# Cardiac Disease Prediction System using Machine Learning

# INTRODUCTION

Cardiac disease remains an enduring challenge in the landscape of public health, casting a profound shadow over individuals, healthcare systems, and societies worldwide. Spanning a spectrum of conditions affecting cardiac and vascular function, heart disease encompasses coronary artery disease, arrhythmias, congenital defects, and other afflictions that compromise cardiovascular integrity. Its pervasive influence transcends geographic, socioeconomic, and demographic boundaries, standing as a formidable adversary on the global health stage.

The repercussions of cardiac disease extend far beyond physiological impairment, encompassing broader socio-economic ramifications. Globally, it stands as the leading cause of mortality, exacting a toll of millions of lives annually. Moreover, its impact extends to morbidity, imposing substantial healthcare expenditures, diminishing productivity, and eroding quality of life for affected individuals and their communities.

Adding complexity to this multifaceted issue is the intricate interplay of numerous factors influencing heart health. Lifestyle choices, including sedentary behavior, dietary habits, tobacco use, and alcohol consumption, have emerged as prominent determinants of cardiovascular risk. Coupled with demographic shifts, rising obesity rates, and the prevalence of diabetes, these modifiable risk factors have fueled a surge in heart disease incidence.

Traditional methods of heart disease detection primarily rely on well-established risk assessment protocols, such as the Framingham Risk Score (a gender-based algorithm that estimates an individual's 10-year risk of developing cardiovascular disease) [16], which utilize readily available clinical data like blood pressure, cholesterol levels, and smoking status. These methods are grounded in clinical familiarity, making them easy to integrate into routine healthcare practice. However, they are often limited in scope, focusing mainly on demographic and clinical variables, and potentially overlooking other important predictors of heart disease, such as genetic predispositions and lifestyle factors. Furthermore, traditional approaches may lack personalization, offering generalized risk estimates that do not reflect the unique characteristics and circumstances of each patient.

Pros and Cons of Existing Methods

**Pros of Traditional Methods:**

- Established Protocols: Traditional methods rely on well-established risk assessment protocols like the Framingham Risk Score [16], utilizing readily available clinical data.
- Clinical Familiarity: Healthcare professionals are often familiar with traditional risk assessment tools, facilitating ease of use and integration into routine clinical practice.

**Cons of Traditional Methods:**

- Limited Scope: Traditional methods may focus primarily on demographic and clinical variables, potentially overlooking other important predictors of heart disease.
- Lack of Personalization: Traditional methods may not account for individual variations in risk factors, providing generalized risk estimates.

In response to these challenges, innovative approaches are imperative to mitigate the burden of heart disease and fortify preventative efforts. In this context, the burgeoning field of machine learning (ML) holds considerable promise. By leveraging vast datasets, ML algorithms can discern intricate patterns and relationships that elude conventional analytical methods. Through predictive modeling, ML offers the potential to revolutionize risk assessment, early detection, and personalized interventions in cardiovascular health.

Pros and Cons of Advanced Techniques:

**Pros of Advanced Techniques:**

- Enhanced Predictive Power: Advanced techniques, like machine learning algorithms, offer superior predictive power by analysing large datasets and identifying complex patterns.
- Personalized Risk Stratification: Advanced models can provide personalized risk estimates by integrating a wide range of data sources, enabling precise risk stratification and tailored interventions.

**Cons of Advanced Techniques:**

- Complexity and Resource Intensiveness: Implementing advanced techniques may require specialized expertise, computational resources, and infrastructure, posing challenges for healthcare systems.
- Interpretability and Transparency: Some advanced models may lack interpretability, making it challenging for healthcare professionals to understand underlying factors driving risk predictions.

Furthermore, the transformative potential of ML extends beyond clinical practice to encompass public health initiatives and policy formulation. By identifying high-risk populations and informing resource allocation, ML-driven approaches can facilitate proactive interventions at the population level, thereby mitigating the overall burden of heart disease and fostering healthier communities.

In conclusion, the convergence of heart disease as a global health crisis and the transformative potential of machine learning heralds a new era in cardiovascular medicine. By harnessing the analytical prowess of ML algorithms, we can transcend traditional boundaries and forge innovative pathways toward a future where heart disease is not merely managed but prevented altogether. As we can see and understand the working of a ML model by the workflow diagram below which shows the generalized steps involved in the implementation the of Machine Learning techniques in various dataset to use it various places required.

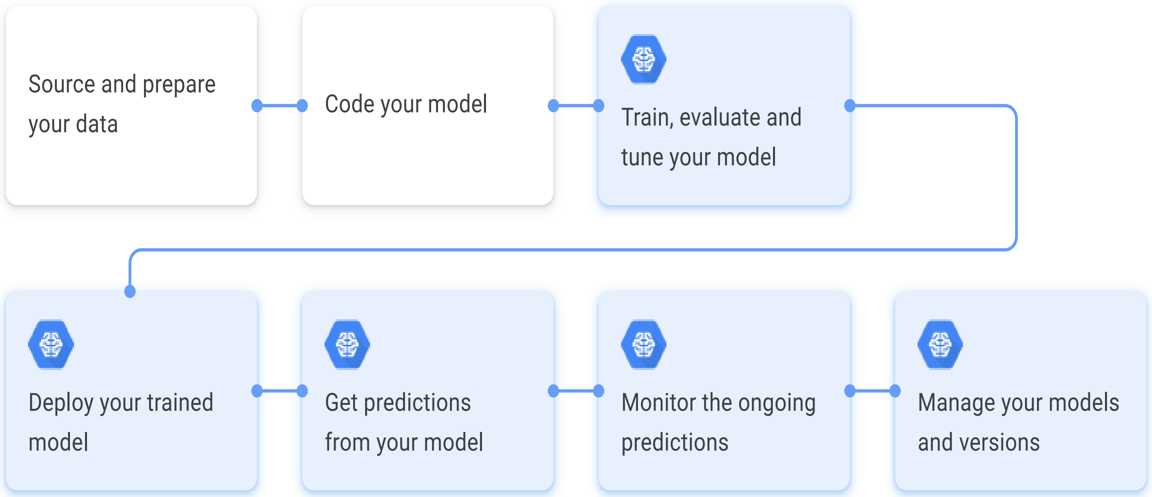

Source and prepare your data
Code your model
Train, evaluate and tune your model
Deploy your trained model
Get predictions from your model
Monitor the ongoing predictions
Manage your models and versions

Fig.1 Machine Learning Workflow

## 1.1 Motivation

The impetus behind developing an advanced cardiac disease risk prediction model stems from the necessity to enhance diagnostic accuracy and preventative measures in cardiovascular healthcare. Traditional methods of risk assessment, which typically consider factors such as age, sex, blood pressure, cholesterol levels, and smoking status, often fall short in capturing the full spectrum of risk determinants. These conventional approaches, while foundational, may overlook the intricate interplay of various health parameters that contribute to the onset of heart disease.

Machine learning algorithms offer a transformative solution by enabling the analysis of extensive datasets to uncover latent patterns and correlations. These algorithms can integrate a broader array of variables, including medical history, lifestyle factors, and biometric data, thus providing a more nuanced risk profile for everyone. The application of machine learning in this context allows for the identification of novel risk factors that traditional methods might miss, thereby improving the precision of risk stratification.

By leveraging machine learning techniques, this project aims to develop a sophisticated heart disease risk predictor that surpasses the limitations of traditional models. The enhanced predictive capabilities of these algorithms can facilitate early detection and personalized intervention strategies. This, in turn, empowers healthcare professionals to implement targeted preventative measures, ultimately reducing the incidence and burden of heart disease. Such advancements not only promise to improve individual patient outcomes but also contribute to the broader public health objective of mitigating the impact of cardiovascular diseases on society.

## 1.3 Objectives

The main objectives for this project are to provide enhancement to traditional methods of heart disease detection with the help of new emerging technology like machine learning:

- To develop a machine learning model that can classify/predict whether a patient is prone of having a cardiovascular failure by taking in account multiple attributes.

- To compare high performing machine learning algorithms and find out which one works the best among the algorithms applied to build and train models on the same dataset.

## 2.2 Summary of the Literary Works

Heart disease remains a significant global health concern, necessitating accurate and timely detection for effective management and prevention of adverse outcomes. Recent advancements in machine learning (ML), deep learning (DL), and data science offer promising avenues for improving heart disease detection and diagnosis.

Studies such as Pagrut et al. (2022) [1], Rahman et al. (2023) [5], and Saikumar et al. (2022) [11] demonstrate the effectiveness of ML and DL techniques in analyzing medical data and predicting heart disease risk. These approaches leverage various algorithms, including random forest, deep graph convolutional neural networks, and linear quadratic discriminant analysis, to achieve high accuracy in detection.

Moreover, techniques like oversampling methods (Albert et al., 2022) [13], sequential feature selection (Assegie et al., 2022) [14], and IoT-enabled sensor data (Lokhande & Chinnaiah, 2023) [9] contribute to improving the robustness and accuracy of heart disease detection models, particularly in scenarios with imbalanced datasets or limited access to healthcare resources.

The findings from these studies underscore the importance of leveraging advanced computational techniques in healthcare to address the challenges associated with heart disease diagnosis. ML and DL models, when trained on comprehensive medical datasets and combined with innovative approaches like IoT sensor data and feature selection methods, offer a powerful tool for early detection and personalized treatment of heart disease.

Overall, this literature review highlights the potential of interdisciplinary collaborations between computer science, medicine, and public health to develop novel solutions for heart disease detection and management, ultimately contributing to improved patient outcomes and reduced mortality rates.

## 2.3 Outcome of Literature Review

The literature review provides a comprehensive overview of recent research endeavors in heart disease detection using machine learning and related techniques. It demonstrates the effectiveness of these approaches in improving diagnostic accuracy and highlights the potential for further advancements in the field. By synthesizing findings from diverse studies, this review underscores the importance of interdisciplinary collaboration and innovative methodologies in addressing the multifaceted challenges posed by heart disease. Furthermore, the successful implementation of machine learning models, deep learning architectures, and data-driven approaches showcases the transformative potential of computational techniques in healthcare. Beyond the technical aspects, the integration of these technologies into clinical practice holds promise for enhancing patient outcomes, optimizing resource allocation, and ultimately reducing the global burden of heart disease. As such, this literature review serves as a roadmap for future research directions, encouraging continued exploration and refinement of data-driven solutions for heart disease detection and management.

## 2.5 Research Objectives

Cognizant of the imperative to bolster predictive healthcare analytics and enable early detection of heart-related ailments, the project aims to:

- Explore the dataset.
- Conduct feature selection via statistical methods.
- Implement diverse machine learning algorithms.
- Optimize hyperparameters using GridSearchCV.
- Evaluate model performance rigorously.
- Perform comparative analysis of model performance.

