International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

Cardiovascular Disease Prediction using Classification Algorithms of Machine Learning

Yash Jayesh Chauhan

[1]Second Year Btech.CSE, Parul University, Parul Institute of Engineering and Technology, Vadodara, Gujarat, India

Abstract: ***Cardiovascular disease is a major health burden worldwide in the 21st century. Human services consumptions are overpowering national and corporate spending plans because of asymptomatic infections including cardiovascular ailments. Consequently, there is an urgent requirement for early location and treatment of such ailments. The information which is gathered by data analysis of hospitals is utilizing by applying different blends of calculations and algorithms for the early-stage prediction of Cardiovascular ailments. Machine Learning is one of the slanting innovations utilized in numerous circles far and wide including the medicinal services application for predicting illnesses. In this research, we compared the accuracy of machine learning algorithms that could be used for predictive analysis of heart diseases and predicting the overall risks. The proposed experiment is based on a combination of standard machine learning algorithms such as Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), support vector machine (SVM) and Decision Tree. Most of the entities in this world are related in one way or another, at times finding a relationship between entities can help you make valuable decisions. Likewise, I will attempt to utilize this information as a model that predicts the patient whether they are having a Cardiovascular disease or on the other hand not. Moreover, the data analysis is carried out in Python using Jupyter Lab in order to validate the accuracy of all the Algorithm.***

Keywords: Machine learning, Data Analysis, Classification algorithms, Heart diseases

1. Introduction

Cardiovascular disease is presently the leading problem of death worldwide. An expected 3.8 million men and 3.4 million women die each year from cardiovascular disease. Diastolic Blood Pressure and Systolic Blood Pressure are related to cardiovascular risk. Thus, a feasible and accurate prediction of heart-related diseases is very important. Medical organizations, all around the world, collect data on various health-related issues. These data can be exploited using various machine learning techniques to gain useful insights [1]. But the data collected is very massive and, many times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart-related diseases accurately [4]. To begin with, the work we are using different types of techniques and algorithms. In this paper, the classification of machine learning techniques and algorithms are used to increase the accuracy rate. In Machine learning, classification algorithms are supervised learning approach in which the computer learns from the input data and learn from it. This data collection may basically be bi-class (like recognizing whether the individual is male or female or that the mail is spam or non-spam) or it might be multi-class. Here are the names of classification algorithms which we are going to implement and compare the accuracy in this research:

1) Linear Classifiers: Logistic Regression
2) K-Nearest Neighbor
3) Support Vector Machine (SVM)
4) Decision Trees
5) Random Forest

2. Background of the Study

The heart is the most important organ of the human body because it pumps our blood and circulates to the entire body. The heart is protected by a rib cage and it is surrounded by two-layered tissue membranes. It is a four-chambered organ that separates oxygenated and deoxygenated blood. The heart is having the five types of blood vessels: arteries, veins, capillaries, arterioles, venules and the size of the human heart is about the size of the fist. The dataset used for the logistic regression analysis is available on the Kaggle website, from an ongoing cardiovascular study of Framingham, Massachusetts. The classification goal of this study is to predict whether the patient has a 10-year risk of future heart diseases. The Framingham dataset consists of 4238 records of patients' data and 14 attributes. The data analysis is carried out in Python programming by using Jupyter Lab which is a more flexible and powerful data science application software.

3. Methodology

3.1 Workflow of building Machine Learning Model

Figure 1 indicates the steps followed in order to build the model in machine learning.

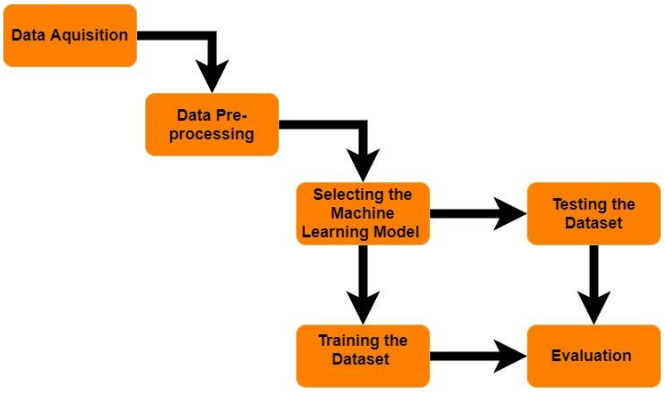


Figure 1: Workflow of building Machine Learning Model

3.2 Data Acquisition

The dataset is collected from Kaggle website.

