**Heart Attack Classification System using machine learning algorithms**

|  |  |  |
| --- | --- | --- |
| Student Name : Sumit Chakraborty  ID : **20-43472-1** *Dept Name : CSE* *Institute Name : American International University Bangladesh* Dhaka , Bangladesh Email address :  **sumitchakraborty218@gmail.com**  Phone Number : 01942646726 | Student Name : Shanto Kumar Basak  ID : **20-42945-1** *Dept Name : CSE* *Institute Name : American International University Bangladesh* Dhaka , Bangladesh Email address : **Shantobasakaiub@gmail.com**  Phone Number : 01621000817 | Student Name : Md Musfiqur Rahman  ID: **20-43482-1** *Dept Name ; CSE* *Institute Name : American International University Bangladesh* Dhaka , Bangladesh Email address :  **musfiqeon@gmail.com**  Phone Number : 01872829950 |
|  | Student Name : Salima Farha  ID : **20-42497-1** *Dept Name :CSE* *Institute Name : American International University Bangladesh* Dhaka , Bangladesh Email address : **farhatuktuki@gmail.com**  Phone Number : 01742082148 |  |

# **Abstract**

This research abstract presents a novel framework for classifying heart attacks utilizing advanced machine learning algorithms. The proposed system integrates feature extraction methodologies to construct robust patient representations by leveraging an extensive dataset comprising diverse patient profiles, clinical variables, and cardiac markers. Rigorous evaluation of machine learning algorithms, encompassing ensemble methods and deep learning architectures, showcases the system's efficacy in accurately categorizing heart attack types. The system's potential to provide rapid and precise classifications holds substantial promise for enhancing clinical decision-making, ultimately improving patient outcomes and streamlining medical interventions.

**Keywords:** Heart attack classification, machine learning algorithms, diagnostic accuracy, feature engineering, clinical data integration, personalized healthcare.

# **Introduction**

The practical and timely classification of heart attacks is a critical challenge in modern healthcare. Integrating machine learning algorithms into this domain offers a promising avenue for revolutionary advancements. By harnessing the power of artificial intelligence, a heart attack classification system can sift through a multitude of patient data, including clinical records, diagnostic tests, and medical imaging, to discern nuanced patterns that might elude traditional diagnostic approaches. This synergy of medical knowledge and computational prowess can enhance diagnostic accuracy and enable rapid interventions, ultimately elevating the standard of care and potentially saving lives.

# **Motivation of the Project**

The project aims to develop a heart attack classification system using three machine learning algorithms: random forest, naive Bayes, K-nearest neighbors (KNN) and Logistic Regression. This project aims to create an accurate and reliable system that can assist in the early detection and classification of heart attack risk in individuals. Nowadays, heart attacks are a leading cause of death worldwide, and early identification of potential threats can remarkably improve patient outcomes and save lives. By employing these three diverse algorithms, the project aims to compare their performances and identify the most effective approach for predicting and classifying heart attack risks based on relevant medical data. This system could be integrated into healthcare systems to provide timely interventions and personalized care for individuals at risk of heart attacks.

# **Objective of the project**

The project aims to create a reliable heart attack classification system employing four machine learning techniques: random forest, naive bayes, logistic regression and K-nearest neighbors (KNN). The approach tries to precisely divide people into various groups according to their risk of having a heart attack. The research intends to develop a comprehensive and trustworthy model that can aid medical practitioners in detecting possible heart attack patients early by utilizing each algorithm's capabilities. The system's effectiveness will be determined by its capacity to accurately anticipate and categories people as high, moderate, or low risk, which will improve patient care and encourage preventive measures. In our proposed machine-learning model, we got 98% accuracy by using Random Forest algorithm.

# **Methodology**

This project shows the analysis of various machine-learning algorithms, such as K nearest neighbors (KNN), Naive Bayes classifier and Random Forest classifier and Logistic Regression that can be helpful for practitioners or medical analysts for accurately diagnose heart disease. For this project, the dataset collected from “Heart Attack Dataset”, Mendeley Data is used to detect heart attack systems. The proposed methodology (Figure 0.) includes steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the third is the processing stage where we explore the data. Data processing deals with the missing values, cleaning of data and normalization depending on algorithms used**.** After processing of data, various feature or model was selected and tested. Classifier is used to classify the processed data the classifier used in the proposed model are KNN, Naïve Bayes classifier, Random Forest classifier and Logistic Regression.

The models were trained with deep learning to further improve the prediction accuracy. Finally, we evaluated our models based on accuracy and performance using various performance metrics.



Figure 1: Methodology

## **Data Collection**

The dataset used for this paper was collected from the "Zheen Hospital in Erbil", Iraq, from January 2019 to May 2019. It was uploaded in Mendeley data published on 23 November 2022 by Tarik A. Rashid and Bryar Hassan.

## **Data processing**

Before processing the data, the result column was assigned negative and positive values. In the processing portion, we appointed the positive results as one and the negative results as 0. Furthermore, we used X and y variables, respectively, where y consists of the Result attribute and X consists of the rest.

