A close up of a flower

Description automatically generated

**Predicting the Coronary Heart disease chances in Logistic Regression using Jypter lab**

**Presented by-**

**Nirvi Khiyani [khyiyanin@lsbu.ac.uk]**

**Student Id: 3901390**

**Module Code: CSI\_7\_SAM\_2021**

**MSc Data Science**

**Guided By -Zhu Yongxu**

Table of Contents

1. Abstract………………………………………………………………………………………………………………………………3
2. [Introduction 3](#_Toc56816019)

3. Logistic Regression……………………………………………………………………………………………………….3

4. Literature review…………………………………………………………………………………………………………………..6

5. Methodology…………………………………………………………………………………………………………………………7

5.1 Data Acquisition………………………………………………………………………………………………………..7

5.2 Data Pre-processing…………………………………………………………………………………………………..7

5.3:Machine Learning model selection …………………………………………………………………………..7

5.4 Evaluating the data …………………………………………………………………………………………………..7

6. Implementation:

6.1 Implementing Logistic Regression……………………………………………………………………………….8

6.2: Modelling……………………………………………………………………………………………………………….11

6.3 Results analysis and discussion……………………………………………………………………………………14

7 Conclusion………………………………………………………………………………………………………………………..15

8 References……………………………………………………………………………………………………………………….15

9 Appendix………………………………………………………………………………………………………………………….

1. **Abstract:**

The traditional way of analyzing binary classification problem is Logistic regression, however, it comprises of many iterations and huge amount of data. In this report, we will study the logistic regression model, its mathematical representation and the implementation. In order to actualize the predictive and probabilistic approach , this report will adhere logistic regression as machine learning the most significant heart diseases predicators and hence, using logistic regression for predicting the overall risks. Therefore, one of the classification algorithms like logistic regression is implemented to identify the predicators .For this we have used the Framingham dataset that comprises of 4240 patient’s data and 15-attributes. Finally, the analysis of data and implementation of code is carried out in Jypter in python

1. **Introduction:**

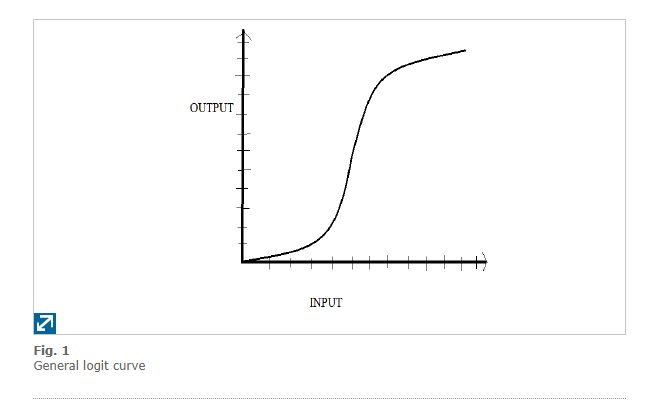
The world of technology for automation is advancing at the vast pace and the one of the most striking features in the field of automation is, Artificial Intelligence. The concept of simulating human like intelligence in computer and to make machine thinks like humans that they intrigue with developers and scientist to achieve the goals is Artificial Intelligence. The aim was to work with man along with the intelligence combined to result more revelations in the era of technology.

AI goes back to the time of the inventions of computers and from that time AI is almost there in each field starting from Banking to Hospitality. Moreover, from the past few decades, researchers have complete understanding that results of developing the model and then implementing that to the actual or real problems. There are various domains where AI is used and they are neural networks, Fuzzy logic, Machine learning, expert systems and NLU (Natural Language Processing).

