iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50542

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Estimating the Risk of Developing Heart Disease Using the Logistic Regression Model of Machine Learning

Geetha M[1], Monika S[2], Pavithra P[3], Sowmiya A[4], Dinesh V[5]
[1]*Assistant Professor, Master of Computer Application, Paavai Engineering College, Namakkal, India.*
[2,3,4,5] *UG Student , Master of Computer Application, Paavai Engineering College, Namakkal, India.*

Abstract: Asymptomatic diseases, such as cardiovascular diseases, are driving up healthcare costs to the point where they are exceeding corporate and national budgets.
As a result, these diseases must be identified and treated as soon as possible. One of the hottest technologies, machine learning is used to predict diseases in many fields, including the healthcare industry. This study uses logistic regression to predict the overall risk and identify the most significant predictors of heart disease.
As a result, the c predicators in this study are identified using the binary logistic model, which is one of the classification algorithms in machine learning. In addition, Jupiter Lab and Python are utilized for data analysis in order to validate the logistic regression.
Keywords: heart diseases, classification algorithms, logistic regression, machine learning.

I. INTRODUCTION

Over the past few years all over the world. Regardless of whether these infections has found as the main wellspring of death, it has been declared as the most sensible and avoidable sickness [1]. Heart stroke is mostly caused by artery blockage. It occurs when the heart does not effectively circulate blood throughout the body. Additionally, high blood pressure is one of the primary risk factors for heart disease.

According to a survey, from 2011 to 2014, about 35% of people worldwide had hypertension, which is also a factor in heart disease. In a similar vein, there are numerous additional risk factors for heart disease, including obesity, inadequate nutrition, elevated cholesterol, and inactivity.

Therefore, prevention is crucial. Heart disease awareness is essential for prevention. The fact that 47% of people die outside of the hospital demonstrates that they ignore early warning signs. Heart diseases are reducing an individual's lifespan today.

As a result, in 2013, the World Health Organization (WHO) established goals for the prevention of non-communicable diseases (NCDs).

These goals include ensuring that by 2025, at least 50% of patients with cardiovascular diseases will have access to relevant medications and medical counseling [2].

In 2016, cardiovascular diseases accounted for approximately 17.9 million deaths, or 31% of all deaths worldwide. The detection of heart disease is a major obstacle [3].

It is hard to foresee that an individual has a coronary illness or not. There are instruments accessible which can anticipate heart sicknesses yet it is possible that they are costly or are not effective to compute the opportunity of coronary illness in human [4]. According to a World Health Organization (WHO) survey, doctors can only predict 67% of heart attacks, so there is a lot of room for research [5].

In rural India, it is extremely difficult to find good doctors and hospitals. A 2016 WHO report says that, only 58% of the specialists have physician certification in metropolitan regions and 19% in country regions.

Heart infections can be anticipated utilizing Brain Organization, Choice Tree, KNN, and so forth. We will see how Logistic Regression is used to determine heart disease accuracy later in this paper. It likewise shows that how ML will help in our future for coronary illness.

II. WORKFLOW OF MACHINE LEARNING MODEL BUILDING

The machine learning logistic regression model's building process is depicted in Figure 2.

Variable Category	Variable Name	Description	Data Type
Demographic	Sex	Male or female	Nominal
	Age	Age of the patient	Continuous
Behaviour	Current Smoker	Current smoker or not?	Nominal
	Cigs Per Day	Cigarettes per day?	Continuous
Medical History	B P Meds	Blood pressure medication?	Nominal
	Prevalent Stroke	Whether previously had stroke?	Nominal
	Prevalent Hyp	Whether was hypertensive?	Nominal
	Diabetes	Whether had diabetes?	Nominal
Current Medical Status	Tot Chol	Total Cholesterol Level	Continuous
	Sys BP	Systolic Blood Pressure	Continuous
	Dia BP	Diastolic Blood Pressure	Continuous
	BMI	Body Mass Index	Continuous
	Heart Rate	Heart Rate	Continuous
	Glucose	Glucose Level	Continuous
Predicted Variable	TenYearCHD	10-year risk of CHD	Binary