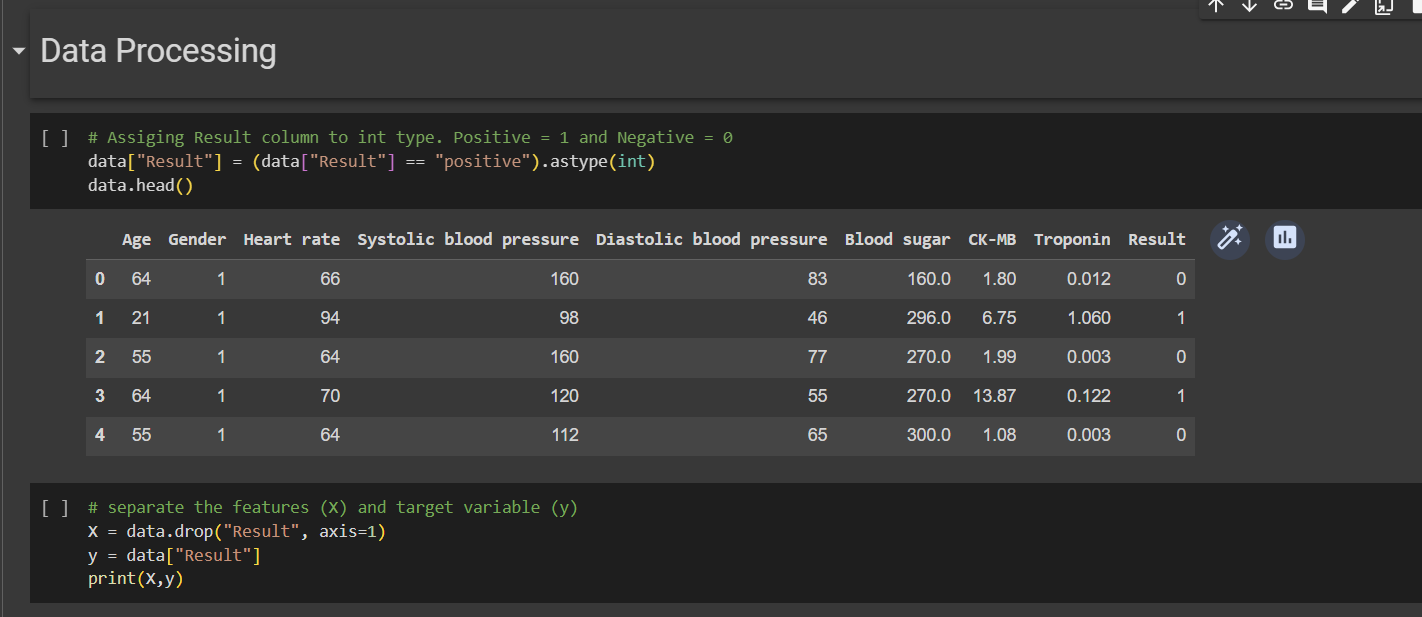


Figure 2: Data Processing

## **Dataset description**

The dataset for this research was collected from Mendeley. The dataset in on csv format. The provided dataset has a size of 1319 and has 8 variables, including age, gender, pulse, blood pressure systolic, blood pressure diastolic, glucose, CK-MB, and troponin. There is only one output, which is either positive or negative depending on whether a heart attack has occurred. You may access the actual data set here.

Data Attributes & Their Descriptions

|  |  |
| --- | --- |
| Attributes | Descriptions |
| Age | Age of the patient in the year |
| Gender | Gender of the patient (male or female) |
| Heart rate | Maximum heart rate achieved 10 to 160-180 |
| Systolic blood pressure | Resting systolic blood pressure (in mm Hg on ad mission to the hospital) (70-190) |
| Diastolic blood pressure | Resting diastolic blood pressure (in mm Hg on ad mission to the hospital) (40-100) |
| Blood sugar | (Blood sugar > 120 mg/dl) (1-900mg/dl) |
| CK-MB | Enzyme CK-MB (male upto-6.22 female upto-4.88) It is Creatine Kinase Enzymes |
| Troponin | Enzyme Troponin (0.0-0.014) |

The purpose of the medical dataset is to identify whether or not it is a heart attack based on the information presented. The following encoding has been used to store the data in an Excel sheet: The setting for the male is 1, while the setting for the female is 0. If the glucose level is greater than 120, the glucose column is set to 1; otherwise, it is set to 0. The output's positive and negative values are set to 1 and 0, respectively.