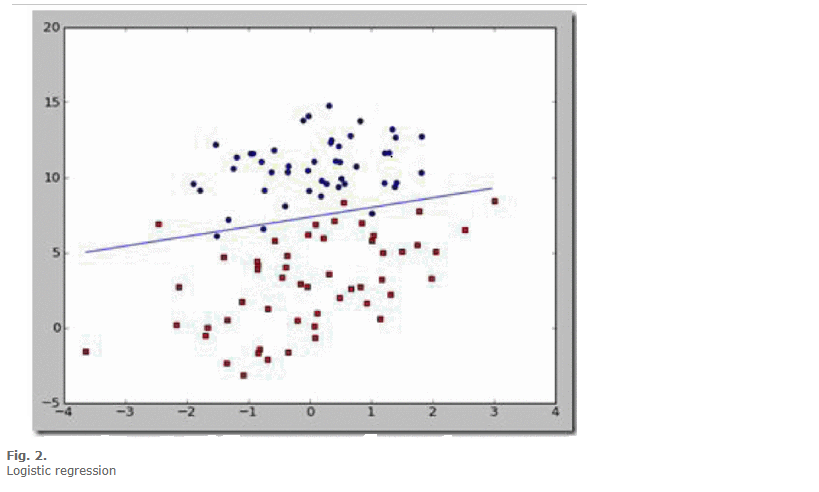
In this, machine learning will be exploited and would be implemented actual problems. ML is the study of computer algorithms that automatically improve with experience and is a subset of [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence). Also, ML tool facilitates analytical models’ developments and that too without explicit programming. There are various algorithms in machine learning and are used in to utilize machine learning capabilities to gain high-sales and statistical analysis gives the results in the positive way. Moreover, many intuitions are generating much more data and hence, due to huge data its manual exploitation id very difficult. Then, here machine learning helps to sough and find a solution with the analytical modelling capability.

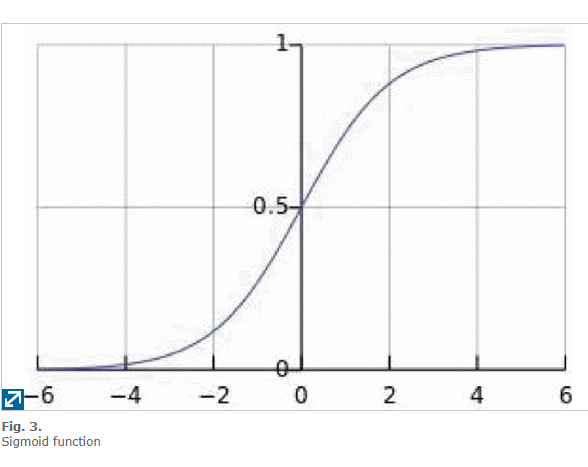
1. **Logistic Regression:**

Logistic regression is one of the predictive modelling techniques is mainly for classification and is also termed as logit regression or logistic model. The inputs to this are independent features and output will be categorical. In logistic regression model, the probability of occurrence of categorical output is achieved by fitting feature in logistic curve. The below mentioned figure is the logistic curve:



It is classification-based algorithm which consist of different input data and then the output is obtained by a function called sigmoid function. For an instance, the output of the data should be binary as 0 and 1’s. Hence, as per my data analysis and need, the range of the function argument can be from positive infinity to negative infinity and the range of dependent variable is either 0 or 1.Moreover, there are various functions which satisfies these conditions and the more inherent one’s is 0-1 step but it is not steerable. Hence, most probably Sigmoid function is use





From fig 3, it can be stated that the value of σ=0.5

 If  z is greater than 0, then functions value will be 1 and will be class 1 as the z’s value increases. Moreover, when z is less than 0, then functions value will be 1 and the class will 0, that satisfy the classification requirements for the function.

The process of classification of logistic regression can be stated as under

It can be presumed that input data characteristics which can be demonstrated that ranges from xo till xn where all the features are multiplied by their coefficient of regression which ranges from wo to wn and then we need to sum z to make the sigmoid function





1. **Literature review:**

In Machine Learning, classification is the classical problem, and, in this report, I would be describing the binary classification for logistic regression . There were various models developed and implemented in various fields. A logistic regression model was developed  that was used to actualize the probabilistic predictive approaches for a loan approval prediction problem (Vaidya, Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval, 2017)

A logistic model was designed that detects HIV/AIDS epidemic prevalence in Maseno Community where they used Cronbach’s alpha for the reliability of the testing (KING’ORA, 2017) In this they came to the conclusion that HIV was associated with TB, marital\_status, origin and age.