# 3. METHODOLOGY

The methodology outlined below delineates a systematic approach to developing and evaluating heart disease prediction models using machine learning algorithms. Beginning with the comprehensive collection of pertinent health parameters, the process emphasizes data integrity through meticulous preprocessing, including addressing missing values and outliers. Subsequently, a diverse set of classification algorithms is selected and fine-tuned to optimize performance using hyperparameter tuning techniques. Model training ensues, followed by rigorous evaluation utilizing established metrics to gauge predictive accuracy and efficacy. The comparison of model performances culminates in the selection of the most proficient algorithm for heart disease prediction. Visualizations further elucidate the results, facilitating interpretation and potential implications for clinical practice. Comprehensive documentation ensures transparency and reproducibility, underpinning the reliability and validity of the methodology.

## 3.1 Dataset Description

### 3.1.1 Attribute Information

Table 1: Attributes of the dataset

| **Feature** | **Description** |
|---|---|
| Age | Age of the patient (years). |
| Sex | Sex of the patient (M: Male, F: Female). |
| ChestPainType | Type of chest pain experienced [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]. |
| RestingBP | Resting blood pressure (mm Hg). |
| Cholesterol | Serum cholesterol level (mg/dl). |
| FastingBS | Fasting blood sugar (1: if FastingBS > 120 mg/dl, 0: otherwise). |
| RestingECG | Results of the resting electrocardiogram [Normal: Normal, ST: ST-T wave abnormality, LVH: Left ventricular hypertrophy]. |
| MaxHR | Maximum heart rate achieved (numeric value between 60 and 202). |
| ExerciseAngina | Exercise-induced angina (Y: Yes, N: No). |
| Oldpeak | ST depression induced by exercise relative to rest (numeric value). |
| ST_Slope | The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]. |
| HeartDisease | Output class (1: heart disease, 0: Normal). |

### 3.1.2 Source and Composition

The dataset was constructed by combining five pre-existing heart disease datasets, each contributing to the robustness and diversity of the data:

Table 2: Combination of datasets

| **Dataset** | **Number of Observations** |
|---|---|
| Cleveland | 303 observations |
| Hungarian | 294 observations |
| Switzerland | 123 observations |
| Long Beach VA | 200 observations |
| Stalog (Heart) Data Set | 270 observations |

From these sources, a total of 1,190 observations were compiled, with 272 duplicated entries subsequently removed, resulting in a final dataset of 918 unique observations.

## 3.2 Proposed Approach

### 3.2.1 Dataset Exploration:

- Begin the project by conducting a thorough exploration of the dataset to gain insights into its structure, dimensions, and content. This initial step is crucial for understanding the data and its potential implications for heart disease prediction.
- Employ descriptive statistics, visualizations, and exploratory data analysis (EDA) techniques to uncover patterns, distributions, and relationships within the data. Descriptive statistics such as mean, median, standard deviation, and quartiles provide summaries of numerical variables, while frequency tables and histograms help analyze categorical variables.
- Visualize the data using scatter plots, correlation matrices, and box plots to identify potential correlations and dependencies between predictor variables and the target variable (presence or absence of heart disease). These visualizations aid in understanding the underlying data structure and guide subsequent analysis.

### 3.2.2 Conduct Extensive Exploratory Data Analysis (EDA):

- Dive deeper into the data by conducting extensive exploratory data analysis (EDA) to examine bivariate relationships between predictor variables and the target variable.
- Explore the distribution of predictor variables across different classes of the target variable to identify potential predictive patterns and trends. Utilize visualizations such as heatmaps and stacked bar charts to visualize the distribution of predictor variables across different classes.
- Conduct hypothesis testing, such as t-tests or chi-square tests, to assess the statistical significance of relationships between predictor variables and the target variable. This step helps identify statistically significant predictors of heart disease.

### 3.2.3 Preprocessing Steps:

a. Remove irrelevant features:

- Identify and eliminate features that are redundant, irrelevant, or contain no meaningful information for heart disease prediction. This step helps streamline the model building process and improve model performance.

b. Address missing values:

- Implement appropriate strategies to handle missing values in the dataset, such as mean imputation, median imputation, or predictive modeling. Addressing missing values ensures the integrity and completeness of the data.

c. Treat outliers:

- Detect and address outliers using methods such as z-score normalization, interquartile range (IQR) method, or transformations like logarithmic or Box-Cox transformations. Treating outliers helps mitigate their impact on model performance and ensures robustness.

d. Encode categorical variables:

- Convert categorical variables into numerical format using techniques such as one-hot encoding or label encoding. This step enables the inclusion of categorical variables in the model training process and ensures compatibility with machine learning algorithms. e. Transform skewed features:
- Apply transformations such as logarithmic, square root, or Box-Cox transformations to achieve normal-like distributions for skewed numerical features. Transforming skewed features helps improve model performance and interpretability.

### 3.2.4 Model Building:

- Establish pipelines for models that require feature scaling, such as Support Vector Machines (SVM) and K-Nearest Neighbours (KNN), to ensure consistency in preprocessing steps across different models. Feature scaling helps improve the convergence and stability of these models.
- Implement and tune classification models including KNN, SVM, Decision Trees, and Random Forest to predict the presence or absence of heart disease. Hyperparameter tuning using techniques such as grid search or random search helps optimize model performance.
- Emphasize achieving high recall for class 1 (presence of heart disease) to ensure comprehensive identification of heart patients, as early detection is crucial for timely intervention and treatment.

### 3.2.5 Evaluate and Compare Model Performance:

- Utilize precision, recall, and F1-score metrics to evaluate the effectiveness of classification models in predicting heart disease. These metrics provide insights into the model's performance in terms of precision (accuracy of positive predictions), recall (sensitivity), and F1-score (harmonic mean of precision and recall).
- Compare the performance of different models using appropriate evaluation metrics and statistical tests, such as paired t-tests or McNemar's test, to identify the most effective model for heart disease prediction. This comparative analysis helps determine the model that best balances predictive performance and generalization capability.
- Interpret and analyze model results to gain insights into the predictive capabilities of the models and their implications for clinical decision-making and patient care. Understanding the strengths and limitations of each model helps inform healthcare professionals in utilizing predictive models for heart disease detection effectively.

By following this comprehensive methodology, we aim to develop robust and accurate predictive models for heart disease detection, ultimately contributing to improved patient outcomes and healthcare decision-making.

## 3.3 Design Implementation

The design and implementation of this project follow a structured and methodical approach to develop a reliable heart disease risk prediction model using machine learning algorithms. Initially, comprehensive data collection is undertaken to gather relevant health parameters, ensuring the integrity of the dataset. This is followed by an extensive data preprocessing phase, which involves handling missing values, detecting and managing outliers, encoding categorical variables, and scaling numerical features. Subsequently, a selection of classification algorithms is made, including Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbors Classifier. To enhance the performance of these algorithms, hyperparameter tuning is conducted using GridSearchCV. Each algorithm is then meticulously trained on the preprocessed data. The next phase involves rigorous model evaluation, employing metrics such as accuracy, cross-validation score, and ROC AUC score, alongside visualizations like confusion matrices and ROC curves. A comparative analysis of performance metrics across models is performed to identify the most effective algorithm. The best-performing model is then selected and deployed for heart disease prediction. The project concludes with thorough documentation and reporting, detailing the entire process and presenting comprehensive findings and insights. This systematic design ensures the development of an accurate and robust heart disease prediction model. Figure below shows the same.

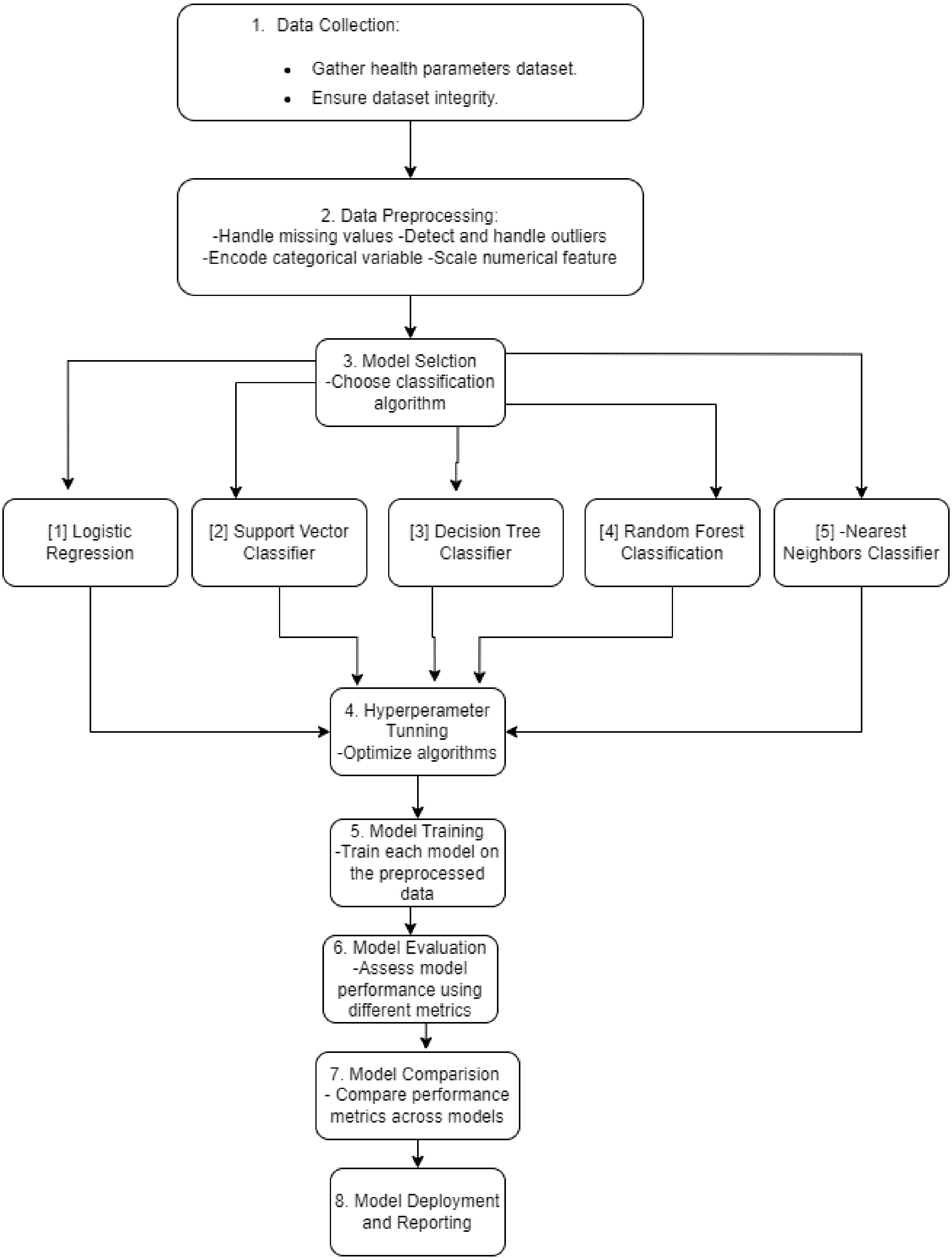

1. Data Collection:
• Gather health parameters dataset.
• Ensure dataset integrity.
2. Data Preprocessing:
-Handle missing values -Detect and handle outliers
-Encode categorical variable -Scale numerical feature
3. Model Selction
-Choose classification algorithm
[1] Logistic Regression
[2] Support Vector Classifier
[3] Decision Tree Classifier
[4] Random Forest Classification
[5] -Nearest Neighbors Classifier
4. Hyperperameter Tunning
-Optimize algorithms
5. Model Training
-Train each model on the preprocessed data
6. Model Evaluation
-Assess model performance using different metrics
7. Model Comparision
- Compare performance metrics across models
8. Model Deployment and Reporting

