SISTER NIVEDITA UNIVERSITY

HEART DISEASE PREDICTION USING MACHINE LEARNING

A project submitted in partial fulfilment of the requirements for the degree of
Bachelor of Technology

PRESENTED BY

SHUVRA CHAKRABARTY

ROLL NO.- 1811100002032
REG. NO. - 180020328417

Under the Supervision of

PROF. SWARUP KR. GHOSH

Assistant Professor, Department of CSE, SNU

CERTIFICATE

To Whom It May Concern

This is to certify that the project "**HEART DISEASE PREDICTION USING MACHINE LEARNING**" is a work done by **SHUVRA CHAKRABARTY** student of 5th Semester in B.Tech (CSBS), department of Computer Science and Engineering, Sister Nivedita University, in partial fulfilment of requirement for the degree of "***Bachelor of Technology in Computer Science and Engineering/ Computer Science and Business System*** ".

………………………………….. ………………………………..

Chairman, **[NAME OF GUIDE]**

Board of Examiner **(**Project Guide)

Department of Computer Sc. and Engineering

Sister Nivedita University **Sister Nivedita University**

……………………………..

Signature of External Examiner

Contents:

1. INTRODUCTORY CONCEPTS

1.1. INTRODUCTION

Over the last decade, heart disease remains the primary basis of death worldwide. An estimate by the World Health Organization, that over 17.9 million deaths occur every year worldwide because of heart disease, and of these deaths, 80% are because of coronary artery disease and cerebral stroke. The heart pumps blood through the circulatory system of the body. In all body part the blood, oxygen is circulated by the circulatory system of the body and if the heart does not work properly then the whole human blood system will be collapsed. So, if the heart does not function properly then it will lead to a serious health condition, it could even lead to death.

The symptoms of the Heart Attack:

1. Chest Pain: This is the most common sign of a heart attack is chest pain. It mainly happens cause of the blockage of the coronary vessel of the body due to the plaque.

2. Arms pain: The pain normally starts in the chest and move towards the arm mainly left arm.

3. Low in oxygen: Because of the plaque the level of oxygen drops in the body and causes the dizziness and loss of balance.

4. Tiredness: this cause for fatigues means simple chores become harder to do.

5. Excessive Sweating: Another common symptom is sweating.

6. Diabetics: In this, the patients have a heart rate of ~ 100 bpm and also occasionally having a heart rate of 130bpm.

7. Bradycardia: In this, the patient will have a slower heartbeat of 60 bpm.

In India, more than 2 lake open heart surgeries are done per year. The patients affected by the heart attack is growing in India is 20% to 30 % every year. The development of the sensor network in the human monitoring system is more applicable from recent years. In this project I use some machine leering Algorithm to predict heart Disease.

