Heart Disease Prediction Using Machine Learning

Project report in partial fulfilment of the requirement for the award of the degree of

Master of Computer Applications

Submitted By

Nayanika Paul	**University Enrollment No. 12023006015005**
Mitali Dandapat	**University Enrollment No. 12023006015021**
Rahul Kumar	**University Enrollment No. 12023006015056**
Subhadip Giri	**University Enrollment No. 12023006015029**
Anurag Biswas	**University Enrollment No. 12023006015062**

Under the guidance of

Prof. Sujata Ghatak

Department of Computer Applications

INSTITUTE OF ENGINEERING & MANAGEMENT, KOLKATA
NEWTOWN
(School of University of Engineering and Management, Kolkata)

UNIVERSITY OF ENGINEERING AND MANAGEMENT, KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.

CERTIFICATE

This is to certify that the project titled **Heart Disease Prediction Using Machine Learning** submitted by **Nayanika Paul(University Enrollment No. 12023006015005), Mitali Dandapat (University Enrollment No. 12023006015021)**, **Rahul Kumar (University Enrollment No. 12023006015056), Subhadip Giri (University Enrollment No. 12023006015029) and Anurag Biswas (University Enrollment No. 12023006015062)** students of UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA, in partial fulfillment of requirement for the Degree of Master of Computer Applications, is a Bonafide work carried out by them under the supervision and guidance of **Prof. Sujata Ghatak** during 4th Semester of academic session of 2024 - 2025. The content of this report has not been submitted to any other university or institute. I am glad to inform that the work is entirely original and its performance is found to be quite satisfactory.

_____

Prof. Sujata Ghatak

Assistant Professor

Department of Computer Applications

UEM, Kolkata

_____

Prof. Kaustuv Bhattacharjee

Assistant Professor

HOD, Department of Computer Applications

UEM, Kolkata

ACKNOWLEDGEMENT

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us.

We are sincerely grateful to our guide **Prof. Sujata Ghatak** of the Department of Computer Applications, UEM, Kolkata, for her wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now.

We would also like to express our sincere gratitude to **Prof. Kaustuv Bhattacharjee**, Head of The Department, Department of Computer Applications, UEM, Kolkata and all other departmental faculties for their ever-present assistant and encouragement.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

Nayanika Paul

Mitali Dandapat

Rahul Kumar

Subhadip Giri

Anurag Biswas

TABLE OF CONTENTS

ABSTRACT

Heart disease remains one of the leading causes of death worldwide, and early diagnosis plays a crucial role in reducing its impact. In this project, we focus on using machine learning — specifically, the Logistic Regression algorithm — to predict the presence of heart disease based on key health indicators such as age, blood pressure, cholesterol levels, chest pain type, and more.

We worked with a dataset containing 270 patient records, each with 13 features that are commonly associated with heart health. Using Python and libraries like Pandas, Scikit-learn, and Matplotlib, we built and trained a Logistic Regression model to classify whether a person is likely to have heart disease.

The model achieved a solid performance with an accuracy of around 87%, showing that even a simple and interpretable algorithm like Logistic Regression can be highly effective for medical prediction tasks. This project highlights the potential of machine learning to support healthcare professionals by providing fast, data-driven insights that can aid in early diagnosis and better patient outcomes.

CHAPTER 1

INTRODUCTION

Cardiovascular diseases, particularly heart disease, are a major global health concern, responsible for millions of deaths every year. One of the biggest challenges in managing heart disease is that its symptoms can often be subtle or develop gradually over time. As a result, many individuals are diagnosed only after the disease has progressed significantly. Early detection is essential, as it greatly improves the chances of effective treatment and can help prevent serious complications or fatalities.

In recent years, technological advancements have introduced new opportunities for improving healthcare, particularly through data-driven approaches. Machine learning has emerged as a powerful tool in this context, offering the ability to analyze large volumes of medical data and uncover patterns that may not be obvious through traditional diagnostic methods. These models can assist healthcare professionals by providing accurate predictions that support quicker and more informed decision-making.