3.3 Data Pre-Processing

In order to build up a more accurate Machine Learning model, data preprocessing is required. Data pre-processing is the process of cleaning the data. It will remove all the NAN values from our data. This process is also known as Data Wrangling. This includes the identification of missing data, noisy data and inconsistent data.

3.4 Proposed System

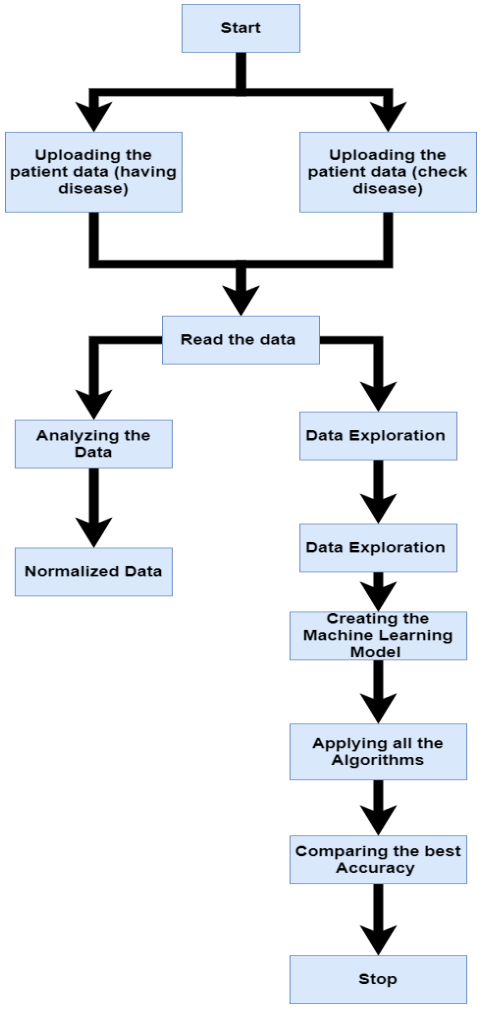


Figure 2: Proposed System

3.5 Select Machine Learning Model

Then the pre-processed data are identified using machine learning algorithms. We will be using the Classification Algorithm to compare the best accuracy from all.

a) Input Variables of the study
The data set consists of 14 IVS. Machine Learning model is based on the identification of DV.

Table 1: Input Variables

Variable Category	Variable Name	Description	Data Type
Demographic	Sex	Male or female	Nominal
	age	Age of the patient	Continuous
Behavior	currentSmoker	Current smoker or not?	Nominal
	cigsPerDay	Cigarettes per day?	Continuous
Medical History	BPMeds	Blood pressuremedication?	Nominal
	prevalentStroke	Whether previously had stroke?	Nominal
	prevalentHyp	Whether was hypertensive?	Nominal
	diabetes	Whether had diabetes?	Nominal
Current Medical Status	totChol	Total Cholesterol Level	Continuous
	sysBP	Systolic Blood Pressure	Continuous
	diaBP	Diastolic Blood Pressure	Continuous
	BMI	Body Mass Index	Continuous
	heartRate	Heart Rate	Continuous
	glucose	Glucose Level	Continuous
Predicted Variable	TenYearCHD	10-year risk of CHD	Binary

Figure 3: [2] Input Variables

4. Data Analysis

Data Analysis was carried out on the Jupyter Notebook for further classification using Python 3.7.

