Early Heart Attack Diagnosis

1st Anmol Kumar
Roll Number: 2022081

2nd Devyansh Chaudhary
Roll Number: 2022156

I. Problem Statement

Heart attacks have become increasingly common due to the fast-paced nature of our lives imposed by modern lifestyles and nutritional issues. World Health Organization estimates that 1 million global deaths take place from (Cardiovascular diseases) CVDs.

Our project focuses on early detection of potential heart attacks in the future with the help of their medical history.
It analyzes a range of medical parameters such as high blood pressure, cholesterol, or chest pain and can help in early diagnosis, thus making a leap towards a reduced mortality rate.

This prediction system helps medical caretakers curate effective treatments as early diagnosis is quite important to prevent even an occurrence of heart attacks.

II. Methods

This Project uses three Machine Learning Models to get our desired output, Logistic Regression, K-NN, and Random Forrest. Each of these ML models is trained on the test dataset, and then we measure their score using various set metrics for binary classification tasks like precision, accuracy, F1-Score, and F-Beta Score on our dataset. The model with the best overall performance score across all the metrics will be selected as our final model for output.

A. *K-Nearest Neighbours[KNN]*

KNN is an ML algorithm used for classification tasks. In this algorithm, predictions are made based on the majority class of k-nearest data points in the space. In diagnosing early heart attacks, KNN can be employed to identify patterns in the dataset that indicate the likelihood of a patient experiencing a heart attack By considering 11 different attributes in our selected dataset.

B. *Random Forrest*

Random Forrest is an ML algorithm that constructs multiple decision trees during training. And outputs the mode of the classes as the prediction. Random forests can capture relationships and interactions among 11 attributes(fig-3.1) from our selected dataset.

C. *Logistic Regression*

We use Binary logistic regression to predict whether a patient has a disease or not. Linear regressive models are prone to outliers, which may be present in our dataset, too due to a certain patient's underlying conditions. This model facilitates the identification and removal of irrelevant parameters, streamlining the feature selection process to focus on factors that significantly influence the prediction outcome.

III. Dataset Details

The dataset was taken from an open-source website, Kaggle. The dataset is in the form of a CSV file. The dataset consists of 11 columns, each representing a specific lifestyle or health attribute associated with heart attacks. Figure 1.1 illustrates the names and datatypes of these columns. With a total of over 950+ rows, the dataset offers a diverse set of observations for analysis.

Attributes	Meaning
Age	Age of the Person, Age >= 28.
Sex	Sex of the Person M: male F: female
Chest Pain Type	TA: Typical Angina ATA: Atypical Angina NAP: Non Anginal pain ASY: Asymptomatic
Resting BP	Resting Blood Pressure (mm Hg)
Cholesterol	Serum Cholesterol (mm/dl)
Fasting BS	Fasting Blood Sugar level If FBS>120 mm/dl = 1 Else = 0
Resting ECG	Resting Elecardiogram Normal: normal ST: having ST-T wave abnormality. LVH: showing probable or definite left ventricular hypertrophy.
MaxHR	Max Heart Rate Between [60, 202]
ExerciseAngina	Exercise Induce Angina Y: Yes N: No
Oldpeak	Oldpeak: ST[Numeric value measured in depression]
ST_Slope	The Slope of the peak exercise ST segment UP: upsloping Flat: flat Down: downsloping
Heart Disease	Output Class 1: Heart Disease 0: Normal

Fig 3.1 Dataset Attributes and their meaning.

IV. Evaluation Metrics

We will employ the following evaluation metrics to measure the correctness of our prediction algorithm.

Accuracy: It is the ratio of the number of correctly predicted instances to the total number of instances.

Precision: It is calculated as the ratio of true positives to the sum of true positives and false positives.

Recall (Sensitivity): It is calculated as the ratio of true positives to the sum of true positives and false negatives.

F1-Score: The F1-Score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is calculated as

$$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F-Beta Score: The F-Beta Score is more general and allows to assign weights to precision and recall.

$$\frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$

V. RESULTS

In below graphs 1: Having heart disease and 0: Not having heart disease.

Pie(1) and Pie(2) in (fig-4.1) depict the segregation of patients by their genders into the binary classes of presence or absence of heart disease. We used the existing dataset to produce these results, and a run of our selected algorithms on the test dataset by the end of this project will depict whether our selected model performs with accuracy or not.

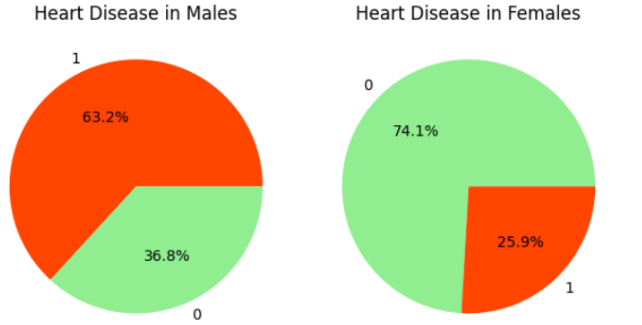

Fig-5.1 Heart Disease percentage in different sex. [code]

Chest pain is the first physical indicator for a layman and thus serves as an important feature to judge whether a person is undergoing a heart stroke. We examined our dataset to segregate each type of chest pain into the binary classes of the presence or absence of a true heart stroke.

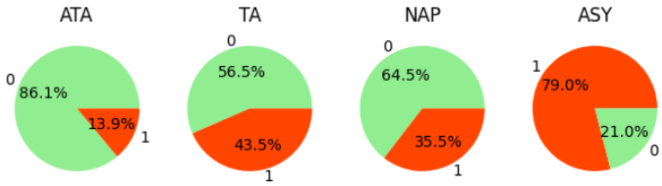

Fig 5.2 Heart Disease rate in people suffering from different chest pains. [code]