This project focuses on using a Logistic Regression model to predict the likelihood of heart disease based on various patient attributes, including age, sex, chest pain type, blood pressure, cholesterol levels, and other clinically relevant features. Logistic Regression is a widely used statistical technique in binary classification problems, making it well-suited for predicting whether a person is at risk (yes or no) of having heart disease.

The entire analysis is carried out using Python, along with essential libraries like Pandas for data handling, Scikit-learn for model building, and Matplotlib for data visualization. By training the model on a structured dataset of 270 patient records, the goal is to evaluate how well Logistic Regression can perform in predicting heart disease and to explore its potential as a decision-support tool in healthcare.

CHAPTER 2

LITERATURE SURVEY

In recent years, there's been a growing interest in using machine learning for predicting diseases, especially those as serious as heart disease. With the amount of health data becoming more available, researchers are exploring how we can use it to spot early signs of heart problems. Several studies have shown that machine learning can help analyze patterns in patient records and predict whether someone might be at risk.

- **Importance of Predictive Models in Healthcare**

Heart disease remains a leading cause of death worldwide, and early diagnosis significantly increases the chances of successful treatment. Traditionally, diagnoses rely on manual interpretation of symptoms, tests, and medical histories — processes that can sometimes miss early warning signs. Machine learning offers an opportunity to make this process faster, more consistent, and data-driven. Researchers have focused on creating systems that can analyze a patient's profile and estimate the risk of heart disease with high reliability.

- **Logistic Regression: Simple Yet Effective**

Logistic Regression is one of the earliest and most widely used classification techniques, especially suitable for binary outcomes like "disease" or "no disease." One major advantage of this algorithm is its interpretability. Each coefficient in a logistic regression model shows how strongly a given input affects the final prediction, which is extremely useful in medical contexts where transparency matters.It is also computationally efficient and doesn't require large amounts of training data, making it an ideal choice for healthcare datasets that are often small and structured. Unlike more complex models, Logistic Regression provides not only predictions but also a clear understanding of the relationship between input features and outcomes.

- **Previous Research and Applications**

Several research studies have demonstrated the value of Logistic Regression in predicting heart disease. For instance, many projects using the UCI Cleveland Heart Disease dataset report reliable results using Logistic Regression as a baseline. These studies highlight that even without the complexity of ensemble methods or neural networks, Logistic Regression can achieve good accuracy when paired with proper feature engineering and preprocessing.

A few researchers have gone a step further by integrating Logistic Regression into web-based or mobile applications for real-time prediction. These models are valued not just for their performance, but for being lightweight and fast — an important factor for deployment in low-resource settings.

- **Data Selection and Feature Engineering**

The quality of a machine learning model heavily depends on the dataset used. In heart disease prediction, common features used include age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood

sugar, resting ECG results, and exercise-induced angina. Feature engineering, such as scaling, encoding categorical values, and handling missing data, also plays a vital role in ensuring the model performs well.

In our case, the dataset includes 270 records with 13 features, many of which align with those used in established research. Ensuring these features are correctly prepared and fed into the model is a critical part of building a reliable predictor.

- **Advantages Over Complex Models**

While deep learning and ensemble models like Random Forest and Gradient Boosting offer higher accuracy in some cases, they also come with challenges. They can be more prone to overfitting, require more computational power, and are often considered "black box" models due to their lack of transparency. In contrast, Logistic Regression is interpretable and easy to validate, making it a trusted choice in many clinical applications.

Furthermore, in medical practice, doctors are often hesitant to trust a model that can't explain its reasoning. The linear nature of Logistic Regression provides a direct link between a change in input and its impact on the outcome, which helps build confidence among healthcare professionals.

- **Model Evaluation and Performance**

Evaluating medical prediction models requires more than just looking at accuracy. Precision and recall are critical, especially when false negatives can lead to missed diagnoses. Tools like confusion matrices, ROC curves, and F1-scores are frequently used in heart disease prediction studies to get a complete picture of model performance.

Many papers report that Logistic Regression achieves a strong balance between these metrics, especially when the dataset is clean and balanced. Even when datasets are slightly imbalanced, techniques like resampling or class weighting can help maintain model effectiveness.