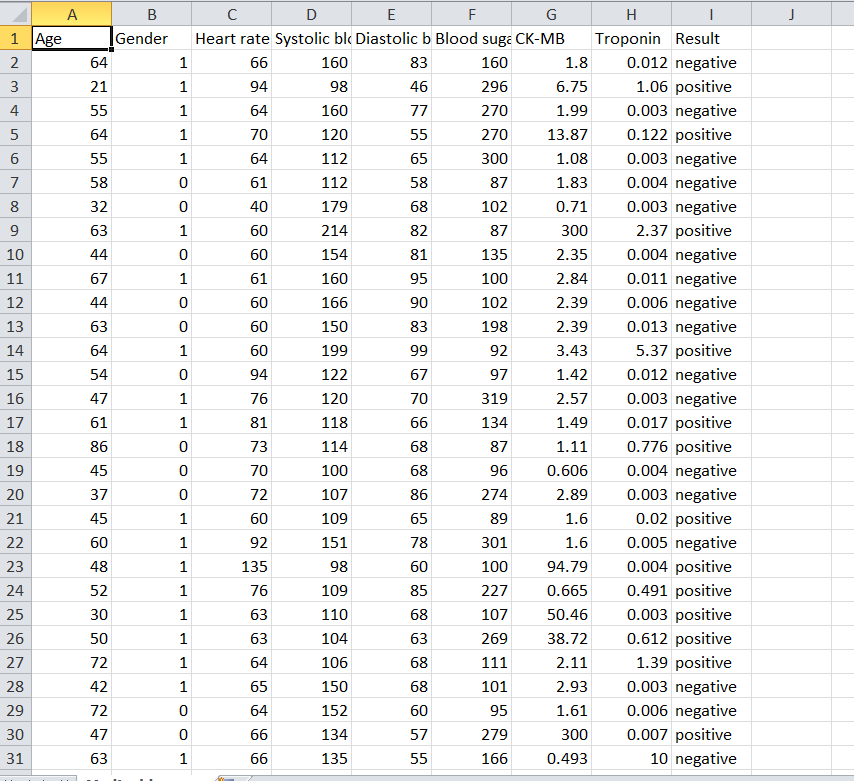


Figure 3: A snapshot of the Medical Dataset in Excel (30 out of 1319)

**Data Analysis**

For data analysis, Random Forest accuracy, Logistic Regression, Naïve Bayes, K-nearest algorithms were used. These are all machine learning algorithms. Random forest consists of large number of decision trees and are known to be very accurate and robust. Naïve Bayes is based on Bayes theorem. Although it is sometimes inaccurate, it is very fast and easy to train. KNN is a non-parametric algorithm. It is very accurate, but can be slow for large datasets. .Logistic regression easier to implement, interpret and very efficient to train.

## **Machine Learning model development and evaluation**

To analyze and predict the presence of heart attack several machine-learning algorithms were used.

1. Random Forest: Combines the output of multiple decision trees to reach a single most accurate result.
2. KNN: One of the simplest forms of machine learning algorithms mostly used for classification.
3. Naïve Bayes: It does not require as much training data so much faster and easy to train although it can be inaccurate.
4. Logistic Regression: Statistical and machine learning algorithm used for binary and multi-class classification tasks.