4.1 Importing the Libraries

Here we have loaded the data into Jupiter Lab to build a machine learning model. In accession to that, the required libraries used as supportive applications are loaded. It has removed the education field 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 sns
from sklearn.metrics import confusion_matrix
import matplotlib.mlab as mlab
%matplotlib inline
```

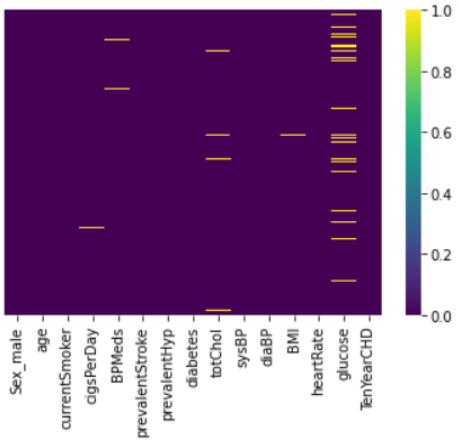
4.2Reading the Dataset

	0	1	2	3	4	5	6	7	8	9	...	4226	4227	4228	4231	4232	4233	4234	4237
Sex_male	1.00	0.00	1.00	0.00	0.0	0.0	0.00	0.00	1.00	1.00	...	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00
age	39.00	46.00	48.00	61.00	46.0	43.0	63.00	45.00	52.00	43.00	...	58.00	43.00	50.00	58.00	68.00	50.00	51.00	52.00
currentSmoker	0.00	0.00	1.00	1.00	1.0	0.0	0.00	1.00	0.00	1.00	...	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00
cigsPerDay	0.00	0.00	20.00	30.00	23.0	0.0	0.00	20.00	0.00	30.00	...	0.00	20.00	0.00	0.00	0.00	1.00	43.00	0.00
BPMeds	0.00	0.00	0.00	0.00	0.0	0.0	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prevalentStroke	0.00	0.00	0.00	0.00	0.0	0.0	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prevalentHyp	0.00	0.00	0.00	1.00	0.0	1.0	0.00	0.00	1.00	1.00	...	0.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00
diabetes	0.00	0.00	0.00	0.00	0.0	0.0	0.00	0.00	0.00	0.00	...	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
totChol	195.00	250.00	245.00	225.00	285.0	228.0	205.00	313.00	260.00	225.00	...	233.00	187.00	260.00	187.00	176.00	313.00	207.00	269.00
sysBP	106.00	121.00	127.50	150.00	130.0	180.0	138.00	100.00	141.50	162.00	...	125.50	129.50	190.00	141.00	168.00	179.00	126.50	133.50
diaBP	70.00	81.00	80.00	95.00	84.0	110.0	71.00	71.00	89.00	107.00	...	84.00	88.00	130.00	81.00	97.00	92.00	80.00	83.00
BMI	26.97	28.73	25.34	28.58	23.1	30.3	33.11	21.68	26.36	23.61	...	26.05	25.62	43.67	24.96	23.14	25.97	19.71	21.47
heartRate	80.00	95.00	75.00	65.00	85.0	77.0	60.00	79.00	76.00	93.00	...	67.00	80.00	85.00	80.00	60.00	66.00	65.00	80.00
glucose	77.00	76.00	70.00	103.00	85.0	99.0	85.00	78.00	79.00	88.00	...	76.00	75.00	260.00	81.00	79.00	86.00	68.00	107.00
TenYearCHD	0.00	0.00	0.00	1.00	0.0	0.0	1.00	0.00	0.00	0.00	...	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00

15 rows × 3751 columns

4.3 Data Pre-processing

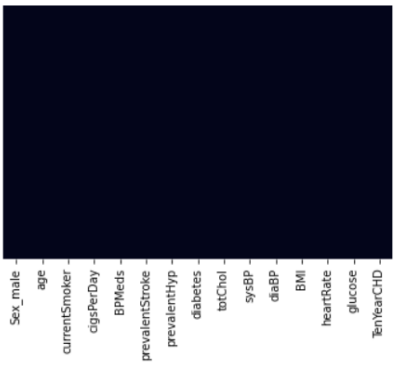
Moreover, the number of missing values has identified for cleaning an existing dataset. The summarized total number of missing values based on the attributes are given below.



The total percentage of missing values in the column was identified using the Pandas Data Frame. The total number of rows with missing values is 489 since it is only 12 percent of the entire dataset the rows with missing values are excluded. It has used the Pandas dropna() method which was used to analyze the drop rows/columns with Null values.

```
heart_df.dropna(axis=0,inplace=True)
```

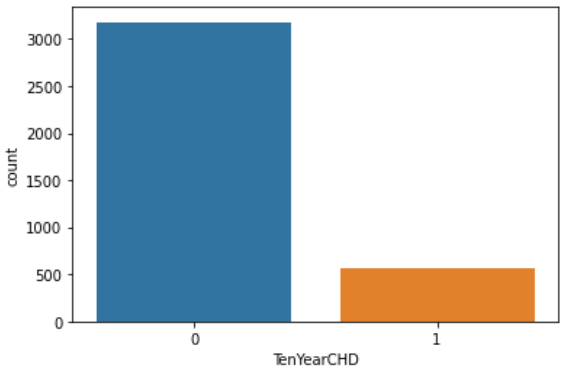
After applying the Pandas dropna() method which was used to analyze the drop rows/columns with Null values we can note by the below graph.



The representative figures related to the 10year risk of Coronary Heart Disease has shown below.

