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. Literature Review

Our course SML provided a basic introduction to the techniques used most widely across the industry. However, researchers also employ image models like CNN and often use hybrid techniques according to the problem's requirements.

A. Deep Learning Models

Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown promising results in analyzing medical data for heart disease prediction.

1) Convolutional Neural Networks: CNNs have been widely used for analyzing medical images, such as ECG signals, for heart disease diagnosis. Badnjevic et al. proposed a CNN model for classifying heart disease data and demonstrated its effectiveness on various datasets.

2) Recurrent Neural Networks: RNNs and LSTMs are well-suited for processing sequential data, making them suitable for analyzing time-series data like ECG signals or patient medical records. Mohan et al. employed a hybrid model combining CNN and LSTM for effective heart disease prediction using electronic health records.

B. Hybrid Models

Hybrid models that combine multiple machine learning techniques have been explored to leverage the strengths of different algorithms. Mohan et al. proposed a hybrid model integrating CNN, LSTM, and support vector machines, achieving improved performance compared to individual models. Fatima and Pasha surveyed various machine learning algorithms and their hybrid combinations for disease diagnosis, highlighting their potential in this domain.

C. Advanced Feature Engineering

Feature engineering plays a crucial role in improving the performance of machine learning models. Ghayoumi and Ghousideveloped an efficient feature selection approach based on genetic algorithms and support vector machines for heart disease diagnosis. Tan et al. employed techniques like particle swarm optimization and representation learning for automated feature engineering in heart disease diagnosis using unstructured datasets.

D. Transfer Learning

Transfer learning, which involves leveraging knowledge from pre-trained models on large datasets, has shown promise in heart disease prediction. Murat et al. explored a transfer learning approach for heart disease detection, demonstrating its effectiveness compared to traditional machine learning models. Wang et al. employed transfer learning techniques for diagnosing heart diseases using electronic medical records.

III. Proposed Architecture

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.

IV. 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.

V. Visualisations

Feature	Mean	Std	Min	Max
Age	53.51	9.43	28	77
RestingBP	132.40	18.51	0	200
Cholesterol	198.80	109.38	0	603
FastingBS	0.23	0.42	0	1
MaxHR	136.81	25.46	60	202
Oldpeak	0.89	1.07	-2.60	6.20
HeartDisease	0.55	0.50	0	1

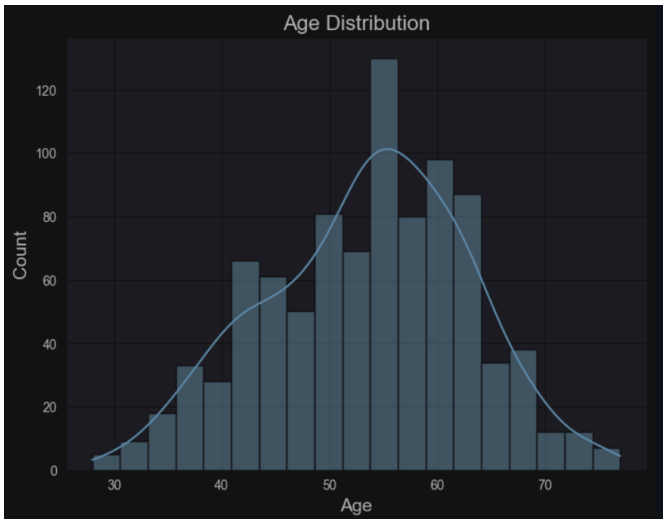

fig-5.1 Age distribution of the dataset, visualized using a histogram with kernel density estimation (KDE).