Fig.2 Proposed System Flow

# 1. WORK DONE

## 4.1 Proposed Work Plan

The proposed work plan for the project can be seen in the Gantt chart below:

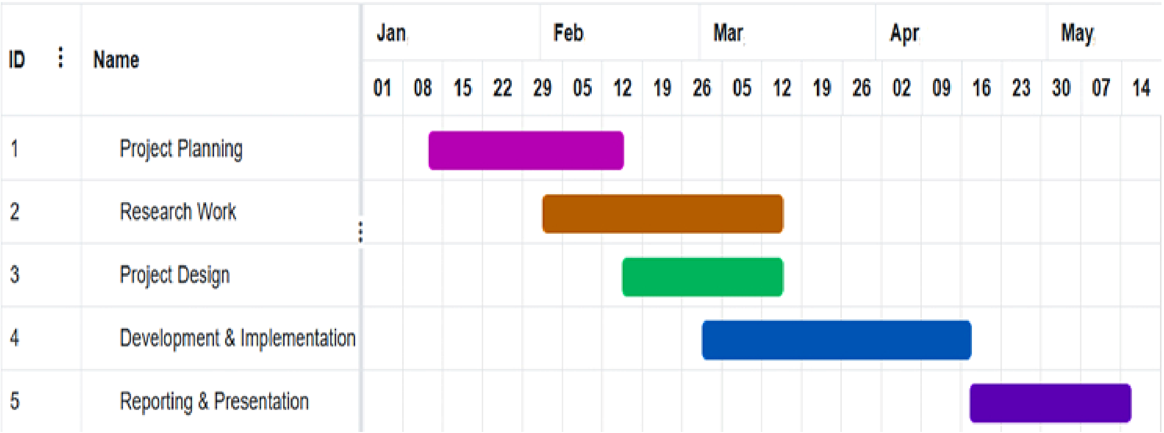


Fig.3 Gantt Chart

The initial phase of the project involved meticulous planning throughout February, with the project framework solidifying as specific details became available. Recognizing the inherent complexity of the chosen topic, a comprehensive research phase commenced in February and extended through March. To optimize workflow and leverage emerging project requirements, design and rudimentary implementation began in late February, running concurrently with the latter stages of research. The subsequent period, encompassing most of the project timeline from mid-April onwards, was dedicated to the implementation of the web application itself. Integration of a local blockchain presented a significant challenge, necessitating substantial time and rigorous testing procedures. The project culminated in the authoring of this report and the development of a final presentation, ensuring timely completion of all deliverables.

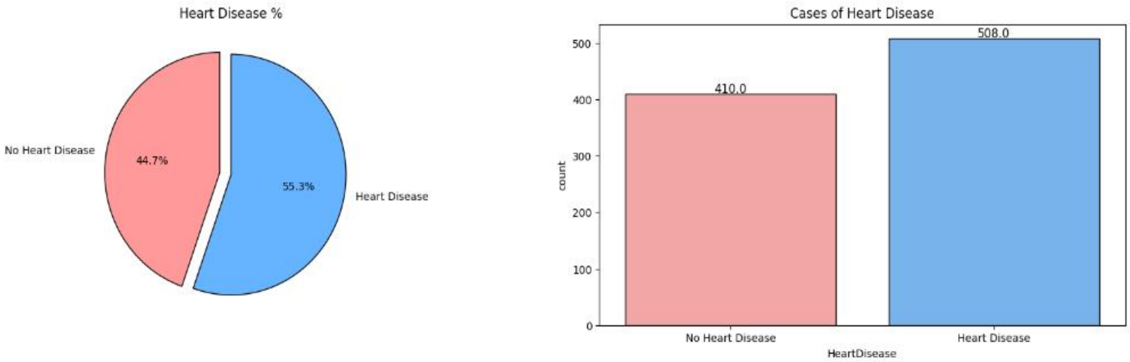
## 4.2 Implementation and Results

### 4.2.1 Data Analysis

Exploratory Data Analysis (EDA) serves as a foundational step in data analysis, aiming to understand the structure, patterns, and relationships within a dataset. In this report, we conduct a comprehensive EDA on a dataset focused on heart disease prediction. Our analysis encompasses various aspects, including the distribution of categorical and numerical features, their relationships with the target variable (heart disease), and inter-feature relationships.

- **Dividing Features into Numerical and Categorical:** In this section, we categorized the dataset's features into numerical and categorical types based on their unique values. Numerical features, such as age, resting blood pressure (RestingBP), cholesterol levels, maximum heart rate (MaxHR), and oldpeak, were identified for their continuous nature. Categorical features, including sex, chest pain type, fasting blood sugar (FastingBS), resting electrocardiographic results (RestingECG), exercise-induced angina (ExerciseAngina), and ST segment slope (ST_Slope), were identified for their discrete values. This division provides a structured understanding of the dataset's composition and facilitates subsequent analyses.

- **Encoding Categorical Features:** To facilitate visualization and modeling tasks, we encoded categorical features using LabelEncoder, transforming text data into numerical representations. This step ensures uniformity in data representation without altering the original dataset. By performing encoding on a deep copy of the dataset, we preserve the integrity of the original data while enabling seamless analysis.
- **Distribution of Categorical Features:** Analyzing the distribution of categorical features revealed insights into their prevalence within the dataset. Notably, all categorical features exhibited near-normal distributions, indicating balanced representations across different categories. This observation lays the groundwork for understanding the relative frequencies of various categorical attributes and their implications for heart disease prediction.
- **Distribution of Numerical Features:** Examining the distribution of numerical features shed light on their variability and potential predictive power regarding heart disease. Notably, oldpeak exhibited a right-skewed distribution, while cholesterol displayed a bimodal pattern. These distributions offer valuable insights into the range and spread of numerical attributes, which are pivotal for subsequent analyses.
- **Target Variable Visualization (Heart Disease):** Visualization of the target variable (Heart Disease) highlighted its balanced distribution within the dataset. With approximately equal proportions of heart disease and non-heart disease cases, the dataset presents an opportunity for robust predictive modeling without class imbalance issues.



The dataset is pretty much evenly balanced!

Fig.4 Target Variable Distribution

- **Categorical Features vs. Target Variable (Heart Disease):** Analyzing the relationship between categorical features and the target variable revealed intriguing patterns. Male individuals showed a higher prevalence of heart disease compared to females. Chest pain type, fasting blood sugar, and exercise-induced angina emerged as significant indicators of heart disease likelihood. Additionally, ST segment slope exhibited distinct patterns, with flat slope values indicating a higher probability of heart disease diagnosis.

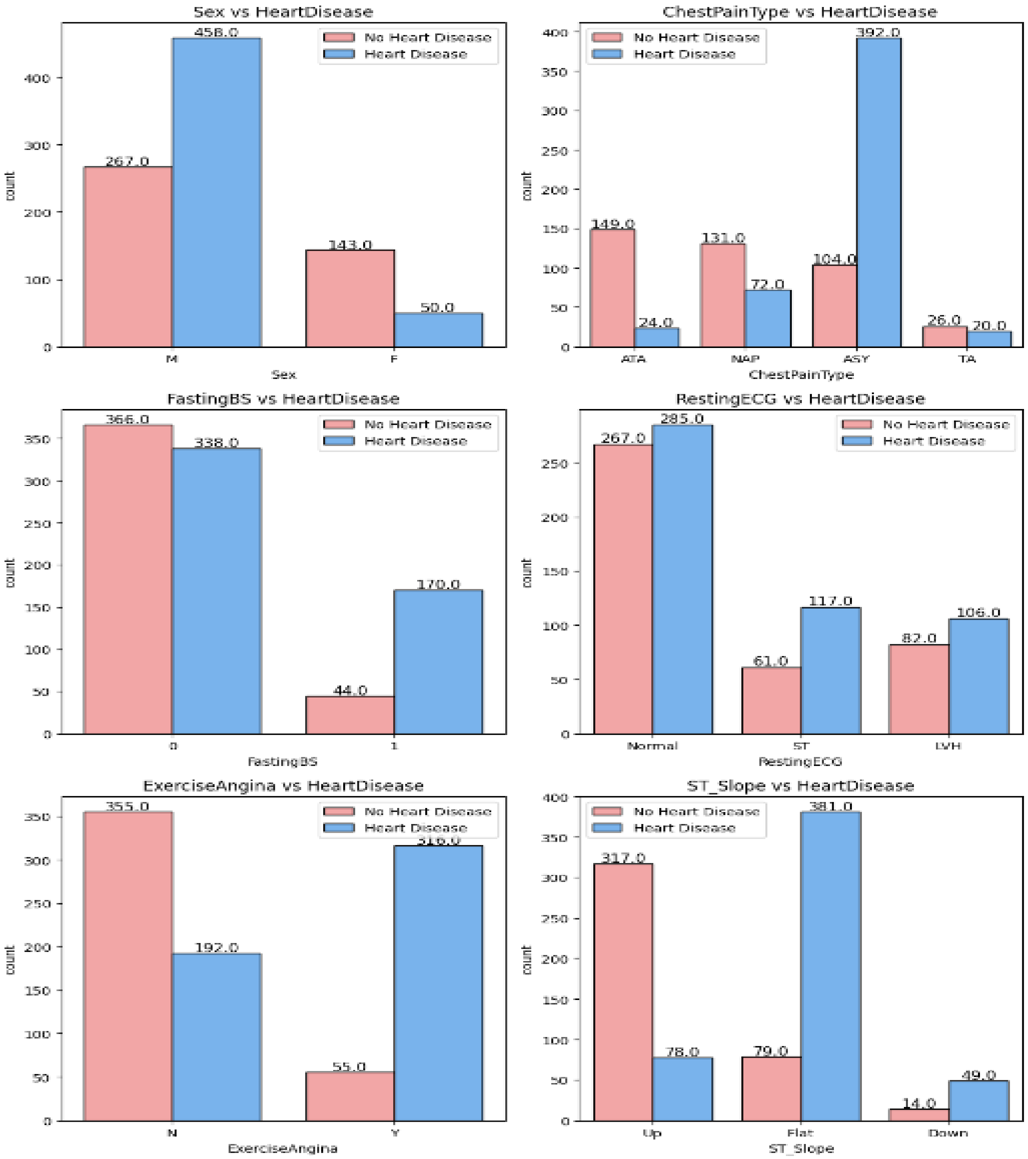


Fig.5 Categorical Features vs. Target Variable

- **Numerical Features vs. Target Variable (Heart Disease):** Exploring the association between numerical features and the target variable elucidated critical insights into potential risk factors for heart disease. Notably, specific ranges of age, resting blood pressure, cholesterol levels, maximum heart rate, and oldpeak values were associated with heightened probabilities of heart disease diagnosis. These findings underscore the importance of numerical attributes in predictive modeling and risk assessment.