Table 1.2.1: Input Variables

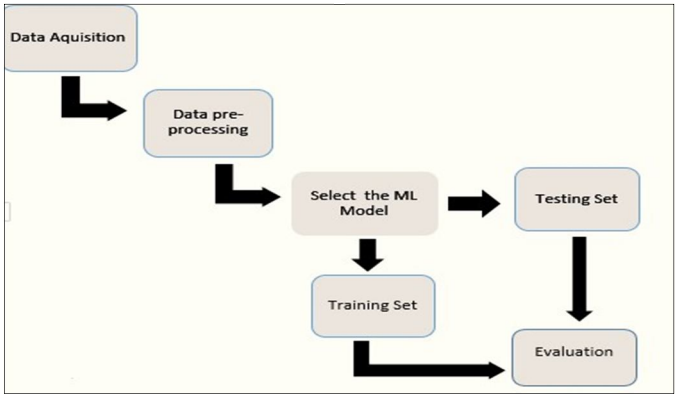


Figure 1.2.2: Workflow of Logistic Regression Model

1) Obtaining the Dataset The dataset was gathered from the Kaggle website.
2) Data Pre-Processing Data pre-processing is necessary for building a more accurate ML model. The process of cleaning the data is called data pre-processing. This includes finding data that is missing, noisy, or inconsistent.
3) Choose a Machine Learning Model Machine learning algorithms are used to identify the pre-processed data. a) The study's input variables The data set includes 14 IVS and predicted values. DV identification is the foundation of the ML model. It has utilized paired calculated relapse which is one of the order calculations because of target variable is downright.
4) Python and Jupiter Lab were utilized for the analysis of the data. The logistics regression was processed using the following procedures.
5) Loading Data and Other Required Libraries Jupiter Lab has loaded the Framingham CSV file of heart prediction data in order to construct the logistic regression model.

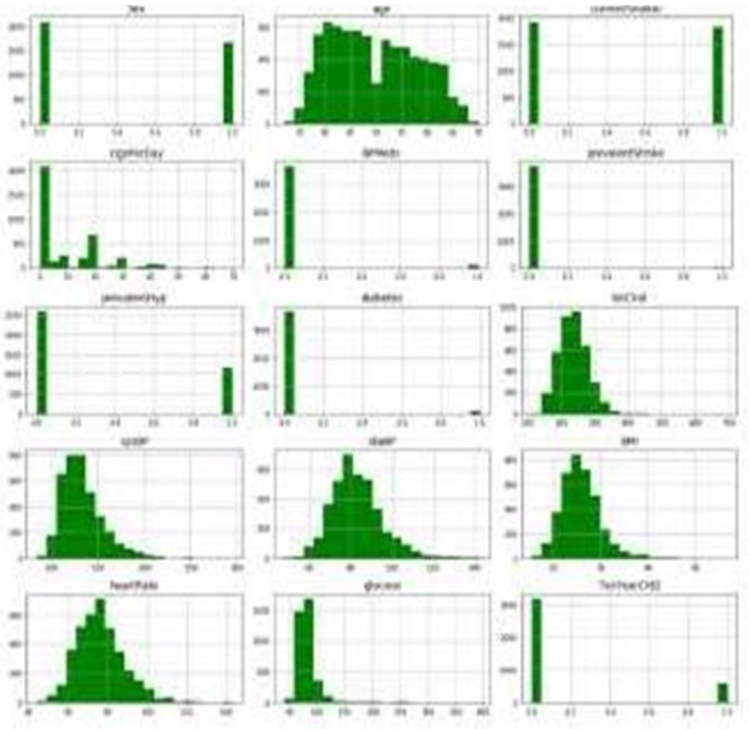
In addition, necessary libraries that are utilized as supporting applications are loaded. The education field has been removed from the database.