A traditional logistic model was developed that predicts whether consumer accepts certain car and provided the references for the binary classification problems (Zou, Hu, Tian, & Shen, 2019).They got the accuracy rate of near about 95.15%

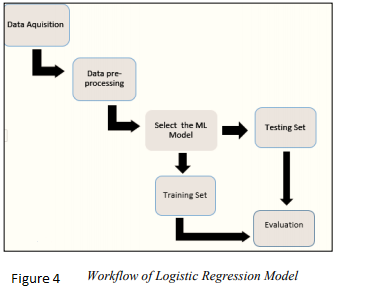
A logistic regression analysis model was implemented in Python on the dataset that is imbalanced and determined various thresholds for classifications problem as the proportion of data on imbalanced dataset (Zhang, et al., 2019).In this they implemented SMOTE algorithm and obtained accuracy rate upto 95%.

Logistic model for predicting heart diseases was implemented in python and the accuracy of the model was 0.87 (Nishadi). This was implemented dataset of Framingham consisting of 4238 records patient’s data and15 attributes.

1. **Methodology:**

Machine learning Model building workflow:

The below mentioned figure states the various steps involved in the machine learning in logistic regression.



**5.1 Data Acquisition:**

I have collected the heart diseases’ data from Kaggle website that has 4240 row and 15 attributes that defined whether the person is suffering from heart disease.

**5.2 Data Pre-processing:**

To have the accurate machine learning model the data should be preprocessed first that includes the cleaning of data. This consists of outliers, missing values , inconsistent data as well as noisy data.

**5.3 Machine Learning model selection:**

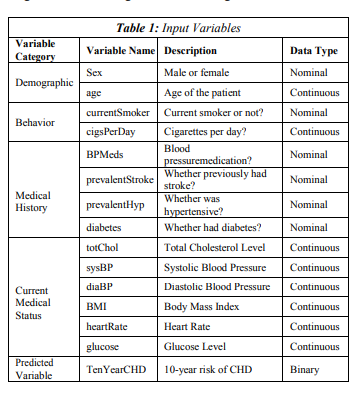
In this step, the machine learning model needs to be decided as per the research and data. So, in this I have selected Logistic regression because the output of my model is supposed to be binary i.e.; yes or no and 1 or 0.

**5.4 Evaluating the data:**

In evaluation we need to divide the data into test and train set because it helps us to minimize the effects of the data and results in better understanding of the model.

**a)Understanding the Dataset and the study of the dataset variables:**

In the Framingham dataset there is one variable that needs to be predicted and 14 IVS. The machine learning model depends on the DV identification. Here, in this binary classification algorithm is used because datatype of the target attribute is categorical.



1. **Implementation:**

**6.1 Implementing Logistic Regression**

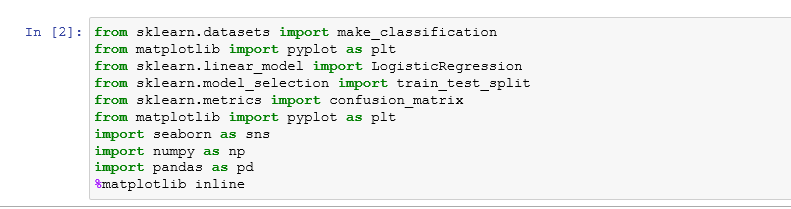
For implementing the logistic regression, I have implemented it in Python in Jypter notebook and have used some of the relevant machine learning libraries in the Python coding. As a part of implementation of model, first the dataset needs to be analyzed and needs to be pre-processed if required. Then, next is selecting the appropriate to ML model and finally, analyzing the results to predict the accuracy.