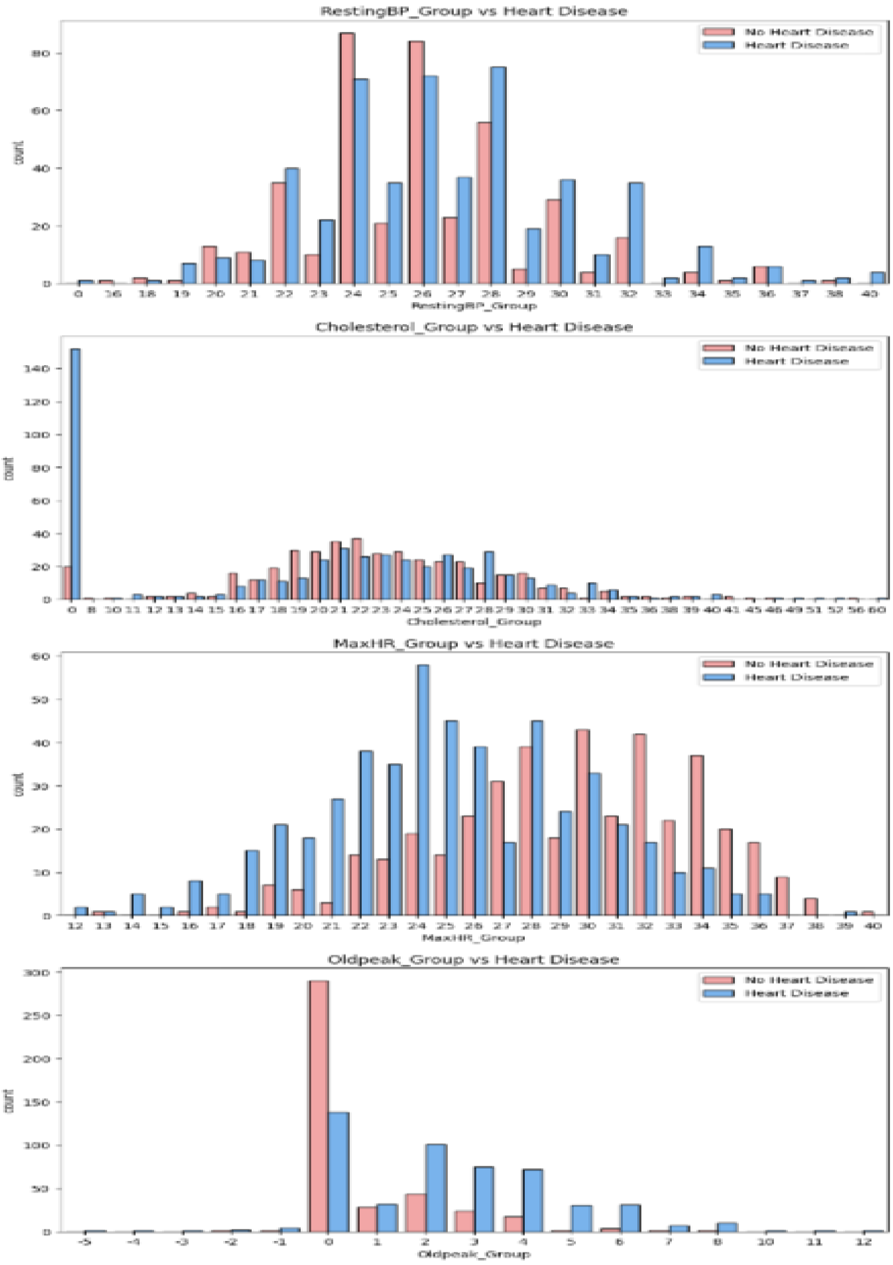


Fig.6 Numerical Features vs. Target Variable (HeartDisease)

- **Numerical Features vs. Numerical Features w.r.t. Target Variable (Heart Disease):** Analyzing the relationships between pairs of numerical features provided nuanced insights into their joint influence on heart disease prediction. Noteworthy observations include correlations between age, blood pressure, cholesterol levels, and heart disease incidence, emphasizing the multifaceted nature of risk factors. Identifying such inter-feature relationships enhances our understanding of the complex dynamics underlying heart disease prediction.

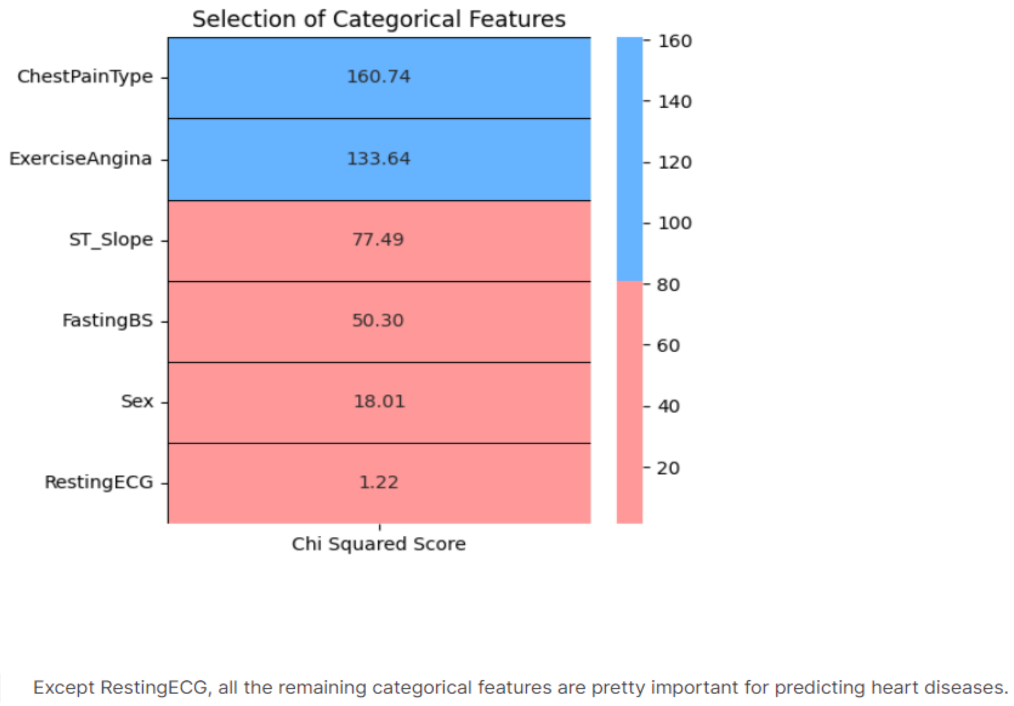
In conclusion, our analysis highlights key trends in features associated with heart disease. Males, those aged 50+, and individuals with ASY chest pain, normal resting ECG, and angina are more susceptible. Numerically, resting BP between 95-170, cholesterol levels of 160-340, MaxHR of 70-180, and oldpeak values of 0-4 correlate with increased risk. These insights inform cardiovascular risk assessment and management.

### 4.2.2 Feature Engineering

Feature scaling is a crucial preprocessing step in machine learning pipelines aimed at standardizing or normalizing the range of features to facilitate better model training. This section delves into the application of feature scaling techniques, namely Min-Max Scaling (Normalization) and Standardization, to ensure uniformity and compatibility among feature values.

**Chi-squared Test for Categorical Features:**

- The Chi-squared test is a statistical method used to determine the independence between categorical variables.
- In the provided code block, the Chi-squared test is applied to categorical features using the SelectKBest function with the Chi-squared score function.
- The resulting Chi-squared scores indicate the degree of association between each categorical feature and the target variable ('Heart Disease').
- Features with higher Chi-squared scores are considered more relevant for predicting heart disease, as they exhibit stronger associations with the target variable.



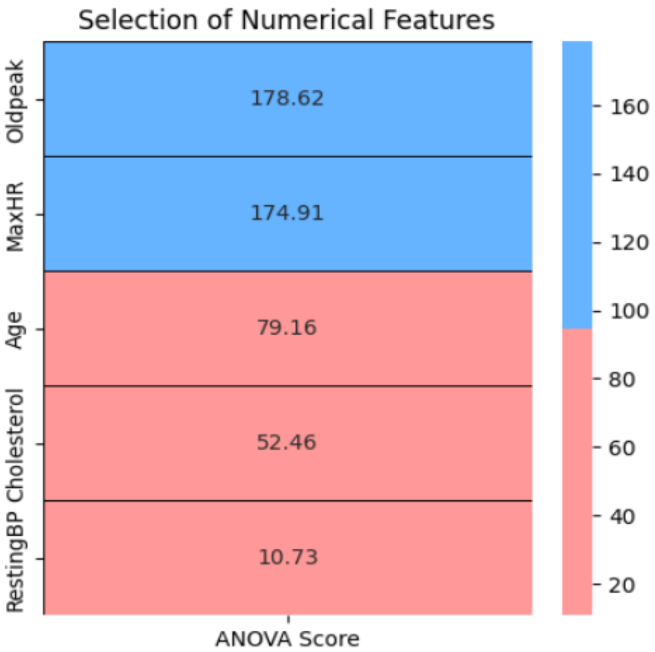
> Except RestingECG, all the remaining categorical features are pretty important for predicting heart diseases.

Fig.7 Feature Selection for Categorical Features

**ANOVA Test for Numerical Features:**

- Analysis of Variance (ANOVA) is a statistical technique used to determine whether there are significant differences between the means of two or more groups.
- In the given code snippet, ANOVA testing is conducted on numerical features using the SelectKBest function with the ANOVA score function.
- The ANOVA scores represent the degree of variance explained by each numerical feature concerning the target variable.

- Features with higher ANOVA scores are deemed more significant in predicting heart disease, as they exhibit greater variability across different levels of the target variable.



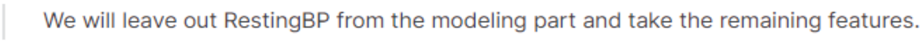

> We will leave out RestingBP from the modeling part and take the remaining features.

Fig.8 Feature Selection for Numerical Features

In conclusion, feature scaling through normalization and standardization plays a pivotal role in enhancing the robustness and performance of machine learning models. By ensuring uniformity and compatibility among feature values, feature scaling contributes to more accurate and reliable model predictions, thus serving as an indispensable preprocessing step in the machine learning pipeline. Additionally, the application of Chi-squared and ANOVA tests for feature selection provides valuable insights into the relevance of categorical and numerical features in predicting heart disease. By identifying and retaining the most informative features, these methods contribute to the development of more accurate and interpretable machine learning models, thereby facilitating better decision-making in clinical and healthcare settings.