**Random Forest**

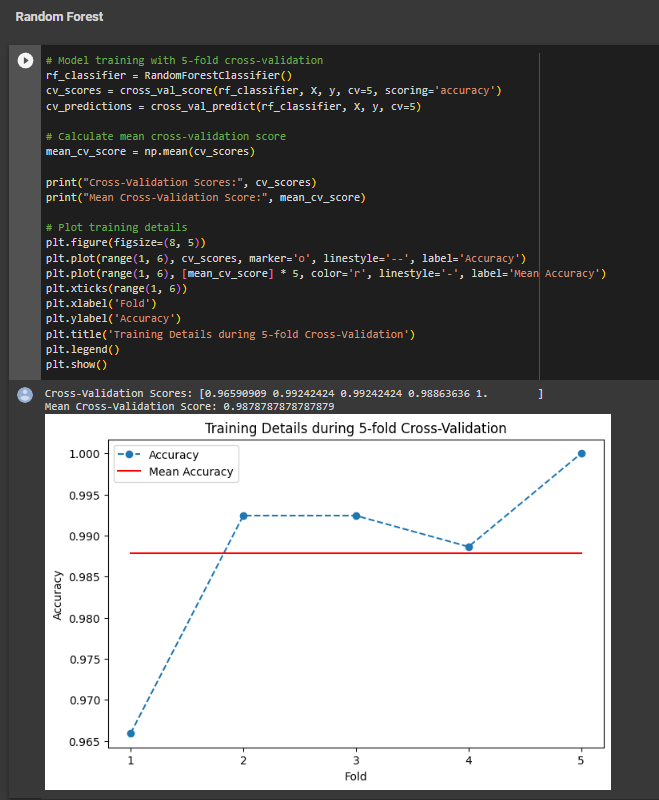


Figure 4: Random Forest 5-fold Cross-Validation

**KNN**

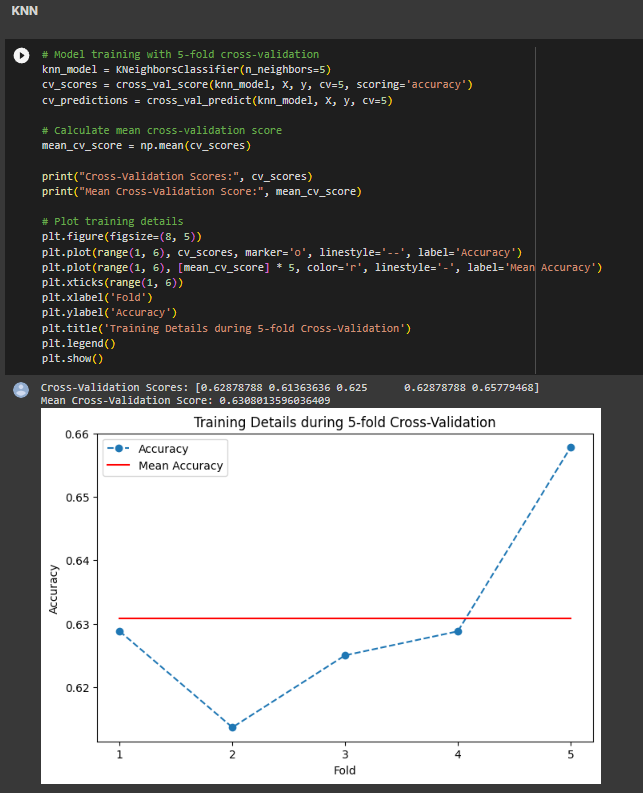


Figure 5: KNN 5-fold Cross-Validation

**Naive Bayes**

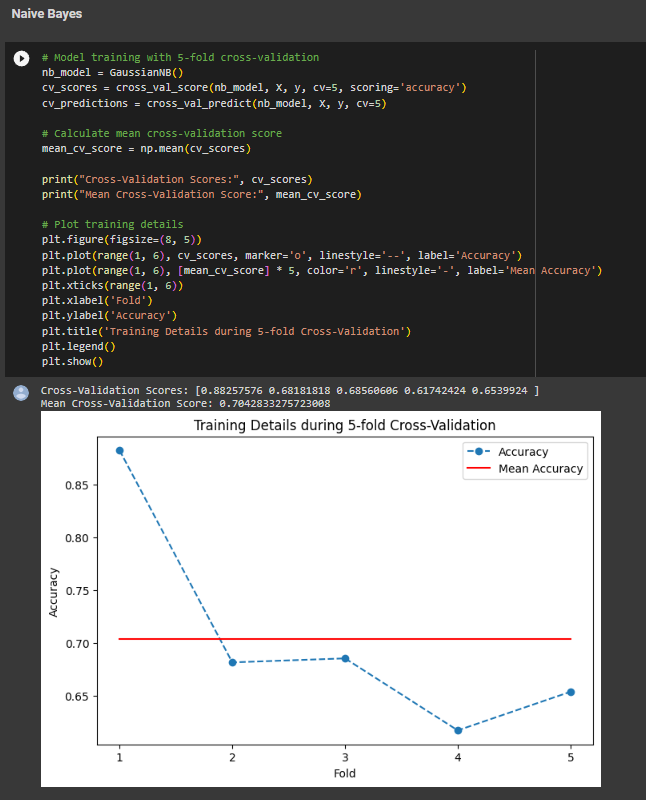


Figure 6: Naïve Bayes 5-fold Cross-Validation

**Logistic Regression:**

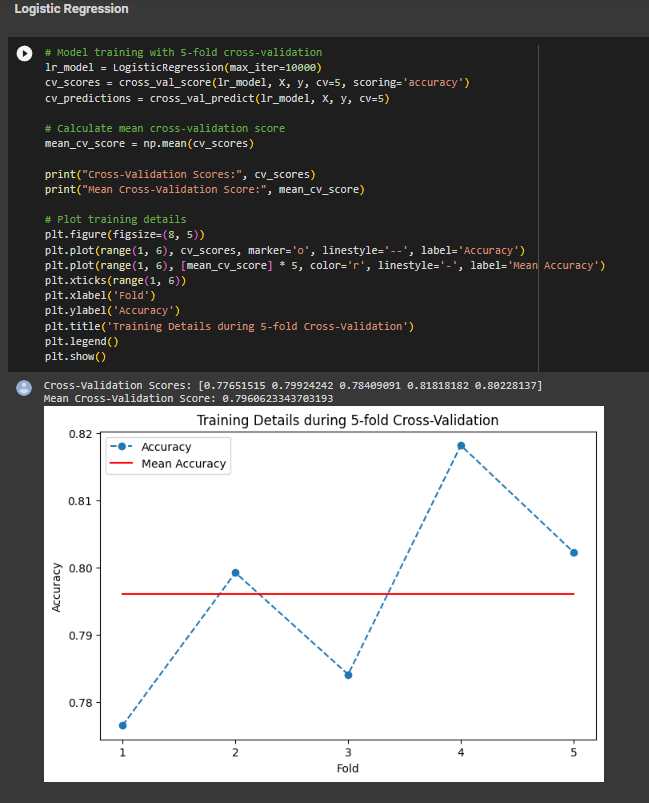


Figure 7: Logistic Regression 5-fold Cross-Validation

## **Results**

**Data Analysis**

For data analysis, Random Forest accuracy, Naïve Bayes, K-nearest algorithms were used. These are all machine learning algorithms. Random forest consists of large number of decision trees and are known to be very accurate and robust. Naive Bayes is based on Bayes theorem. Although it is sometimes inaccurate, it is very fast and easy to train. KNN is a non-parametric algorithm. It is very accurate, but can be slow for large datasets. Logistic regression easier to implement, interpret and very efficient to train.

