SHRI RAMDEOBABA COLLEGE OF ENGINEERING AND MANAGEMENT, NAGPUR.

Machine Learning

(7TH SEM-B, SESSION 2022-2023, ECT-453-3)

“Prediction of Heart Disease using Machine Learning”

Submitted By

Huzefa Essaji (Roll No.-42)
Raj Singh (Roll No.-52)
Stephen Anthony (Roll No.-74)
Vivek Kaushik (Roll No.-82)

Department of Electronics and Communication Engineering

1. Introduction:

Healthcare is one of the primary focuses for humanity. According to WHO guidelines, good health is a fundamental right for individuals. It is considered that appropriate healthcare services should be available for regular checkups of one's health. Almost 31% of all deaths are due to heart-related diseases all over the world. Early detection and treatment of several heart diseases are very complex, especially in developing countries, because of the lack of diagnostic centers and qualified doctors and other resources that affect the accurate prognosis of heart disease. With this concern, in recent times computer technology and machine learning techniques are being used to make medical aid software as a support system for early diagnosis of heart disease.

Several different symptoms are associated with heart disease, which makes it difficult to diagnose it quicker and better. Working on heart disease patient databases can be compared to real-life applications. Doctors' knowledge to assign weight to each attribute. More weight is assigned to the attribute having a high impact on disease prediction. Therefore, it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the Diagnosis process. It also provides healthcare professionals an extra source of knowledge for making decisions.

This project aims to check whether the patient is likely to be diagnosed with any cardiovascular heart disease based on their medical attributes such as gender, age, chest pain, fasting sugar level, etc. A dataset is selected from the UCI repository with the patient's medical history and attributes. By using this dataset, we predict whether the patient can have heart disease or not. To predict this, we use 14 medical attributes of a patient and classify him if the patient is likely to have heart disease. These medical attributes are trained under algorithms like : Logistic regression, KNN and Random Forest Classifier. The most efficient of these algorithms is Decision Tree classifier which gives us an accuracy of 90.46%. And, finally, we classify patients that are at risk of getting a heart disease or not and also this method is totally cost efficient.

2. Existing Approaches or algorithms:

Due to lack of health care infrastructure, a lot of people are not able to get an idea of heart related problem, and if someone undergoes for preliminary tests the cost can skyrocket.
Currently medical specialists suggest many different tests to diagnose heart disease. Besides blood tests and a chest X-ray, tests to diagnose heart disease can include:

- Electrocardiogram (ECG or EKG)
- Holter monitoring
- Echocardiogram
- Exercise tests or stress tests
- Cardiac catheterization
- Heart (cardiac) CT scan
- Heart (cardiac) magnetic resonance imaging (MRI) scan.

3. Implemented approach or Algorithm:

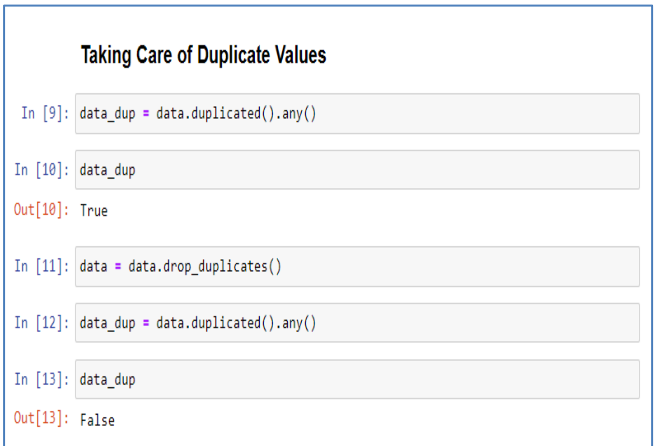
In India, huge mortality occurs due to cardiovascular diseases (CVDs) as these diseases are not diagnosed in early stages. Machine learning (ML) algorithms can be used to build efficient and economical prediction system for early diagnosis of CVDs in India.

a. Collecting Data:

Data collection is the process of gathering and measuring information from countless different sources. Here we used a data set from the UCI Machine learning archive, this database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date.
We used pandas to import the data in data frame for mathematical manipulation and easy handling of data.

b. Preparing the Data:

After importing the data, we implemented EDA (Exploratory Data Analysis) to understand and deduce important features from the data, also this step is implemented to remove duplicate and unwanted data, at the end we Visualized the data using matplotlib and seaborn. we had used a Correlation matrix to understand the correlation between all the features, through a scatter plot we tried to understand the relation between age and heart rate and heart disease.