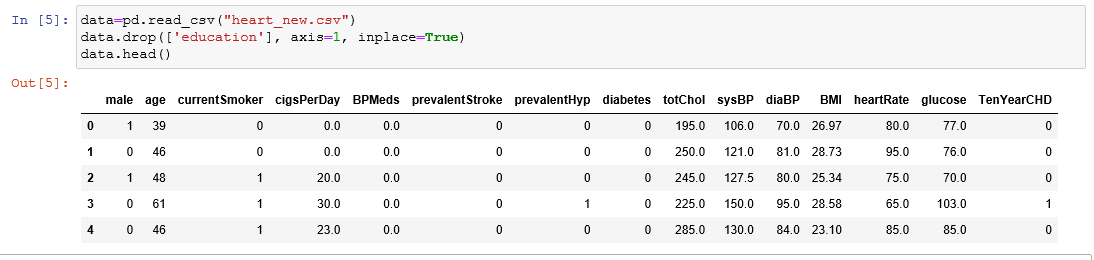
**a) Data Analysis:**

The Framingham dataset is used in this report which is exported from Kaggle.com site and consist of 4240 and 15 attributes. The data analysis part is done in Python by using Jypter lab. Below mentioned steps are implemented for developing a logistic regression model.

Step1: Importing the libraries and data:

In this firstly, we have imported and loaded all the required libraries that are used as supportive application. Next, the prediction heart data is loaded and then finally eliminated the education column because education does not have any significance on heart disease.





**b) Exploration of Data:**

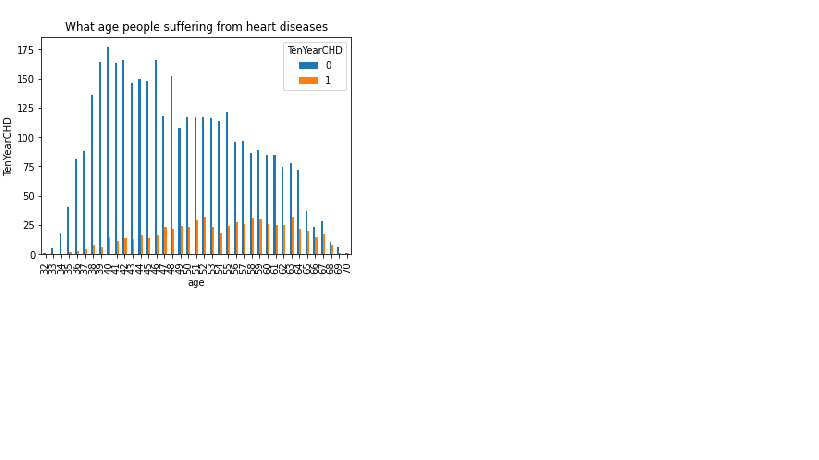
There are several attributes which are significant and are mandatory to predict the heart disease in an individual and they are age, sex, BMI, heartrate, etc. Hence, there exists the relation between the above-mentioned attributes which can be stated from the below graphs taken from Jypter notebook:

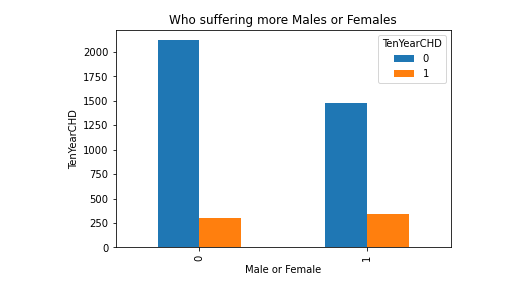
Graph 1 : It states the relationship the between the age and at what age people suffer from heart disease and it can be predicted that the younger individuals are more likely to suffer.

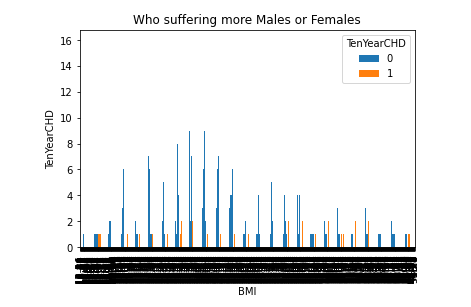
Graph 2: : It states the relationship the between the gender and which gender suffer’s more from heart disease and it can be predicted that the mostly males suffer more from heart diseases than females.