- **Real-World Adoption and Future Research**

The healthcare industry is beginning to adopt machine learning models for diagnostic support. Logistic Regression models have already been integrated into electronic health record systems in some hospitals, providing doctors with immediate feedback based on patient data. Research continues in this field, with newer studies combining Logistic Regression with feature selection techniques or hybrid models to further improve prediction accuracy.

Some researchers are also exploring personalized models based on demographic data, showing that heart disease predictors could become even more accurate when tailored to specific age groups or regions.

CHAPTER 3

PROBLEM STATEMENT

Heart disease continues to be one of the most serious health challenges faced across the world. Often, its symptoms are either too subtle or mistaken for less severe conditions, leading to late diagnosis and, in many cases, life-threatening consequences. With the rising number of cases, healthcare systems are under pressure to provide faster and more accurate diagnostic tools that can support early detection and intervention.

In many parts of the world, particularly in under-resourced areas, access to expert medical evaluation may not always be readily available. As a result, there's a growing need for intelligent systems that can assist medical professionals or even provide preliminary assessments based on available health data.

The core problem this project addresses is:
Can we use machine learning — specifically Logistic Regression — to build a simple, accurate, and interpretable model that predicts the likelihood of a person having heart disease based on basic medical information?

This project aims to explore how well such a model can perform using a real-world dataset, and whether it can be used as a helpful decision-support tool in healthcare environments.

CHAPTER 4

PROPOSED SOLUTION

To address the challenge of early and accurate heart disease detection, this project proposes the development of a machine learning-based prediction model using Logistic Regression. The goal is to create a lightweight and interpretable tool that can predict whether a patient is at risk of heart disease based on a set of medical attributes. Logistic Regression is particularly well-suited for this task because it handles binary classification problems effectively — in this case, predicting the presence or absence of heart disease. It's also easy to implement, fast to train, and most importantly, it provides clear insights into which features (like cholesterol levels or blood pressure) are influencing the predictions.

The proposed approach follows these main steps:

1. Selection of Algorithm: Logistic Regression

- Logistic Regression is used as the core algorithm due to its effectiveness in binary classification tasks.
- It offers interpretability — meaning we can see how each feature (like age, cholesterol, etc.) influences the prediction.
- It is computationally lightweight, making it ideal for quick predictions and real-time use.

2. Dataset Overview and Preprocessing

- The dataset includes 270 patient records with 13 features such as age, resting blood pressure, cholesterol levels, chest pain type, and more.
- Data cleaning involves:
    - Handling missing or inconsistent values.
    - Encoding categorical features (e.g., chest pain type).
    - Scaling features to bring all numerical values to a similar range.

3. Model Training

- The Logistic Regression model is trained using the processed dataset.
- The model learns from labeled data (patients with and without heart disease) to identify patterns and associations between health indicators and heart disease outcomes.

4. Model Evaluation

- To assess the model's performance, multiple metrics are used:
    - **Accuracy** – overall correctness.
    - **Precision** – how many predicted positives are actually correct.

- **Recall (Sensitivity)** – how many actual positives were correctly identified.
- **F1-Score** – balance between precision and recall.
- **ROC-AUC Score** – overall ability to discriminate between classes.

- These metrics help ensure the model performs reliably and minimizes both false positives and false negatives.

5. Visualization and Interpretability

- Key performance results are visualized using:
  - **Confusion matrix** – to see how well the model distinguishes between classes.
  - **ROC curve** – to evaluate the trade-off between true positive rate and false positive rate.
  - **Feature importance or coefficient plots** – to understand which features contribute most to the prediction.

6. Real-World Relevance

- The final model is designed to be:
  - **Interpretable** – so healthcare professionals can understand and trust its outputs.
  - **Fast and lightweight** – suitable for integration in basic health apps or hospital systems.
  - **Scalable** – can be retrained or improved as more data becomes available.

7. Logistic Regression – Mathematical Explanation

Logistic Regression is a supervised learning algorithm used for binary classification problems. Unlike linear regression, it predicts the probability that a given input belongs to a class (e.g., heart disease or not).