Analysis's findings Different machine learning algorithms produced different outcomes. Even though Random Forest and Naive Bayes have the best accuracy rates. Furthermore, the accuracy of KNN is 0.62, Naive Bayes is 0.89, Random Forest is 0.98 and Logistic Regression 0.79, respectively.

The study's findings imply that random forests are the most reliable method for predicting Heart attacks. The others, though, are also reasonably correct. The ideal algorithm, however, will depend on the problem that must be solved.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Methods** | **Result of Accuracy** | **Result of Precision** | **Result of Recall** |
| **Random Forrest** | 0.98 | 0.98 | 0.99 |
| **KNN** | 0.62 | 0.69 | 0.71 |
| **Naïve Bayes** | 0.89 | 0.99 | 0.83 |
| **Logistic Regression** | 0.79 | 0.81 | 0.86 |

Figure 8: Performance Comparison analysis of all models

**Random Forest:**

Random forest is a commonly-used machine learning algorithm, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

The dataset comprising data relating to the prediction of heart attack diseases was applied to the Random Forest algorithm. An 80/20 split was kept between the training and test sets while dividing the dataset. A combination of decision trees was used to train the Random Forest model on the training set. On the test set, the model's effectiveness was evaluated. The Random Forest outperformed random guessing with an accuracy of 98.10%, highlighting its potential for precise heart disease prediction. This outcome demonstrates how the Random Forest method may be used to increase prediction performance in heart disease classification tasks by utilizing the collective wisdom of several decision trees.

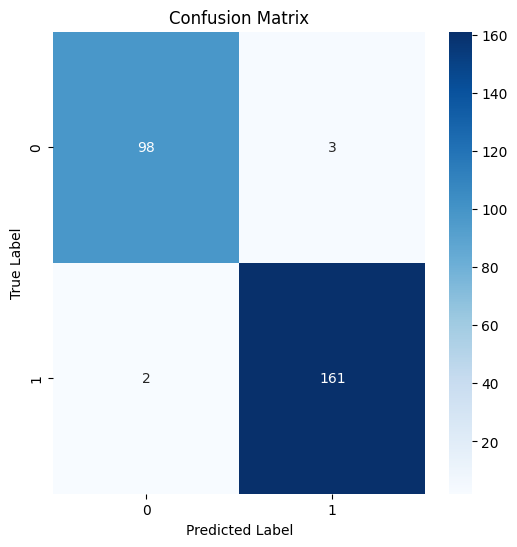
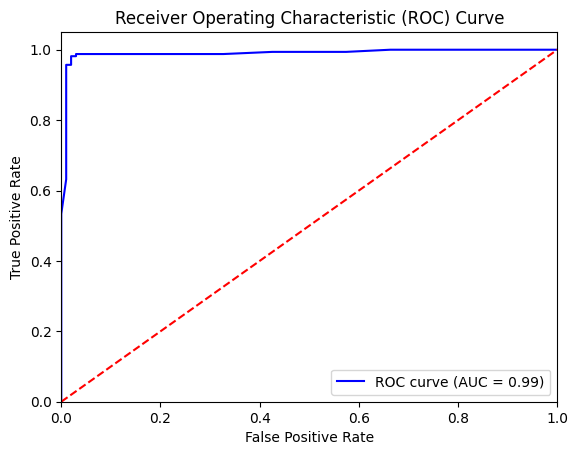
** **

Figure 1: Confusion Matrix of Random Forest Figure 2: ROC Curve of Random Forest

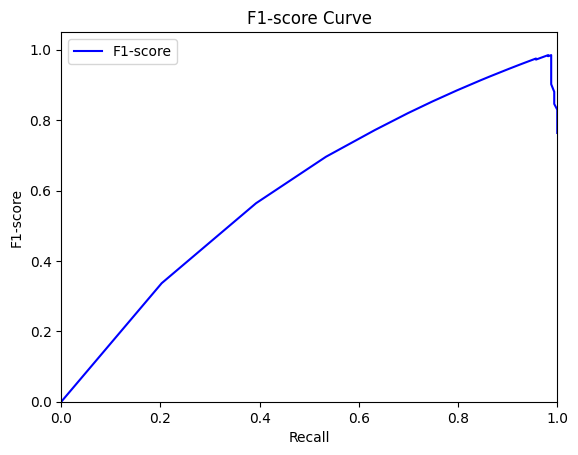
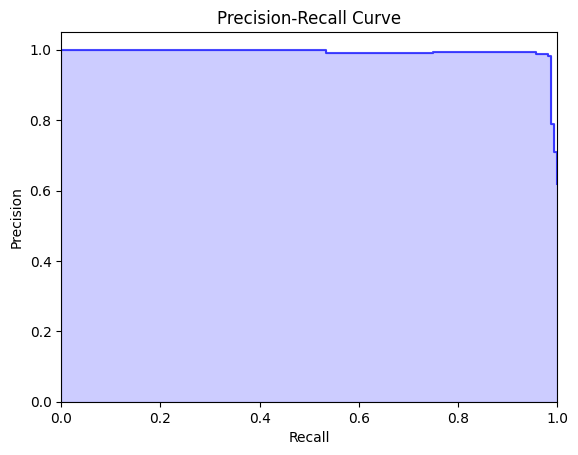
** **