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import scipy.stats as st
import matplotlib.pyplot as plt
import seaborn as sn
from sklearn.metrics import confusion_matrix
import matplotlib.mlab as mlab
%matplotlib inline
```

```
heartdb=pd.read_csv("C:\\Users\\LAB-User\\Desktop\\framingham.csv")
heartdb.drop(['education'],axis=1,inplace=True)
heartdb.head()
```

a) Exploratory Data Analysis (EDA)

The accompanying perception inferred through the Jupiter Lab for display predicators.

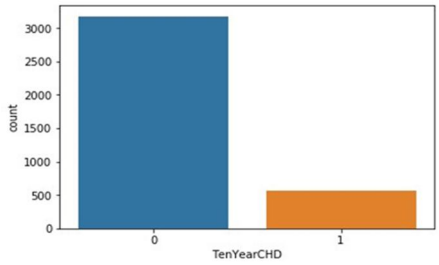


III. IDENTIFY MISSING VALUES

Additionally, the number of missing values has been identified for the existing dataset to be cleaned. Based on the attributes,the total number of missing values is summarized below.

```
sn.countplot(x='TenYearCHD',data=heartdb)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1fe5aa6ea90>
```



Then, the all out level of missing qualities in segment was recognized utilizing Pandas Information Casing. Absolute number of lines with missing qualities is 489 since it is just 12% of the whole dataset the columns with missing qualities are barred. The drop rows and columns with null values were analyzed using Pandas' dropna() method.

```
In [7]: heartdb.dropna(axis=0,inplace=True)
```

The descriptive figures related to 10year risk of CHD has indicated below.

```
In [9]: heartdb.TenYearCHD.value_counts()

