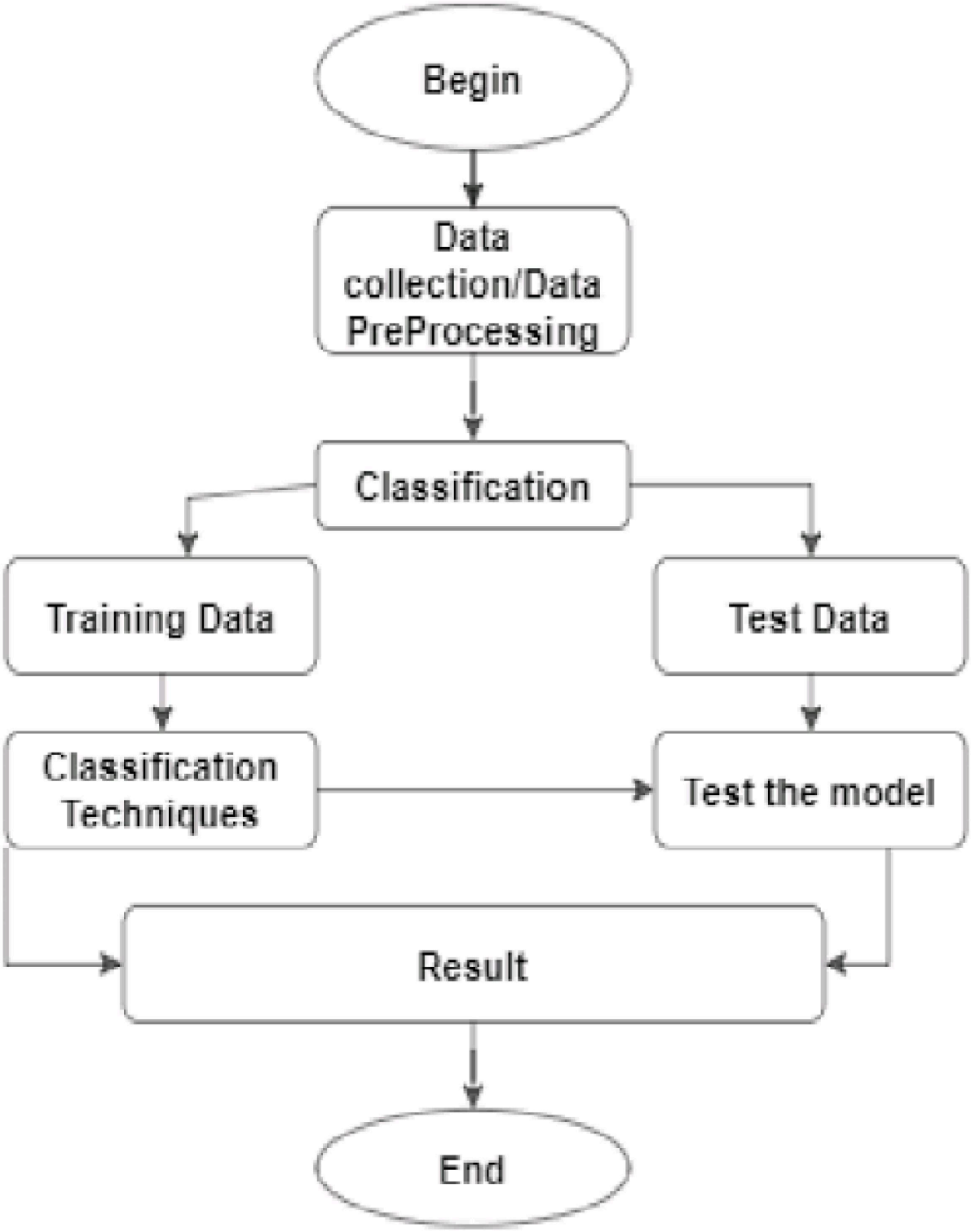
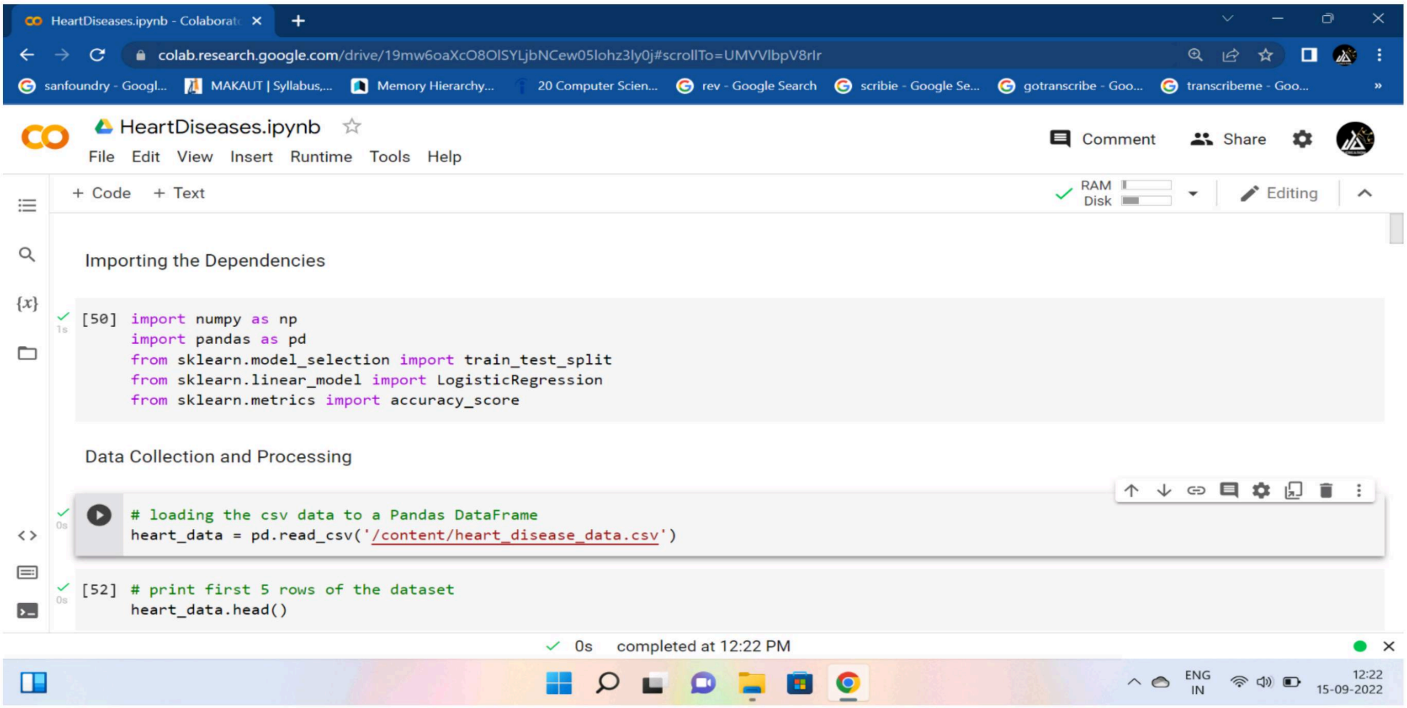


