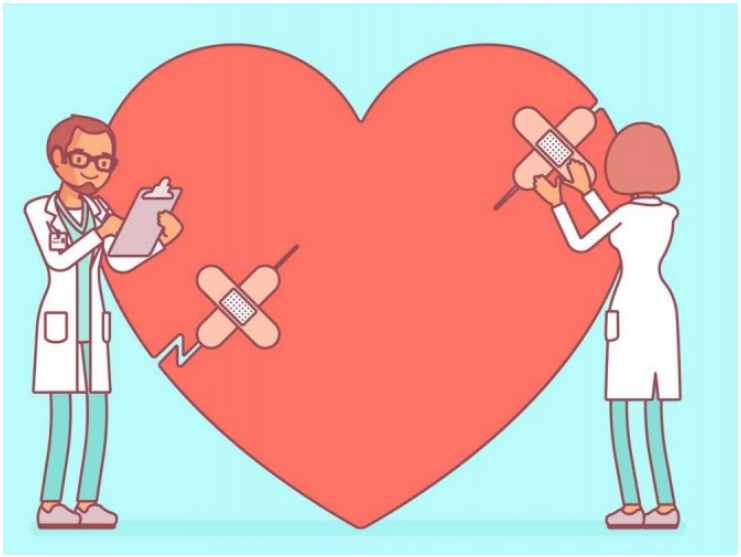


# CLEAVELAND HEART DISEASE ANALYSIS

Assignment Report 1- CBD2214


## ABSTRACT

Report on building a performing machine learning model to predict and analyse the data for a person having CVDs



Akshay Goswami- C0780364
Jayasree Reddy Manda – C0790067
Navneet Kaur - C0783511
Salma Sheikh - C0787129
Suhas Bonthala – C0774427

# Table of Contents

# Introduction

Cardiovascular illnesses (CVDs) or heart illness are the main sources of death worldwide, with 17.9 million passing cases every year. CVDs are purposefully contributed by hypertension, diabetes, overweight and undesirable ways of life. This undertaking covers manual exploratory information investigation and utilizing pandas profiling in Jupyter Notebook.

The dataset used is The UCI Heart Disease dataset.

## Motivation

Exploratory Data Analysis (EDA) is a pre-processing step to understand the data. There are different methods and steps in performing EDA; in any case, most of them are express, focusing on either observation or allocation, and are lacking. Thusly, here, we will see step by step to understand, research, and eliminate the information from the data to react to requests or assumptions. There are no coordinated advances or method to follow.

## Data Set Explanations

This information base contains 76 Attributes. However completely distributed investigations allude to utilizing a subset of 14 of them. Specifically, the Cleveland information base is the one in particular that has been utilized by ML analysts to this date. The "objective" field alludes to the presence of coronary illness in the patient. It is a number esteemed from 0 (no presence) to 4. At first, the dataset contains 76 highlights or qualities from 303 patients; nonetheless, distributed examinations picked just 14 highlights that are significant in anticipating coronary illness. Thus, here we will utilize the dataset comprising of 303 patients with features set.

## Libraries

1. Pandas- is a software library written for the python programming language for data manipulation and analysis.
2. Numpy- is powerful library used when working with Arrays
3. Scikit-Learn- is a library for Python that was first evolved by David Cournapeau in 2007. It contains a scope of helpful calculations that can undoubtedly be actualized and changed for the motivations behind grouping and other machine learning task.

## Variables or features explanations:

| |
|---|
| 1. age (Age in years) |
| 2. sex : (1 = male, 0 = female) |
| 3. cp (Chest Pain Type): [0: Typical Angina, 1: Atypical Angina, 2: Non-Anginal Pain, 3: Asymptomatic] |
| 4. trestbps (Resting Blood Pressure in mm/hg ) |
| 5. chol (Serum Cholesterol in mg/dl) |
| 6. fps (Fasting Blood Sugar > 120 mg/dl): [0 = no, 1 = yes] |
| 7. restecg (Resting ECG): [0: normal, 1: having ST-T wave abnormality , 2: showing probable or definite left ventricular hypertrophy] |
| 8. thalach (maximum heart rate achieved) |
| 9. exang (Exercise Induced Angina): [1 = yes, 0 = no] |
| 10. oldpeak (ST depression induced by exercise relative to rest) |
| 11. slope (the slope of the peak exercise ST segment) |
| 12. ca [number of major vessels (0–3)] |
| 13. thal (Thallium heart scan): [1 = normal, 2 = fixed defect, 3 = reversible defect] |
| 14. target: [0 = disease, 1 = no disease] |