Figure 3: F1-score curve of Random Forest Figure 4: Precision-Recall Curve Random Forest

**KNN:**

The K-nearest neighbors (KNN) model is a simple, supervised machine learning (ML) algorithm classifying models based on the attributes of the input dataset. The dataset contains 1319 samples, each with eight features. Furthermore, divided into two parts, the dataset was used in training and test sets in an 80/20 ratio. Using different values of k, ranging from 1 to 100, denoting the number of nearest neighbors taken into consideration, the model was trained on the training set. 5-fold cross-validation was used to evaluate the model's performance, and the best k value was chosen based on the highest score obtained. The KNN model slightly outperformed random guessing with an accuracy of 62.12%.

The accuracy of 62.12% obtained serves as a starting point for studying and examining the dataset. From these results, it can be suggested that the KNN model has the potential to predict heart attack disease based on the given dataset. Future studies can focus on improving the model's accuracy through feature selection and size reduction strategies, as well as researching other classification algorithms. These efforts are intended to improve the accuracy and reliability of heart disease prediction models.

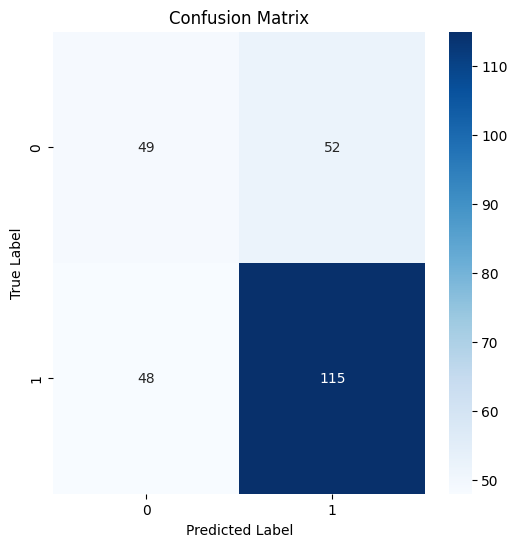
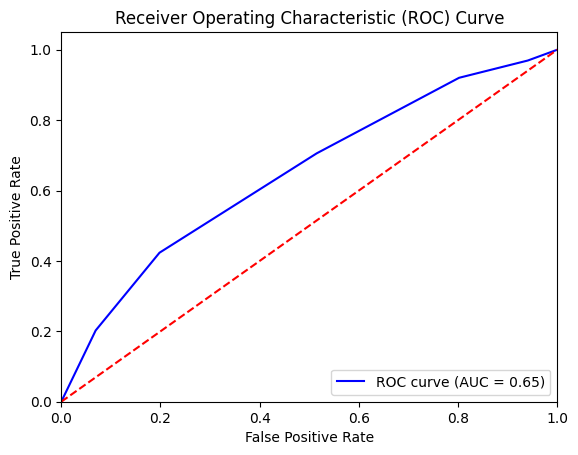


Figure 1: ROC Curve of KNN Figure 2: Confusion Matrix of KNN

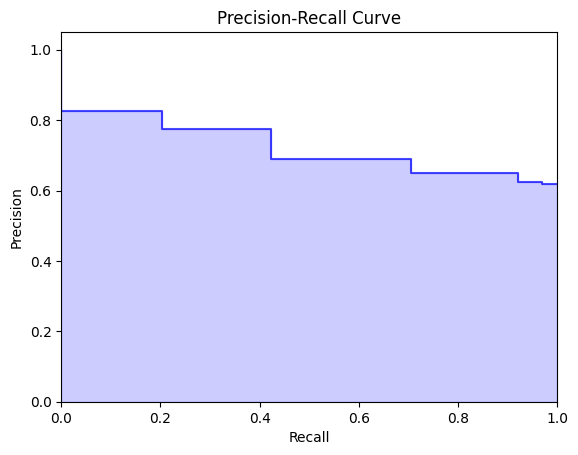
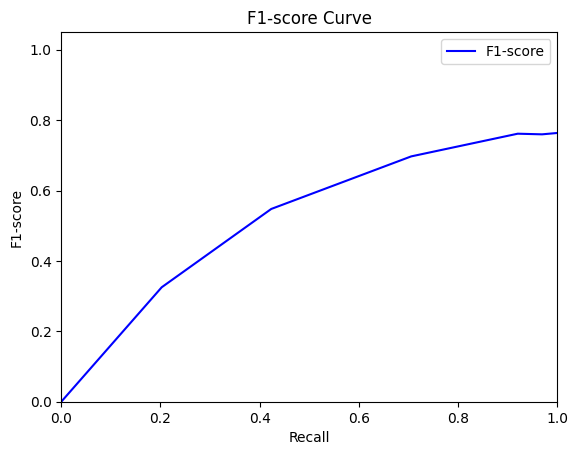