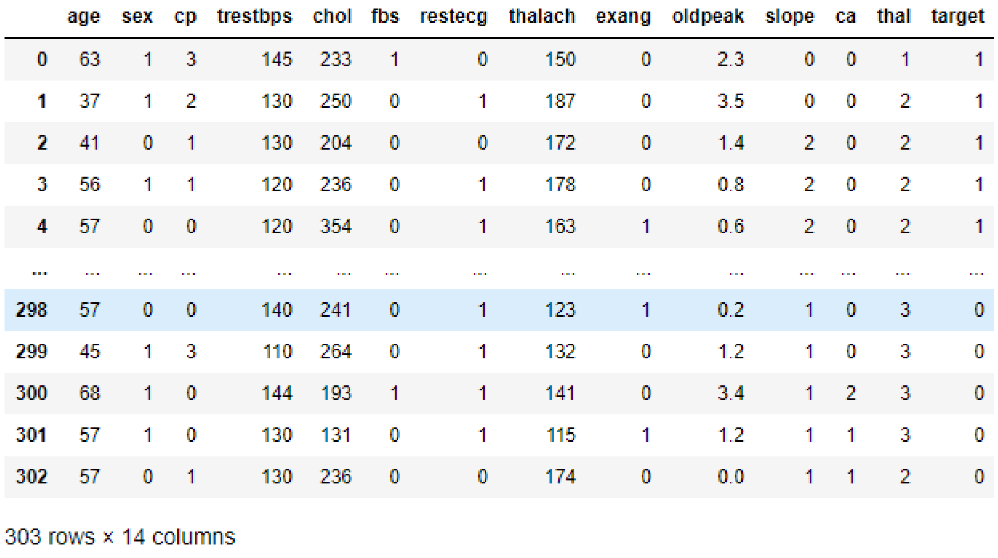
1.2. APPROACH METHODOLOGY

My Project aims to foresee the odds of having heart disease as probable cause of computerized prediction of heart disease that is helpful in the medical field for clinicians and patients. To accomplish the aim, we have discussed the use of various machine learning algorithms on the data set and dataset analysis is mentioned in this research paper. This paper additionally depicts which attributes contribute more than the others to anticipation of higher precision. This may spare the expense of different trials of a patient, as all the attributes may not contribute such a substantial amount to expect the outcome.

2. PROPOSEDMETHODOLOGY

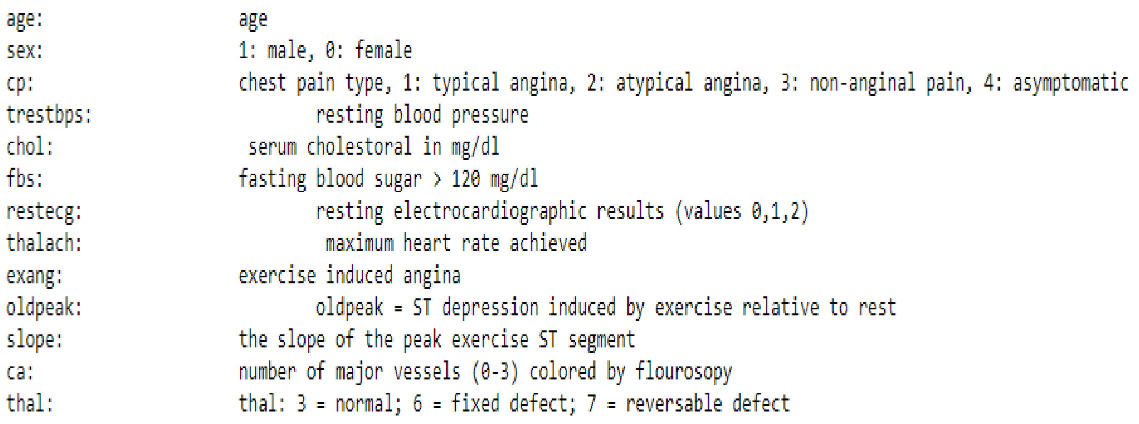
2.1. Data Source

For this project I have collected the dataset from Kaggle. It comprises a real dataset of 300 examples of data with 14 various attributes (13 predictors; 1 class) like blood pressure, type of chest pain, electrocardiogram result, etc. IN this research, we have used four algorithms to get reasons for heart disease and create a model with the maximum possible accuracy.



	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
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

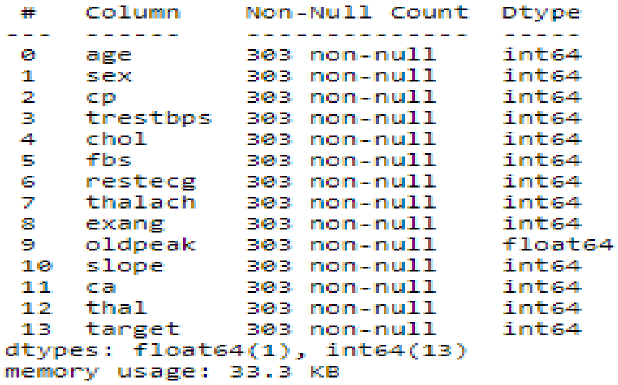
303 rows × 14 columns



age:	age
sex:	1: male, 0: female
cp:	chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
trestbps:	resting blood pressure
chol:	serum cholestoral in mg/dl
fbs:	fasting blood sugar > 120 mg/dl
restecg:	resting electrocardiographic results (values 0,1,2)
thalach:	maximum heart rate achieved
exang:	exercise induced angina
oldpeak:	oldpeak = ST depression induced by exercise relative to rest
slope:	the slope of the peak exercise ST segment
ca:	number of major vessels (0-3) colored by flourosopy
thal:	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