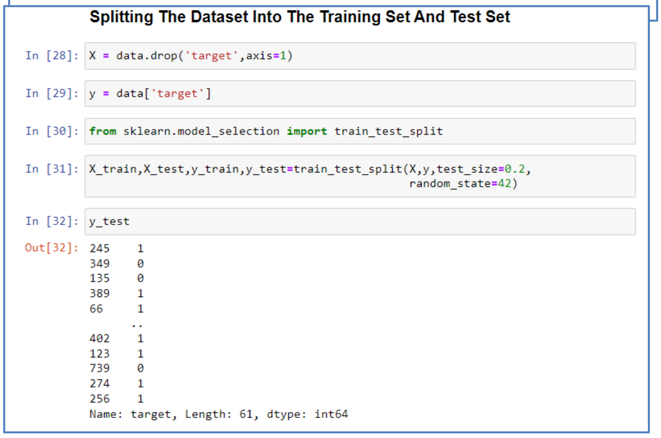


Taking Care of Duplicate Values

```
In [9]: data_dup = data.duplicated().any()

In [10]: data_dup
Out[10]: True

In [11]: data = data.drop_duplicates()

In [12]: data_dup = data.duplicated().any()

In [13]: data_dup
Out[13]: False
```

c. Splitting Data

In this step we split the data for training and testing. Almost 80% of data is for training and 20% for testing is a basic rule in machine learning.

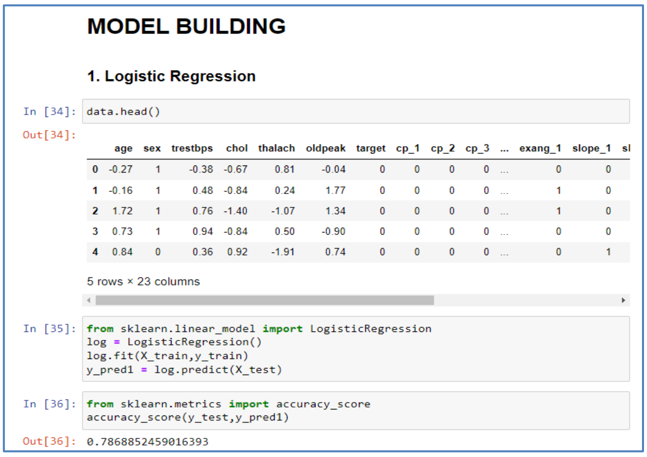


Splitting The Dataset Into The Training Set And Test Set

```
In [28]: X = data.drop('target',axis=1)

In [29]: y = data['target']

In [30]: from sklearn.model_selection import train_test_split

In [31]: X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,
                                                        random_state=42)

In [32]: y_test
Out[32]: 245    1
         349    0
         135    0
         389    1
         66     1
               ..
         402    1
         123    1
         739    0
         274    1
         256    1
         Name: target, Length: 61, dtype: int64
```

d. Choosing a Model:

A machine learning model determines the output you get after running a machine learning algorithm on the collected data. It is important to choose a model which is relevant to the task at hand. Over the years, scientists and engineers developed various models suited for different tasks, here we had tried to implement a classification model and thus experimented with 6 different models for understanding the behavior.



MODEL BUILDING

1. Logistic Regression

```
In [34]: data.head()
Out[34]:
```

	age	sex	trestbps	chol	thalach	oldpeak	target	cp_1	cp_2	cp_3	...	exang_1	slope_1	sl
0	-0.27	1	-0.38	-0.67	0.81	-0.04	0	0	0	0	...	0	0	
1	-0.16	1	0.48	-0.84	0.24	1.77	0	0	0	0	...	1	0	
2	1.72	1	0.76	-1.40	-1.07	1.34	0	0	0	0	...	1	0	
3	0.73	1	0.94	-0.84	0.50	-0.90	0	0	0	0	...	0	0	
4	0.84	0	0.36	0.92	-1.91	0.74	0	0	0	0	...	0	1	

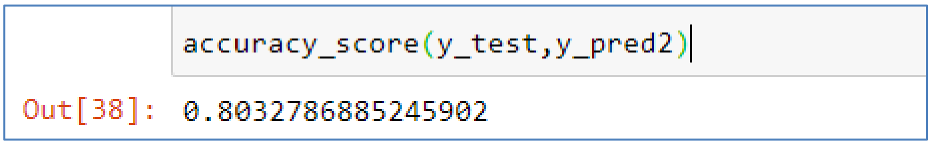
5 rows × 23 columns

```
In [35]: from sklearn.linear_model import LogisticRegression
         log = LogisticRegression()
         log.fit(X_train,y_train)
         y_pred1 = log.predict(X_test)

In [36]: from sklearn.metrics import accuracy_score
         accuracy_score(y_test,y_pred1)
Out[36]: 0.7868852459016393
```