Figure 3: F1-score curve of KNN Figure 4: Precision-Recall curve of KNN

**Naive Bayes**

Naive Bayes is a probabilistic machine learning algorithm commonly used for classification tasks. It's based on the principles of Bayes' theorem, which describes the probability of an event occurring given certain prior knowledge or evidence.

The Naive Bayes algorithm was used to predict heart attack with the given dataset. A 80/20 split of the dataset was used to create training and test sets. Based on the feature independence assumption, the Naive Bayes model was trained on the training data. A test set was used to evaluate the model's performance, and the results showed accuracy of 89.39% surpassing the random guessing. To enhance the model's performance, further exploration and analysis may be carried out. For example, feature selection can be improved, and new methods can be included to account for feature dependencies.

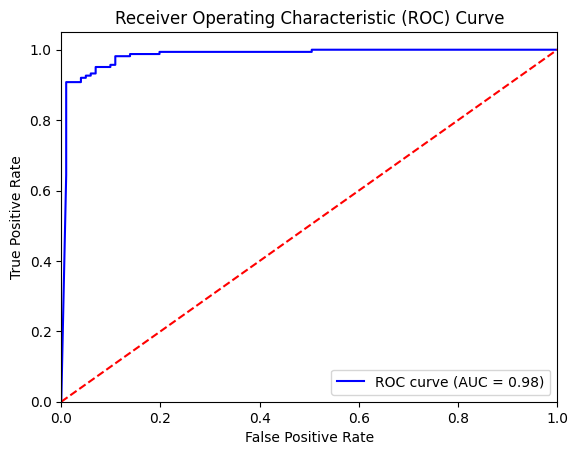
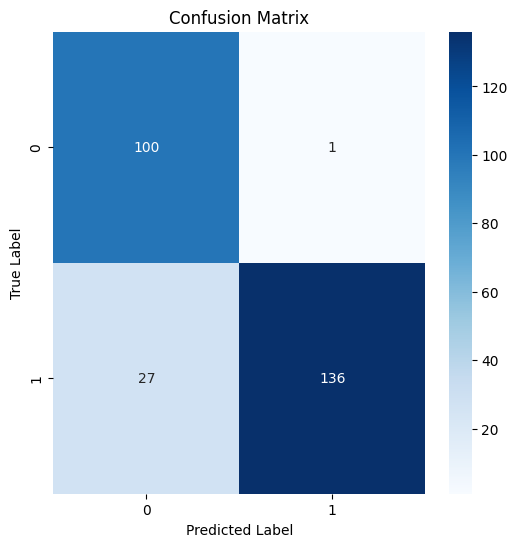
 

Figure 1: ROC Curve of Naïve Bayes Figure 2: Confusion Matrix of Naïve Bayes

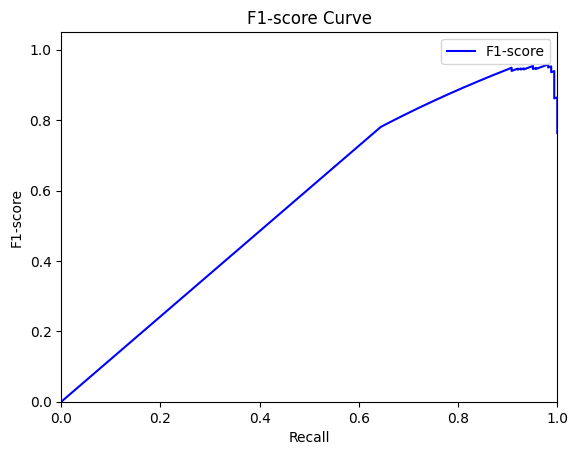
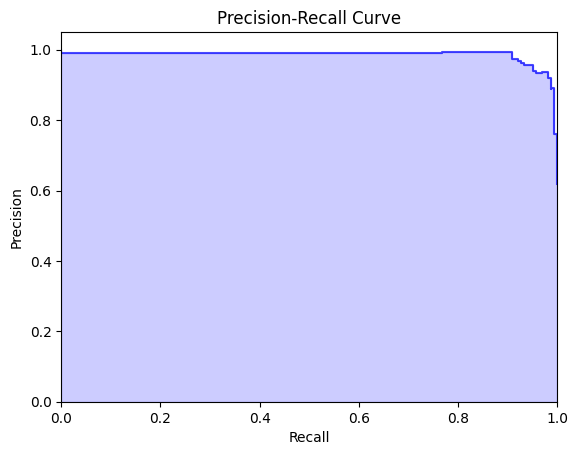
 

Figure 3: F1-score Curve of Naïve Bayes Figure 4: Precision-Recall Curve of Naïve Bayes

**Logistic Regression**

Logistic Regression is a statistical and machine learning algorithm used for binary and multi-class classification tasks. Despite its name, it's actually a method for classification rather than regression. It's based on the logistic function, which transforms any input into a value between 0 and 1. This makes it suitable for modeling probabilities and making binary decisions.