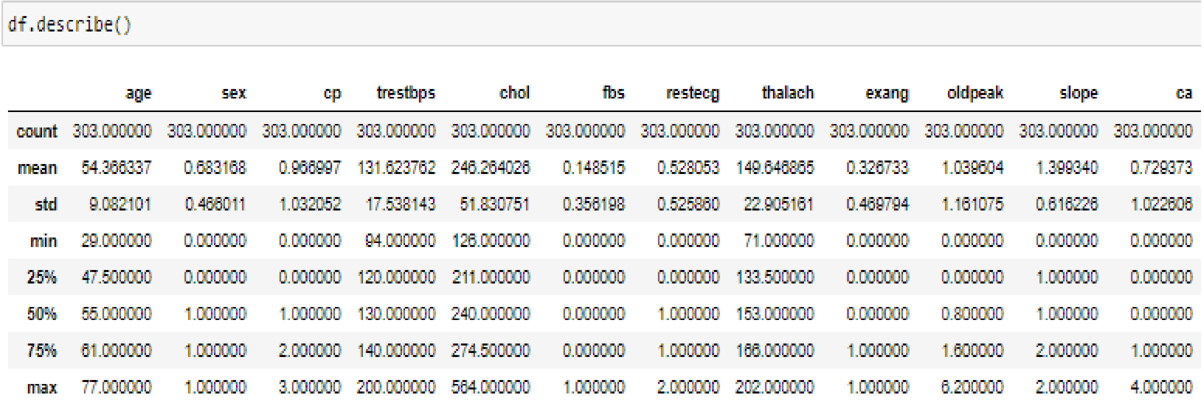
2.2. UNDERSTANDING THE DATA

As you can see from the dataset, there are a total of 13 features and 1 target variable. Also, there are no missing values so we don't need to take care of any null values.



```
 #   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  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Next, I used describe() method. The method revealed that the range of each variable is different. The maximum value of age is 77 but for chol it is 564. Thus, feature scaling must be performed on the dataset.

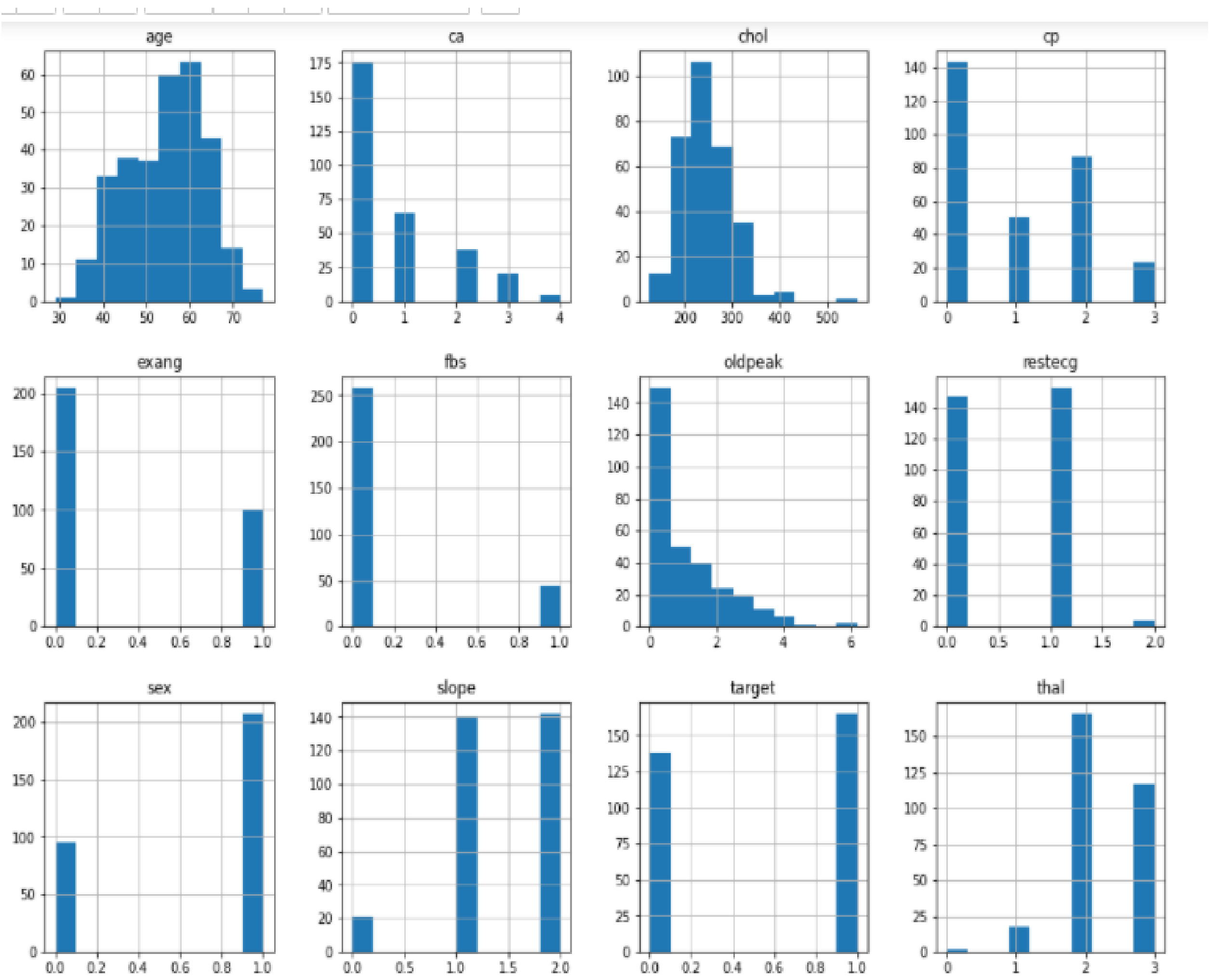