Training the Model:

Training is the most important step in machine learning. In training, we pass the prepared data to our machine-learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting.

e. Evaluating the Model:

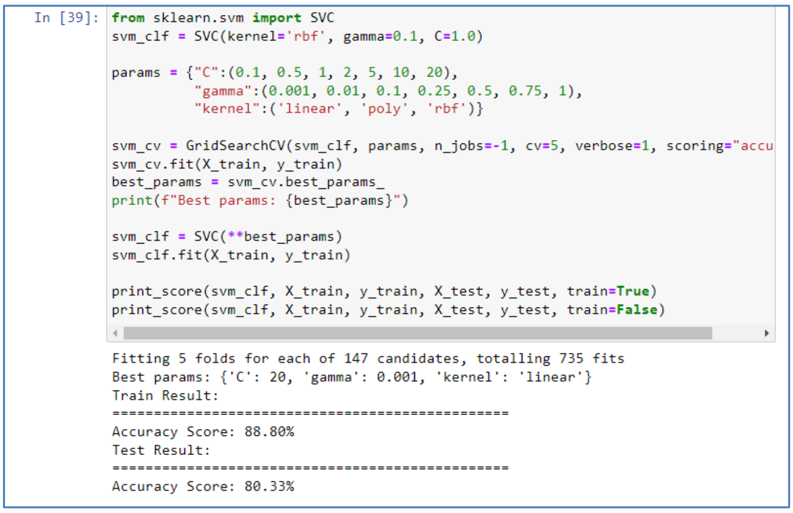


```
         accuracy_score(y_test,y_pred2)
Out[38]: 0.8032786885245902
```

After training your model, you have to check to see how it’s performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that you split our data into earlier. If testing was done on the same data which is used for training, you will not get an accurate measure, as the model is already used to the data, and finds the same patterns in it, as it previously did. This will give you disproportionately high accuracy.

f. Parameter Tuning:

Once we have created and evaluated our model, we tried to observe if its accuracy can be improved in any way. This is done by tuning the parameters present in your model. Parameters are the variables in the model that the programmer generally decides. At a particular value of your parameter, the accuracy will be the maximum. Parameter tuning refers to finding these values.

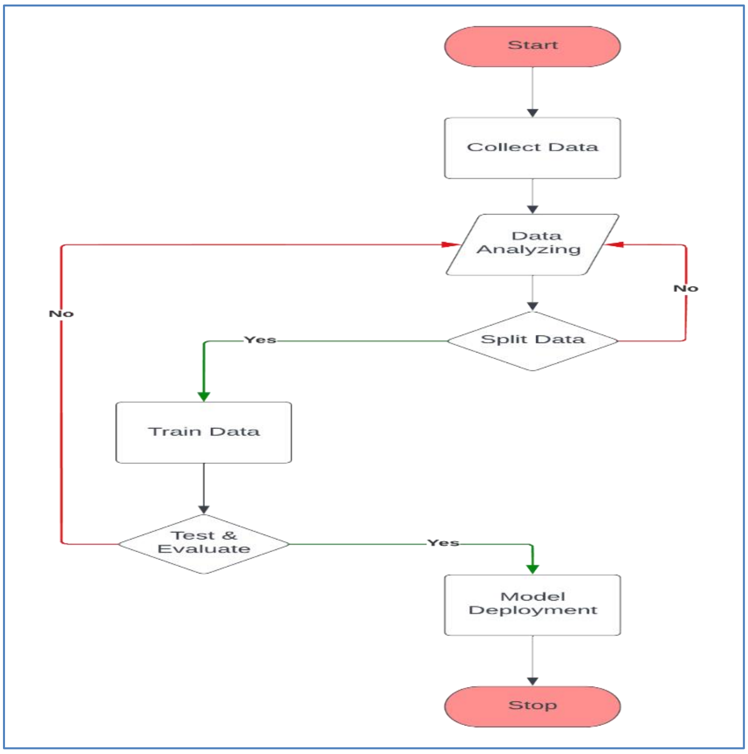


```
In [39]: from sklearn.svm import SVC
         svm_clf = SVC(kernel='rbf', gamma=0.1, C=1.0)

         params = {"C":(0.1, 0.5, 1, 2, 5, 10, 20),
                   "gamma":(0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 1),
                   "kernel":('linear', 'poly', 'rbf')}

         svm_cv = GridSearchCV(svm_clf, params, n_jobs=-1, cv=5, verbose=1, scoring="accu
         svm_cv.fit(X_train, y_train)
         best_params = svm_cv.best_params_
         print(f"Best params: {best_params}")

         svm_clf = SVC(**best_params)
         svm_clf.fit(X_train, y_train)

         print_score(svm_clf, X_train, y_train, X_test, y_test, train=True)
         print_score(svm_clf, X_train, y_train, X_test, y_test, train=False)
```

```
Fitting 5 folds for each of 147 candidates, totalling 735 fits
Best params: {'C': 20, 'gamma': 0.001, 'kernel': 'linear'}
Train Result:
================================================
Accuracy Score: 88.80%
Test Result:
================================================
Accuracy Score: 80.33%
```