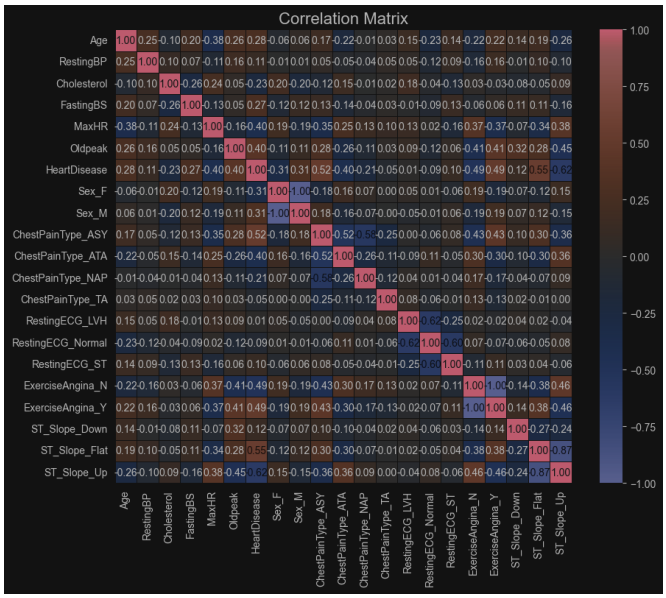

fig-5.2 Correlation matrix of the dataset, visualized using a heatmap with annotations.

Pie(1) and Pie(2) in (fig-5.3) 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.

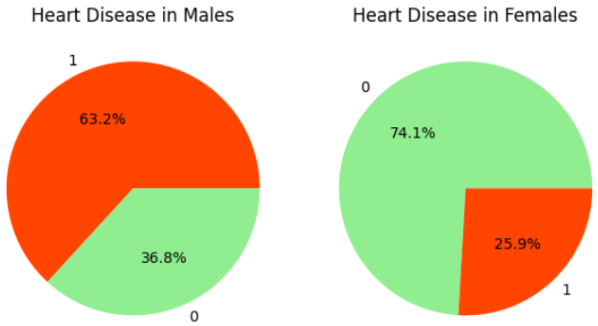

Fig-5.3 Heart Disease percentage in different sex. [code]

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.

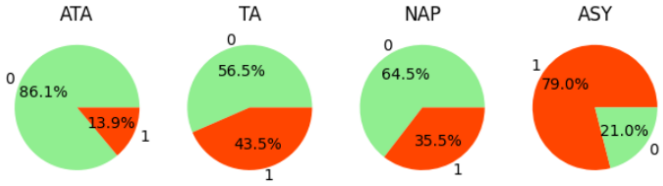

Fig-5.4 Heart Disease rate in people suffering from different chest pains. [code]

VI. 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

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

VII. Results

A. *Random Forrest*

The Random Forest model, trained with specified parameters, constructs multiple decision trees during training and outputs the mode of the classes as predictions. After training, the model's performance was evaluated using metrics such as accuracy and the classification report, which provides insights into precision, recall, and F1-score for each class.

TABLE I
Accuracy and Classification Report for Random Forest Model

Metric	Class 0	Class 1
Accuracy	88.04%	-
Precision	85%	90%
Recall	87%	89%
F1-Score	86%	90%
Support	77	107
Macro Avg	88%	88%
Weighted Avg	88%	88%

The accuracy and classification report metrics for the Random Forest model are presented below. The accuracy metric represents the percentage of correctly predicted instances out of the total number of instances. For Class 0, the precision, recall, and F1-score are 85%, 87%, and 86%, respectively. Similarly, for Class 1, the precision, recall, and F1-score are 90%, 89%, and 90%, respectively.

The support column in the table represents the number of instances in each class, providing further context for the evaluation metrics.

Overall, the Random Forest model demonstrates strong performance, achieving an accuracy of 88.04% on the test dataset. These results indicate the model's effectiveness in predicting the presence or absence of heart disease based on the input features.

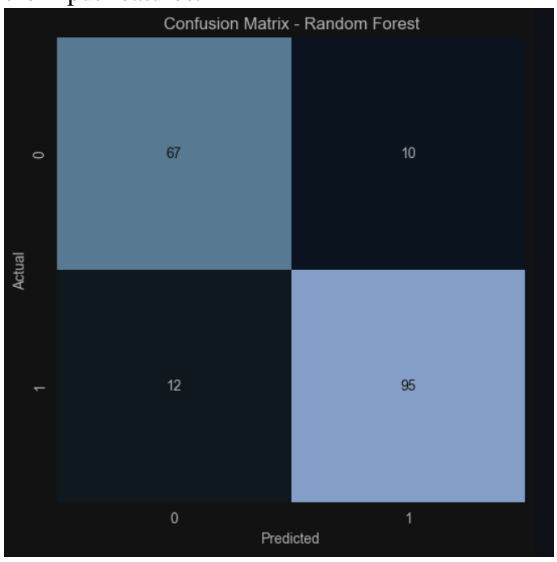

fig-6.1 Confusion matrix for the Random Forest model.