```
df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
count	303.000000	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	0.729373
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	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	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000

2.3. HISTOGRAM

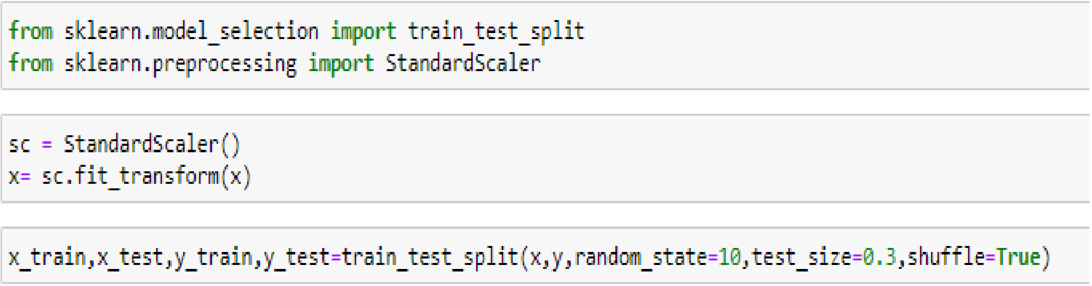
The best part about this type of plot is that it just takes a single command to draw the plots and it provides so much information in return. Just use dataset.hist() . Let's take a look at the plots. It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling. Next, wherever you see discrete bars, it basically means that each of these is actually a categorical variable. We will need to handle these categorical variables before applying Machine Learning. Our target labels have two classes, 0 for no disease and 1 for disease.



2.4. DATA PRE-PROCESSING

There are many projects have the real-life information contains large numbers with missing and noisy data. These data are pre-processed to overcome such issues and make predictions vigorously.

You can see in this dataset there are no missing value so I no need to done any operation to fix this. So, the data are then classified and split into training data set and test data set which is run on various algorithms to achieve accuracy score results. To work with this data set I split this dataset into 70 training data and 30% testing data.



```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