```
0    3179
1     572
Name: TenYearCHD, dtype: int64
```

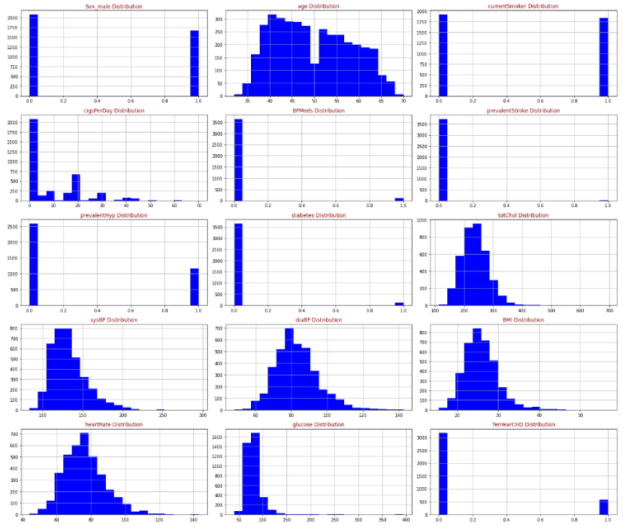
Visualization in bar graph for better understanding.



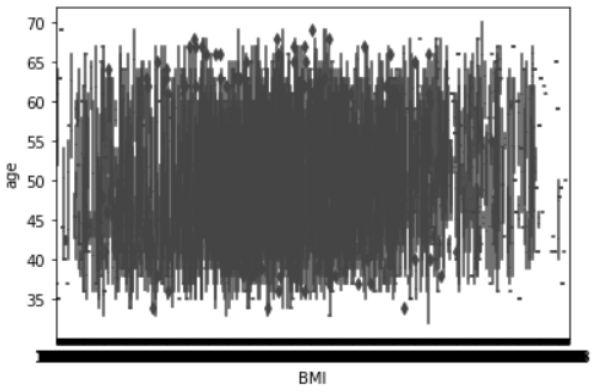
According to the above data, there are 3179 patents with no heart disease and 572 patients with risk of heart disease.

4.4 Visualization of data by Scatter Plot

The following visualization derived through the JupyterLab for display predicators.



Visualization of Body Mass Index according to Age.

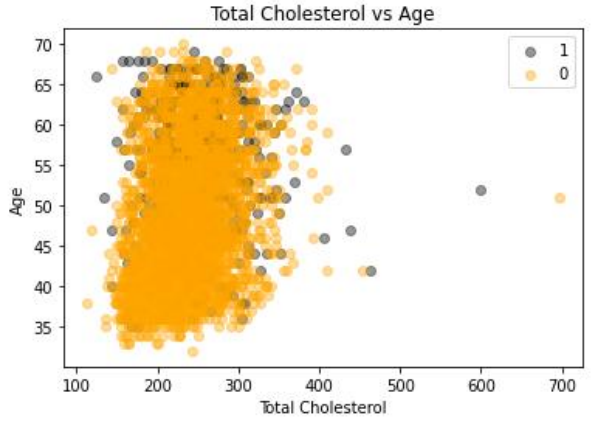


4.5 Visualization of data by Scatter Plot

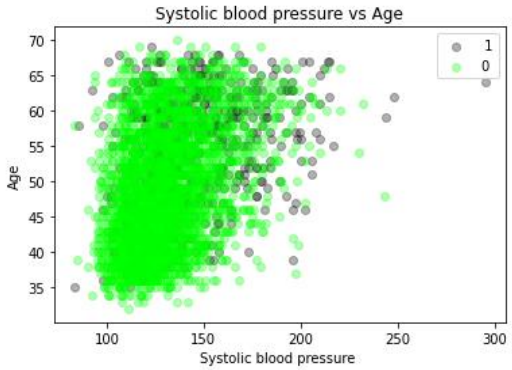
Here we are splitting the data for the better visualization.

```
#Spliting the Data for visualization
A = heart_df[heart_df.TenYearCHD == 1]
B = heart_df[heart_df.TenYearCHD == 0]
```

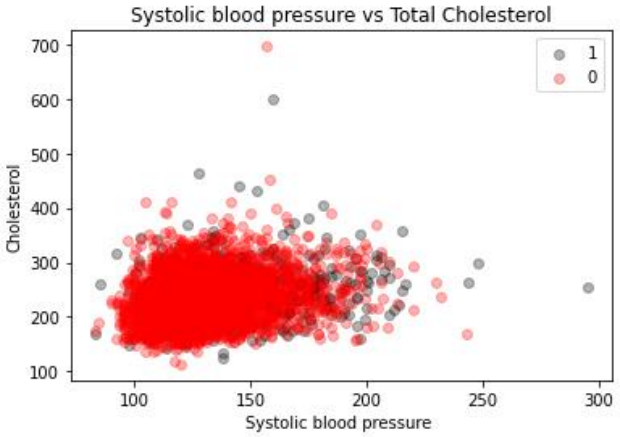
Visualization of Total Cholestrol by Age with Scatter plot.



Visualization of Total Cholestrol by Age with Scatter plot.



Visualization of Systolic Blood Pressure by Total Cholestrol with Scatter plot.



4.6 Training and Testing the Datasets

The data set was separated into training and testing sets for the evaluation process. We have used a sci-kit learn library.