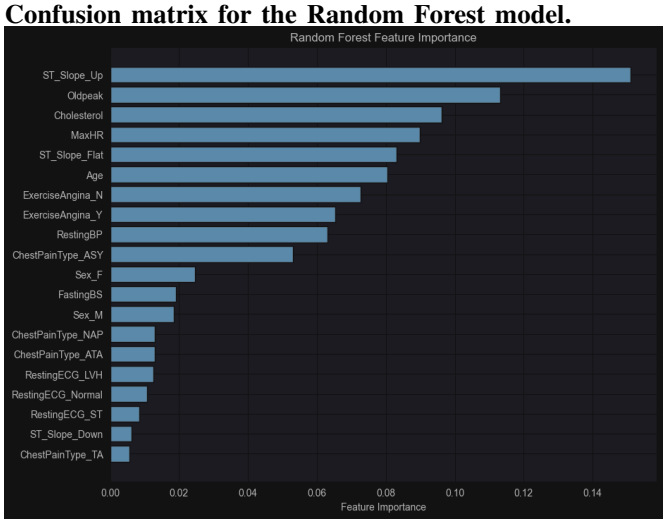

fig-6.2 Random Forest feature importance analysis.

B. *K-NearestNeighbour*

The KNN model achieved an accuracy of 70.65% on the test dataset. The precision, recall, and F1-score for Class 0 are 63%, 71%, and 67%, respectively. Similarly, for Class 1, the precision, recall, and F1-score are 77%, 70%, and 74%, respectively. These metrics provide insights into the model's

performance in predicting the presence or absence of heart disease based on the input features.

TABLE II
CLASSIFICATION REPORT FOR KNN MODEL

Class	Precision	Recall	F1-score	Support
0	0.63	0.71	0.67	77
1	0.77	0.70	0.74	107
Accuracy		0.71		
Macro avg	0.70	0.71	0.70	184
Weighted avg	0.71	0.71	0.71	184

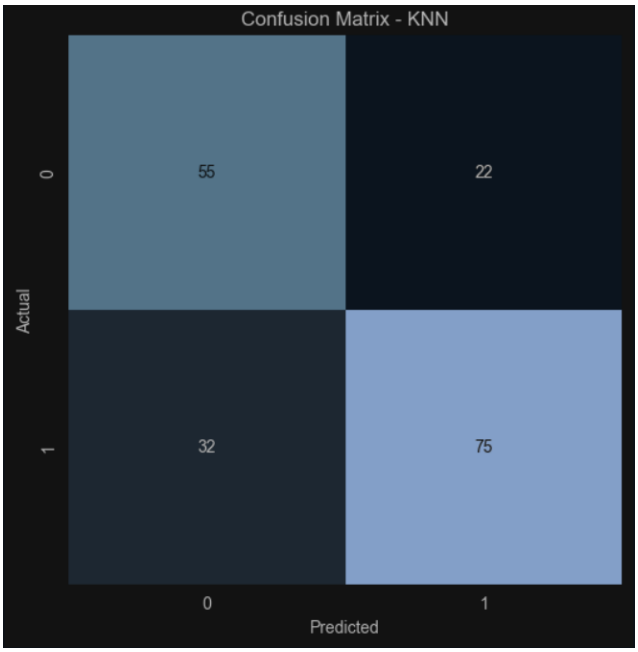

Fig-6.3: Confusion matrix of KNN classifier.

TABLE III
PERFORMANCE METRICS OF LOGISTIC REGRESSION MODEL

Metric	Value
Accuracy	0.6685
Classification Report	
Precision (Class 0)	0.59
Recall (Class 0)	0.66
F1-score (Class 0)	0.63
Precision (Class 1)	0.73
Recall (Class 1)	0.67
F1-score (Class 1)	0.70
Macro avg	0.66
Weighted avg	0.67