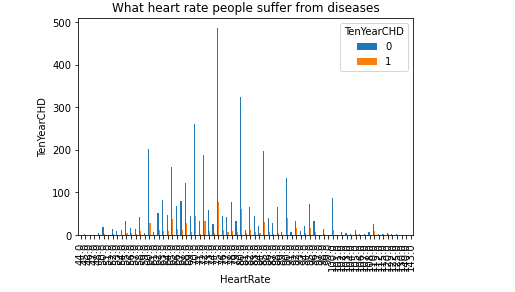
Graph 3: It states the relationship the between the BMI and who suffer more from heart disease.

Graph 4: It states the relationship between heartrate and people who suffer from heart disease who suffer from heart disease and it can be predicted that people whose heart rate is less than 60 and greater than 100 mostly suffer from heart disease and mostly in this people have heart rate less than 60 and hence, suffering from heart disease.



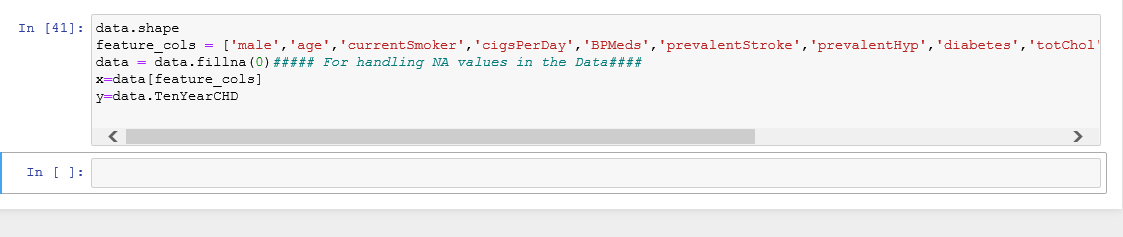




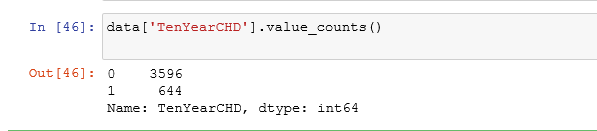


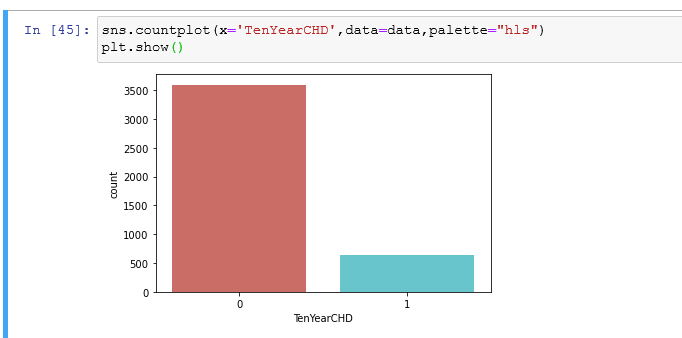
**c)Data Cleansing:**

Data Cleansing includes finding and replacing outliers, missing values , inconsistent data and noisy data. The missing values in the dataset was identified by the data frame Pandas. For this in Jypter first I have created one variable named feature\_cols in which I have taken all the attributes and y as the target variable. Next, for missing values and NA values in data I have used fillna() function and kept those values as 0. This can be stated in the below mentioned snip:



Next, we need to calculate the count in Jypter, which is done by value\_count() and can be done in the below mentioned way:

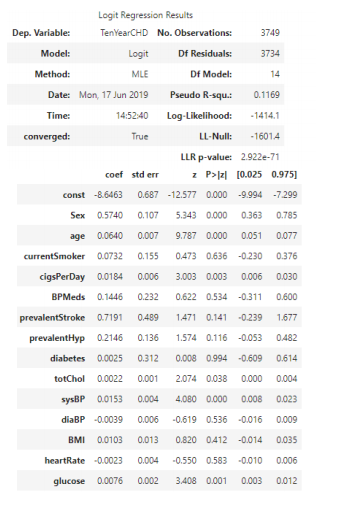




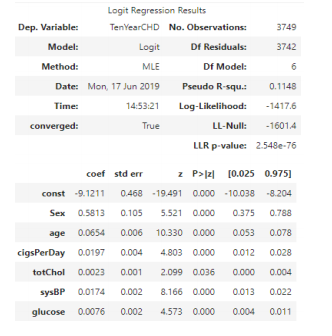
From the above graph, we can say that the count is 3596, with no heart disease and 644 with heart diseases.