Fig. 1: Generic Model Predicting Heart Disease

DATA COLLECTION AND PROCESSING

S/N	Attribute	Description	Values
1	Age	Age in years	Continuous
2	Gender	Male or Female	1=Male, 0=Female
3	Cp	Chest Pain Type	1= Typical angina 2=Atypical angina 3=Non-anginal pain 4= Asymptomatic
4	Trestbps	Resting blood pressure (in mm Hg)	Continuous
5	Chole	Serum Cholesterol (in mg/dl)	Continuous
6	FBS	Fasting Blood Sugar	1 >= 120 mg/dl 0 <= 120 mg/dl
7	Restecg	Resting electrocardiographic results	0=Normal 1=Having ST_T wave abnormality 2=Left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	Continuous
9	Exang	Exercise induced angina	1 = Yes 0 = No
10	Old peak	ST depression induced by exercise relative to rest	Continuous
11	Slope	The slope of the peak exercise segment	1 = Up sloping 2 = Flat 3 = Down sloping
12	Ca	Number of major vessels colored by fluoroscopy	(1 – 4)
13	Thal	Thallium scan	3 = Normal 6 = Fixed defect 7 = Reversible defect

The Source Code



HeartDiseases.ipynb - Colaborat

colab.research.google.com/drive/19mw6oaXcO8OlSYLjbNCew05lohz3ly0j#scrollTo=UMVVlbpV8rlr

HeartDiseases.ipynb

File Edit View Insert Runtime Tools Help

Comment Share

+ Code + Text

RAM Disk Editing

Importing the Dependencies

```
[50] import numpy as np
     import pandas as pd
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
     from sklearn.metrics import accuracy_score
```

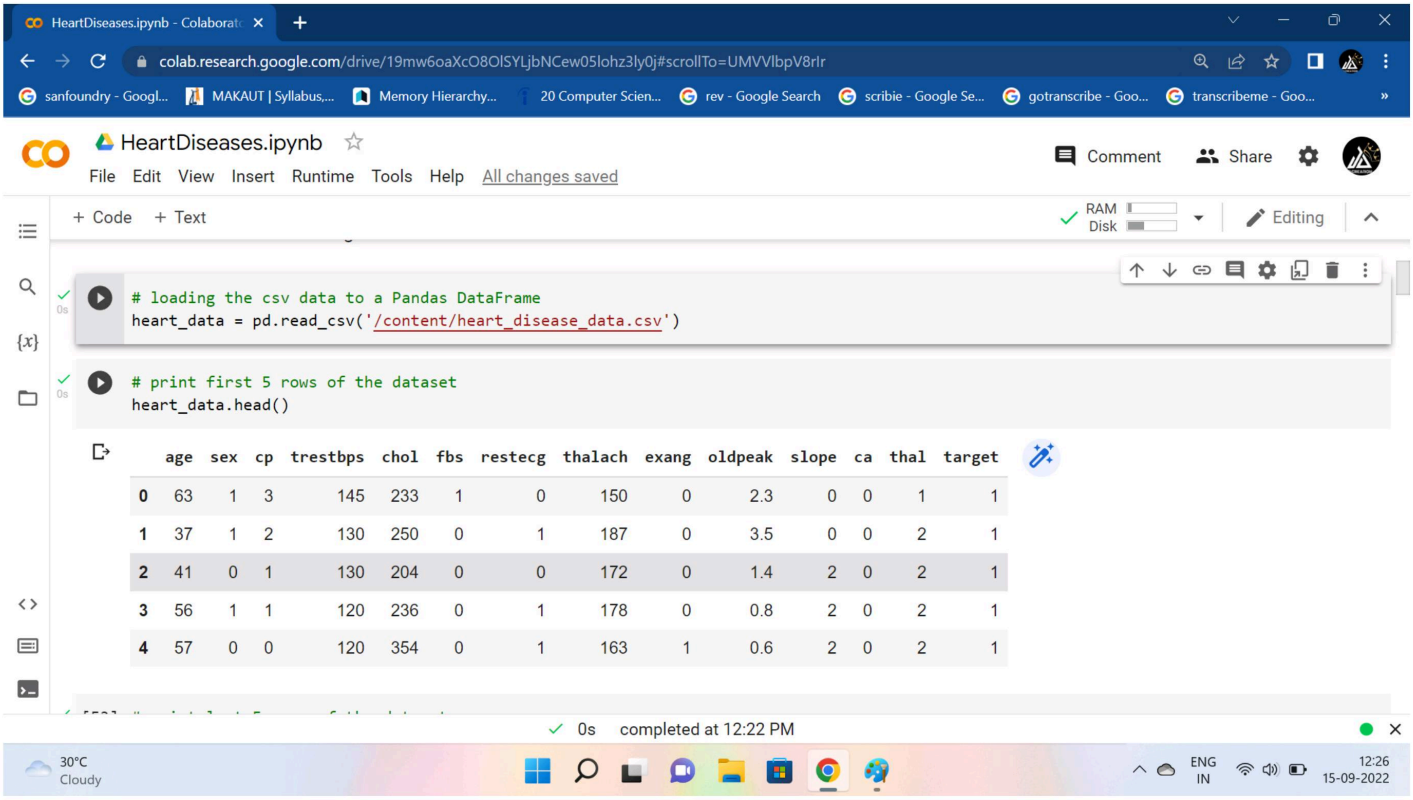
Data Collection and Processing

```
# loading the csv data to a Pandas DataFrame
heart_data = pd.read_csv('/content/heart_disease_data.csv')
```

```
[52] # print first 5 rows of the dataset
     heart_data.head()
```

0s completed at 12:22 PM

ENG IN 12:22 15-09-2022



HeartDiseases.ipynb - Colaborat

colab.research.google.com/drive/19mw6oaXcO8OlSYLjbNCew05lohz3ly0j#scrollTo=UMVVlbpV8rlr

HeartDiseases.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

```
# loading the csv data to a Pandas DataFrame
heart_data = pd.read_csv('/content/heart_disease_data.csv')
```

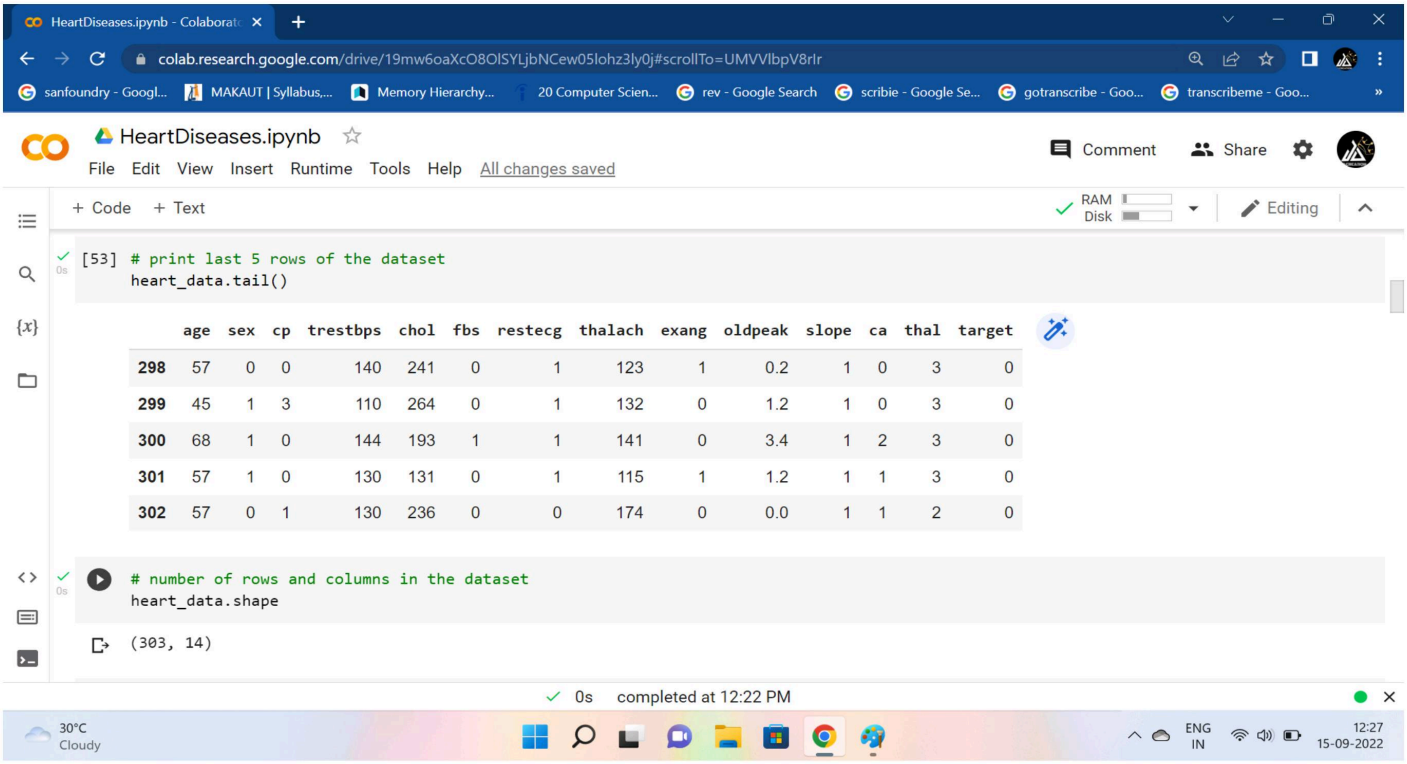
```
# print first 5 rows of the dataset
heart_data.head()
```