### 4.2.3 Modeling

In the modeling phase, we first split our dataset into training and testing sets. Then, we train and evaluate our classification models using key metrics such as accuracy and ROC-AUC score. Additionally, we visualize the ROC curve and generate a confusion matrix and classification report to assess model performance comprehensively. Through these steps, we aim to develop accurate models for detecting heart disease.

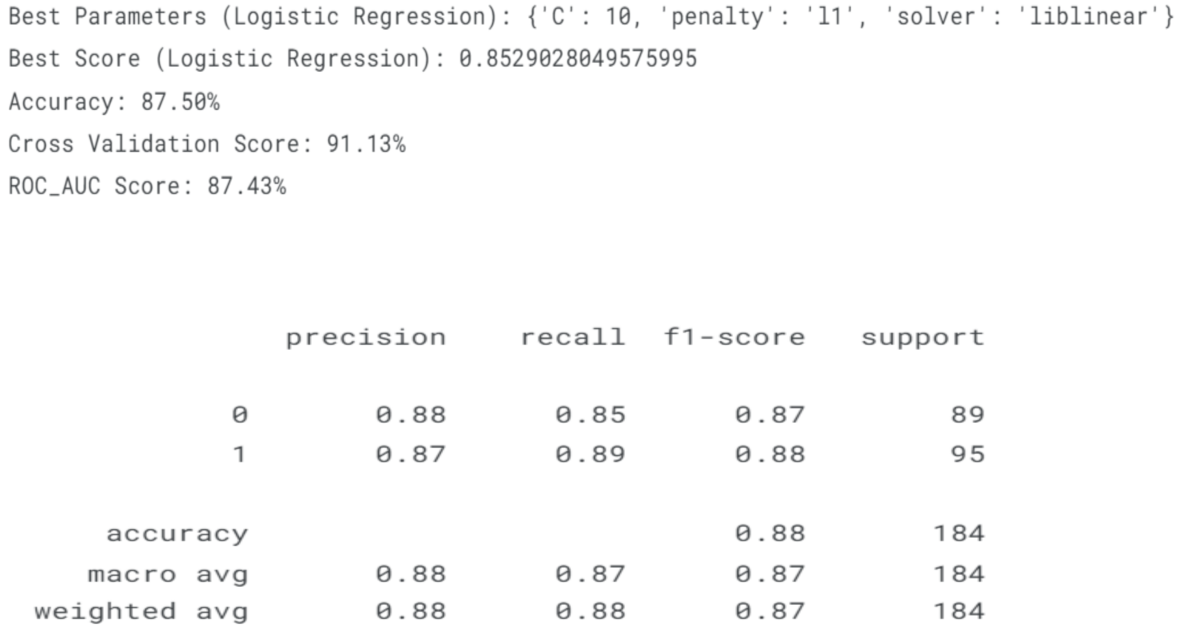
**Logistic Regression:**

Logistic regression is a statistical method for classification tasks with binary outcomes (yes/no, 0/1). It uses a logistic function to model the probability of an observation belonging to a specific class based on independent variables. Unlike linear regression, it predicts probabilities between 0 and 1 instead of continuous values. This allows the model to classify new data points by comparing the predicted probability to a chosen threshold. It is a powerful tool for understanding relationships between features and binary outcomes.

In the context of machine learning, logistic regression is a widely used algorithm for binary classification tasks. Despite its name, logistic regression is a classification algorithm, not a regression algorithm. It models the probability that an instance belongs to a particular class by applying a logistic function to a linear combination of the input features. The logistic function, also known as the sigmoid function, maps any real-valued number into the range [0, 1], representing the probability of the positive class.

Now, let's delve into how the logistic regression model was built and optimized. The hyperparameters of the logistic regression model were tuned using a grid search technique, implemented through the GridSearchCV class from the sklearn.model_selection module. The hyperparameters considered for tuning were the regularization strength (C), penalty type (penalty), and solver algorithm (solver). The grid search involved exhaustively searching over the specified parameter values to identify the combination that yielded the highest accuracy score during cross-validation.

The best parameters obtained from the grid search were C=10, penalty='l1', and solver='liblinear'. This indicates that a regularization strength of 10 with L1 penalty (Lasso regularization) using the 'liblinear' solver algorithm produced the optimal logistic regression model.



```
Best Parameters (Logistic Regression): {'C': 10, 'penalty': 'l1', 'solver': 'liblinear'}
Best Score (Logistic Regression): 0.8529028049575995
Accuracy: 87.50%
Cross Validation Score: 91.13%
ROC_AUC Score: 87.43%
```

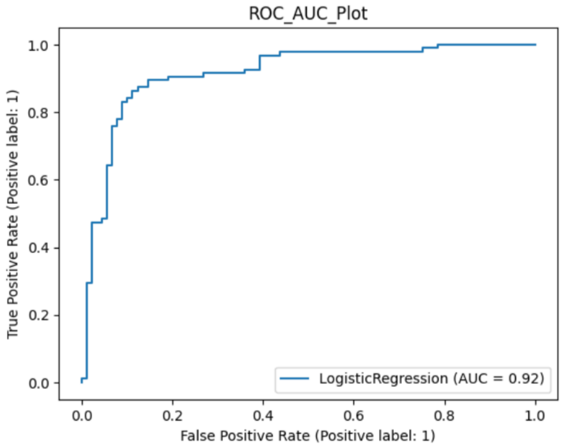
```
              precision    recall  f1-score   support

           0       0.88      0.85      0.87        89
           1       0.87      0.89      0.88        95

    accuracy                           0.88       184
   macro avg       0.88      0.87      0.87       184
weighted avg       0.88      0.88      0.87       184
```

Fig.9 Results of Logistic Regression Algorithm

The accuracy of the tuned logistic regression model on the test set was found to be 87.50%, indicating that the model correctly classified 87.50% of the instances. The cross-validation score, which provides an estimate of the model's generalization performance, was 91.13%. The ROC AUC score, which measures the model's ability to distinguish between the positive and negative classes, was 87.43%.

Furthermore, the classification report provides insights into the precision, recall, and F1-score for both classes (0 and 1). The precision denotes the proportion of true positive predictions among all positive predictions, while recall represents the proportion of true positive predictions among all actual positive instances. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance across both classes.

In summary, the optimized logistic regression model demonstrates strong performance in classifying instances of heart disease based on the provided dataset, achieving high accuracy, cross-validation score, and ROC AUC score. The classification report further confirms the model's effectiveness in terms of precision, recall, and F1-score for both classes, highlighting its potential utility in clinical decision-making processes.



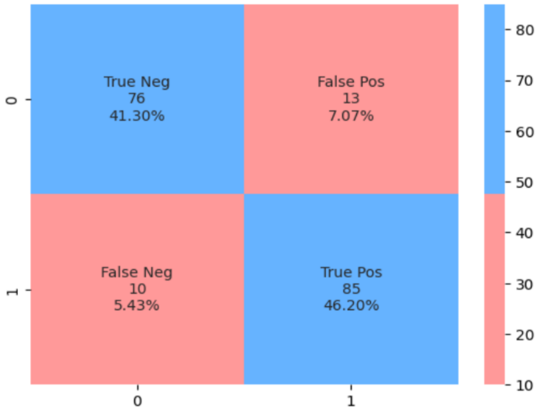


Fig.10 ROC curve

Fig.11 Confusion Matrix

**Support Vector Classifier:**

The Support Vector Classifier (SVC) is a powerful machine learning algorithm used for classification tasks. It works by finding the optimal hyperplane that separates data points of different classes with the maximum margin, thereby enhancing the model's generalization and robustness.

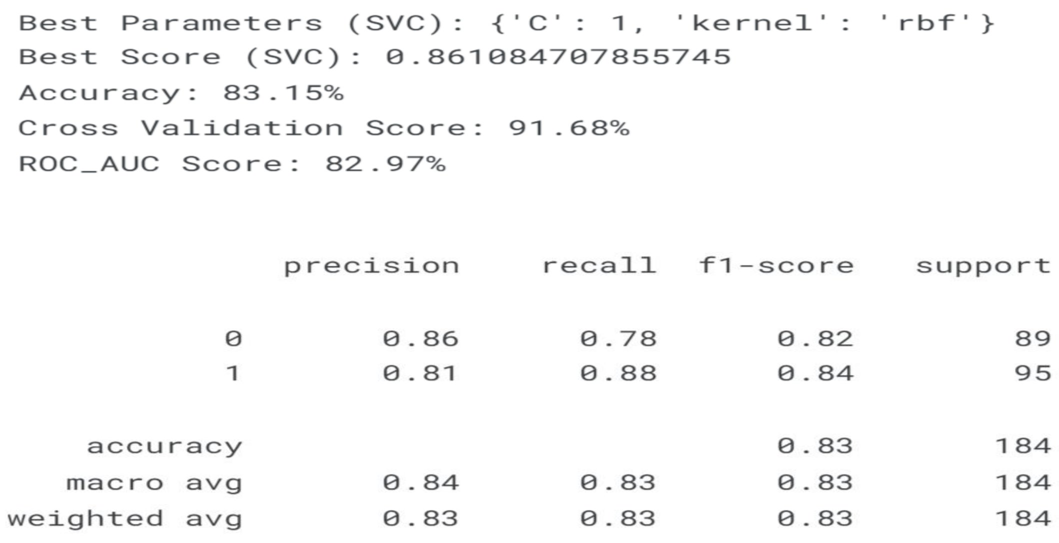
The process of constructing and optimizing the SVC model involves several key steps:

Hyperparameter Tuning: Hyperparameters like the regularization parameter C and the choice of kernel function (kernel) significantly impact the performance of the SVC model. In the provided code, a grid search technique is employed to systematically explore various combinations of hyperparameters and identify the optimal configuration that maximizes model performance.

Grid Search Procedure: The grid search traverses through a predefined grid of hyperparameter values, evaluating the performance of the SVC model with each combination using cross-validation. Different values for C and kernel functions ('linear', 'rbf', 'poly', and 'sigmoid') are tested to determine the optimal configuration that yields the highest mean cross-validated accuracy.

Selection of Best Model: After the grid search process, the hyperparameters corresponding to the highest mean cross-validated accuracy are identified as the optimal configuration. In this case, setting C = 1 and utilizing the radial basis function kernel (kernel = 'rbf') resulted in the highest accuracy.

Model Evaluation: The performance of the best SVC model is evaluated using various metrics, including accuracy, cross-validation score, and ROC_AUC score. These metrics provide insights into the model's ability to accurately classify instances of heart disease and distinguish between different classes.