g. Making Predictions:

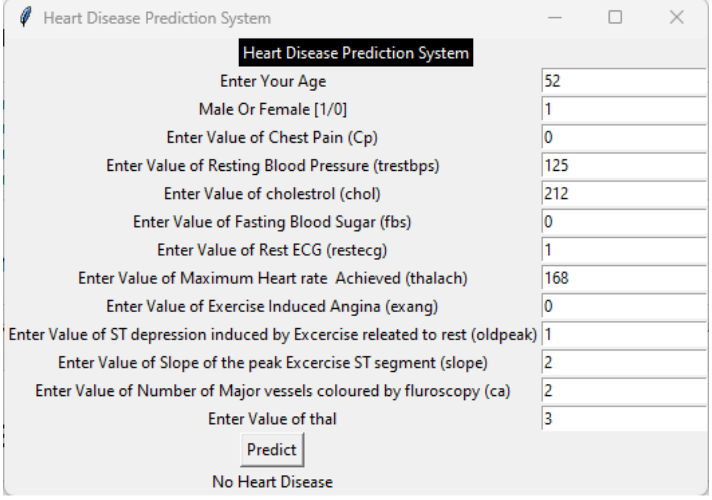
In the end, we used Tkinter to present a User interface to take input from the used and predicted.



4. Results:

Sno.	Model	Accuracy before Tuning	Accuracy after Tuning
1	Logistic Regression	78.68%	86.31%
2	Support vector Classifier	80.32%	88.80%
3	KNeighbors Classifier	73.77%	82.16%
4	Decision Tree Classifier	73.77%	90.46%
5	Random Forest Classifier	85.24%	88.38%
6	Gradient Boosting Classifier	80.32%	85.89%

At the end of our experiment, the result shows that SVC and Decision tree classifiers performed very well in comparison to Logistic regression, KNeighbour classifier, Random Forest Classifier and Gradient Boosting classifier. The model developed with SVM gives 88.80% accuracy and the Decision tree classifier 90.46%.



Heart Disease Prediction System

Enter Your Age	52
Male Or Female [1/0]	1
Enter Value of Chest Pain (Cp)	0
Enter Value of Resting Blood Pressure (trestbps)	125
Enter Value of cholestrol (chol)	212
Enter Value of Fasting Blood Sugar (fbs)	0
Enter Value of Rest ECG (restecg)	1
Enter Value of Maximum Heart rate Achieved (thalach)	168
Enter Value of Exercise Induced Angina (exang)	0
Enter Value of ST depression induced by Excercise releated to rest (oldpeak)	1
Enter Value of Slope of the peak Excercise ST segment (slope)	2
Enter Value of Number of Major vessels coloured by fluroscopy (ca)	2
Enter Value of thal	3

No Heart Disease

5. Conclusion:

A cardiovascular disease detection model has been developed using three ML classification modelling techniques. This project predicts people with cardiovascular disease by extracting the patient's medical history that leads to fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. By using these computer-aided techniques we can predict the patient fast and better and the cost can be reduced very much. There are a number of medical databases that we can work on as these Machine learning techniques are better and they can predict better than a human being which helps the patient as well as the doctors. Therefore, in conclusion, this project helps us predict the patients who are diagnosed with heart diseases by cleaning the dataset and applying Gradient Boost Classifier to get an accuracy of an average of 90.46%.

6. References:

a. *https://www.mayoclinic.org/diseases-conditions/heart-disease/diagnosis-treatment/drc-20353124#:~:text=Besides%20blood%20tests%20and%20a,too%20fast%20or%20too%20slowly*.

b. *https://archive.ics.uci.edu/ml/datasets/Heart+Disease*

c. *https://www.kaggle.com/code/cdabakoglu/heart-disease-classifications-machine-learning*

d. *https://scikit-learn.org/stable/*

e. *https://www.jeremyjordan.me/hyperparameter-tuning/*