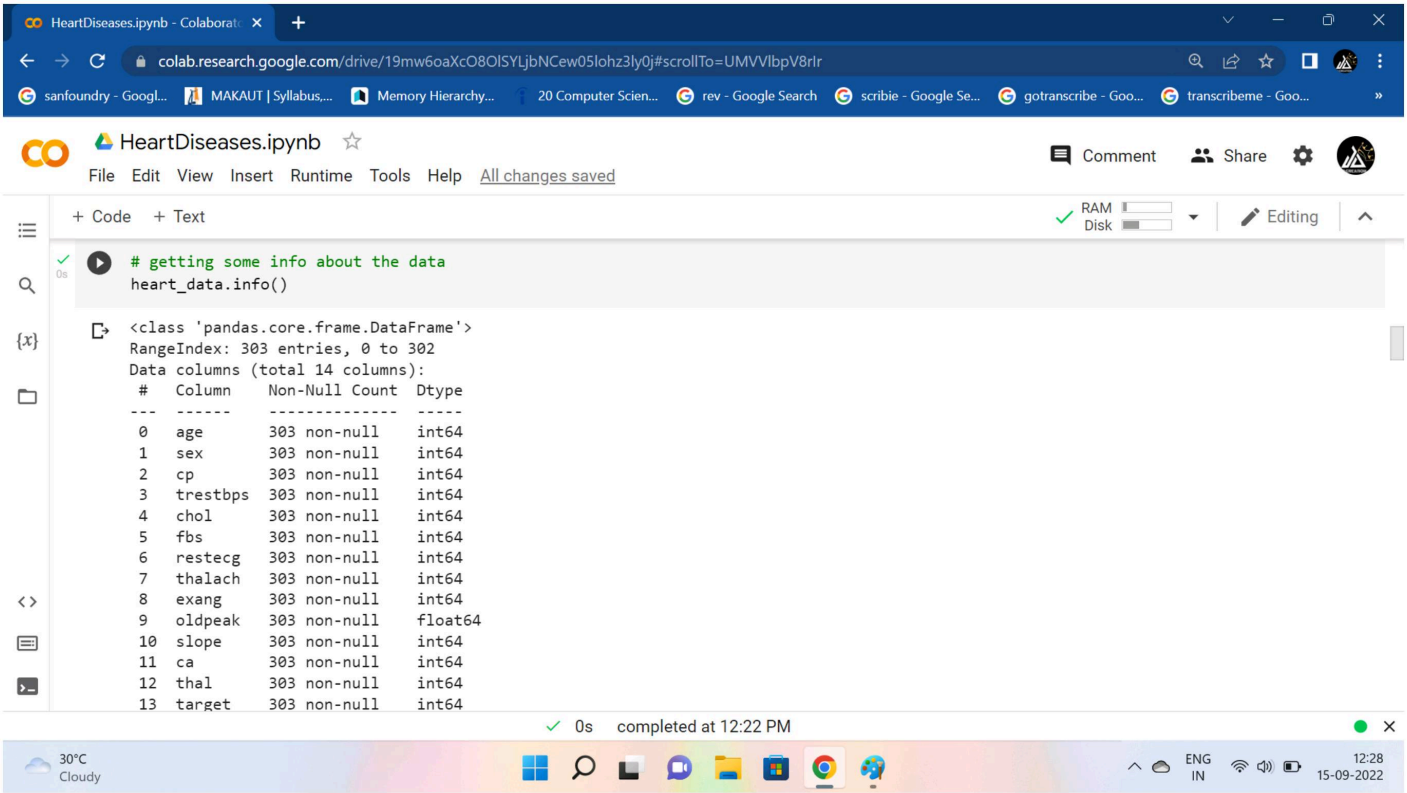
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