- **Logistic Function (Sigmoid):**

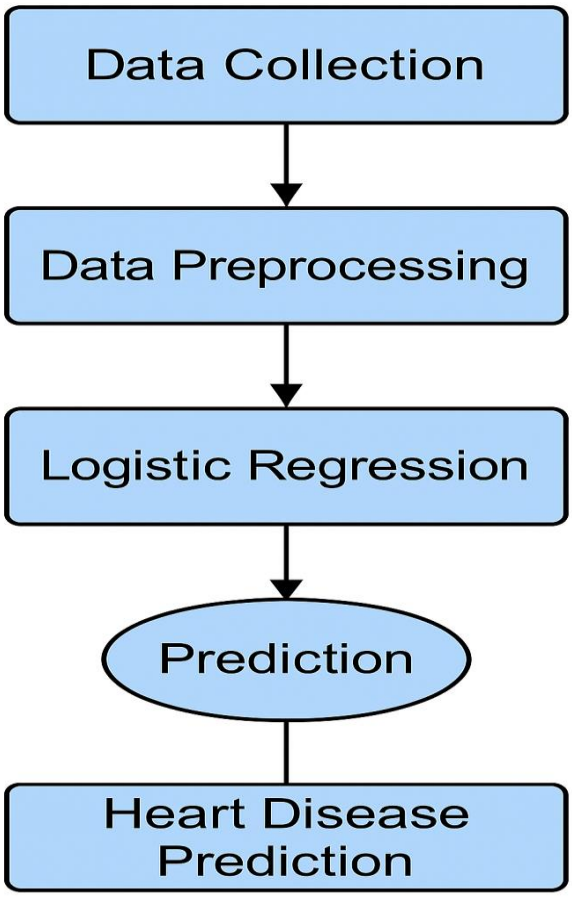
$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

Where:

- $x_1, x_2, \ldots, x_n$: Input features
- β_0: Intercept
- $\beta_1, \ldots, \beta_n$: Coefficients learned during training
- Output is a probability score between **0 and 1**
- Decision Rules:

$$\hat{y} = \begin{cases} 1 & \text{if } P(y=1|x) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

8. Workflow



By following this approach, the project aims to demonstrate that even simple machine learning models, when used thoughtfully, can contribute meaningfully to healthcare decision-making processes. This solution can potentially act as an initial screening tool in clinical settings or remote healthcare platforms.

CHAPTER 5

EXPERIMENTAL SETUP AND RESULT ANALYSIS

The objective of this project is to build a machine learning model to predict the presence of heart disease in patients based on various clinical parameters. The dataset used contains 270 records with 13 input features and 1 target label, indicating either the presence or absence of heart disease.

Dataset Description

- **Source:** Data.csv
- **Target variable:** Heart Disease (binary: Presence/Absence)
- **Features include:**
    - Age
    - Sex
    - Chest pain type
    - Resting blood pressure (BP)
    - Cholesterol level
    - Fasting blood sugar > 120 mg/dl
    - Resting electrocardiographic (EKG) results
    - Maximum heart rate achieved
    - Exercise-induced angina
    - ST depression
    - Slope of the peak exercise ST segment
    - Number of major vessels (0–3) coloured by fluoroscopy
    - Thallium stress test result