### Variable types:

Binary: sex, fbs, exang, target

Categorical: cp, restecg, slope, ca, thal

Continuous: age, trestbps, chol, thalac, oldpeak

## Data Pre-processing:

### Data Preparation

In this step we import all the libraries such as pandas, matplotlib, numpy and seaborn and then load the dataset clevelanddataset.csv. and process the data files and then print the data frames.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Age                303 non-null    int64
 1   Sex                303 non-null    int64
 2   ChestpainType      303 non-null    int64
 3   RestingBP          303 non-null    int64
 4   Cholestrol         303 non-null    int64
 5   FastingBloodSugar  303 non-null    int64
 6   RestingECG         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    object
 12  Thal               303 non-null    object
 13  Target             303 non-null    int64
dtypes: float64(1), int64(11), object(2)
memory usage: 33.3+ KB
```

### Data Cleaning

In this step, we process the data by comparing the attribute values from the dataset info and removing all the null and invalid values present in the dataset. Such value can create errors for machine learning.

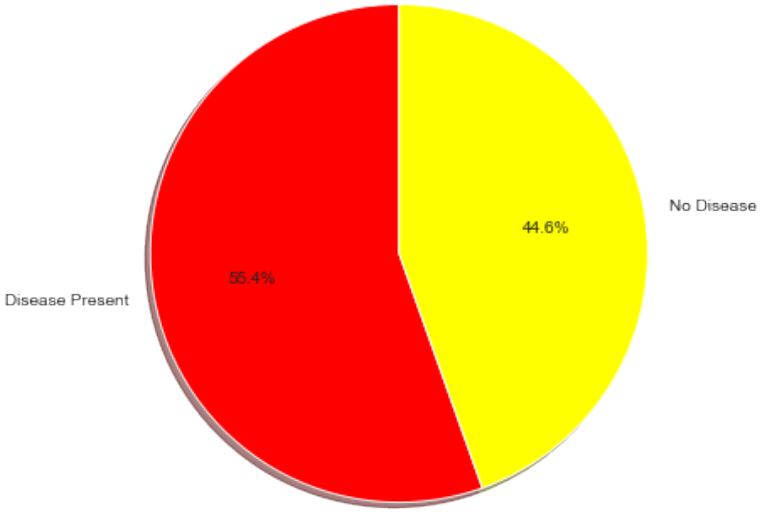
```
Out[420]: Age                 0
          Sex                 0
          ChestpainType       0
          RestingBP           0
          Cholestrol          0
          FastingBloodSugar   0
          RestingECG          0
          Thalach             0
          Exang               0
          Oldpeak             0
          Slope               0
          Ca                  4
          Thal                2
          Target              0
          dtype: int64
```

## Data Visualisation

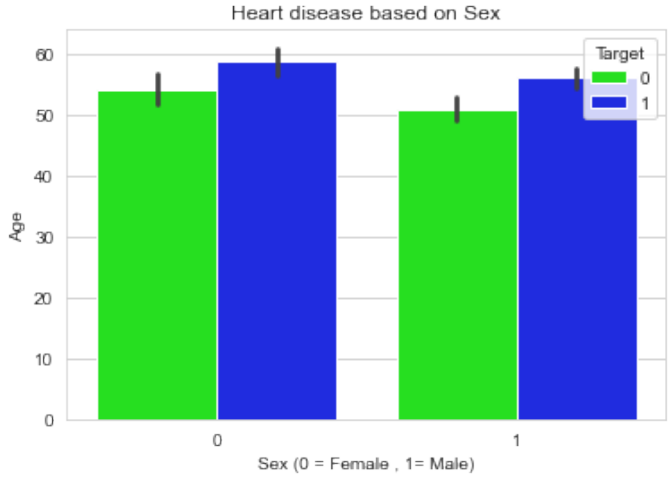
### Histogram for all the variables:

- Heart disease rate is higher in the age group 55-60
- Cholesterol is also higher in the age group 55-60
- The highest chest pain reported ranges 120-140

Pie chart and bar graph for the presence of heart disease:




The pie chart visualise the count of number of patients with absence or presence of stages of heart disease.




The bar graph visualise the heart disease based on the sex

## Machine Learning Algorithms or Classifiers

Scikit-Learn provides easy access to numerous different classification algorithms. Among these classifiers are:

### K-Nearest Neighbors-

K-Nearest Neighbors works by checking the good ways from some test guide to the known estimations of some preparation model. The gathering of information focuses/class that would give the littlest separation between the preparation focuses and the testing point is the class that is chosen.

**Algorithm:-**Import the dataset and the next step is to split the dataset into its attributes and tables. The final step is to make predictions on our test data.

### Logistic Regression-

Logistic Regression yields expectations about test information focus on a parallel scale, zero or one. On the off chance that the benefit of something is 0.5 or above, it is classified belonging to class 1, while beneath 0.5 on the off chance that it is named having a place with 0. Every one of the highlights additionally has a mark of just 0 or 1. Calculated relapse is a straight classifier and hence utilized when there is a type of direct connection between the information.

**Algorithm:-**Logistic regression is a powerful machine learning algorithm that utilizes a sigmoid function and works best on binary classification problems, although it can be used on multi-class classification problems through the “one vs. all” method

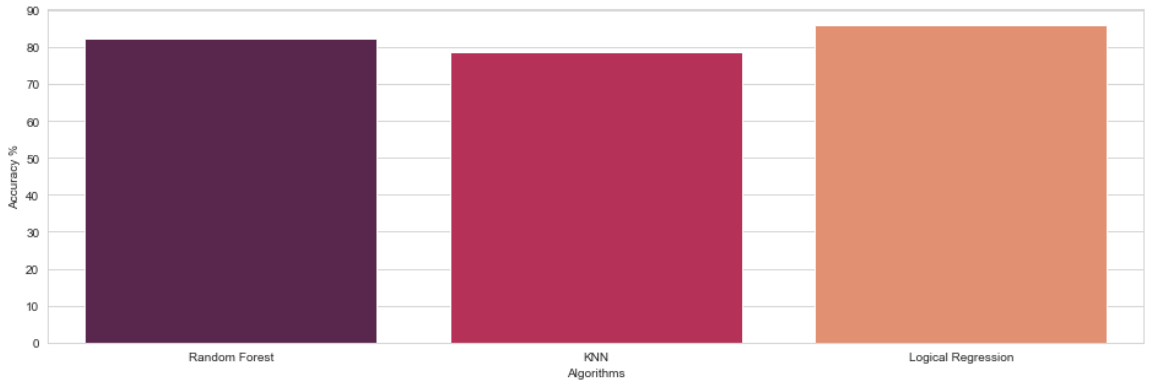
### Random Forest

Random forest is a kind of regulated AI calculation dependent on ensemble learning. Ensemble learning is a sort of realizing where you join various kinds of calculations or a similar calculation on numerous occasions to frame an all the more impressive expectation model. The random forest algorithm consolidates numerous calculations of a similar kind, i.e., various choice trees, bringing about a backwoods of trees, henceforth the name "Random Forest." The random forest algorithm can be utilized for both relapse and characterization errands.

**Algorithm:-** Pick N random records from the dataset. Build a decision tree based on these N records. Choose the number of trees you want in your algorithm and repeat steps 1 and 2

## Prediction Result

- All the data classified through multiple machine learning techniques have been depicted into single graph.
- Graph shows the accuracy in different classifiers of the same set of data.



## References


- https://www.kaggle.com/ronitf/heart-disease-uci/kernels
- https://towardsdatascience.com/exploratory-data-analysis-on-heart-disease-uci-data-set-ae129e47b323
- https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/#:~:text=Among%20these%20classifiers%20are%3A,Decision%20Tree%20Classifiers%2FRandom%20Forests