```
import sklearn
new_features=heart_df[['age','Sex_male','cigsPerDay','totChol','sysBP','glucose','TenYearCHD']]
x=new_features.iloc[:,:-1]
y=new_features.iloc[:,-1]
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.11,random_state=5)
```

5. Implementation of Classification Algorithms

5.1 Linear Regression: Logistic Regression

The logistic regression is also known as the sigmoid function which helps in the easy representation in graphs. In this algorithm first, the data should be imported and then trained. By using equation the logistic regression algorithm is represented in the graphs showing the difference between the attributes. From the training data, we have to estimate the best and approximate coefficient and represent [3]. It also provides high accuracy by applying different techniques.

The resulting outcomes are used to prove the logistic regression. Here Logistic regression Algrothim is mainly used for prediction and also calculating the probability of success through the mathematical equation.

Logit Regression Results

Dep. Variable:	TenYearCHD	**No. Observations:**	3751
Model:	Logit	**Df Residuals:**	3736
Method:	MLE	**Df Model:**	14
Date:	Fri, 24 Apr 2020	**Pseudo R-squ.:**	0.1170
Time:	16:09:52	**Log-Likelihood:**	-1414.3
converged:	True	**LL-Null:**	-1601.7
Covariance Type:	nonrobust	**LLR p-value:**	2.439e-71

	coef	std err	z	P>|z|	[0.025	0.975]
const	-8.6532	0.687	-12.589	0.000	-10.000	-7.306
Sex_male	0.5742	0.107	5.345	0.000	0.364	0.785
age	0.0641	0.007	9.799	0.000	0.051	0.077
currentSmoker	0.0739	0.155	0.478	0.633	-0.229	0.377
cigsPerDay	0.0184	0.006	3.000	0.003	0.006	0.030
BPMeds	0.1448	0.232	0.623	0.533	-0.310	0.600
prevalentStroke	0.7193	0.489	1.471	0.141	-0.239	1.678
prevalentHyp	0.2142	0.136	1.571	0.116	-0.053	0.481
diabetes	0.0022	0.312	0.007	0.994	-0.610	0.614
totChol	0.0023	0.001	2.081	0.037	0.000	0.004
sysBP	0.0154	0.004	4.082	0.000	0.008	0.023
diaBP	-0.0040	0.006	-0.623	0.533	-0.016	0.009
BMI	0.0103	0.013	0.827	0.408	-0.014	0.035
heartRate	-0.0023	0.004	-0.549	0.583	-0.010	0.006
glucose	0.0076	0.002	3.409	0.001	0.003	0.012

As per the above logistic regression results, P >= 0.05 shows a low statistically significant relationship with the probability of heart disease. Hence, a backward elimination approach has been used to remove the attributes with the highest P values. The process will be continued until all the attributes of P values less than 0.05.

Logit Regression Results

Dep. Variable:	TenYearCHD	No. Observations:	3751
Model:	Logit	Df Residuals:	3744
Method:	MLE	Df Model:	6
Date:	Fri, 24 Apr 2020	Pseudo R-squ.:	0.1149
Time:	13:56:00	Log-Likelihood:	-1417.7
converged:	True	LL-Null:	-1601.7
Covariance Type:	nonrobust	LLR p-value:	2.127e-76

	coef	std err	z	P>|z|	[0.025	0.975]
const	-9.1264	0.468	-19.504	0.000	-10.043	-8.209
Sex_male	0.5815	0.105	5.524	0.000	0.375	0.788
age	0.0655	0.006	10.343	0.000	0.053	0.078
cigsPerDay	0.0197	0.004	4.805	0.000	0.012	0.028
totChol	0.0023	0.001	2.106	0.035	0.000	0.004
sysBP	0.0174	0.002	8.162	0.000	0.013	0.022
glucose	0.0076	0.002	4.574	0.000	0.004	0.011

The above output indicates the result after using backward elimination. The logistic regression equation for the heart prediction data as follows.

$$logit(P) = \log\left(\frac{P}{1-P}\right) \qquad (1)$$

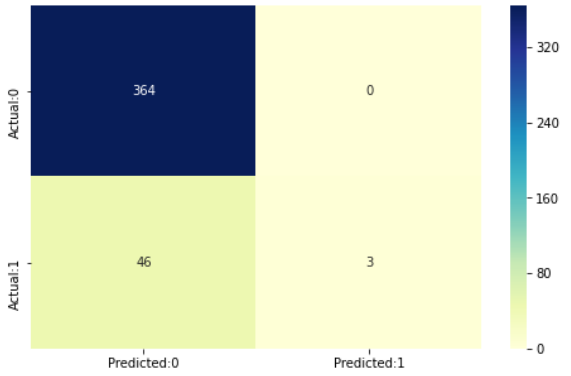
$$= \beta_0 + \beta_1 * Sex + \beta_2 * age + \beta_3 * cigsPerDay + \beta_4 * totChol + \beta_5 * sysBP + \beta_6 * glucose$$

	CI 95%(2.5%)	CI 95%(97.5%)	Odds Ratio	pvalue
const	0.000043	0.000272	0.000109	0.000
Sex_male	1.455242	2.198536	1.788687	0.000
age	1.054483	1.080969	1.067644	0.000
cigsPerDay	1.011733	1.028128	1.019897	0.000
totChol	1.000158	1.004394	1.002273	0.035
sysBP	1.013292	1.021784	1.017529	0.000
glucose	1.004346	1.010898	1.007617	0.000

Confidence Intervals(CI):

Moreover, the accuracy of OR is estimated by using a 95% confidence interval (CI)(2.5%). A large CI(97.5%) represents the low level of precision of OR and also small CI represents the higher precision of OR. However, 95% CI does not indicate the statistical significance, unlike the p-value.