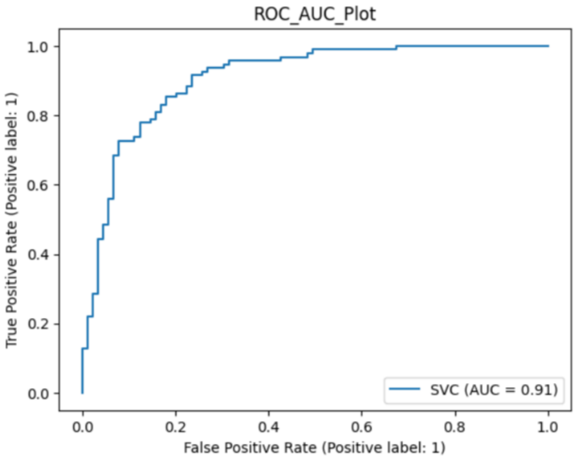


```
 Best Parameters (SVC): {'C': 1, 'kernel': 'rbf'}
 Best Score (SVC): 0.861084707855745
 Accuracy: 83.15%
 Cross Validation Score: 91.68%
 ROC_AUC Score: 82.97%


              precision    recall  f1-score   support

           0       0.86      0.78      0.82        89
           1       0.81      0.88      0.84        95

    accuracy                           0.83       184
   macro avg       0.84      0.83      0.83       184
weighted avg       0.83      0.83      0.83       184
```

Fig.12 Results of SVC

Analysis of Results: The obtained results offer a comprehensive overview of the SVC model's performance. With an accuracy of 83.15% on the test dataset and a mean cross-validation score of 91.68%, the model demonstrates robust predictive capabilities. Furthermore, the ROC_AUC score of 82.97% indicates the model's efficacy in distinguishing between positive and negative instances of heart disease.

In summary, through meticulous hyperparameter tuning and model evaluation, the SVC model exhibits promising performance in predicting heart disease based on the provided dataset. Leveraging principles such as optimal margin separation and kernel methods, SVC emerges as a potent tool for binary classification tasks in healthcare and medical diagnostics.



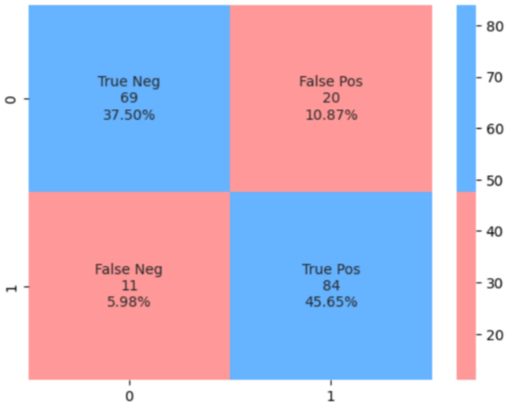


Fig.13 ROC curve

Fig.14 Confusion Matrix

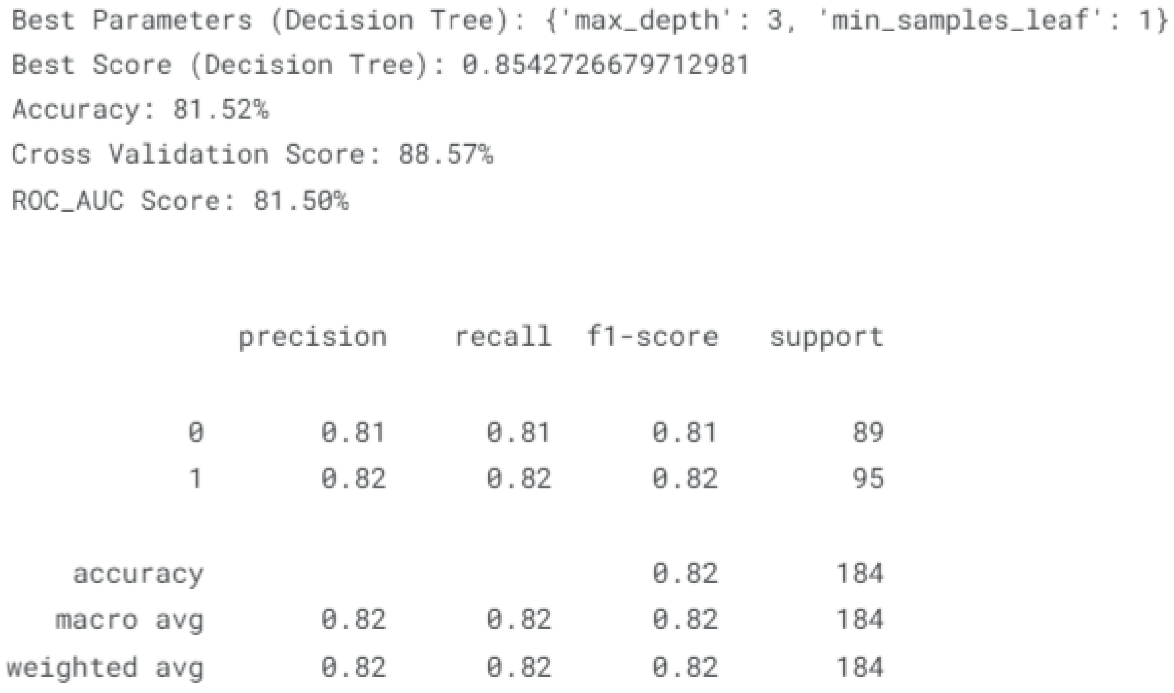
**Decision Tree:**

The Decision Tree model, a fundamental algorithm in machine learning, offers a transparent and interpretable approach to classification tasks. It constructs a tree-like structure where each internal node represents a decision based on a feature value, ultimately leading to the assignment of a class label or a numerical value at the leaf nodes. In the context of heart disease prediction, Decision Trees provide valuable insights into the factors influencing cardiovascular health.

The Decision Tree Classifier was implemented using the scikit-learn library in Python. Through a process of hyperparameter tuning, specifically targeting parameters such as max_depth and min_samples_leaf, the model's performance was optimized. The max_depth parameter controls the maximum depth of the tree, thereby regulating its complexity and potential for overfitting, while min_samples_leaf determines the minimum number of samples required to be at a leaf node, influencing the smoothness of decision boundaries.

The optimization process, conducted via grid search cross-validation, systematically explored various combinations of hyperparameters to identify the configuration that maximized model performance. The resulting best parameters for the Decision Tree model were found to be max_depth=3 and min_samples_leaf=1, yielding a best score of approximately 85.43%.

These parameters indicate a moderately shallow tree with minimal samples per leaf, striking a balance between complexity and generalization.



```
Best Parameters (Decision Tree): {'max_depth': 3, 'min_samples_leaf': 1}
Best Score (Decision Tree): 0.8542726679712981
Accuracy: 81.52%
Cross Validation Score: 88.57%
ROC_AUC Score: 81.50%


              precision    recall  f1-score   support

           0       0.81      0.81      0.81        89
           1       0.82      0.82      0.82        95

    accuracy                           0.82       184
   macro avg       0.82      0.82      0.82       184
weighted avg       0.82      0.82      0.82       184
```

Fig.15 Results of Decision Tree

Upon evaluation on the test set, the optimized Decision Tree model demonstrated promising performance metrics. With an accuracy of 81.52%, a cross-validation score of 88.57%, and a ROC AUC score of 81.50%, the model exhibits robust predictive capability. Additionally, the classification report reveals balanced precision, recall, and F1-score across both classes, underscoring the model's effectiveness in accurately classifying instances into the correct target class.

In conclusion, the Decision Tree model emerges as a valuable tool for heart disease prediction, offering both transparency and predictive power. By leveraging interpretable decision rules and optimizing hyperparameters, Decision Trees provide actionable insights into cardiovascular risk assessment, facilitating informed decision-making and preventive interventions in clinical practice.

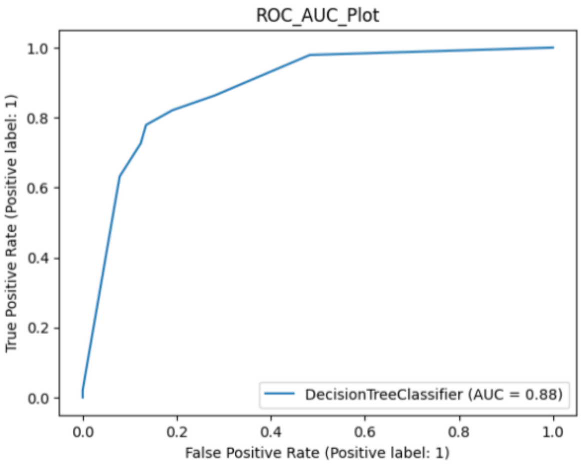


Fig.16 ROC curve

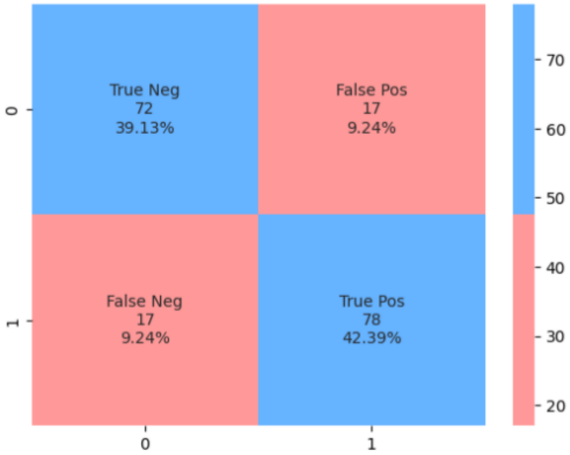


Fig.17 Confusion Matrix

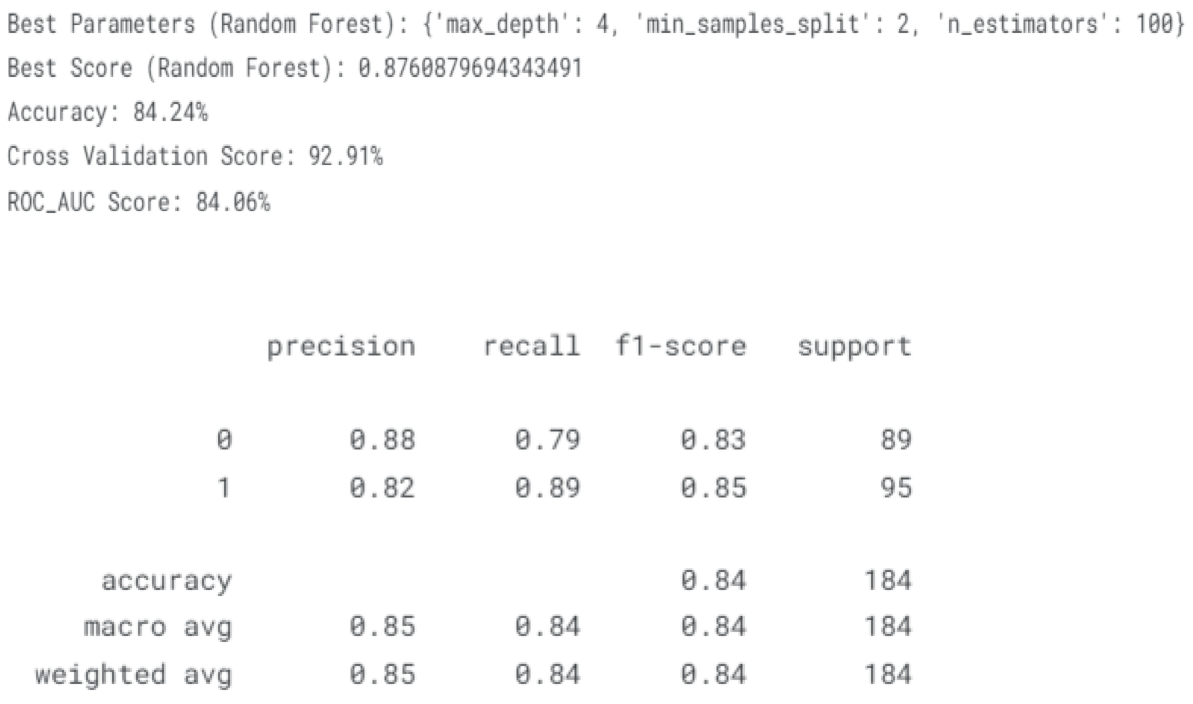
**Random Forest:**

The Random Forest model, a powerful ensemble learning technique, operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or the mean prediction (regression) of individual trees. Each tree is built from a bootstrapped sample of the training data and a subset of features, introducing randomness and reducing variance. This ensemble approach leverages the wisdom of crowds, combining the predictions of multiple trees to produce a robust and accurate model.