```
sc = StandardScaler()
x= sc.fit_transform(x)
```

```
x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=10,test_size=0.3,shuffle=True)
```

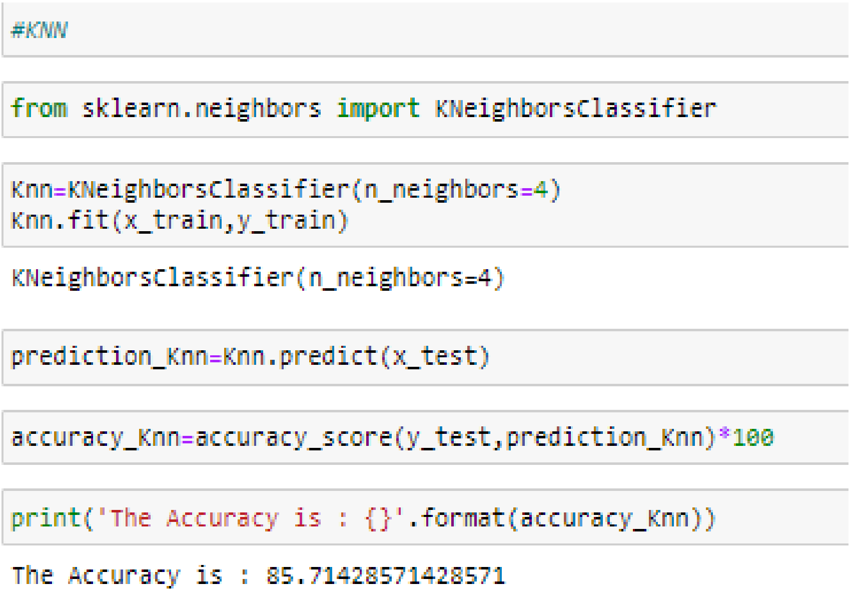
dataset for which we will use the StandardScaler.

2.5. ALGORITHM USE:

2.5.1. K-NEAREST NEIGHBORS:

The K-nearest neighbors algorithm is a supervised classification algorithm method. It classifies objects dependant on nearest neighbor. It is a type of instance-based learning. The calculation of distance of an attribute from its neighbors is measured using Euclidean distance. It uses a group of named points and uses them on how to mark another point. The data are clustered based on similarity amongst them, and is possible to fill the missing values of data using K-NN. Once the missing values are filled, various prediction techniques apply to the data set. It is possible to gain better accuracy by utilizing various combinations of these algorithms. K-NN algorithm is simple to carry out without creating a model or making other assumptions.

In This project I use KNN algorithm to predict the accuracy where K=4 then I got about 85.71% accuracy.



```
#KNN
```

```
from sklearn.neighbors import KNeighborsClassifier
```

```
Knn=KNeighborsClassifier(n_neighbors=4)
Knn.fit(x_train,y_train)
```

```
KNeighborsClassifier(n_neighbors=4)
```

```
prediction_Knn=Knn.predict(x_test)
```

```
accuracy_Knn=accuracy_score(y_test,prediction_Knn)*100
```

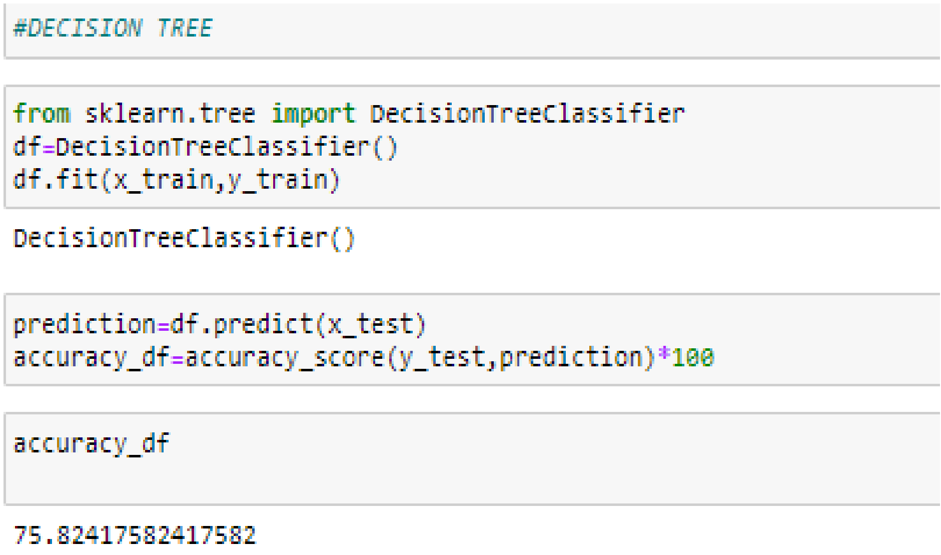
```
print('The Accuracy is : {}'.format(accuracy_Knn))
```