**6.2: Modelling:**

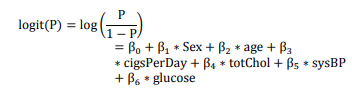
The below mentioned output indicating the logistic regression is majorly used for predicting as well as success probability calculation



From the above logistic regression results, it can be concluded that the attributes whose value is P greater or equal to 0.05 have very less significance and hence, backward elimination method is used to eliminate high values of P and till process is in progress until and unless p range is less than 0.05.



The result for the logistic regression is calculated by the backward elimination method and hence, the equation will be given by:



Logistic Regression Interpretation:

The accuracy measurements are given by the following methods:

Odd ratio interpretation:

This is used to measure the relationship exposure with outcomes. Furthermore, these ratios can be used to detect the risk factor for given outcome and comparing magnitudes of various factors of risk for particular outcome.

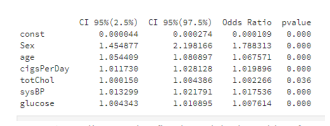
<1 then the exposure is associated lower-odds of values

=1 then exposures will not effect the odd of outcomes.

>1 then the exposure is associated higher-odds of values.

Confidence -Intervals:

The accuracies of the odd’s ratio are estimated if the confidence level is greater than or equal to 95 percent. A small CI value represents OR precisions higher and large CI represents lower OR precision.

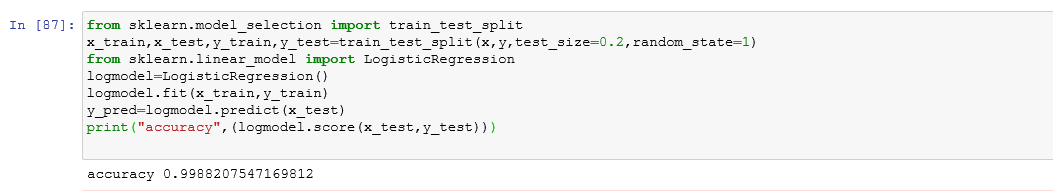


Firstly, the odds of diagnosed with heart disease of males which is 78% and is higher when compared to females, as per the fitted model. Next, there is an increase in CHD which is close to 7% for one year age. Moreover, SYSBP odd’s increase by 1.7, however, the ciggrates have only 2% risk. Furthermore, there is no significant change in glucose and cholesterol levels.

**6.3 Results analysis and discussion:**

**Testing and training the datasets:**

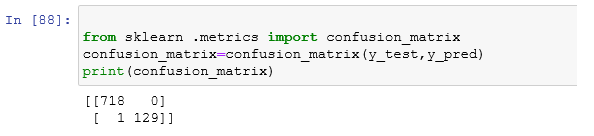
In Machine learning, data set needs to be divided into parts i.e.; test and training for the process of evaluation. Hence, I have also divided my dataset into test and train. This is achieved by the inbuilt library of python called Scikit-learn. The steps in Python are mentioned below:



The logistic regression model built is 0.9988207

**Confusion Matrix Outcomes:**

Confusion matrix is used that indicates the prediction results summary of incorrect and correct values of classification-problems. In my logistic model, confusion matrix is not only used for errors but as well for types of errors. There are four factors in confusion matrix and they are TP(True positive) who have disease and are predicted yes by the model, TN(True negative) the model predicted no and they do not have disease, FP(False Positive) they actually do not have disease but the model predicted that they have disease and FN(False negative)they actually have disease but predicited no by the model. The confusion matrix in python is implemented as sklearn.metrics library and is stated under:



As per the confusion matrix outcome,

correctly predicted are (718+129)= 847

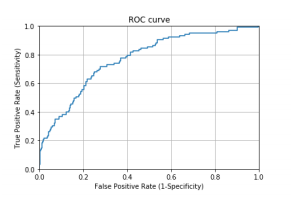
Incorrect: (1+0)= 1

Therefore, the confusion matrix will be



**ROC Curve:**

A simple curve called ROC curve is used to visualize the results of the binary classifier. Furthermore, it the comparison between the true-positive rate and false-positive rate for binary classifier with different probability thresholds.