In the context of heart disease prediction, Random Forests offer several advantages. They excel in handling high-dimensional data with complex interactions, making them well-suited for analyzing the diverse array of health parameters influencing cardiovascular risk. Additionally, their inherent ability to handle missing data and outliers enhances model robustness and reliability.

This Random Forest Classifier was implemented using the scikit-learn library in Python. The model underwent hyperparameter tuning to optimize its performance. Key hyperparameters such as max_depth, min_samples_split, and n_estimators were fine-tuned to strike a balance between model complexity and generalization. The max_depth parameter controls the maximum depth of individual trees, while min_samples_split determines the minimum number of samples required to split an internal node. Moreover, n_estimators specifies the number of trees in the forest.

The optimization process, conducted through grid search cross-validation, systematically explored various combinations of hyperparameters to identify the configuration yielding the highest performance. The resulting best parameters for the Random Forest model were found to be max_depth=4, min_samples_split=2, and n_estimators=100, with a corresponding best score of approximately 87.61%. These parameters indicate a moderately deep forest with optimal splitting criteria and a substantial number of trees, striking a balance between complexity and predictive power.



```
Best Parameters (Random Forest): {'max_depth': 4, 'min_samples_split': 2, 'n_estimators': 100}
Best Score (Random Forest): 0.8760879694343491
Accuracy: 84.24%
Cross Validation Score: 92.91%
ROC_AUC Score: 84.06%
```

```
              precision    recall  f1-score   support

           0       0.88      0.79      0.83        89
           1       0.82      0.89      0.85        95

    accuracy                           0.84       184
   macro avg       0.85      0.84      0.84       184
weighted avg       0.85      0.84      0.84       184
```

Fig.18 Result of Random Forest

Upon evaluation on the test set, the optimized Random Forest model exhibited promising performance metrics. With an accuracy of 84.24%, a cross-validation score of 92.91%, and a ROC AUC score of 84.06%, the model demonstrates robust predictive capability. Furthermore, the classification report reveals balanced precision, recall, and F1-score across both classes, underscoring the model's effectiveness in accurately classifying instances into the correct target class.

In conclusion, the Random Forest model emerges as a potent tool for heart disease prediction, offering both robustness and predictive accuracy. By harnessing the collective wisdom of diverse decision trees, Random Forests provide valuable insights into cardiovascular risk assessment, enabling informed decision-making and personalized interventions in clinical practice.

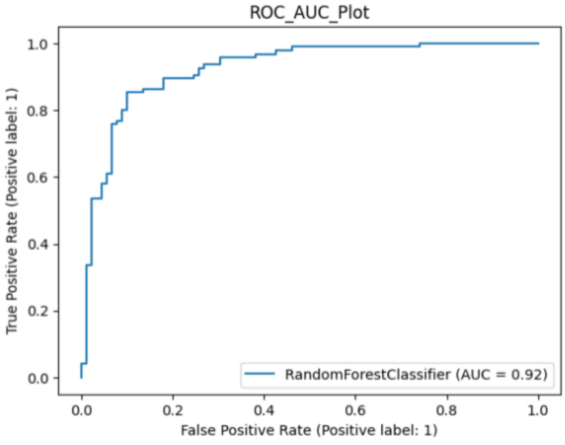


Fig.19 ROC curve

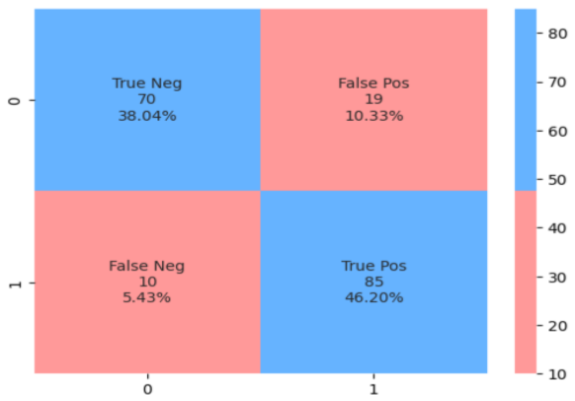


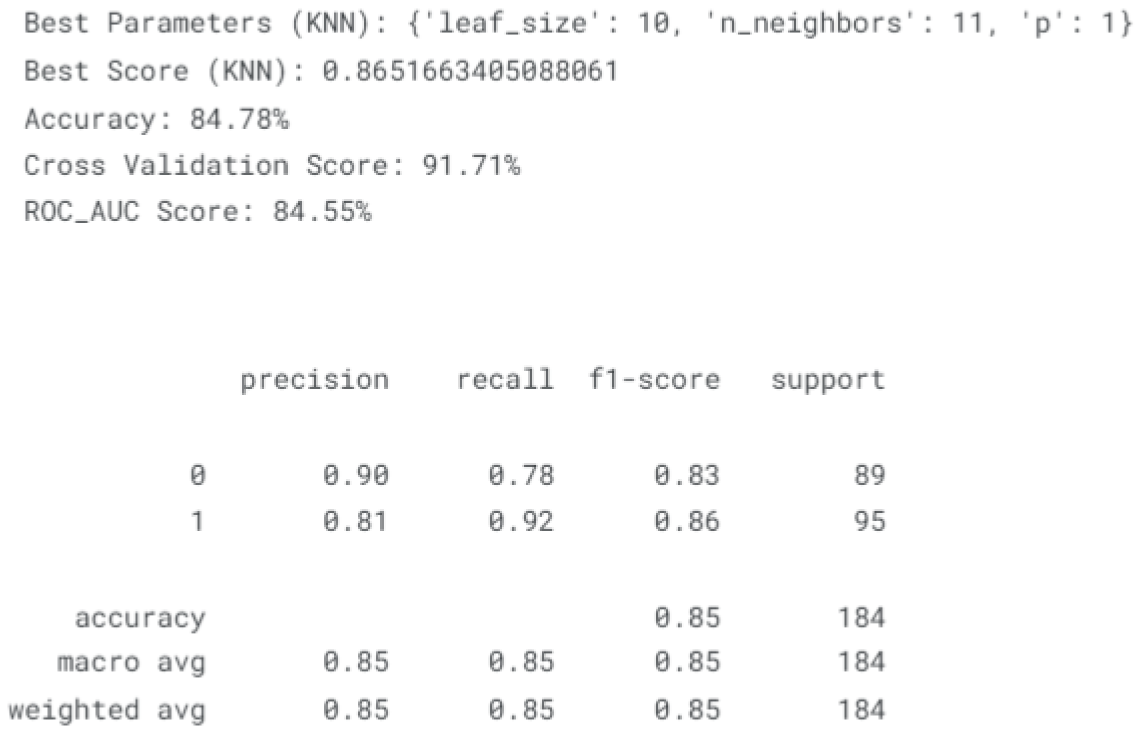
Fig.20 Confusion Matrix

**K-Nearest Neighbors:**

The K-Nearest Neighbors (KNN) model is a simple, yet powerful machine learning algorithm used for both classification and regression tasks. The KNN algorithm operates on the principle that similar data points are likely to have similar outcomes. It classifies a data point based on the majority class among its k-nearest neighbors, where 'k' is a user-defined parameter. This non-parametric method relies on calculating distances (commonly Euclidean) between data points to determine proximity.

For the heart disease prediction task, KNN was chosen due to its simplicity and effectiveness in handling classification problems with clear decision boundaries. The model's strength lies in its ability to adapt to the underlying data distribution without making strong assumptions about the form of the decision boundary.

In this project, the KNN model was built using the scikit-learn library in Python. Hyperparameter tuning was crucial to optimize the model's performance. Key hyperparameters, including n_neighbors (the number of nearest neighbors to consider), leaf_size (the size of the leaf in the tree structure used for nearest neighbor search), and p (the power parameter for the Minkowski distance metric), were fine-tuned.



```
Best Parameters (KNN): {'leaf_size': 10, 'n_neighbors': 11, 'p': 1}
Best Score (KNN): 0.8651663405088061
Accuracy: 84.78%
Cross Validation Score: 91.71%
ROC_AUC Score: 84.55%
```

```
              precision    recall  f1-score   support

           0       0.90      0.78      0.83        89
           1       0.81      0.92      0.86        95

    accuracy                           0.85       184
   macro avg       0.85      0.85      0.85       184
weighted avg       0.85      0.85      0.85       184
```

Fig.21 Result of KNN

Hyperparameter tuning was achieved through grid search cross-validation, a systematic approach to exhaustively search the hyperparameter space for the best combination of parameters. The grid search evaluated multiple configurations to find the optimal set of hyperparameters that maximized the model's performance on the validation set. The best parameters identified were n_neighbors=11, leaf_size=10, and p=1, yielding a best score of approximately 86.52%. These parameters indicate a balanced model with a moderate number of neighbors and a specific distance metric (Manhattan distance when p=1).

Upon evaluating the KNN model on the test set, the results were promising. The model achieved an accuracy of 84.78%, a cross-validation score of 91.71%, and a ROC AUC score of 84.55%. These metrics highlight the model's robustness and its ability to generalize well to unseen data. The classification report further underscores the model's balanced performance across classes, with precision, recall, and F1-score metrics being consistently high for both classes. Specifically, the precision for class 0 (no heart disease) was 90%, with a recall of 78%, while class 1 (heart disease) had a precision of 81% and a recall of 92%.

In conclusion, the KNN model proved to be an effective tool for heart disease prediction. By leveraging the proximity of similar data points, KNN offers a straightforward and interpretable approach to classification, which, when fine-tuned appropriately, can provide accurate and reliable predictions. This makes it a valuable asset in clinical settings, aiding healthcare professionals in assessing cardiovascular risk and making informed decisions for early intervention and treatment.