```
The Accuracy is : 85.71428571428571
```

2.5.2. DECISION TREE:

In Decision tree is a classification algorithm that works on categorical as well as numerical data. Decision tree is used for creating tree-like structures. Decision tree is simple and widely used to handle medical dataset. It is easy to implement and analyse the data in tree-shaped graph. The decision tree model makes analysis based on three nodes.

This algorithm has higher accuracy in comparison to other algorithms as it analyses the dataset in the tree-like graph. However, the data may be over classified and only one attribute is tested at a time for decision-making. An accuracy of **75.82%** has been achieved by the decision tree.

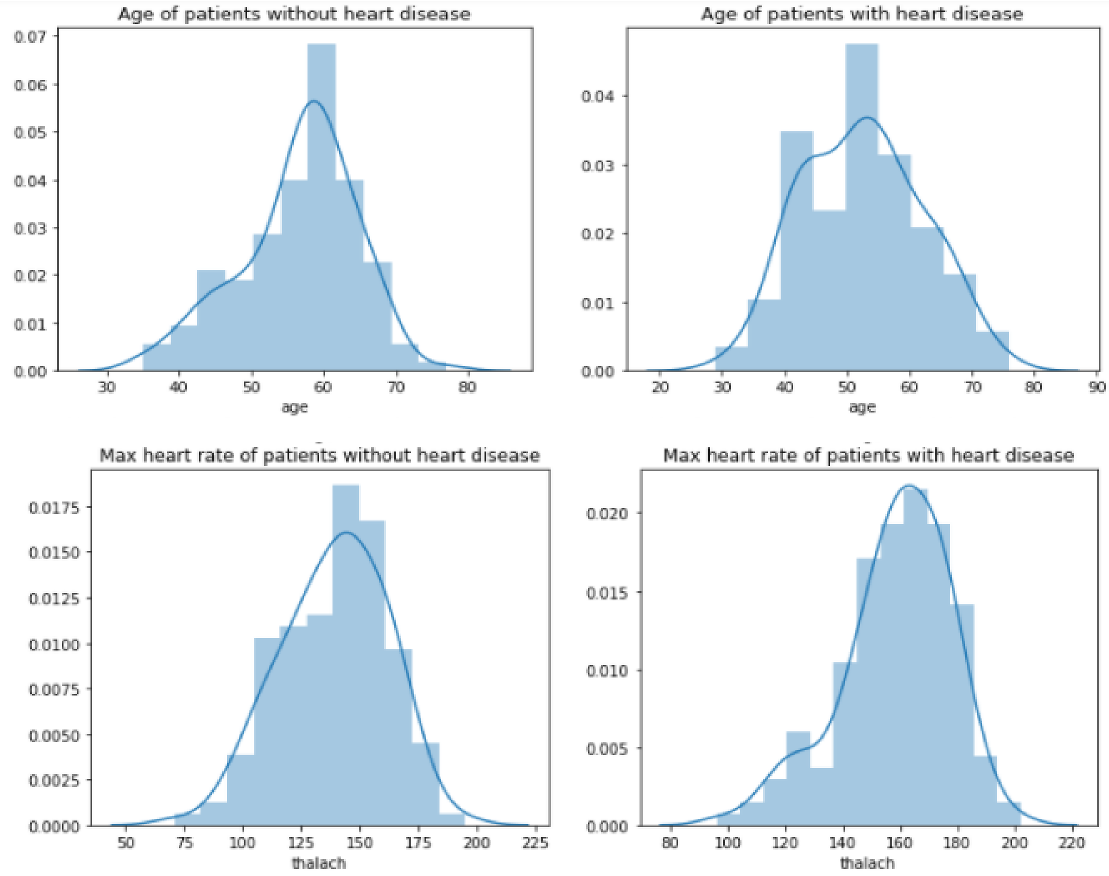


```
#DECISION TREE

from sklearn.tree import DecisionTreeClassifier
df=DecisionTreeClassifier()
df.fit(x_train,y_train)