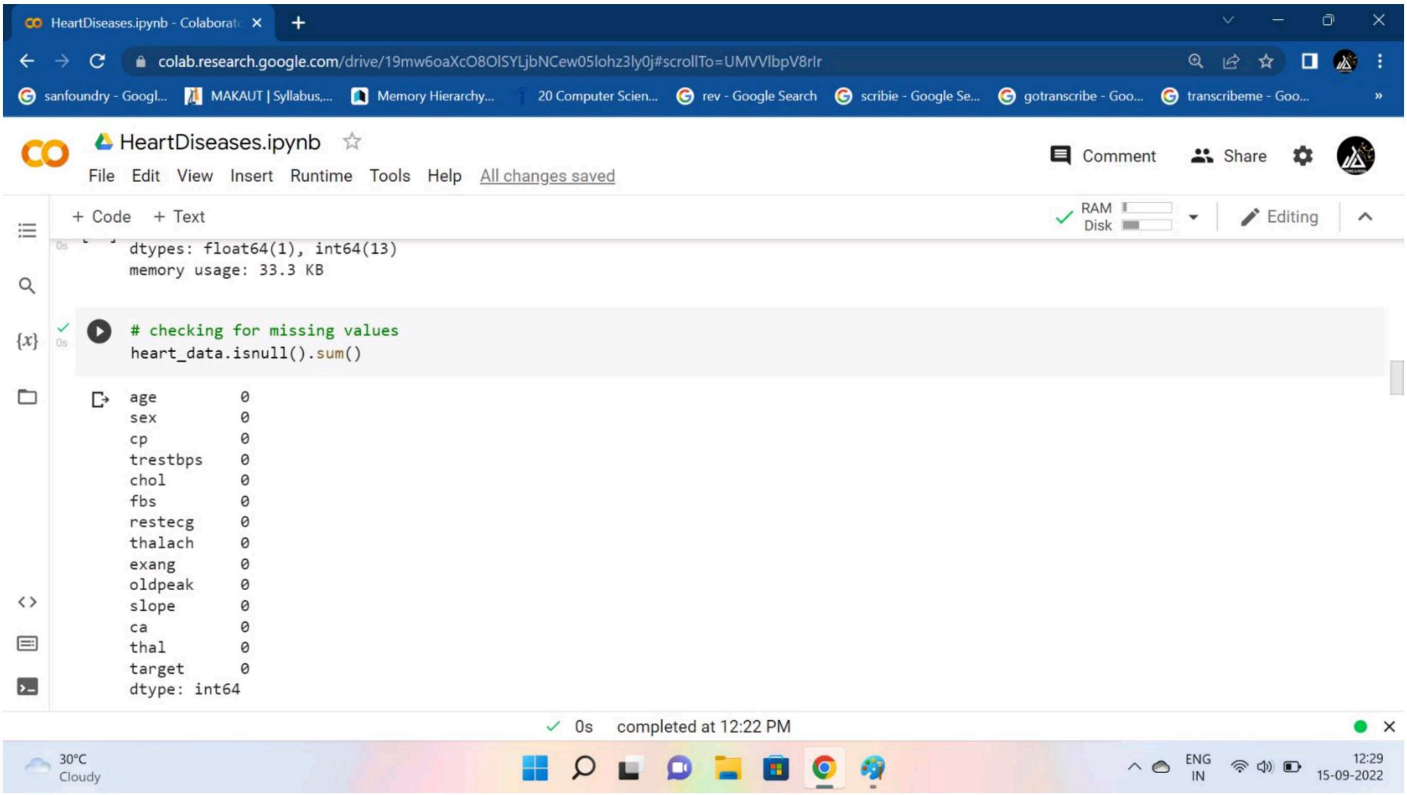
0s completed at 12:22 PM

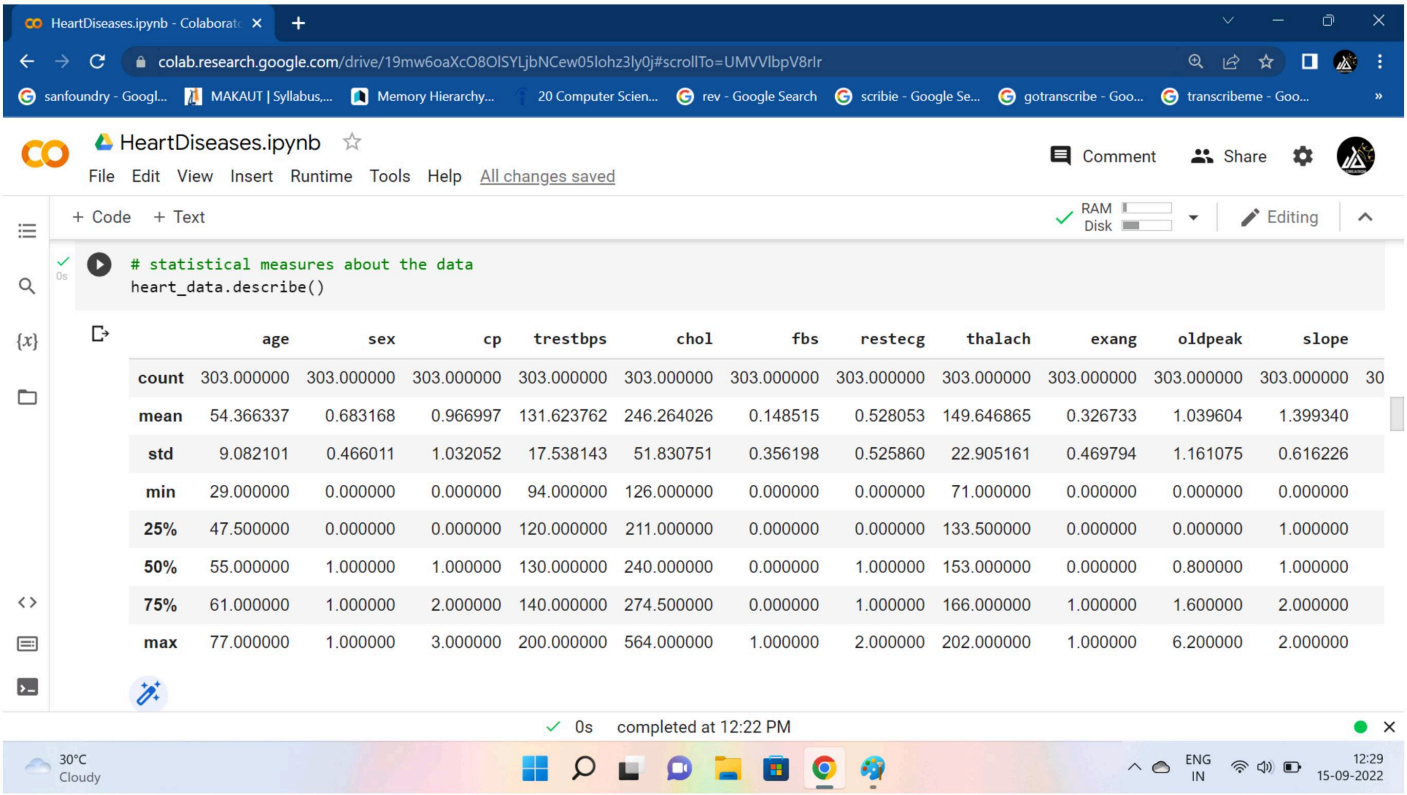
30°C Cloudy

ENG IN 12:26 15-09-2022

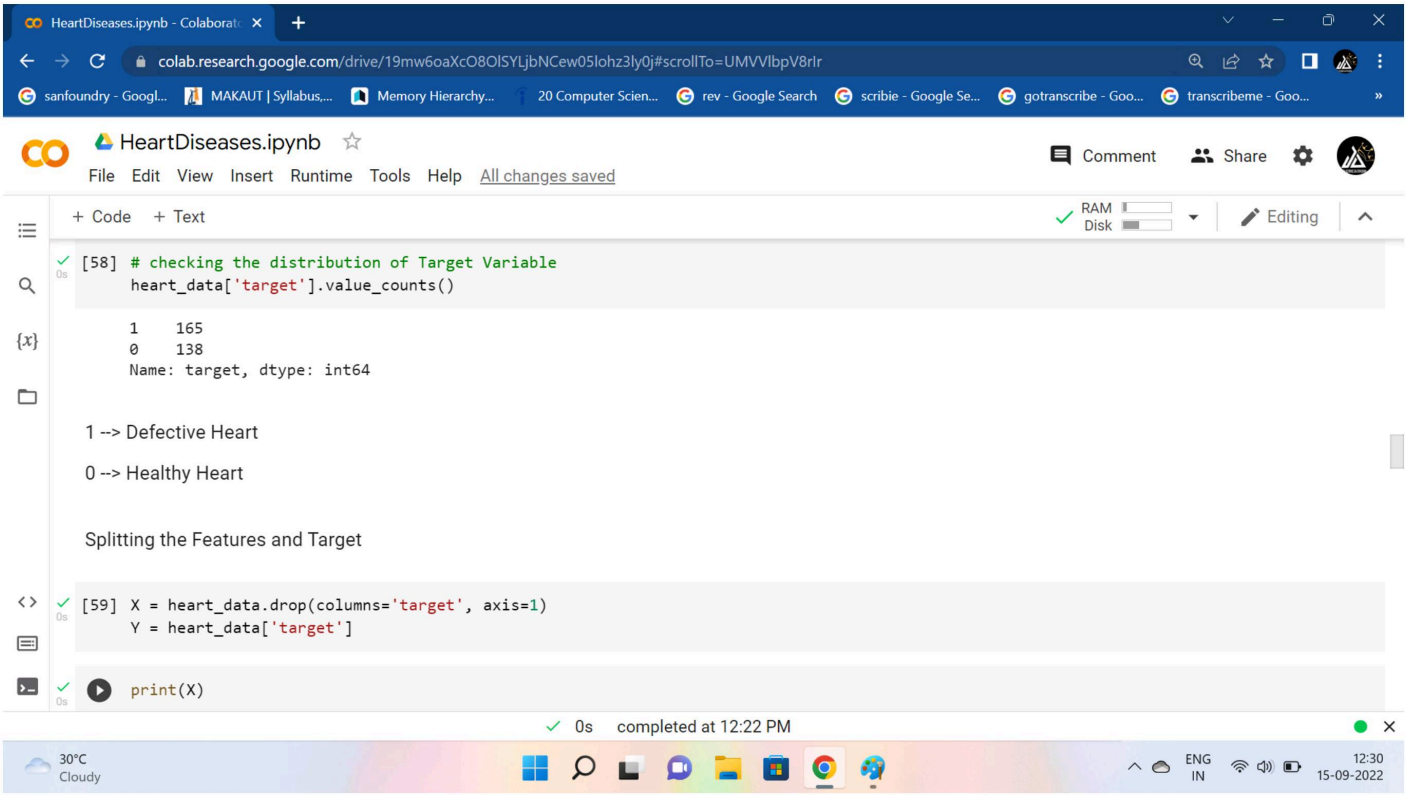

HeartDiseases.ipynb - Colaborat
colab.research.google.com/drive/19mw6oaXcO8OlSYLjbNCew05lohz3ly0j#scrollTo=UMVVlbpV8rIr
HeartDiseases.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
[53] # print last 5 rows of the dataset
heart_data.tail()
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
298 57 0 0 140 241 0 1 123 1 0.2 1 0 3 0
299 45 1 3 110 264 0 1 132 0 1.2 1 0 3 0
300 68 1 0 144 193 1 1 141 0 3.4 1 2 3 0
301 57 1 0 130 131 0 1 115 1 1.2 1 1 3 0
302 57 0 1 130 236 0 0 174 0 0.0 1 1 2 0
number of rows and columns in the dataset
heart_data.shape
(303, 14)
0s completed at 12:22 PM
30°C
Cloudy
ENG
IN
12:27
15-09-2022