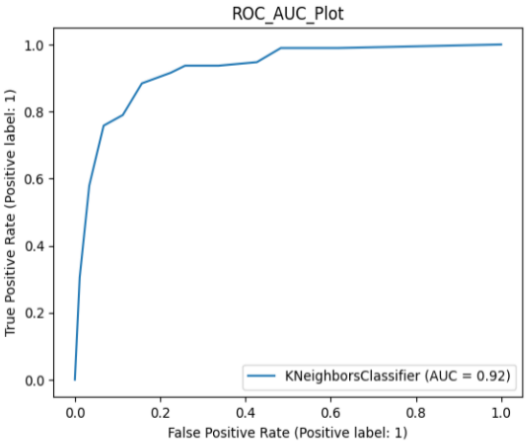


Fig.22 ROC curve

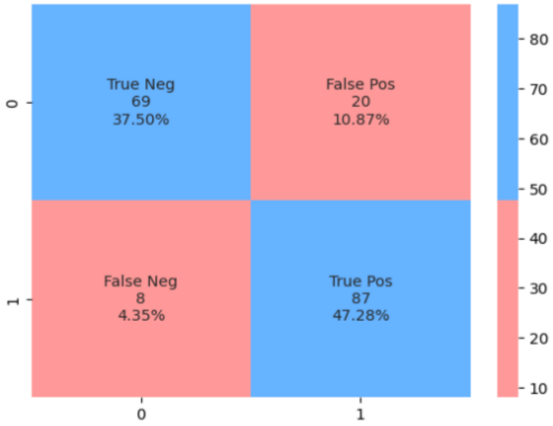


Fig.23 Confusion Matrix

### 4.2.4 Discussion

The results of our heart disease prediction project, summarized in the table, evaluate the performance of Logistic Regression, Support Vector Classifier (SVC), Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbors (KNN) Classifier based on Accuracy, Cross Validation Score, and ROC AUC Score.

Logistic Regression achieved the highest accuracy (87.50%) and strong cross-validation (91.13%) and ROC AUC scores (87.43%), indicating robust performance and excellent discriminatory ability. This aligns with its reputation as a reliable baseline for binary classification problems.

Support Vector Classifier (SVC) demonstrated good performance with an accuracy of 83.15%, a cross-validation score of 91.68%, and a ROC AUC score of 82.97%. While effective in high-dimensional spaces, its slightly lower ROC AUC score suggests it may not rank positive cases as effectively as Logistic Regression.

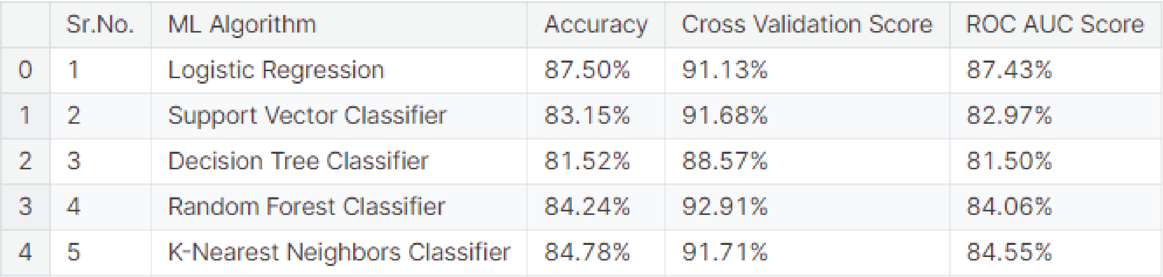
Decision Tree Classifier had an accuracy of 81.52%, a cross-validation score of 88.57%, and a ROC AUC score of 81.50%. Despite being easy to interpret, decision trees can overfit, which is evident in its lower performance metrics. Hyperparameter tuning (max_depth=3, min_samples_leaf=1) was used to address overfitting.

Random Forest Classifier achieved an accuracy of 84.24%, a cross-validation score of 92.91%, and a ROC AUC score of 84.06%. This ensemble method outperformed the single Decision Tree by reducing overfitting through averaging multiple trees. The selected hyperparameters (max_depth=4, min_samples_split=2, n_estimators=100) optimized the model's performance.

K-Nearest Neighbors (KNN) Classifier showed an accuracy of 84.78%, a cross-validation score of 91.71%, and a ROC AUC score of 84.55%. KNN, with hyperparameters (n_neighbors=11, leaf_size=10, p=1), demonstrated competitive performance, indicating effective classification likely due to the dataset's clear class separations.

In summary, Logistic Regression and Random Forest emerged as the top performers, with Logistic Regression slightly ahead in overall accuracy and ROC AUC score. The ensemble approach of Random Forest provided robust generalization, while SVC and KNN also showed strong, reliable results. The Decision Tree, while less effective, highlighted the importance of model complexity and tuning. This analysis underscores the critical role of algorithm selection and optimization in enhancing predictive accuracy for heart disease risk.

The image below shows the result of the models all together.



|  | Sr.No. | ML Algorithm | Accuracy | Cross Validation Score | ROC AUC Score |
|---|---|---|---|---|---|
| 0 | 1 | Logistic Regression | 87.50% | 91.13% | 87.43% |
| 1 | 2 | Support Vector Classifier | 83.15% | 91.68% | 82.97% |
| 2 | 3 | Decision Tree Classifier | 81.52% | 88.57% | 81.50% |
| 3 | 4 | Random Forest Classifier | 84.24% | 92.91% | 84.06% |
| 4 | 5 | K-Nearest Neighbors Classifier | 84.78% | 91.71% | 84.55% |

Fig.24 Cumulative result of all the models

# 2. CONCLUSION

This project aimed to develop a predictive model for heart disease risk using various machine learning algorithms, leveraging a comprehensive dataset and rigorous evaluation metrics. The analysis employed Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbors Classifier to assess their performance in terms of accuracy, cross-validation score, and ROC AUC score. Logistic Regression emerged as the top performer with the highest accuracy and ROC AUC score, demonstrating its reliability as a baseline model for binary classification tasks. Random Forest Classifier also showed strong performance, benefiting from the ensemble approach to mitigate overfitting and enhance generalization.

The Decision Tree Classifier, although less effective compared to other models, provided insights into the importance of model complexity and the necessity of hyperparameter tuning. Support Vector Classifier and K-Nearest Neighbors Classifier exhibited competitive results, underscoring their viability for heart disease prediction with the right parameter configurations.

The project successfully highlighted the potential of machine learning in enhancing the accuracy and reliability of heart disease risk prediction. The findings underscore the importance of selecting appropriate algorithms and optimizing their parameters to achieve the best predictive performance. This work contributes to the growing field of medical data science, offering promising tools for early detection and prevention of heart disease.

## 5.1 Future Work

To extend the utility of this project, it can be proposed to develop a user-friendly web application where individuals can input their health data to receive real-time predictions of their heart disease risk. This platform can be designed to securely collect user data, which can then be anonymized and added to the existing dataset, allowing for continuous improvement of the predictive models. Additionally, incorporating stronger machine learning algorithms, such as XGBoost or neural networks, could further enhance the accuracy and robustness of the predictions. This iterative process of data collection and model refinement will ensure that the application remains up-to-date and effective in providing personalized health insights. Furthermore, the web app could integrate educational resources and preventive measures, empowering users to take proactive steps towards heart health.

# 6. BIBLIOGRAPHY

[1] Prof. S. B. Pagrut, M Areeb Ozair, Suryakant Ingle, Rutika Dharangaonkar, & Apeksha Mundhada. (2023, April 23). Heartcare: Heart Disease Detection using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology, 249–254. https://doi.org/10.48175/ijarsct-9352.

[2] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, 7, 81542–81554. https://doi.org/10.1109/access.2019.2923707.

[3] Salunke, T., Jagade, P., Pawar, S., Rathod, P., & Ghawate, P. N. (2023, April 11). Heart Disease Detection using Hybrid Machine Learning and IoT (Software Based). April-May 2023, 33, 10–13. https://doi.org/10.55529/jaimlnn.33.10.13.

[4] Heart Disease Detection Using Machine Learning and Deep Learning. (2023, April 5). International Journal of Food and Nutritional Sciences, 11(12). https://doi.org/10.48047/ijfans/v11/i12/216.

[5] Rahman, B., Sabarguna, B. S., Warnars, H. S., & Budiharto, W. (2023, June 12). Early Detection of Heart Disease Based on Medical Check-Up Datasets Using Multilayer PerceptronClassifier.https://doi.org/10.21203/rs.3.rs-2992373/v1.

[6] Prof. R. N. Kankrale, Game Arpita, Jadhav Punam, & Kurund Chaitali. (2023, June 24). Heart Disease Detection using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology, 334–341. https://doi.org/10.48175/ijarsct-11652.

[7] Aher, C. N., & Jena, A. K. (2023). Heart Disease Detection from Gene Expression Data Using Optimization Driven Deep Q-Network. Intelligent Data Engineering and Analytics, 601–611.https://doi.org/10.1007/978-981-19-7524-0_53.

[8] Ms. Lolakshi, Gowthami K M, Gowthami K M, Sheekha, & Vaishnavi A S. (2023, January 4). Heart Diseases Detection System. International Journal of Advanced Research in Science, Communication and Technology, 45–49. https://doi.org/10.48175/ijarsct-7833.

[9] Lokhande, P. P., & Chinnaiah, K. (2023). Cardiac Disease Detection Using IoT-Enabled ECG Sensors and Deep Learning Approach. Communications in Computer and Information Science, 195–204. https://doi.org/10.1007/978-3-031-25088-0_16.

[10] Josephine Reenamary, S., & Rani, R. S. A. (2023, February 1). Heart Disease Detection - A Machine Learning Approach. Data Analytics and Artificial Intelligence, 3(2), 59–63. https://doi.org/10.46632/daai/3/2/12.

[11] Saikumar, K., Rajesh, V., Srivastava, G., & Lin, J. C. W. (2022, October 7). Heart disease detection based on internet of things data using linear quadratic discriminant analysis and a deep graph convolutional neural network. Frontiers in Computational Neuroscience, 16. https://doi.org/10.3389/fncom.2022.964686.

[12] Swain, S., Chakravarty, S., Paikaray, B., & Bhoyar, H. (2023). Heart Disease Detection Using Machine Learning Techniques. Lecture Notes in Electrical Engineering, 273–284. https://doi.org/10.1007/978-981-19-9090-8_24.

[13] Albert, A. J., Murugan, R., & Sripriya, T. (2022, December 27). Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology. Research on Biomedical Engineering, 39(1), 99–113. https://doi.org/10.1007/s42600-022-00253-9.

[14] Assegie, T. A., Dr. Tamilarasi, & Kumar, N. K. (2022, September 30). Sequential feature selection for heart disease detection using random forest. Iraqi Journal of Science, 3947–3953. https://doi.org/10.24996/ijs.2022.63.9.26.

[15] fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[16] Jahangiry, L., Farhangi, M.A. & Rezaei, F. Framingham risk score for estimation of 10-years of cardiovascular diseases risk in patients with metabolic syndrome. *J Health Popul Nutr* **36**, 36 (2017). https://doi.org/10.1186/s41043-017-0114-0.

[17] Markelle Kelly, Rachel Longjohn, Kolby Nottingham, The UCI Machine Learning Repository, https://archive.ics.uci.edu.