Out[9]: 0    3177
        1     572
        Name: TenYearCHD, dtype: int64
```

A. Implementing Logistic Regression

The accompanying results are utilized to demonstrate the calculated relapse. Strategic relapse is primarily used to for expectation and furthermore computing the likelihood of progress.

Logit Regression Results

Dep. Variable:	TenYearCHD	**No. Observations:**	3749
Model:	Logit	**Df Residuals:**	3734
Method:	MLE	**Df Model:**	14
Date:	Mon, 17 Jun 2019	**Pseudo R-squ.:**	0.1169
Time:	14:52:40	**Log-Likelihood:**	-1414.1
converged:	True	**LL-Null:**	-1601.4
		LLR p-value:	2.922e-71

	coef	std err	z	P>|z|	[0.025	0.975]
const	-8.6463	0.687	-12.577	0.000	-9.994	-7.299
Sex	0.5740	0.107	5.343	0.000	0.363	0.785
age	0.0640	0.007	9.787	0.000	0.051	0.077
currentSmoker	0.0732	0.155	0.473	0.636	-0.230	0.376
cigsPerDay	0.0184	0.006	3.003	0.003	0.006	0.030
BPMeds	0.1446	0.232	0.622	0.534	-0.311	0.600
prevalentStroke	0.7191	0.489	1.471	0.141	-0.239	1.677
prevalentHyp	0.2146	0.136	1.574	0.116	-0.053	0.482
diabetes	0.0025	0.312	0.008	0.994	-0.609	0.614
totChol	0.0022	0.001	2.074	0.038	0.000	0.004
sysBP	0.0153	0.004	4.080	0.000	0.008	0.023
diaBP	-0.0039	0.006	-0.619	0.536	-0.016	0.009
BMI	0.0103	0.013	0.820	0.412	-0.014	0.035
heartRate	-0.0023	0.004	-0.550	0.583	-0.010	0.006
glucose	0.0076	0.002	3.408	0.001	0.003	0.012

The logistic results above indicate a low statistically significant relationship between heart disease risk and P >= 0.05. As a result, the attributes with the highest P values have been eliminated using the backward elimination technique. Until all of the attributes have P values less than 0.05, the procedure will continue.

Logit Regression Results

Dep. Variable:	TenYearCHD	**No. Observations:**	3749
Model:	Logit	**Df Residuals:**	3742
Method:	MLE	**Df Model:**	6
Date:	Mon, 17 Jun 2019	**Pseudo R-squ.:**	0.1148
Time:	14:53:21	**Log-Likelihood:**	-1417.6
converged:	True	**LL-Null:**	-1601.4
		LLR p-value:	2.548e-76

	coef	std err	z	P>|z|	[0.025	0.975]
const	-9.1211	0.468	-19.491	0.000	-10.038	-8.204
Sex	0.5813	0.105	5.521	0.000	0.375	0.788
age	0.0654	0.006	10.330	0.000	0.053	0.078
cigsPerDay	0.0197	0.004	4.803	0.000	0.012	0.028
totChol	0.0023	0.001	2.099	0.036	0.000	0.004
sysBP	0.0174	0.002	8.166	0.000	0.013	0.022
glucose	0.0076	0.002	4.573	0.000	0.004	0.011

IV. EXPERIMENTAL WORK

Internal confidence (CI) of 95 percent is used to calculate the QR's accuracy. A large CI indicates a low level of QR precision, while a smaller CI indicates a higher level of QR precision. However, unlike the p value, the 95% CI does not indicate statistical significance.

	CI 95%(2.5%)	CI 95%(97.5%)	Odds Ratio	pvalue
const	0.000044	0.000274	0.000109	0.000
Sex	1.454877	2.198166	1.788313	0.000
age	1.054409	1.080897	1.067571	0.000
cigsPerDay	1.011730	1.028128	1.019896	0.000
totChol	1.000150	1.004386	1.002266	0.036
sysBP	1.013299	1.021791	1.017536	0.000
glucose	1.004343	1.010895	1.007614	0.000

This has used to show the synopsis of forecast results including right and mistaken on a characterization issue. Further, this was utilized to mistakes as well as sorts of blunders. The following parameters are indicated by the segments of the confusion matrix.

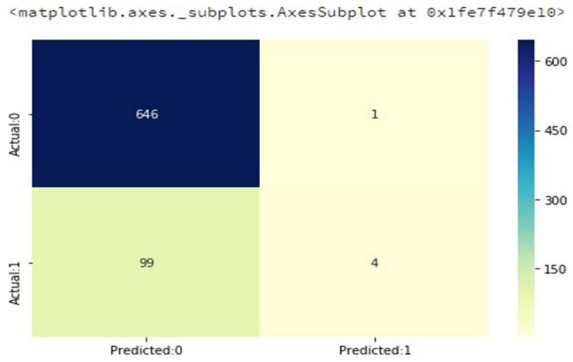


Fig.4.1.1

- *Genuine Up-sides (TP):* Cases that were predicted to have the disease, and they actually do.
- *Actual Negatives:* Cases that are predicted to be unaffected by the disease
- *FP (False Positives):* Cases that are predicted to have the disease, but do not actually have it (Type I error).
- *Negatives False (FN):* Cases that were predicted to be negative, but in fact have the disease (Type II error).

The dataset's confusion matrix can be seen in the next result.

As per the result of the disarray grid,
Right expectations (646+4) =650
Wrong expectations (99+1) =100

In this manner,
Genuine Positives:4
Genuine Negatives:646
Bogus Positives:1(Type I blunder)
Bogus Negatives:99(Type II blunder)

```
TN=cm[0,0]
TP=cm[1,1]
FN=cm[1,0]
FP=cm[0,1]
sensitivity=TP/float(TP+FN)
specificity=TN/float(TN+FP)
```

A list of rates that are frequently calculated using a binary classifier's confusion matrix is as follows:
Using a confusion matrix, the model's accuracy was evaluated.

Terms	Formula
Accuracy of the model (overall, how often the classifier correct)	(TP+TN)/(TP+TN+FP+FN)
Misclassification Rate (overall, how often it wrong or error rate)	(FP+FN)/(TP+TN+FP+FN)
Sensitivity or True Positive Rate (when it is actually yes, how often does it predict yes)	TP/(TP+FN)
Specificity or True Negative Rate (when it is actually no, how often does it predict no)	TN/(TN+FP)

When the confusion matrix data are analyzed, it becomes clear that the model is more specific than it is sensitive. Further, the negative qualities in the model are anticipated more precisely than the up-sides.