```
from sklearn.linear_model import LogisticRegression
logreg=LogisticRegression()
logreg.fit(x_train,y_train)
y_pred=logreg.predict(x_test)
```




```
LRaccuracy = sklearn.metrics.accuracy_score(y_test,y_pred)*100
print(LRaccuracy)
```

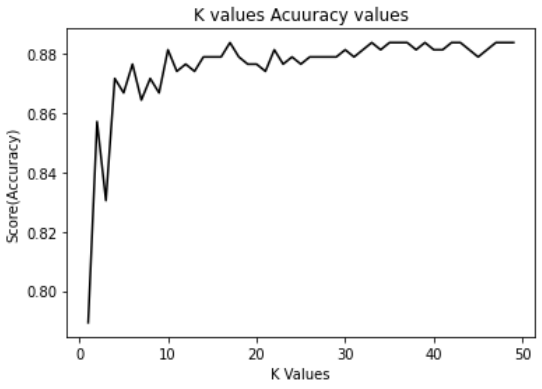
```
88.86198547215496
```

The Accuracy of Logistic Regression Algorithm is 89%.

5.2 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors algorithm is a simple, supervised machine learning algorithm that can be used to solve problems. Here we are using for predicting Cardiovascular diseases using a dataset acquired from Kaggle. Moreover, it has a major drawback of becoming significantly slows as the size of that data in use grows. It is used for classification and regression types of problems. KNN is instance-based learning [5].

```
from sklearn.neighbors import KNeighborsClassifier

scores = []
for each in range(1,50):
    KNNfind = KNeighborsClassifier(n_neighbors = each)
    KNNfind.fit(x_train,y_train)
    scores.append(KNNfind.score(x_test,y_test))

plt.plot(range(1,50),scores,color="black")
plt.title("K values Acuuracy values")
plt.xlabel("K Values")
plt.ylabel("Score(Accuracy)")
plt.show()
```




```
KNNfind = KNeighborsClassifier(n_neighbors = 24) #n_neighbors = K value
KNNfind.fit(x_train,y_train) #Learning model
prediction = KNNfind.predict(x_test)
print("{}-NN Score: {}".format(25,KNNfind.score(x_test,y_test)))
KNNaccuracy = KNNfind.score(x_test,y_test)*100
```

```
25-NN Score: 0.8789346246973365
```

The Accuracy of K-Nearest Neighbor Algorithm is 88%.

5.3 Support Vector Machine

Support Vector Machine is an extremely popular supervised machine learning technique (having a pre-defined target variable) that can be used as a classifier as well as a predictor. A Support Vector Machine model represents the training data points as points in the feature space.