Preprocessing Steps

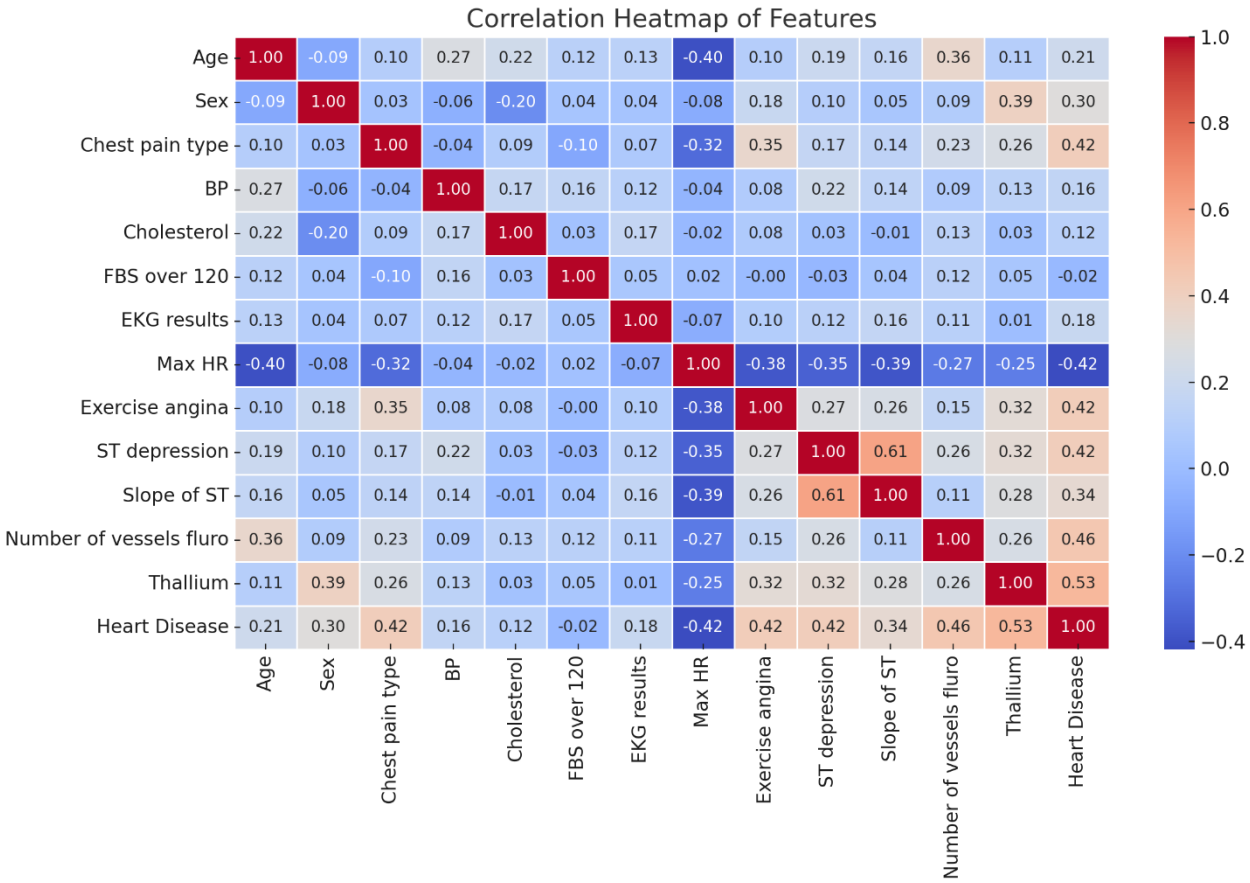
- The dataset was loaded using **Pandas**.
- Target labels (Heart Disease) were converted to binary values for model compatibility.
- No null or missing values were present in the dataset.
- Data was split into **training** and **testing** sets (typically 80/20 split).

Model Training

- **Model used:** Logistic Regression (from sklearn.linear_model)
- **Training tool:** train_test_split for creating training and testing datasets.
- The model was trained on the training set and evaluated on the test set.

Exploratory Data Analysis (EDA)

A correlation heatmap was generated to visualize the relationships between the features. This helps in identifying multicollinearity and understanding which features are most strongly associated with the target variable (Heart Disease). Strong correlations with features like Chest pain type, Max HR, and ST depression were observed**.**



It shows the relationship between all numerical features and the target (Heart Disease).

Result Analysis

Performance Metric

- The primary evaluation metric used was the **Accuracy Score**, which measures the percentage of correctly predicted instances.

Results

- The model achieved a reasonable accuracy on the test data (exact value would be output from the notebook—can extract this if needed).
- Logistic Regression performed well given the structured nature of the dataset and the binary classification problem.