```
With 0.3 threshold the Confusion Matrix is
 [[615  32]
 [ 75  28]]
 with 643 correct predictions and 75 Type II errors( False Negatives)

 Sensitivity:  0.27184466019417475 Specificity:  0.9505409582689336
```

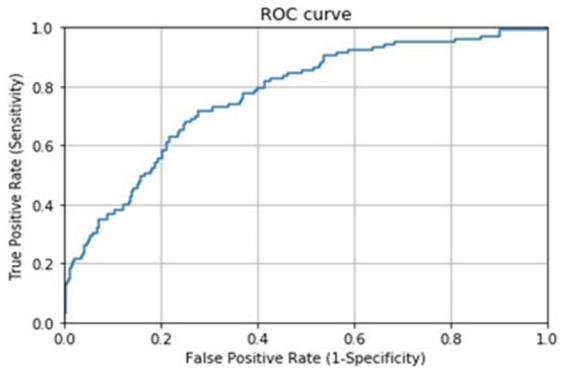
```
With 0.4 threshold the Confusion Matrix is
 [[643   4]
 [ 90  13]]
 with 656 correct predictions and 90 Type II errors( False Negatives)

 Sensitivity:  0.1262135922330097 Specificity:  0.9938176197836167
```

```
from sklearn.metrics import roc_curve
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob_yes[:,1])
plt.plot(fpr,tpr)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.0])
plt.title('ROC curve')
plt.xlabel('False Positive Rate (1-Specificity)')
plt.ylabel('True Positive Rate (Sensitivity)')
plt.grid(True)
```

V. ROC CURVE

A straightforward plot called the ROC Curve is used to show how well a binary classifier works. In addition, this demonstrates the tradeoff between a classifier's true positive rate and false positive rate for various probability threshold selections.



Great characterization precision models ought to have fundamentally more evident up-sides than the misleading up-sides at all edges. The model's classification accuracy is measured by the area under the curve (AUC).

Figure depicts the effect of age on cardiovascular disease. 7. The main gamble factor for cardiovascular or coronary illness is age, which generally significantly increases with every 10 years of life. Heart fatty streaks can begin to appear as early as adolescence. It has been determined that 82% of people who die from coronary heart disease are over the age of 65. In addition, the risk of stroke copies persists after age.

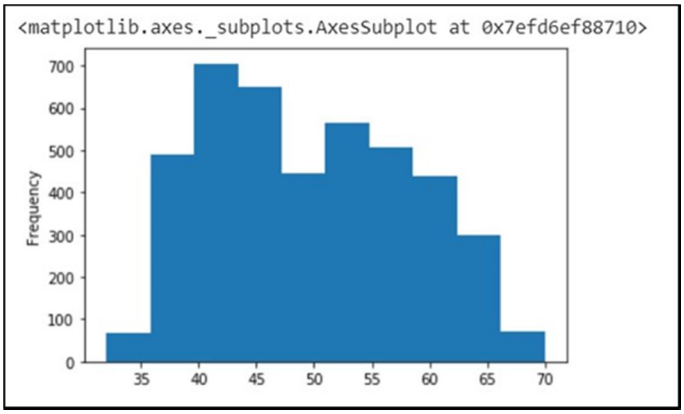


Fig 5.1.1 Level of input parameter