Area Under the Curve quantifies the model classification accuracy. Good classification accuracy models should have significantly more true positives than the false positives at all thresholds.

1. **Conclusion:**

The objective of this study was to implement the Logistic regression model to find the risk of 10 year CHD by using 14-different variables.In this the backward elimination process is used that contains the the value of p which is less than 5% .Hence, logistic model is deprived from the attributes like age, glucose, cigsPerDay, sysBP, totChol. Furtehmore, as per the output of the regression model it can be predicted that men are expected to suffer more than women. The odds of the CHD are age, number of cigrates, systolic blood pressure and .Moreover, no change has been observed for glucose and cholostorol level. Howver glucose levels have very less odd changes and the model is not sensitive rather it is specific. In the end the accuracy of the model is 0.99882. Finally, ROC curve value is 73.5 which is good but can be improved with more data.

**8. References:**

[1] Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8203946.

[2] X. Zou, Y. Hu, Z. Tian and K. Shen, "Logistic Regression Model Optimization and Case Analysis," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 2019, pp. 135-139, doi: 10.1109/ICCSNT47585.2019.8962457.

[3] H. Zhang, Z. Li, H. Shahriar, L. Tao, P. Bhattacharya and Y. Qian, "Improving Prediction Accuracy for Logistic Regression on Imbalanced Datasets," 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 2019, pp. 918-919, doi: 10.1109/COMPSAC.2019.00140.

[4] Nishadi, AS Thanuja. "Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab.

[5] <https://repository.maseno.ac.ke/handle/123456789/1164?show=full>

**9.Appendix:**

**Code for my Logistic Regression Model:**

##### Importing necessary Libraries ########

from sklearn.datasets import make\_classification

from matplotlib import pyplot as plt

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import confusion\_matrix

import numpy as np

import pandas as pd

####Reading the data from the CSV file#####

data=pd.read\_csv("heart\_new.csv")

data.head()

##### Creating variables for accessing data ######

data.shape

feature\_cols = ['male','age','education','currentSmoker','cigsPerDay','BPMeds','prevalentStroke','prevalentHyp','diabetes','totChol','sysBP','diaBP','BMI','heartRate','glucose','TenYearCHD']

######Identifying Missing and NA values and filling them with NA####

data = data.fillna(0)

print(x\_train)

x=data[feature\_cols]

y=data.TenYearCHD

print(x)

print(y)

###### Getting the Count#####

data['TenYearCHD'].value\_counts()

from matplotlib import pyplot as plt

import seaborn as sns

sns.countplot(x='TenYearCHD',data=data,palette="hls")

plt.show()

######Splitting the data in test and train######

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.2,random\_state=1)

##### Predicting the Accuracy######

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.2,random\_state=1)

from sklearn.linear\_model import LogisticRegression

logmodel=LogisticRegression()

logmodel.fit(x\_train,y\_train)

y\_pred=logmodel.predict(x\_test)

print("accuracy",(logmodel.score(x\_test,y\_test)))

#######Printing Confusion Matrix######

from sklearn .metrics import confusion\_matrix

confusion\_matrix=confusion\_matrix(y\_test,y\_pred)

print(confusion\_matrix)

######Plotting the ROC Code######

from sklearn.metrics import roc\_auc\_score

from sklearn.metrics import roc\_curve

logit\_roc\_auc = roc\_auc\_score(y\_test, logmodel.predict(x\_test))

fpr, tpr, thresholds = roc\_curve(y\_test, logmodel.predict\_proba(x\_test)[:,1])

plt.figure()

plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit\_roc\_auc)

plt.plot([0, 1], [0, 1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

plt.title('Receiver operating characteristic')

plt.legend(loc="lower right")

plt.savefig('Log\_ROC')

plt.show()