```
SVM Accuracy: 0.8813559322033898
```

The Accuracy of Support Vector Machine Algorithm is 88%.

5.4 Decision Tree

Decision Tree Algorithm is known as the supervised learning algorithm. Moreover, in supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems. The Decision tree Algorithm is a decision support tool that uses a tree-like model. The goal of using a Decision Tree is to create a training model that can use to predict the target variable by learning simple decision rules inferred from training data [6].

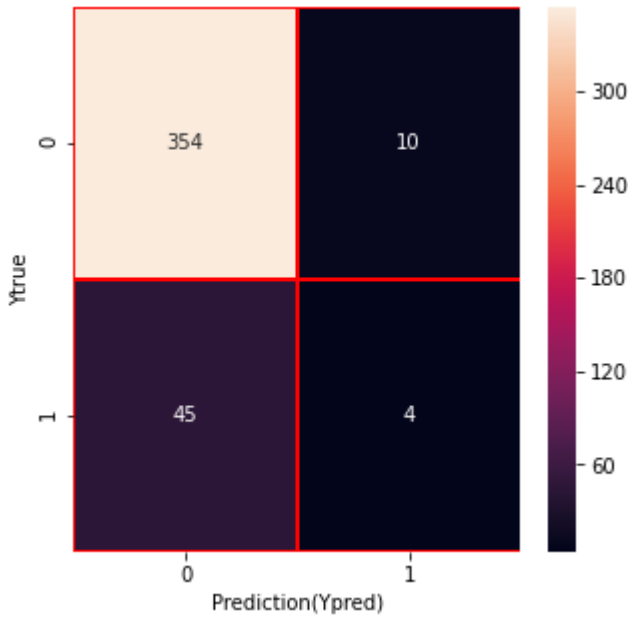
```
Decision Tree accuracy: 0.7990314769975787
```

The Accuracy of Decision Tree Algorithm is 80%.

5.5 Random Forest

Random Forest Algorithm is used for supervised and classification, but mostly it's used for classification problems. It generates decision trees based on data samples and then gets the prediction from each of them. After prediction, it selects the most suitable solution by means of voting. It is an aggregate method that is better than a single decision tree because it decreases the over-fitting by averaging the result [7].

Confusion Matrix for Random Forest




```
Random Forest Score:  0.8668280871670703
```

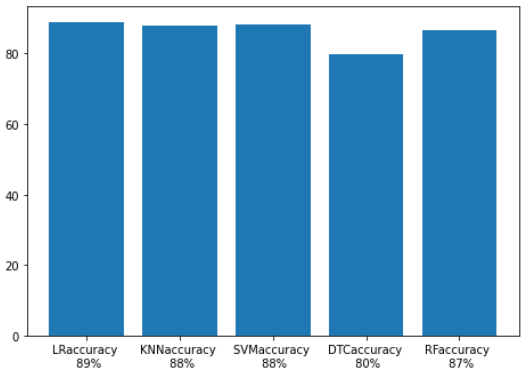
The Accuracy of Random Forest Algorithm is 87%.

6. Comparison of the best Algorithm by Bar Graph

6.1 Accuracy bar graph of all the Algorithms

By analyzing all the five Classification Algorithm of machine learning for the prediction of the cardiovascular disease, It comes to know that logistic regression is the most efficient algorithm out of all five algorithms, as it has 89% accuracy As shown below.

```
LRaccuracy:  88.86198547215496
KNNaccuracy:  87.89346246973365
SVMaccuracy:  88.13559322033898
DTCaccuracy:  79.90314769975787
RFaccuracy:  86.68280871670703
```



After analyzing confusion matrix data, it is apparent that the model is highly specific than sensitive. Moreover, the negative values in the model are predicted more accurately than the positives.

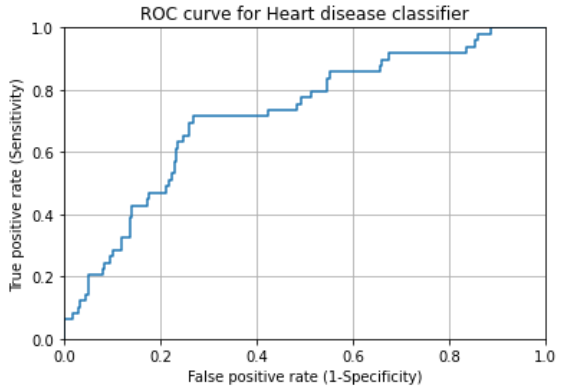
```
With 0.1 threshold the Confusion Matrix is
 [[135 229]
 [  7  42]]
 with 177 correct predictions and 7 Type II errors( False Negatives)
 Sensitivity:  0.8571428571428571 Specificity:  0.3708791208791209