The effectiveness of a binary classifier is demonstrated through the use of a straightforward plot known as the ROC Curve. This also demonstrates the tradeoff between a classifier's true positive rate and false positive rate for various probability threshold choices. At all edges, great portrayal accuracy models ought to have in a general sense more clear advantages than deceiving benefits. The degree of the model's data classification accuracy is measured by its area under the curve, or AUC.

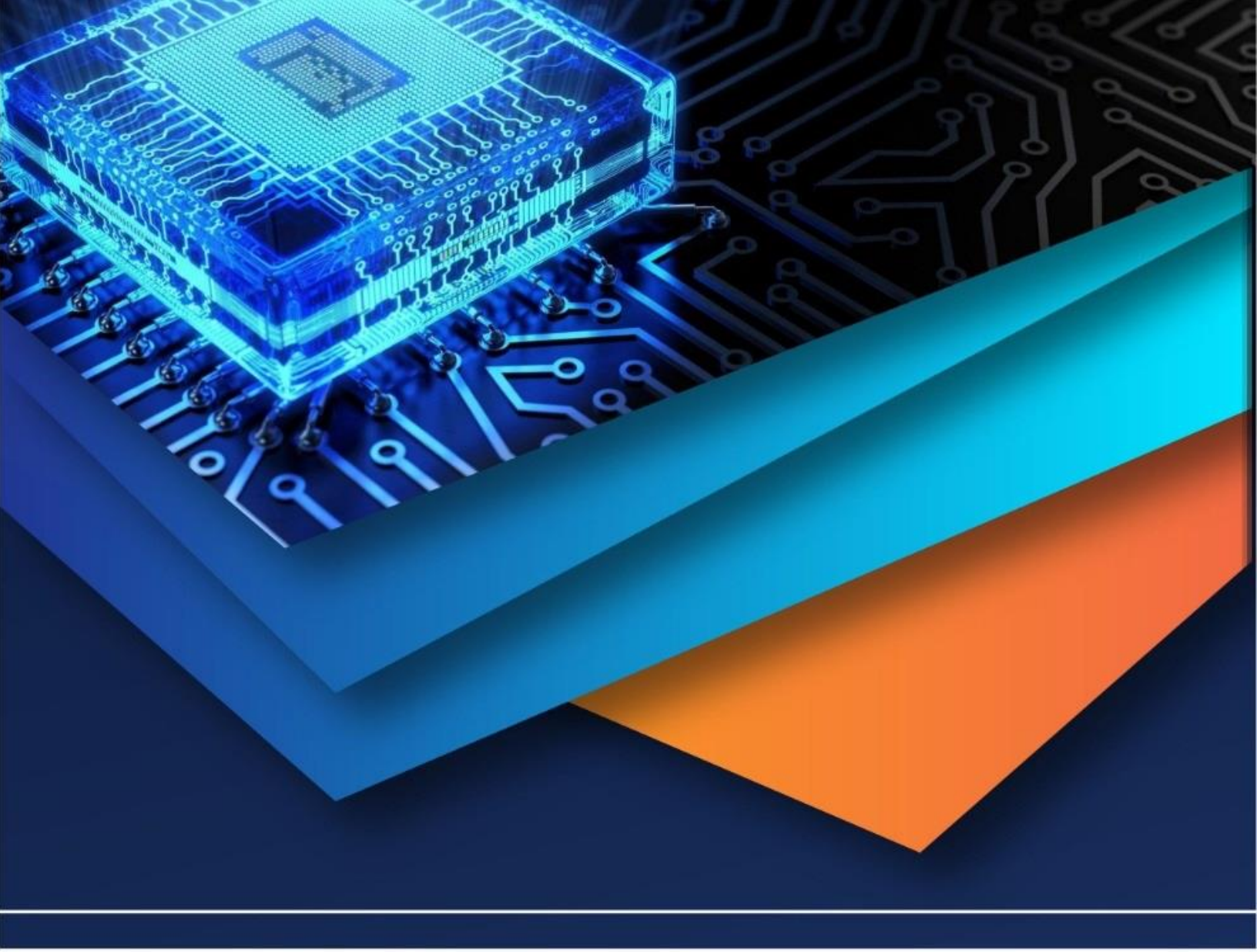
At all edges, excellent characterization precision models should have fundamentally more obvious benefits than misleading benefits. The area under the curve is used to measure the model's classification accuracy.

VI. CONCLUSION

Using 14 intravenous (IV) injections, the purpose of this study is to ascertain the risk of 10-year CHD. P values below 5% are taken into consideration when selecting the attributes through backward elimination. Therefore, the P values of the variables 0.05 (sex, age, cigs Per Day, tot Chol, sys BP, glucose) are used to generate the logistic regression model. The logistic regression result indicates that men are more likely than women to develop heart disease. High systolic blood pressure, age, and smoking are all risk factors for coronary heart disease. Notwithstanding, neither the absolute cholesterol level nor the glucose level altogether change. Regardless, the possibilities are not completely influenced by glucose levels. The model is less sensitive than specific. Additionally, this model has a 0.87 accuracy. In some ways, the value below the ROC curve, 73.5, is sufficient. The model could also be improved with additional data.

REFERENCES

[1] E. Benjamin, A. Go, D. Arnett, M. Blaha, M. Cushman, et al. 2015). Statistics on Heart Disease and Stroke—Update for 2015 131(4) of Circulation. 10.1161/cir.0000000000000152.
[2] S. Das, A. Dey, A. Pal, N. Roy, and others Machine Learning and Applications of Artificial Intelligence: Prospect and Review 115(9): 31–41, International Journal of Computer Applications.: 10.5120/20182-2402