The Logistic Regression algorithm was used with an 80/20 split of the dataset to create training and test sets. Relevant features that are likely to be predictive of heart disease was selected. The logistic regression was trained on the training data for a binary classification. A test set was used to evaluate the model's performance, and the results showed accuracy of 79.16% surpassing the random guessing. To enhance the model's performance, further exploration and analysis may be carried out. For example, feature selection can be improved, and new methods can be included to account for feature dependencies.

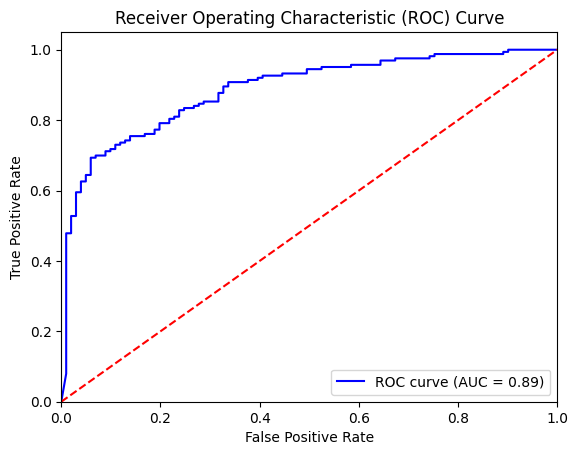
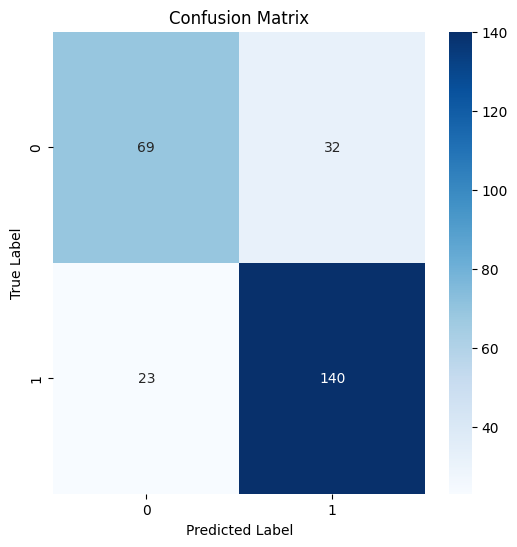
 

Figure 1: ROC Curve of Logistic Regression Figure 2: Confusion Matrix of Logistic Regression

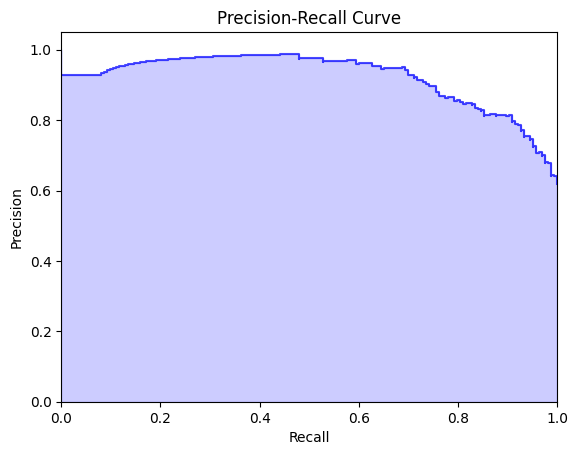
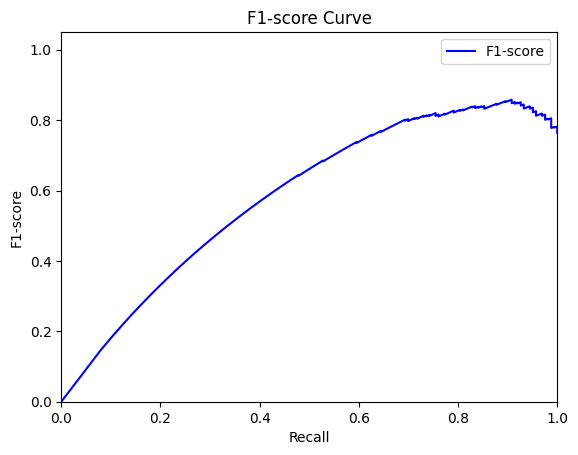


Figure 3: F1-score Curve of Logistic Regression Figure 4: Precision-Recall Curve of Logistic Regression