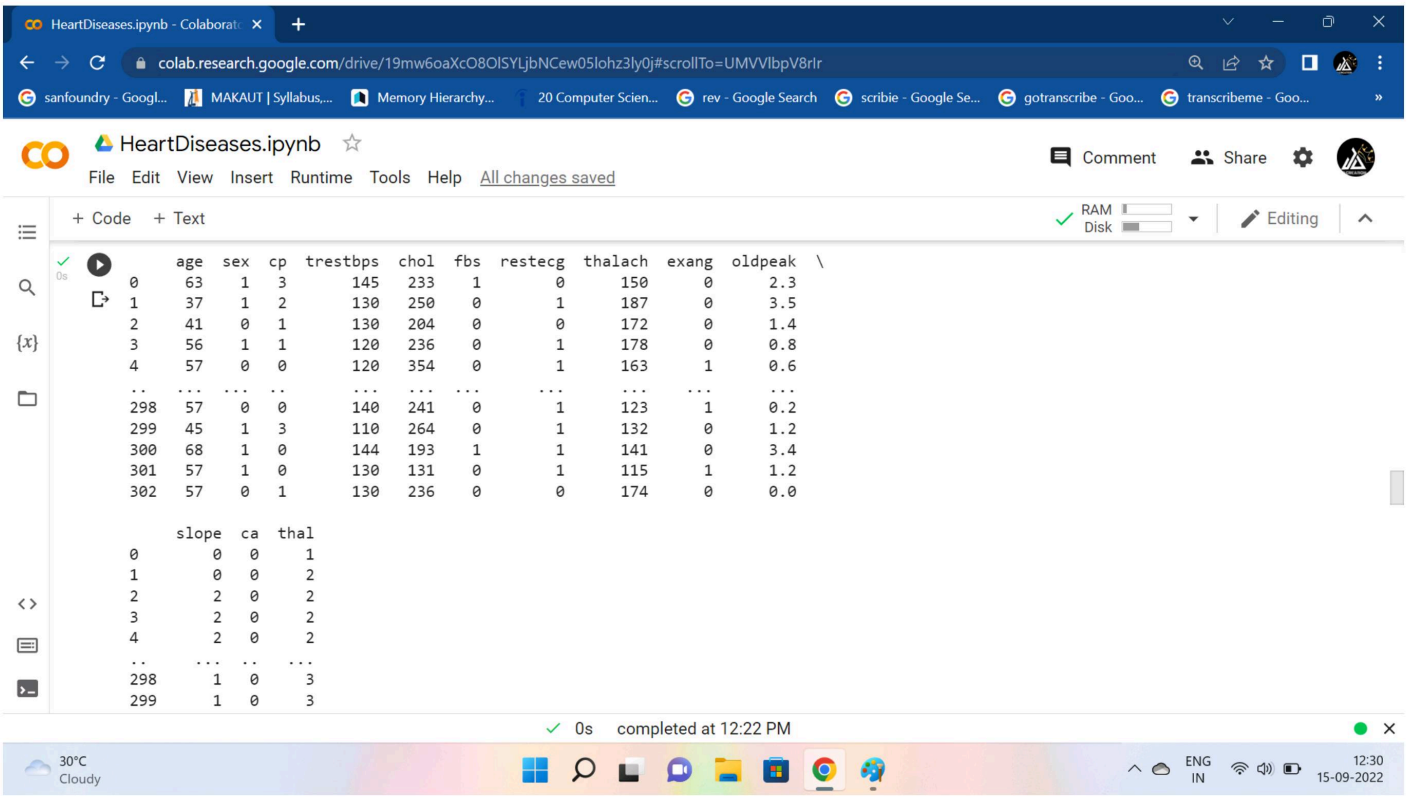

HeartDiseases.ipynb - Colaborat
colab.research.google.com/drive/19mw6oaXcO8OlSYLjbNCew05lohz3ly0j#scrollTo=UMVVlbpV8rIr
HeartDiseases.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
# getting some info about the data
heart_data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
Column Non-Null Count Dtype
--- ------ -------------- -----
0 age 303 non-null int64
1 sex 303 non-null int64
2 cp 303 non-null int64
3 trestbps 303 non-null int64
4 chol 303 non-null int64
5 fbs 303 non-null int64
6 restecg 303 non-null int64
7 thalach 303 non-null int64
8 exang 303 non-null int64
9 oldpeak 303 non-null float64
10 slope 303 non-null int64
11 ca 303 non-null int64
12 thal 303 non-null int64
13 target 303 non-null int64
0s completed at 12:22 PM
30°C
Cloudy
ENG
IN
12:28
15-09-2022