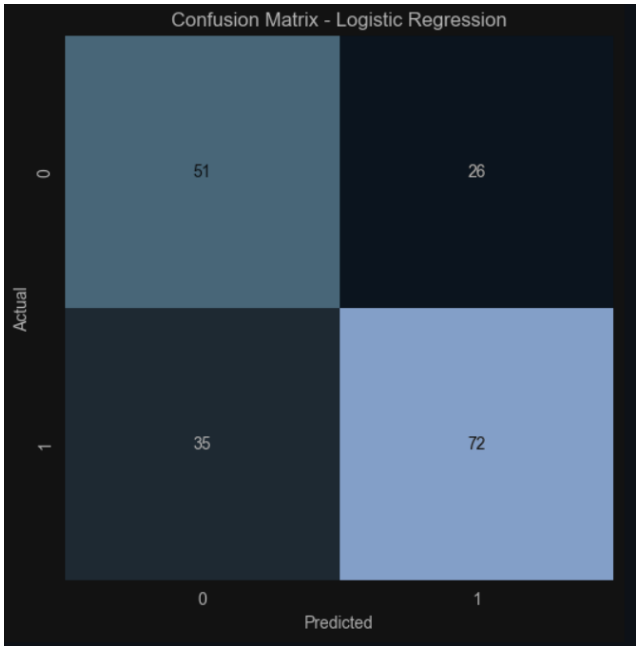

Fig-6.4 Confusion matrix of Logistic Regression.

C. Logistic Regression

The table presents the performance metrics of the Logistic Regression model. The accuracy metric indicates the percentage of correctly predicted instances out of the total number of instances. For each class (0 and 1), precision, recall, and F1-score are provided, evaluating the model's performance in predicting positive and negative instances. Additionally, macro and weighted averages offer aggregated measures of performance across all classes, considering class imbalances in the dataset.

The plot in Figure 6.5 visualizes the feature coefficients of the Logistic Regression model. Feature coefficients represent the weight or importance assigned to each feature by the model during the training process. In the plot, each horizontal bar represents a feature from the dataset. The length and direction of the bar indicate the magnitude and direction (positive or negative) of the coefficient assigned to that feature. Positive coefficients indicate that the corresponding feature positively contributes to the prediction of the target variable, while negative coefficients indicate a negative contribution. The higher the absolute value of the coefficient, the more significant the feature's impact on the model's predictions.

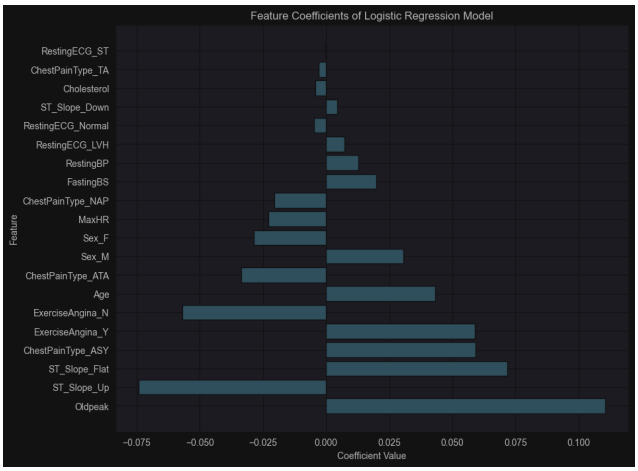

Fig-6.5 Confusion matrix of Logistic Regression.

VIII. Summary

A. Accuracy Comparison

- Random Forest: 88.04%
- KNN: 70.65%
- Logistic Regression: 66.85%

B. Precision, Recall, and F1-Score

Random Forest outperformed KNN and Logistic Regression in terms of precision, recall, and F1-score for both classes.

C. Model Complexity and Interpretability

- Random Forest: More complex with good interpretability through feature importances.
- KNN: Simpler model but lacks interpretability compared to Random Forest.
- Logistic Regression: Simple and interpretable model with coefficient analysis.

D. Scalability

- Random Forest and Logistic Regression: Scale well to larger datasets.
- KNN: Computationally expensive, especially with large datasets.

E. Suitability for the Task

- Random Forest: Well-suited for capturing complex relationships in data.
- KNN: Simple and intuitive but may not perform well with complex data.
- Logistic Regression: Often used as a baseline model for binary classification tasks.

F. Conclusion

While Random Forest achieved the highest accuracy and demonstrated robust performance, the choice of the best model depends on factors such as dataset size, complexity, interpretability, and computational resources. you can visit our online repository at online repository.