With 0.2 threshold the Confusion Matrix is
 [[286  78]
 [ 25  24]]
 with 310 correct predictions and 25 Type II errors( False Negatives)
 Sensitivity:  0.4897959183673469 Specificity:  0.7857142857142857

With 0.3 threshold the Confusion Matrix is
 [[342  22]
 [ 39  10]]
 with 352 correct predictions and 39 Type II errors( False Negatives)
 Sensitivity:  0.20408163265306123 Specificity:  0.9395604395604396

With 0.4 threshold the Confusion Matrix is
 [[359   5]
 [ 46   3]]
 with 362 correct predictions and 46 Type II errors( False Negatives)
 Sensitivity:  0.061224489795918366 Specificity:  0.9862637362637363
```

6.2 Predicting through the probability of total number of Heart Disease.

	Probability of no Heart Disease (0)	Probability of Heart Disease (1)
0	0.867888	0.132112
1	0.939146	0.060854
2	0.787333	0.212667
3	0.809448	0.190552
4	0.881475	0.118525

6.3 ROC curve for Heart disease classifier

The Receiver Operating Characteristic (ROC) curve is a simple plot used to visualize the performance of a binary classifier. Good classification accuracy models should have significantly more true positives than false positives at all thresholds. Area Under the Curve (AUC) quantifies the model classification accuracy.




```
Area Under The Curve: 71.93316887194439
```

7. Conclusion

The main aim of this research is to compare the accuracy of all the classification algorithms to evaluate the risk of 10-year CHD using 14 IVs, we come to know that Logistic Regression is a more appropriate algorithm to predict the risk. The following attributes are selected after the backward elimination process considering the values of P, which are lower than 5%. The primary motive of this research is the prediction of heart disease with a high rate of accuracy by comparing all five classification algorithms. Further, the accuracy of the Logistic Regression model is 0.89 which is best out of all five algorithms. The value under the AUC curve is 72 which is somewhat satisfactory.

8. Future Work

Nowadays most of the data is computerized, the data is distributed everywhere but we're not utilizing it properly. By Analyzing the available data we can also use for unknown patterns. The motive of this future work is to predict heart diseases with high rate of accuracy by using the Classification Algorithms of Machine Learning. For predicting the heart disease with the help of different parameters, we can use Logistic Regression, Support Vector Machine, KNN, Decision Tree, naviebayes, sklearn in machine learning Algorithm. Moreover, the model could be improved by using more data and techniques. The future scope of the paper is the prediction of heart diseases by using advanced techniques, with a high rate of accuracy and algorithms in less time complexity.

References

[1] V.V. Ramalingam*, Ayantan Dandapath, M Karthik Raja; Heart disease prediction using machine learning techniques : a survey – March 2018 https://www.sciencepubco.com/index.php/ijet/article/view/10557
[2] A. S. Thanuja Nishadi University of Colombo, Faculty of Graduate Studies, Sri Lanka, Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab- Volume 3 Issue 8, August 2019. http://www.ijarp.org/published-research-papers/aug2019/Predicting-Heart-Diseases-In-Logistic-Regression-Of-Machine-Learning-Algorithms-By-Python-Jupyterlab.pdf
[3] Reddy Prasad,Pidaparthi Anjali, S.Adil, N.Deepa. Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning - ISSN: 2249 – 8958, Volume-8, Issue-3S, February 2019. https://www.ijeat.org/wp-content/uploads/papers/v8i3S/C11410283S19.pdf
[4] Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. (2019). Applications of Artificial Intelligence in Transport: An Overview. Sustainability, 11(1), 189. doi: 10.3390/su11010189
[5] Cunningham, Padraig & Delany, Sarah. (2007). k-Nearest neighbour classifiers. Mult Classif Syst.
[6] Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5.
[7] Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues (IJCSI). 9.

Author Profile

Yash Jayesh Chauhan is a second year student pursing Computer Science in Parul Institute of Engineering and Technology, Parul University. He is very fascinated with technology and he also represented his university for more than 27 times at National/ International hackathons, summits and conferences. His research Interset is in Machine Learning, Deep Learning, Artificial Intelligence, Virtual reality, Computer Vision, Augmented Reality, Gestural Interaction, Automation, Natural Language Processing. "I'll never stop striving for what I've always wanted. Though at some instances it became difficult to outperform students from IITs, I didn't give up and I never will" said by **Yash**.