DecisionTreeClassifier()

prediction=df.predict(x_test)
accuracy_df=accuracy_score(y_test,prediction)*100

accuracy_df

75.82417582417582
```

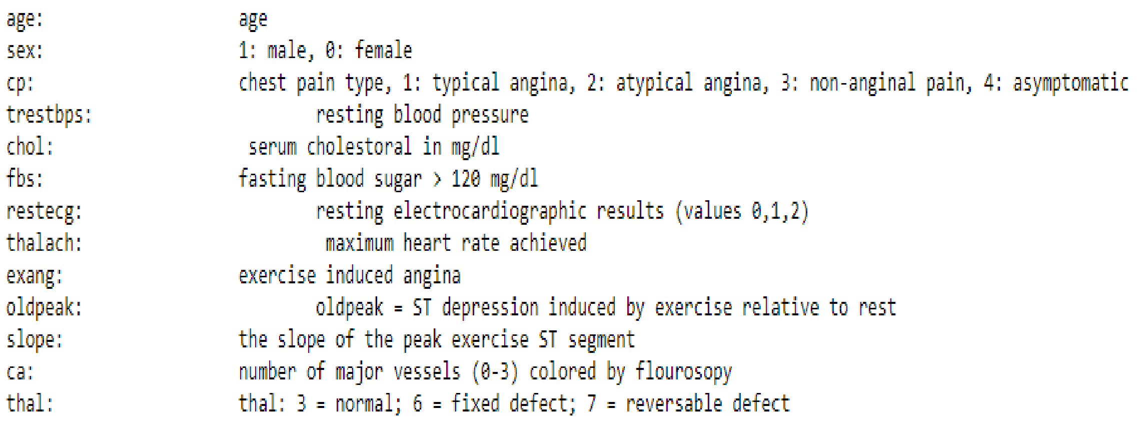
3. EXPREMENTAL RESULT AND DISCUSSION

3.1. Database

The Dataset For this project I have used the heart dataset from the KEGGLE Repository. The dataset consists of 303 individual clinical reports in which 164 did not have any disease. In this dataset there is a total of 97 female patients in which 25 people are the affirmative case, also there are 206 male patients in which 114 are diagnosed with the disease



Thus, we have 13 features that are shown below -------



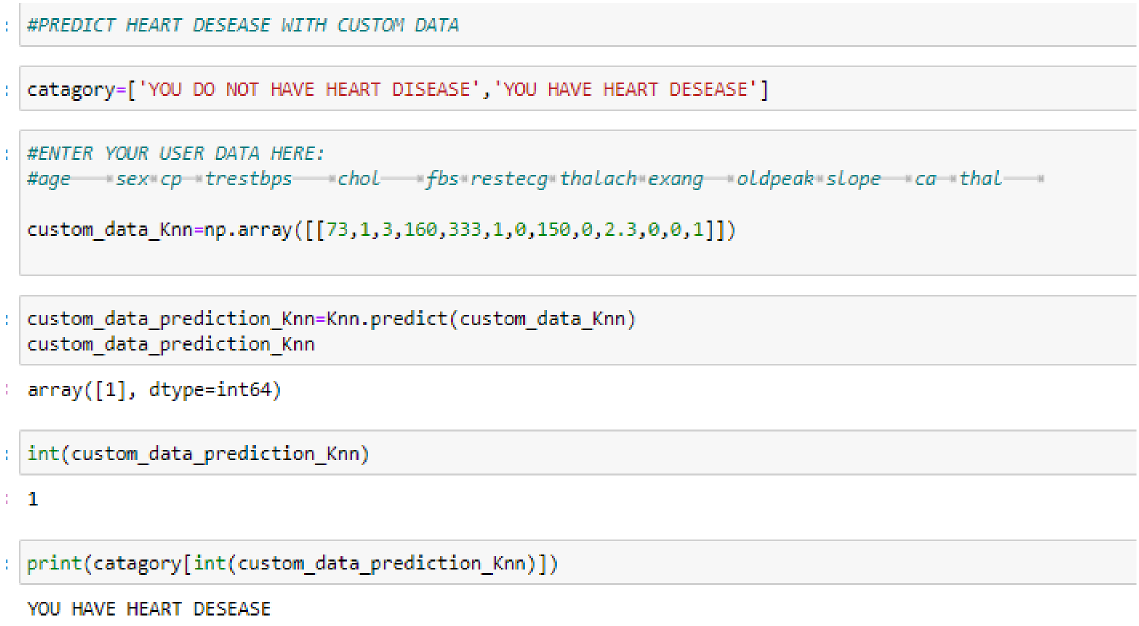
age:	age
sex:	1: male, 0: female
cp:	chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
trestbps:	resting blood pressure
chol:	serum cholestoral in mg/dl
fbs:	fasting blood sugar > 120 mg/dl
restecg:	resting electrocardiographic results (values 0,1,2)
thalach:	maximum heart rate achieved
exang:	exercise induced angina
oldpeak:	oldpeak = ST depression induced by exercise relative to rest
slope:	the slope of the peak exercise ST segment
ca:	number of major vessels (0-3) colored by flourosopy
thal:	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

3.2. Experiment

In this project I use the data set from Kaggle and using two machine learning algorithm I want to predict heart disease. During this process I can find any null value in this data set for that reason I scaled the data set and train it. Using those algorithms, I got different accuracy.

3.3. RESULT

The Aim of this research is to predict whether or not a patient will develop heart disease. This research was done on supervised machine learning classification techniques using decision tree and K-nearest neighbor on Kaggle repository. Various experiments using different classifier algorithms were conducted through the Jupiter notebook tool. Dataset was classified and split into a training set and a test set. Pre-processing of the data is done and supervised classification techniques such as decision tree, K-nearest neighbor are applied to get accuracy score. The accuracy score results of different classification techniques were noted using Python Programming for training and test data sets.



```
#PREDICT HEART DESEASE WITH CUSTOM DATA
```

```
catagory=['YOU DO NOT HAVE HEART DISEASE','YOU HAVE HEART DESEASE']
```

```
#ENTER YOUR USER DATA HERE:
#age    sex cp  trestbps    chol    fbs restecg thalach exang   oldpeak slope   ca  thal    

custom_data_Knn=np.array([[73,1,3,160,333,1,0,150,0,2.3,0,0,1]])
```

```
custom_data_prediction_Knn=Knn.predict(custom_data_Knn)
custom_data_prediction_Knn
```

```
array([1], dtype=int64)
```

```
int(custom_data_prediction_Knn)
```

```
1
```

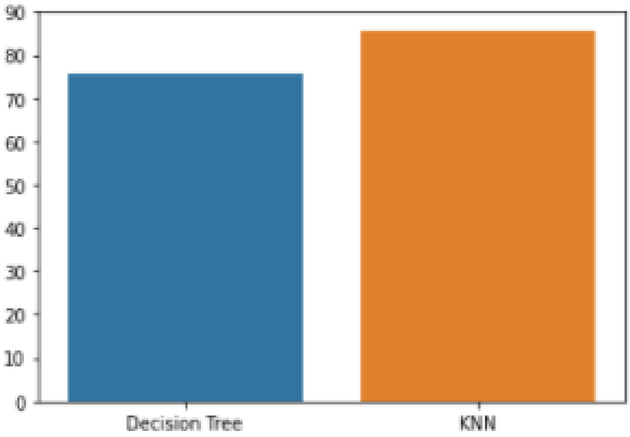
```
print(catagory[int(custom_data_prediction_Knn)])
```

```
YOU HAVE HEART DESEASE
```

3. CONCLUSION AND FUTURE WORK

4.1. CONCLUSION

The overall aim is to define various data mining techniques useful in effective heart disease prediction. Efficient and accurate prediction with a lesser number of attributes and tests is our goal. In this study, I consider only 14 essential attributes. I applied two techniques, K-nearest neighbour, decision tree, the Percentage accuracy results of classification techniques K-nearest neighbour and Decision Tree is 85.15789474 (K=4) 76.68421052631578. The data were pre-processed and then used in the model. K-nearest neighbour algorithms showing the best results in this model. I found the accuracy after implementing two algorithms to be highest in K-nearest neighbours (k = 4).



4.2 Future:

In further I expand this project incorporating other techniques such as Random Forest Classifier, time series, clustering and association rules, support vector machine, and genetic algorithm. Considering the limitations of this study, there is a need to implement more complex and combination of models to get higher accuracy for early prediction of heart disease. I also try to build a platform for patients like, - website or app, so they can easily check their heart condition.

5. REFERENCES

- Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7.
- Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Dis. 2015;7(1):129–37.
- Chaurasia V, Pal S. Data mining approach to detect heart diseases. Int J Adv Comput Sci Inf Technol (IJACSIT). 2014;2:56–66.
- Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease use case. In: 2014 13th international conference on machine learning and applications. IEEE. p. 482–86.