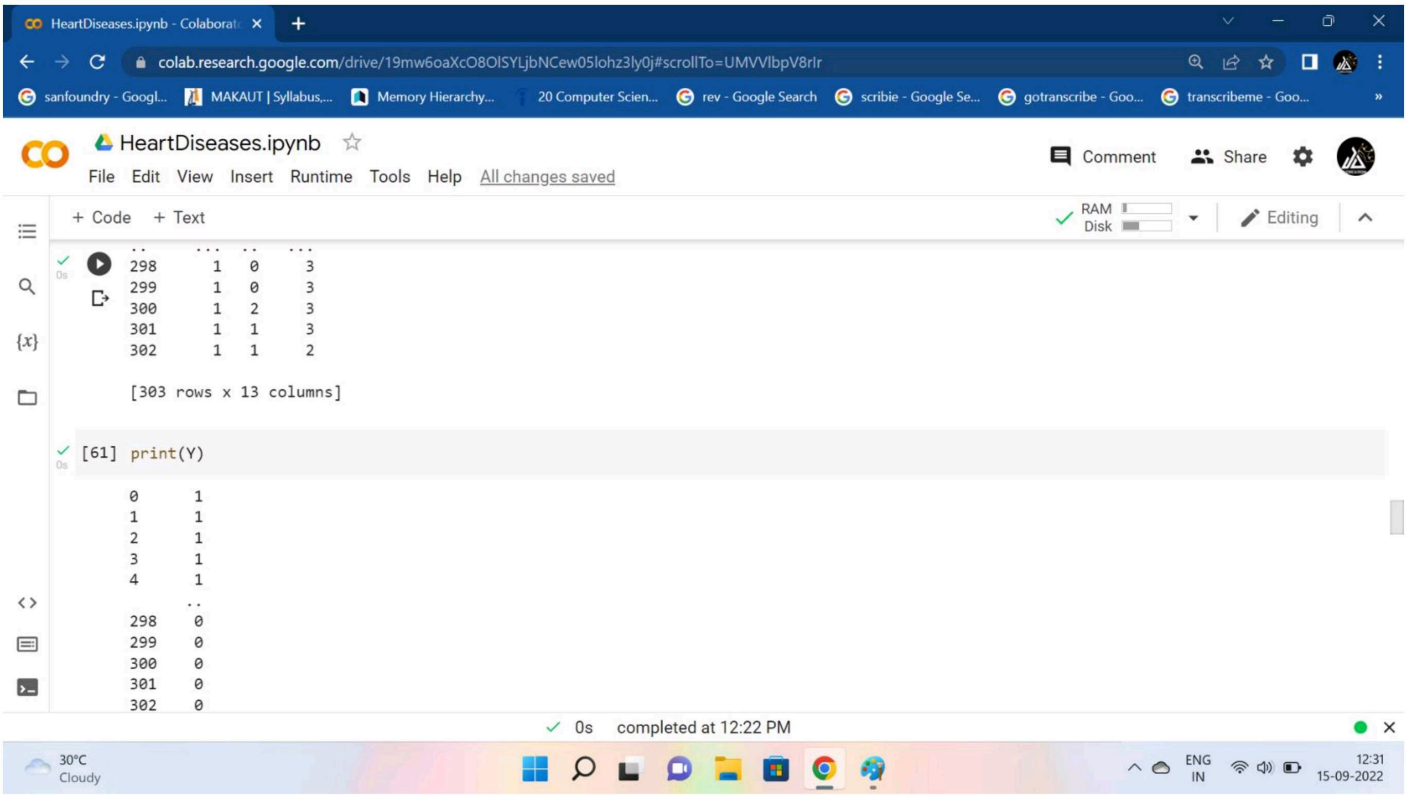


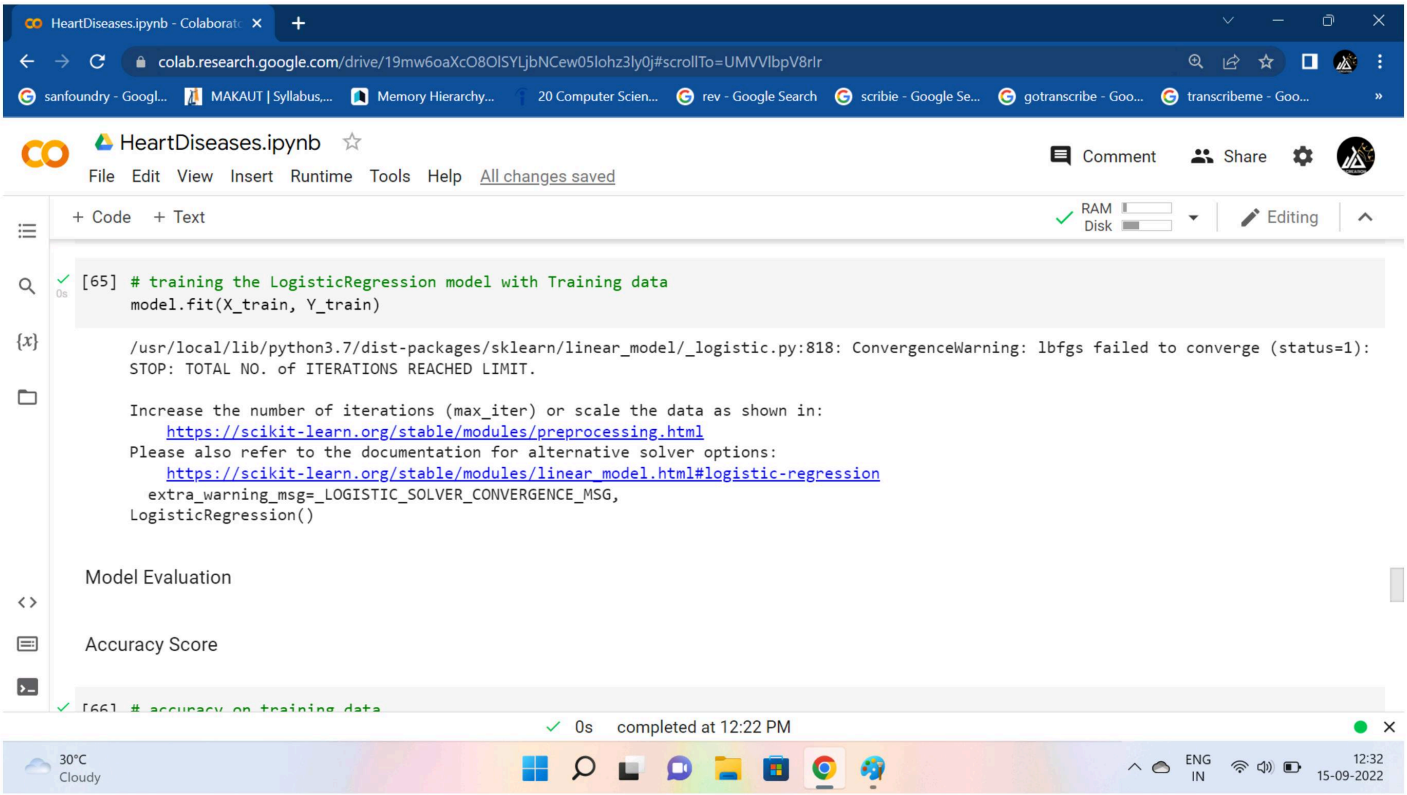


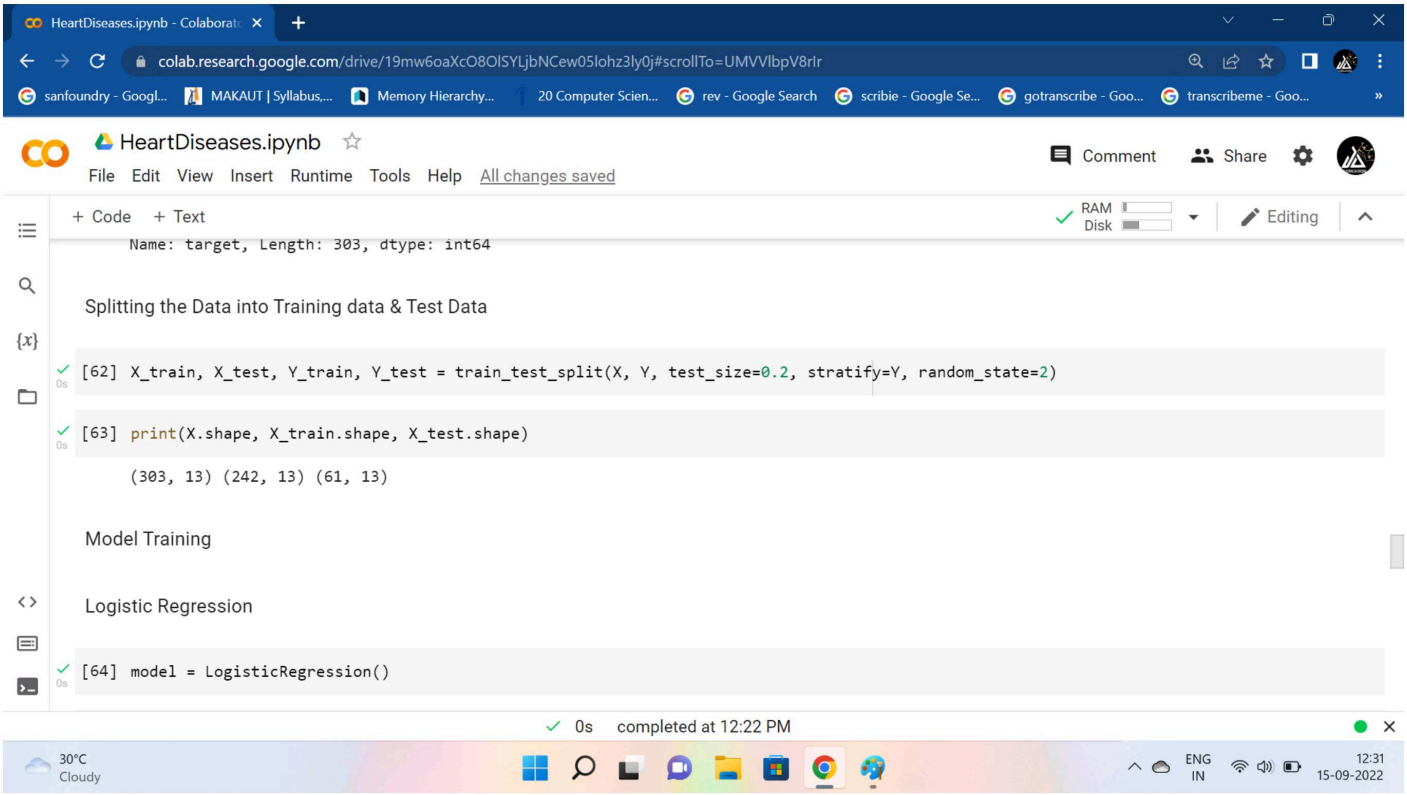
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	30
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	

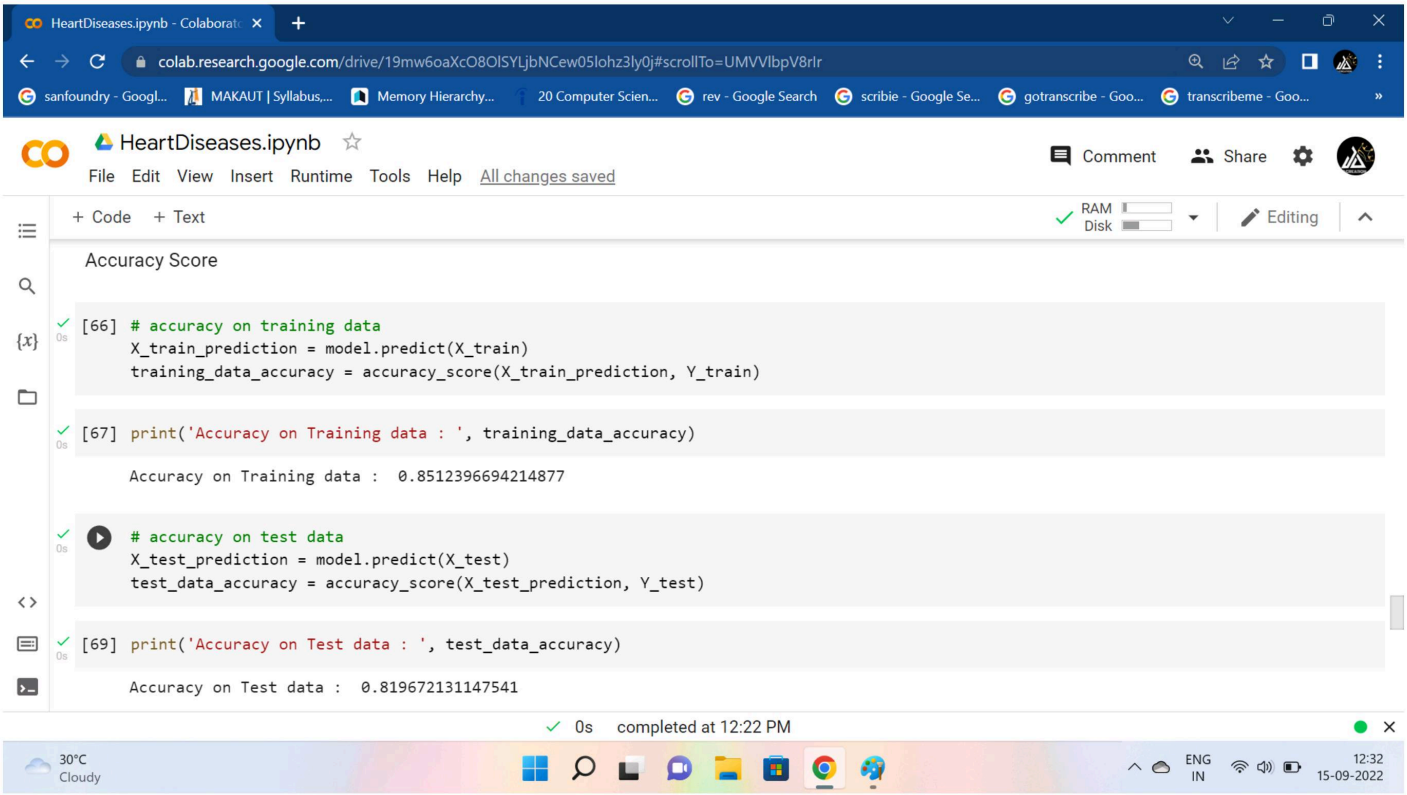

HeartDiseases.ipynb - Colaborat
colab.research.google.com/drive/19mw6oaXcO8OlSYLjbNCew05lohz3ly0j#scrollTo=UMVVlbpV8rlr
HeartDiseases.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
RAM Disk
Editing
[58] # checking the distribution of Target Variable
heart_data['target'].value_counts()
1 165
0 138
Name: target, dtype: int64
1 --> Defective Heart
0 --> Healthy Heart
Splitting the Features and Target
[59] X = heart_data.drop(columns='target', axis=1)
Y = heart_data['target']
print(X)
0s completed at 12:22 PM
30°C Cloudy
ENG IN
12:30 15-09-2022