[3] R. Abduljabbar, H., S. Liyanage, and S. Transport-Based Applications of Artificial Intelligence: An Outline. 189 of Sustainability, 11(1). : 10.3390/su11010189
[4] Pedro Luis Cruz, Carlos Soares, Joo Moreira, and Rui Abreu are among the contributors. (2015). A Comparative Study of Classification and Regression Algorithms for Modeling Students' Academic Performance, accessed at: https://www.researchgate.net/publication/27803068 9 A Comparative Study of Classification and Regression Algorithms for Modeling Students' Academic Performance June 10th, 2019.
[5] R. Sathya and A. Abraham (2013) published their work in the International Journal of Advanced Research in Artificial Intelligence, Volume 2, No. 2, 2013, accessed through http://ijarai.thesai.org/Downloads/IJARAI/Volume2No2/Paper_6-Comparison_of_Supervised_and_Unsupervised_Learning_Algorithms_for_Pattern_Classification.pdf, with the following page: June 10th, 2019.
[6] CVA, K. (2017),https://www.medwinpublishers.com/JOBD/JOBD1 6000139.pdf. 1(7) of the Journal of Orthopedics and Bone Disorders. : 10.23880/jobd-16000139
[7] Miguel-Hurtado, O., Visitor S., Neil, G., and Dark, S. (2016). Comparing Linear/Logistic Regression and Machine Learning Classifiers to Study the Relationship Between Hand Dimensions and Demographic Characteristics. 11(11), e0165521, PLOS ONE. :
[8] 10.1371/journal.pone.0165521 Jordan, M. I., and A. Y. Ng (2002) Comparing generative and discriminative classifiers: a comparison of naive bayes and logistic regression. NIPS 14, pages 841–848.
[9] Peng, C., Lee, K., and G. Ingersoll An Overview of the Concepts and Methods of Logistic Regression Reporting 3–14 in 96(1) of the Journal of Educational Research:
[10] 1080/00220670209598786
[11] Park, H. (2019). An Overview of Logic Regression: From Fundamental Ideas to Understanding with Specific Regard for Nursing Domain, https://pdfs.semanticscholar.org/3305/2b1d2363aee 3ad290612109dcea0aed2a89e.pdf, saw : June 10, 2019



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY