Tech Saksham

Capstone Project Report

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FUNDAMENTALS

HEARTDISEASE PREDICTION

ANNA UNIVERSITY REGIONAL CAMPUS -

TIRUNELVELI

NM ID	NAME
A8BF2E9BCADCA22BC4B4025B15F33CC5	DIVINA H

Ramar Bose

Sr. AI Master Trainer

ABSTRACT

Heart disease symptoms are caused by abnormal heartbeats and diseased heart muscle. There are two cases of prediction of which the correct prediction can help to prevent threats whereas incorrect prediction can lead to fatal. This report is based on heart disease prediction using logistic regression model. Machine Learning is one the trending technology in which various researches around the world is used for predicting diseases. Nowadays it is important for the early detection and for its treatment. The dataset consists of 14 attributes used for performing the analysis. Accuracy is validated and promising results are achieved. Heart disease dataset analyses is used to predict the result whether the patient has heart disease or not i.e., using logistic regression technique. This prediction gives the result in the form of logistic representations which produces efficient and accurate results in healthcare sectors.

INDEX

CHAPTER 1

INTRODUCTION

1.1 Machine Learning :

Machine learning is used to provide the good learning to the machines and analyze some pattern for handling the data in extra efficient manner. Sometimes, it may happen that after viewing the data, we even unable to predict the actual pattern or acquire the valuable information from the data. In this condition, we have to go for machine learning. The motive of machine learning is to grasp some knowledge from the data by themselves. Even, many studies have been terminated which highlights the purpose of machine learning that how do machines learn by its.

Machine Learning Techniques :

The main ML techniques can be classified as follows.

Supervised Learning :

The supervised machine learning algorithms are those which demand some external assistance. The input dataset splits into training and test dataset. The trained dataset composed of output variable which is to be predicted or classified. Each algorithm gets to know a specific pattern from the training dataset and just apply them to the test dataset for prediction or classification purposes. This algorithm is named as supervised learning in view of the fact that the process of algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. Three most prominent supervised learning algorithms are considered below.

a) Logistic Regression

b) K Neighbor Classifier

c) Random Forest Classifier

1.2 Problem Statement:

Modern information technology tools and techniques such as AI, machine learning and data mining could help support healthcare professionals by providing them with the information they need to make decisions that will minimize deaths caused by heart disease at minimal cost. For example, machine learning algorithms can mine large databases to identify frequent patterns that eventually lead to heart disease and death.

1.3 Proposed System:

The proposed system has datasets consist of 14 main attributes for the prediction of heart disease. The system is based on logistic regression method in which efficient algorithms are utilized for heart disease prediction which helps to prevent at the earlier stage. In this system, the data is collected and then converted it into knowledge by data analysis. This proposed system consists of data which determines whether the patient has heart diseases or not with the help of some parameters. Here, we use the sources for datasets from Kaggle. The proposed system consisting of sklearn library which helps in testing and training of the datasets. sklearn also known as Scikit-learn is probably most useful library for implementing machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modelling including classification, regression, clustering and also dimensionality reduction. Logistic Regression is a type of regression analysis which is used in statistics to predict the outcome of a categorical dependent variable from a collection of predictors or independent variables. In logistic regression the dependent variable always represented in binary. Logistic regression method mainly in prediction and calculation of the probability of success. The proposed system splits the data into two parts one for testing and the other for training using sklearn library functions imported. The logistic regression model to call the labels by logistic model and use the accuracy function to predict the labels and find the accuracy of the model. Out of the 14 attributes 13 are of integer data type and the other 1 is of floating data type. The data we are having, is classified into different structured data based on the features of the patient's heart. From the available data, we need to create a model that predicts the patient's disease using a logistic regression algorithm. First, we have to import datasets and then read the datasets; the data should contain different variables such as age, gender, sex, chest pain, slope, target, resting blood pressure, thalach, etc. The data need to be explored so that the information gets verified. After that create a temporary variable and also build a model for logistic regression, K Nearest Neighbor Classifier and Random Forest Classifier. By using these algorithms, the accuracy is increased as compared to the other works done on the existing system.

1.4 Objectives :

The main objective of this project is to explore how Machine Learning algorithms can be used in the diagnosis of heart disease by building an optimized model that can be used to predict heart diseases.

1.5 Advantages :

Using machine learning to classify cardiovascular disease occurrence can help diagnosticians reduce misdiagnosis. This project develops a model that can correctly predict cardiovascular diseases to reduce the fatality caused by cardiovascular diseases..

CHAPTER 2

PROJECT SYSTEM REQUIREMENT

2.1 GENERAL:

Requirements are the basic constrains that are required to develop a system. Requirements are collected while designing the system. The following are the requirements that are to be discussed.

2.1.1 Functional requirements

2.2.2 Non-Functional requirements

2.1.3 Environment requirements

a) Hardware requirements

b) Software requirements

2.2.1 FUNCTIONALREQUIREMENTS:

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details to follow the special libraries like sklearn, pandas, NumPy, mat plot lib and sea born**.**

2.2.2 NON-FUNCTIONAL REQUIREMENTS:

Process of functional steps,

1. Problem definition

2.Preparing data

3.Evaluating algorithms

4. Improving results

5. Prediction the result

2.1.3ENVIRONMENTAL REQUIREMENTS:

a) Software Requirements:

Operating System: Windows

Tool: Google Colaboratory

Internet

b) Hardware requirements:

Processor : Intel (R) Core i3

Hard disk : minimum 80 GB

RAM: minimum 2GB

CHAPTER 3

METHODOLOGY

3.1 Description of the Dataset:

Data contains;

- age - age in years
- sex - (1 = male; 0 = female)
- cp - chest pain type
- trestbps - resting blood pressure (in mm Hg on admission to the hospital)
- chol - serum cholestrol in mg/dl
- fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg - resting electrocardiographic results
- thalach - maximum heart rate achieved
- exang - exercise induced angina (1 = yes; 0 = no)
- 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 - 3 = normal; 6 = fixed defect; 7 = reversable defect
- target - have disease or not (1=yes, 0=no)

3.2 Project Development and design

3.2.1 Technical Approach & Python libraries:

The solution I am proposing is to use Python as the primary language in implementing a machine learning model that will predict whether a person has a heart disease or not. This model will be implemented through Google Colaboratory. I will import several libraries such as:

- Pandas – this library in Python is used for data analysis and machine learning. It is used to manipulate and analyze the csv files and data frame.

- Matplotlib – this library in Python is used for plotting and visualizing the data. It will help in plotting graphs, line plots, histograms, etc. as well define parameters for the charts, and color the charts.

- train_test_split – this scikit-learn library in Python will be used to split the dataset into two parts – training dataset and test dataset.

- StandardScalar – this scikit-learn library in Python will be performed as a preprocessing step in order to standardize the functionality of the input dataset.

Next, using the scikit-learn, or sklearn, library in Python, I import the three machine learning algorithms that will be used to implement the model which will predict heart disease in a person. These algorithms are Logistic Regression, K Nearest Neighbors Classifier and Random Forest Classifier.

3.2.2 PROPOSED WORK AND ALGORITHM:

The collection of data and selection of the most crucial attributes is the first step in the system's operation. The relevant data is then preprocessed into the format needed. After that, the data is split into training and testing data. The algorithms are used, and the training data is used to train the model. By testing the system with test data, the correctness of the system is determined. The modules listed below are used to implement this system.

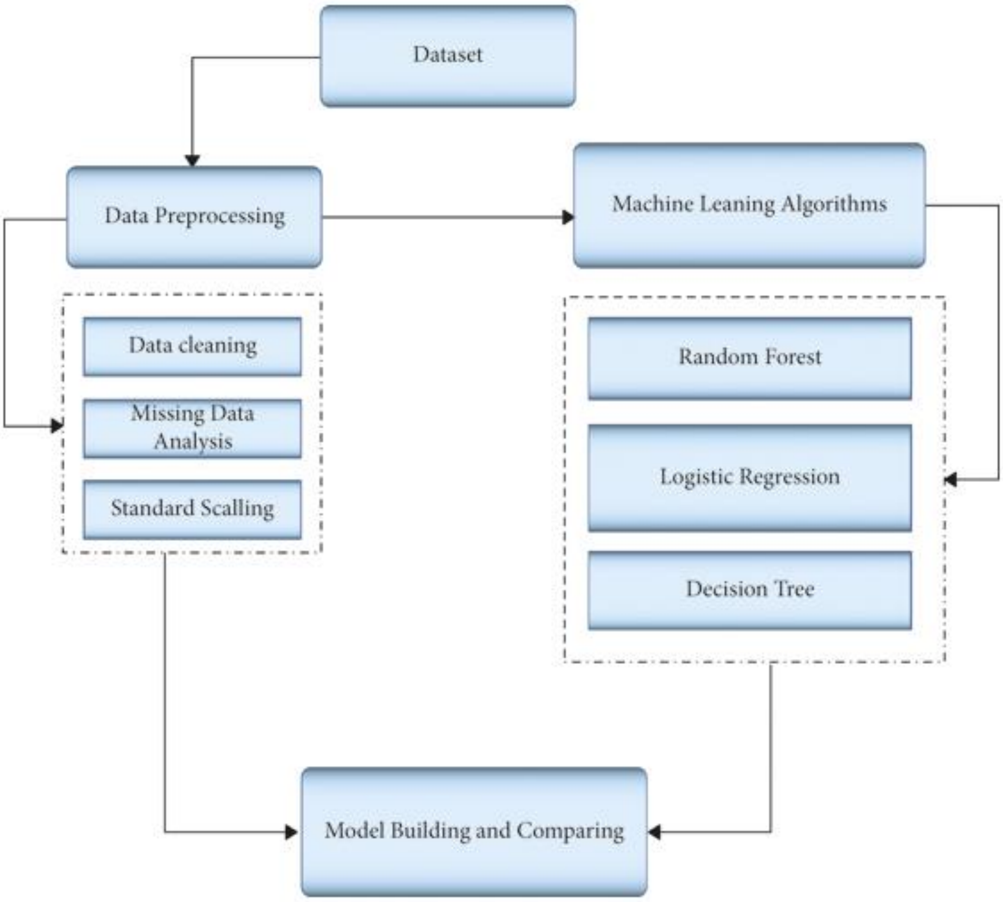
1. Collection of Dataset

2. Selection of attributes

3. Data Pre-Processing

4. Balancing of Data

5. Disease Prediction



FLOWCHART OF METHODOLOGY

a) Logistic Regression:

Logistic Regression is a statistical approach that is often used to solve issues involving binary classification. Rather than fitting a straight line or hyperplane, logistic regression employs the logistic function to constrain the output of a linear equation to the range of 0 to 1. Due to the presence of 13 independent variables, logistic regression is well suited for categorization.

b) K Nearest Neighbor Classifier:

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

c) Random Forest Classifier:

Random Forest Algorithm is a supervised learning algorithm used for both classification and regression. This algorithm works on 4 basic steps

1. It chooses random data samples from a dataset.
2. It constructs decision trees for every sample dataset chosen.
3. At this step every predicted result will be compiled and voted on.
4. At last most voted predictions will be selected and be presented as the result of classification.

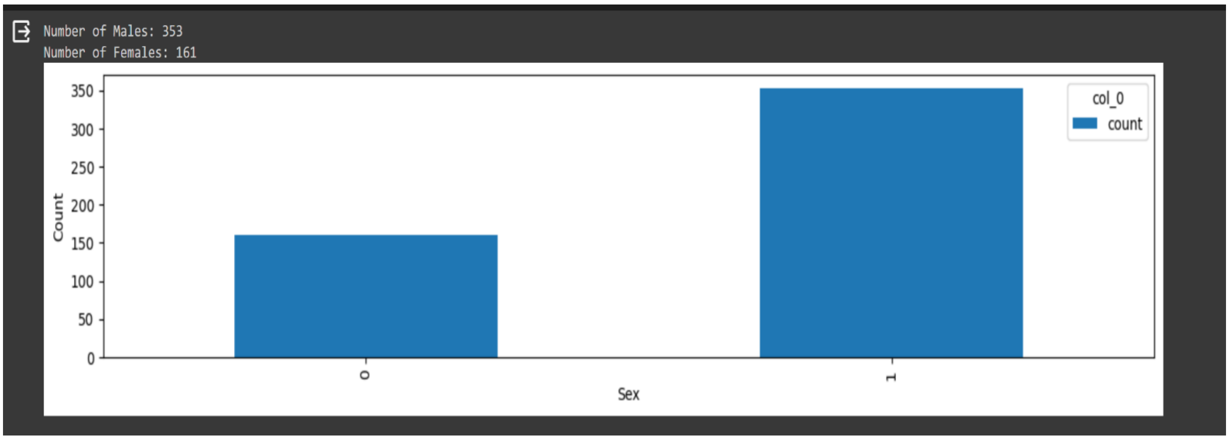
In this project we have used a random forest classifier and the result given is 100% accuracy.

CHAPTER 4

PROJECT OUTCOME

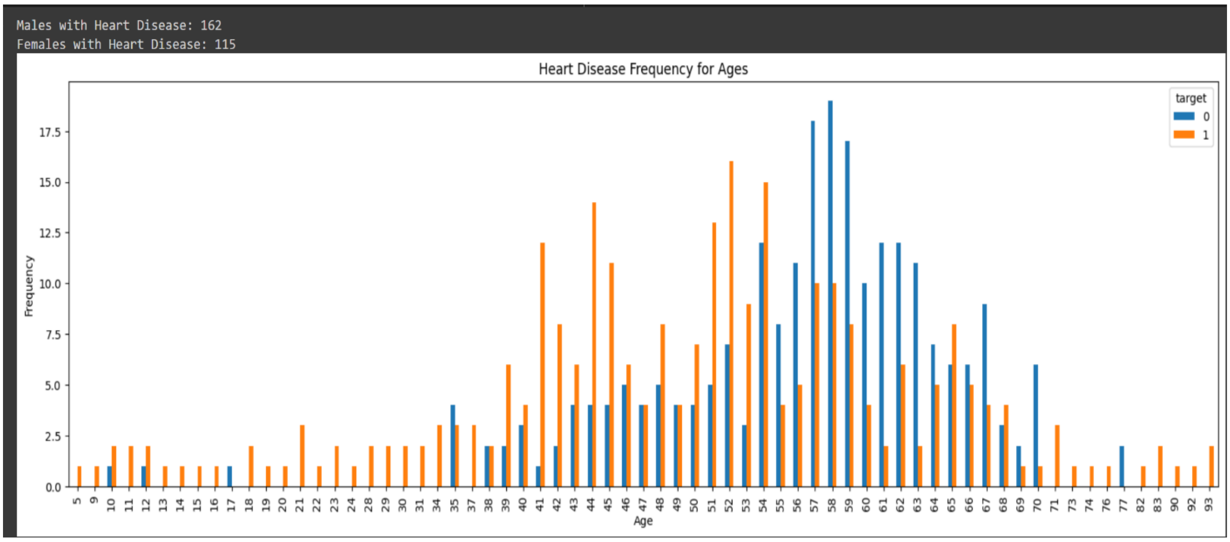
I. EDA:

1. Number of males and females whose heart data is stored in the dataset



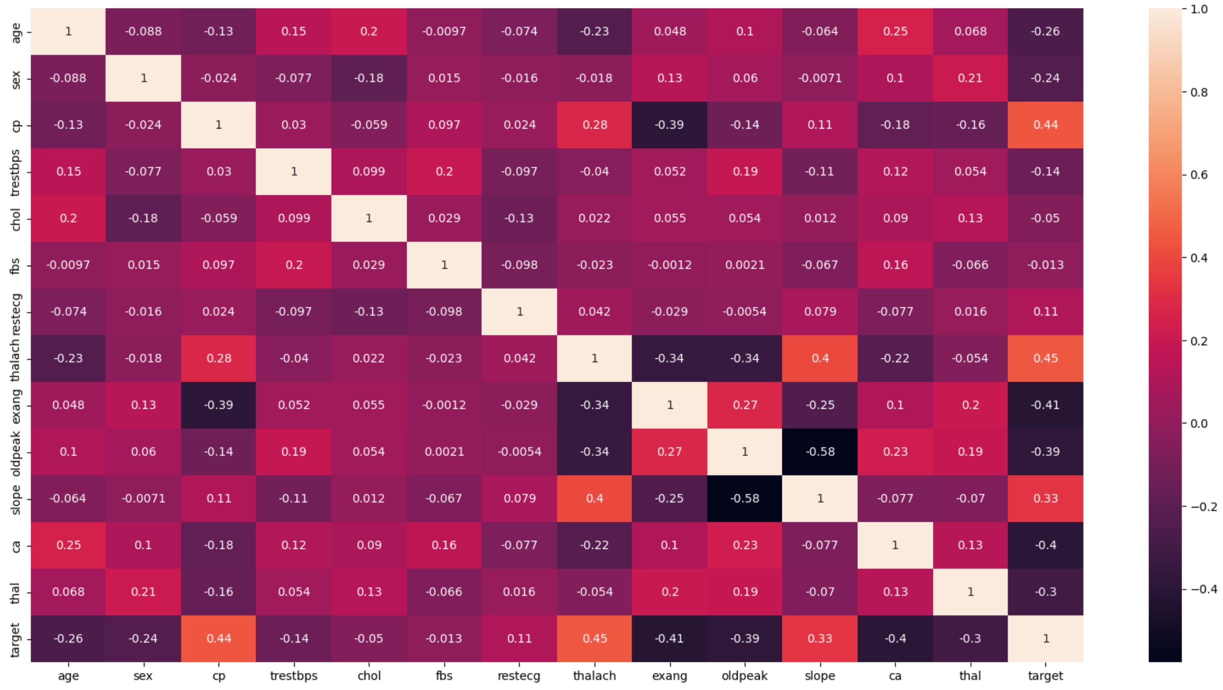
```
Number of Males: 353
Number of Females: 161
```

2. Count of the number of males and females who have heart disease

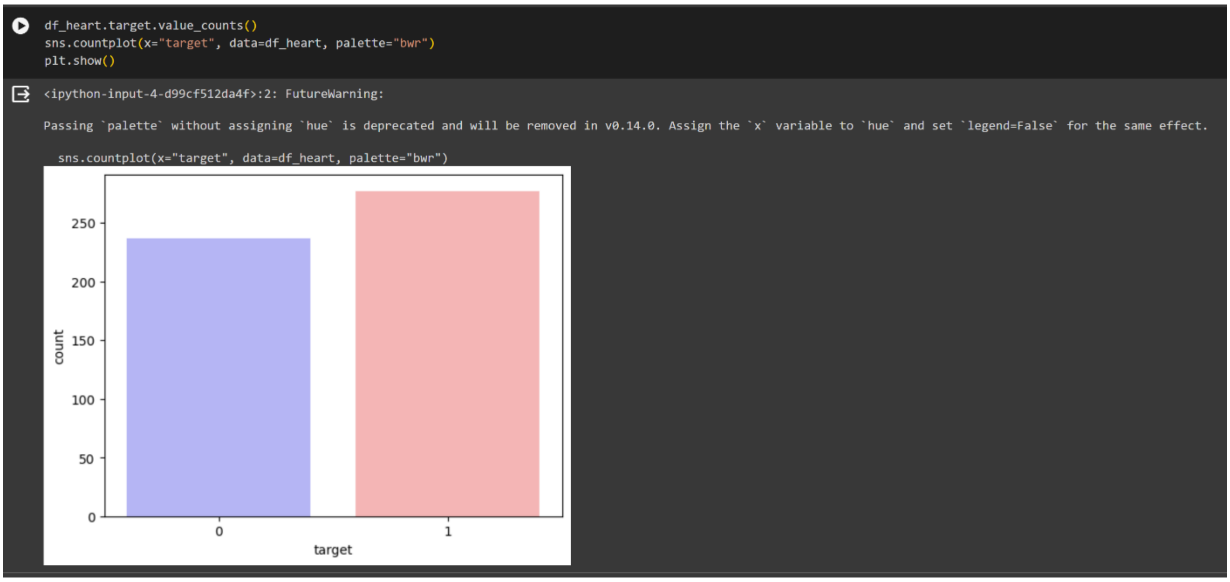


```
Males with Heart Disease: 162
Females with Heart Disease: 115
```

3. Building a Correlation Matrix



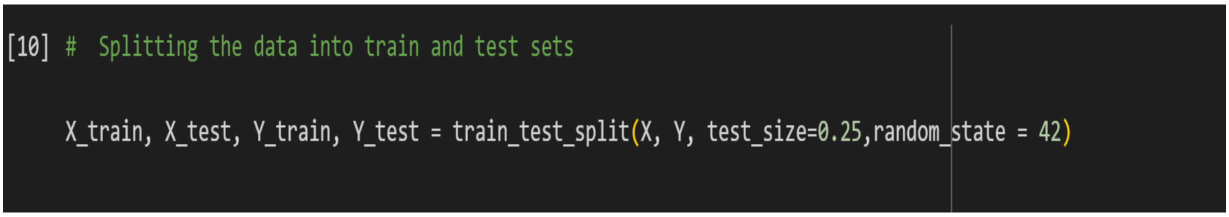
II. Creating Features and Target variable



```
df_heart.target.value_counts()
sns.countplot(x="target", data=df_heart, palette="bwr")
plt.show()
```

```
<ipython-input-4-d99cf512da4f>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(x="target", data=df_heart, palette="bwr")
```

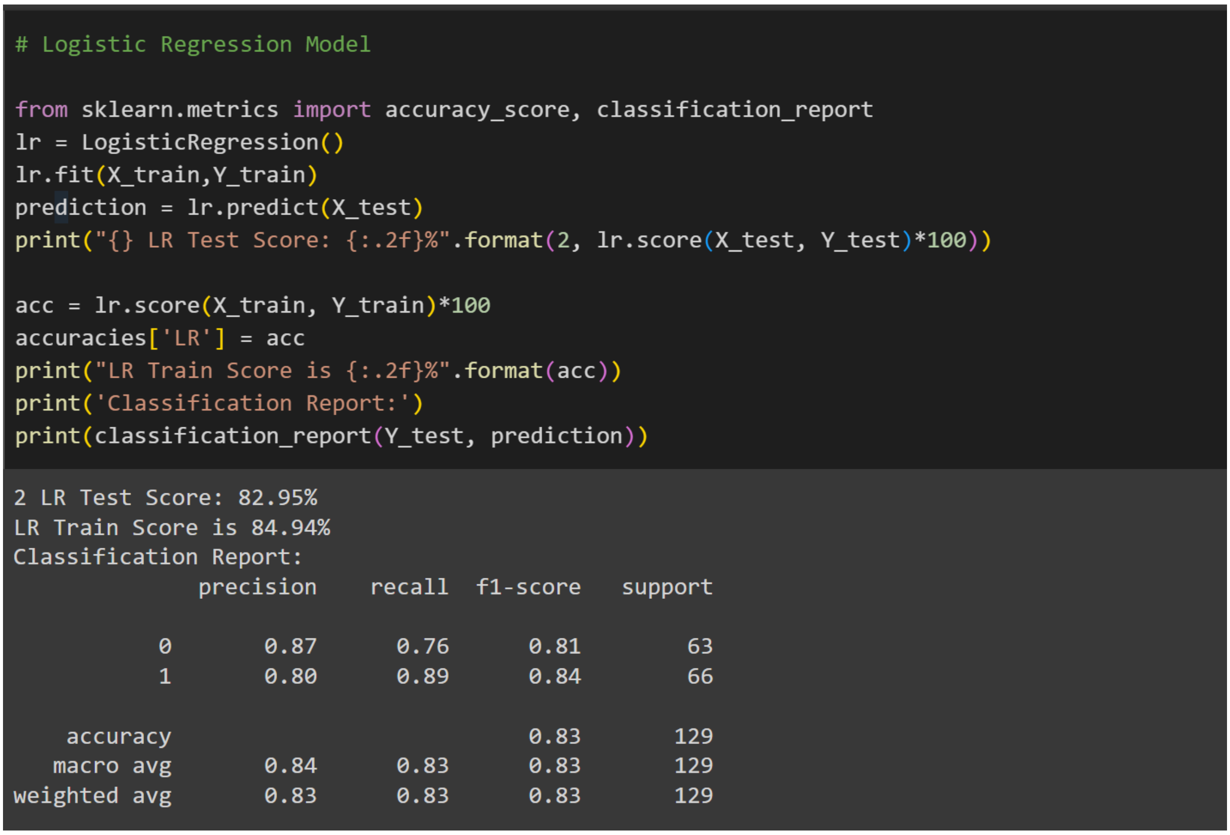
III. Splitting the data into train and test sets



```
[10] #  Splitting the data into train and test sets

     X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.25,random_state = 42)
```

IV. Create a function for evaluating metrics

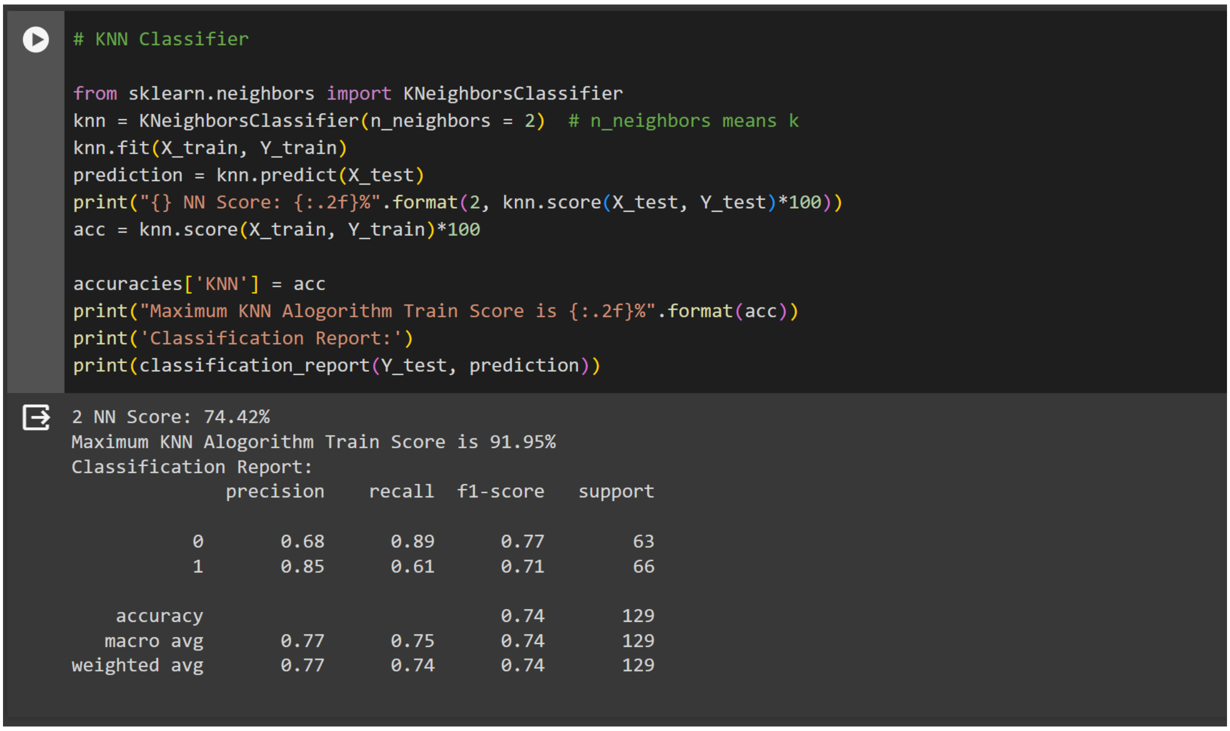
a) Logistic Regression:



```
# Logistic Regression Model

from sklearn.metrics import accuracy_score, classification_report
lr = LogisticRegression()
lr.fit(X_train,Y_train)
prediction = lr.predict(X_test)
print("{} LR Test Score: {:.2f}%".format(2, lr.score(X_test, Y_test)*100))

acc = lr.score(X_train, Y_train)*100
accuracies['LR'] = acc
print("LR Train Score is {:.2f}%".format(acc))
print('Classification Report:')
print(classification_report(Y_test, prediction))
```

```
2 LR Test Score: 82.95%
LR Train Score is 84.94%
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.76      0.81        63
           1       0.80      0.89      0.84        66

    accuracy                           0.83       129
   macro avg       0.84      0.83      0.83       129
weighted avg       0.83      0.83      0.83       129
```

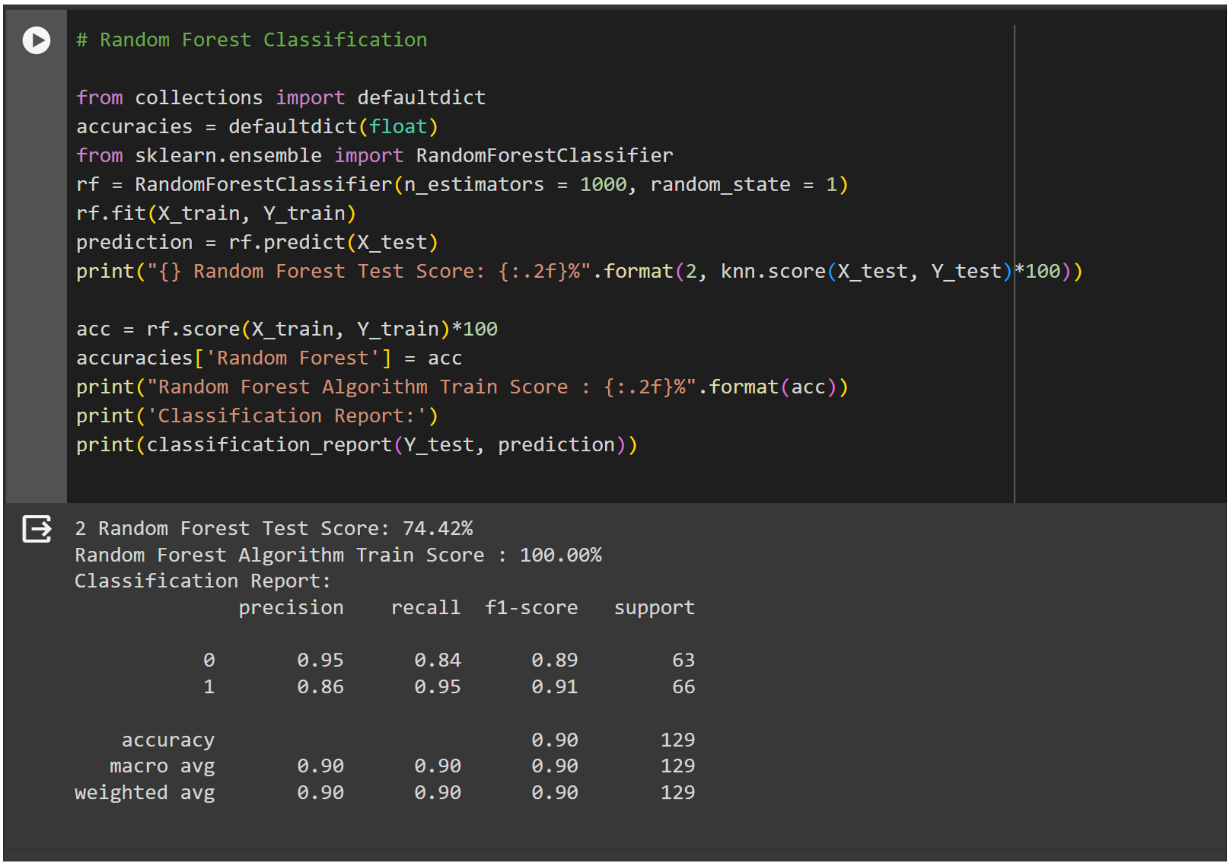
b) K Nearest Neighbor Classifier:



```python
# KNN Classifier

from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors = 2)  # n_neighbors means k
knn.fit(X_train, Y_train)
prediction = knn.predict(X_test)
print("{} NN Score: {:.2f}%".format(2, knn.score(X_test, Y_test)*100))
acc = knn.score(X_train, Y_train)*100

accuracies['KNN'] = acc
print("Maximum KNN Alogorithm Train Score is {:.2f}%".format(acc))
print('Classification Report:')
print(classification_report(Y_test, prediction))
```

```
2 NN Score: 74.42%
Maximum KNN Alogorithm Train Score is 91.95%
Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.89      0.77        63
           1       0.85      0.61      0.71        66

    accuracy                           0.74       129
   macro avg       0.77      0.75      0.74       129
weighted avg       0.77      0.74      0.74       129
```

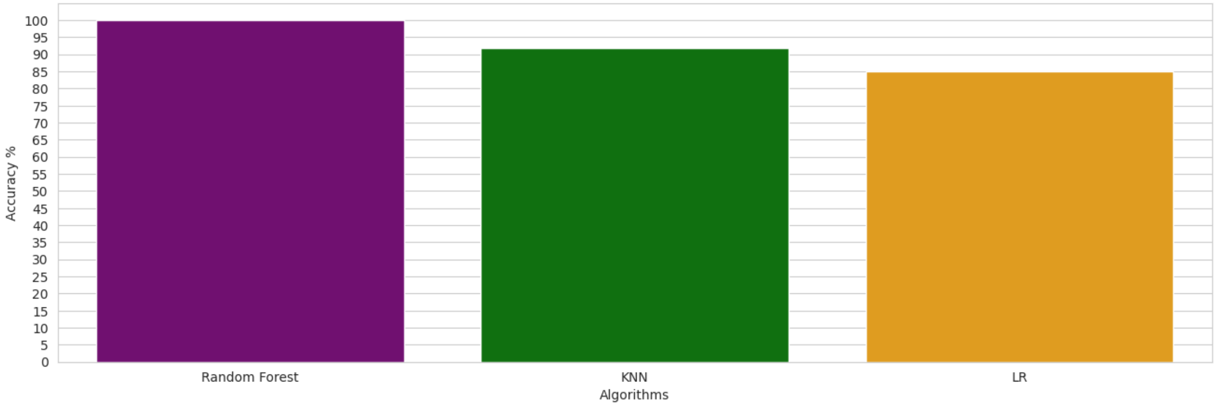
c) Random Forest Classifier :



```python
# Random Forest Classification

from collections import defaultdict
accuracies = defaultdict(float)
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(n_estimators = 1000, random_state = 1)
rf.fit(X_train, Y_train)
prediction = rf.predict(X_test)
print("{} Random Forest Test Score: {:.2f}%".format(2, knn.score(X_test, Y_test)*100))

acc = rf.score(X_train, Y_train)*100
accuracies['Random Forest'] = acc
print("Random Forest Algorithm Train Score : {:.2f}%".format(acc))
print('Classification Report:')
print(classification_report(Y_test, prediction))
```

```
2 Random Forest Test Score: 74.42%
Random Forest Algorithm Train Score : 100.00%
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.84      0.89        63
           1       0.86      0.95      0.91        66

    accuracy                           0.90       129
   macro avg       0.90      0.90      0.90       129
weighted avg       0.90      0.90      0.90       129
```

V. Comparing different Models:



CONCLUSION

Collecting the maximum accuracy scores from each of the three classifiers, it can be seen that Logistic Regression returned the accuracy score of 84.94% when the maximum number of features selected. K-Neighbors Classifier returned with the highest accuracy score of 91.95%, which was achieved when the number of neighbors was selected. Random Forest Classifier returned with the highest accuracy score of 100% when a forest with the size of 3200 decision trees was selected. Therefore, it can be concluded that Random Forest Classifier performed best amongst all three classifiers.

FUTURE SCOPE

In this project we made a which gives us the best accuracy among all the machine learning algorithms and by using that model we have made a web service using streamlit which will connect the back end of machine learning model to the front end that we designed it, with of help of this many people can get their health condition priorly and 50-43might taken care accordingly . We have tried it on our local host. We are going to implement by adding some features to our deployed model like adding all algorithms and by showing the ROC curve to the people who are using it to predict and make some changes and then we can implement it as an online website.

REFERENCES

- **"Logistic Regression"**. En.Wikipedia.Org, 2020, https://en.wikipedia.org/wiki/Logistic_regression.
- **"Random Forest"**. Medium, 2020, https://towardsdatascience.com/understanding-random-forest-58381e0602d2.
- https://www.geeksforgeeks.org/ml-heart-disease-prediction-using-logistic-regression/
- M. I. K. ,. A. I. ,. S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms".
- M. A. K. S. H. K. M. a. V. P. M Marimuthu, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach"

CODE

- https://github.com/DIVINA-012/NM-project---Heart-Disease-Prediction
-
- Google Colab Link: https://colab.research.google.com/drive/1ZRX_SUI5q8sqpIHojMuOe2KQqy329CWr#scrollTo=cixSdmC8mglD