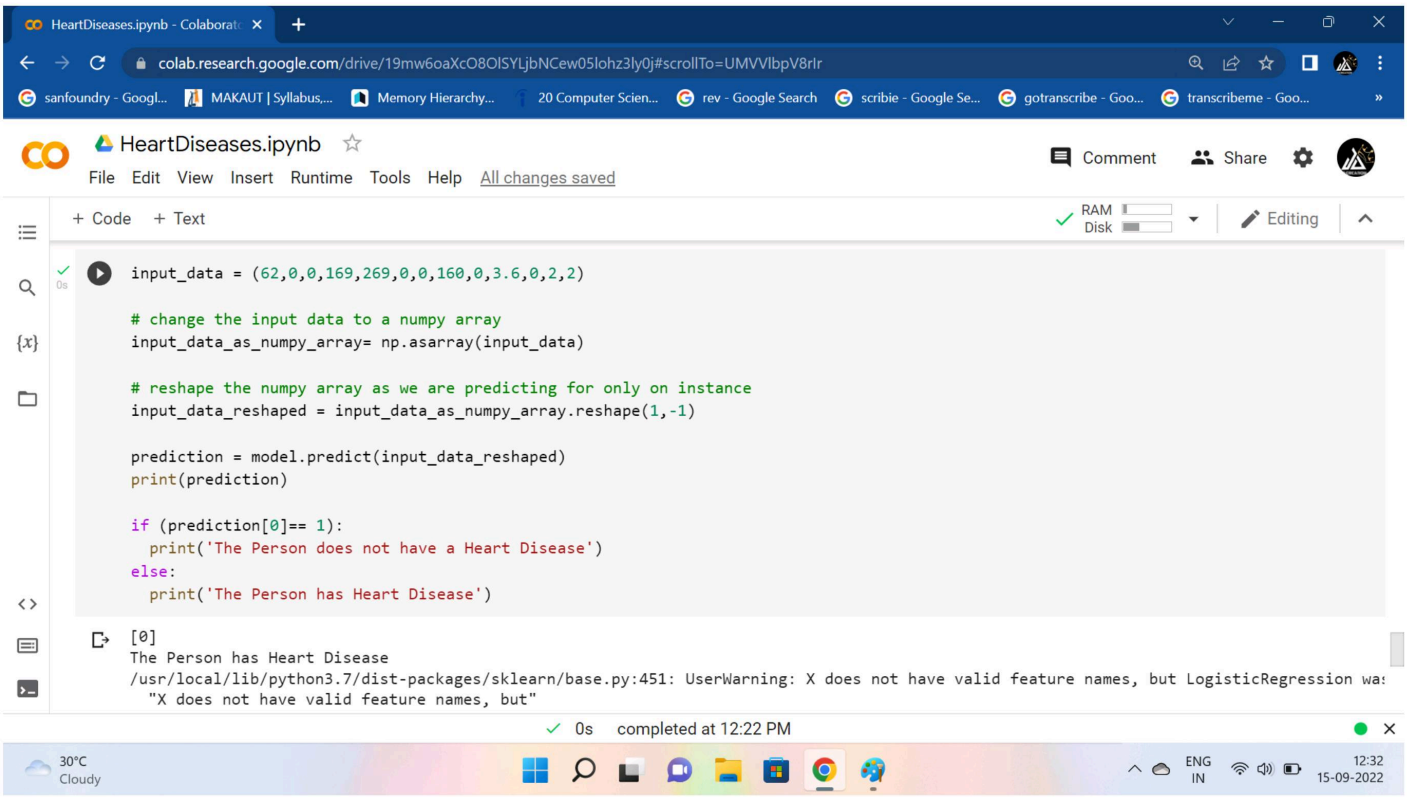

HeartDiseases.ipynb - Colaborat
colab.research.google.com/drive/19mw6oaXcO8OlSYLjbNCew05lohz3ly0j#scrollTo=UMVVlbpV8rlr
HeartDiseases.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
RAM Disk
Editing
age sex cp trestbps chol fbs restecg thalach exang oldpeak \
0 63 1 3 145 233 1 0 150 0 2.3
1 37 1 2 130 250 0 1 187 0 3.5
2 41 0 1 130 204 0 0 172 0 1.4
3 56 1 1 120 236 0 1 178 0 0.8
4 57 0 0 120 354 0 1 163 1 0.6
.. ..
298 57 0 0 140 241 0 1 123 1 0.2
299 45 1 3 110 264 0 1 132 0 1.2
300 68 1 0 144 193 1 1 141 0 3.4
301 57 1 0 130 131 0 1 115 1 1.2
302 57 0 1 130 236 0 0 174 0 0.0
slope ca thal
0 0 0 1
1 0 0 2
2 2 0 2
3 2 0 2
4 2 0 2
.. ..
298 1 0 3
299 1 0 3
0s completed at 12:22 PM
30°C Cloudy
ENG IN
12:30 15-09-2022


HeartDiseases.ipynb - Colaborat
colab.research.google.com/drive/19mw6oaXcO8OlSYLjbNCew05lohz3ly0j#scrollTo=UMVVlbpV8rlr
HeartDiseases.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
298 1 0 3
299 1 0 3
300 1 2 3
301 1 1 3
302 1 1 2
[303 rows x 13 columns]
[61] print(Y)
0 1
1 1
2 1
3 1
4 1
..
298 0
299 0
300 0
301 0
302 0
0s completed at 12:22 PM
30°C
Cloudy
ENG
IN
12:31
15-09-2022


HeartDiseases.ipynb - Colaborat
colab.research.google.com/drive/19mw6oaXcO8OlSYLjbNCew05lohz3ly0j#scrollTo=UMVVlbpV8rlr
HeartDiseases.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
[65] # training the LogisticRegression model with Training data
model.fit(X_train, Y_train)
/usr/local/lib/python3.7/dist-packages/sklearn/linear_model/_logistic.py:818: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max_iter) or scale the data as shown in:
https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG,
LogisticRegression()
Model Evaluation
Accuracy Score
0s completed at 12:22 PM
30°C
Cloudy
ENG
IN
12:32
15-09-2022


HeartDiseases.ipynb - Colaborat
colab.research.google.com/drive/19mw6oaXcO8OlSYLjbNCew05lohz3ly0j#scrollTo=UMVVlbpV8rIr
HeartDiseases.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
Name: target, Length: 303, dtype: int64
Splitting the Data into Training data & Test Data
[62] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2)
[63] print(X.shape, X_train.shape, X_test.shape)
(303, 13) (242, 13) (61, 13)
Model Training
Logistic Regression
[64] model = LogisticRegression()
0s completed at 12:22 PM
30°C
Cloudy
ENG
IN
12:31
15-09-2022


HeartDiseases.ipynb - Colaborat
colab.research.google.com/drive/19mw6oaXcO8OlSYLjbNCew05lohz3ly0j#scrollTo=UMVVlbpV8rIr
HeartDiseases.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
Accuracy Score
[66] # accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
[67] print('Accuracy on Training data : ', training_data_accuracy)
Accuracy on Training data : 0.8512396694214877
# accuracy on test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
[69] print('Accuracy on Test data : ', test_data_accuracy)
Accuracy on Test data : 0.819672131147541
0s completed at 12:22 PM
30°C
Cloudy
ENG
IN
12:32
15-09-2022


HeartDiseases.ipynb - Colaborat
colab.research.google.com/drive/19mw6oaXcO8OlSYLjbNCew05lohz3ly0j#scrollTo=UMVVlbpV8rIr
sanfoundry - Googl...
MAKAUT | Syllabus,...
Memory Hierarchy...
20 Computer Scien...
rev - Google Search
scribie - Google Se...
gotranscribe - Goo...
transcribeme - Goo...
HeartDiseases.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
input_data = (62,0,0,169,269,0,0,160,0,3.6,0,2,2)
# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)
# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)
prediction = model.predict(input_data_reshaped)
print(prediction)
if (prediction[0]== 1):
print('The Person does not have a Heart Disease')
else:
print('The Person has Heart Disease')
[0]
The Person has Heart Disease
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but LogisticRegression was
"X does not have valid feature names, but"
0s completed at 12:22 PM
30°C
Cloudy
ENG
IN
12:32
15-09-2022

Reference:

- YouTube
- Google
- Kaggle

Conclusion:

The aim of machine learning is to automate analytical model building and enable computers to learn from data without being explicitly programmed to do so. Conclusion **Machine learning is a powerful tool for making predictions from data**. This project predicts people with cardiovascular disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc.

Thank You