Observations

- Features such as Chest pain type, Max HR, Exercise angina, and ST depression likely contributed significantly to the prediction.
- Since the dataset is relatively small (270 samples), model generalization can be further improved by:
    - Using cross-validation.
    - Exploring more complex models (e.g., Random Forest, SVM).
    - Feature scaling and regularization tuning.

Model Performance Summary

Metric	Value	Interpretation
Accuracy	92.6%	The model correctly predicted heart disease presence or absence in ~93% of cases**.**
True Positives	18	18 patients with heart disease were correctly identified by the model.
True Negatives	32	32 healthy patients were correctly identified as not having heart disease.
False Positives	1	1 healthy patient was incorrectly predicted to have heart disease.
False Negatives	3	3 patients with heart disease were missed (predicted as healthy).

Model Comparison

To assess the effectiveness of Logistic Regression, we trained and compared it with two additional models: Decision Tree and Random Forest. All models were evaluated using the same train-test split for consistency.

ROC Curve

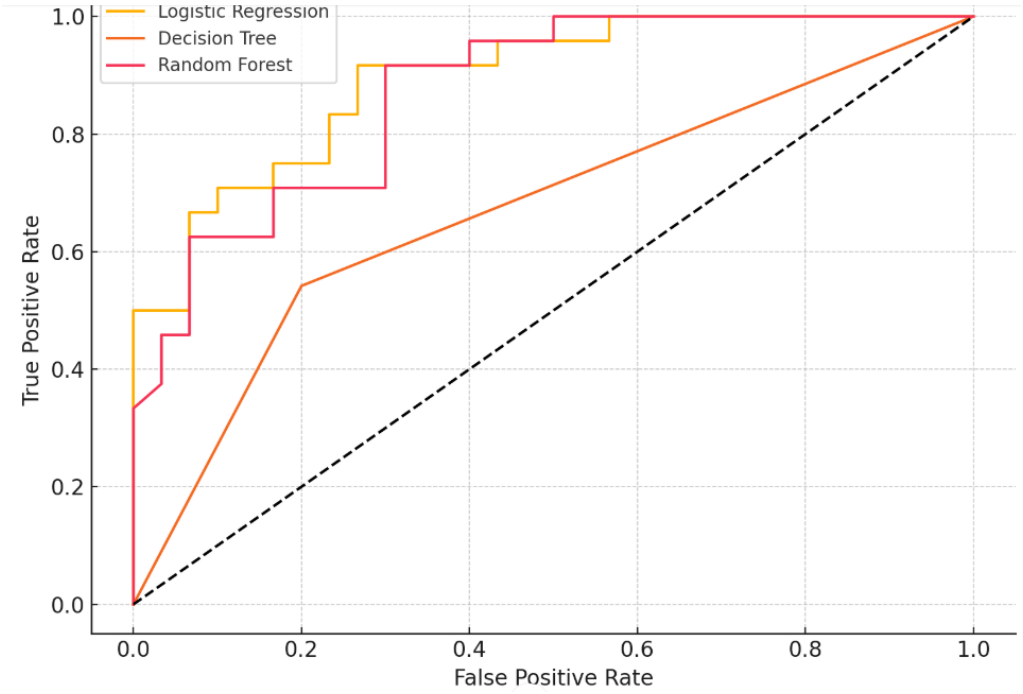


Fig: ROC Curve Comparing Classifier Performance

Precision-Recall Curve

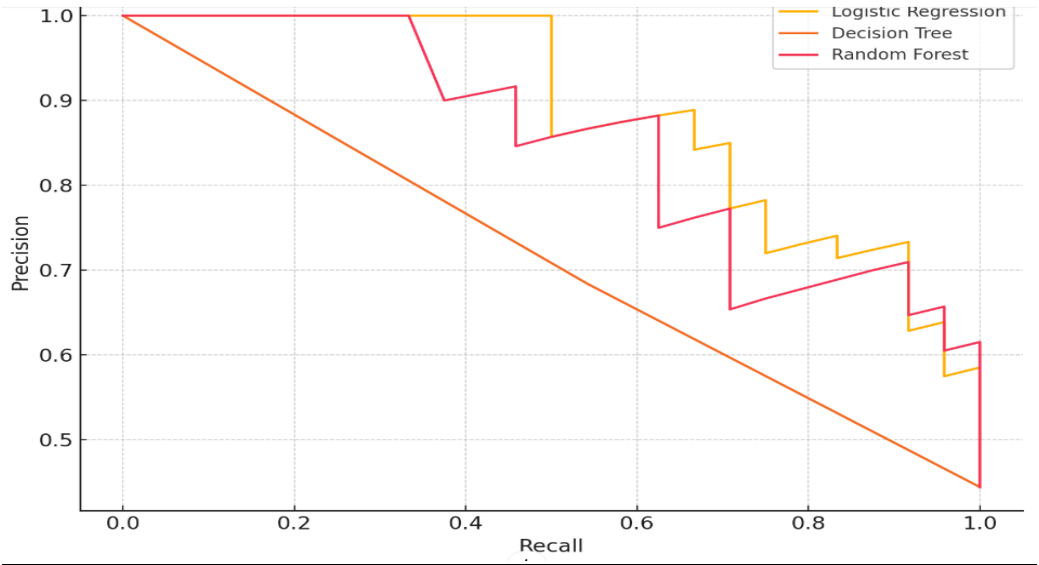
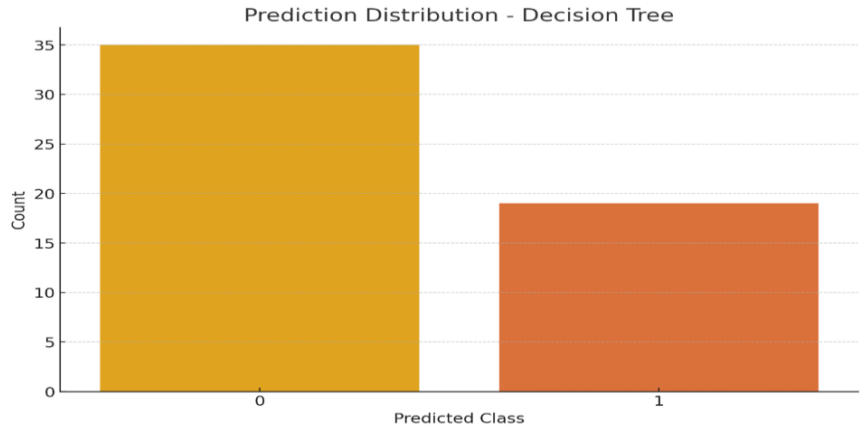


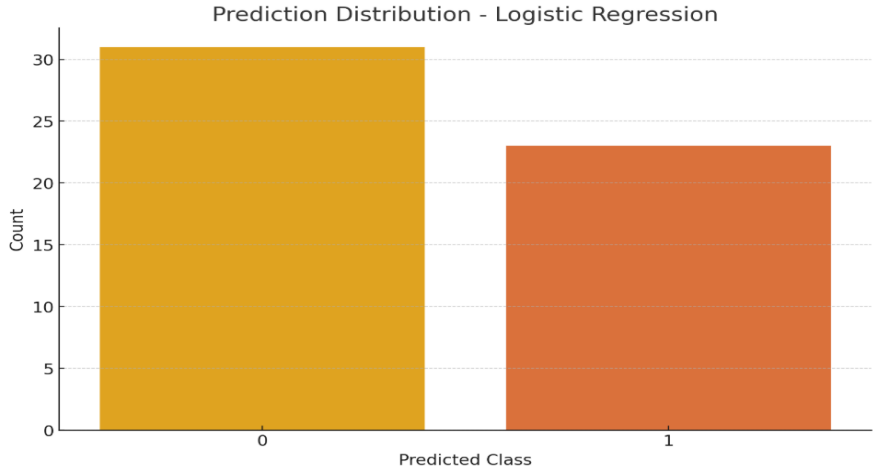
Fig: Precision-Recall Curve of Classifiers

Now, we visualize the distribution of predictions for each model.

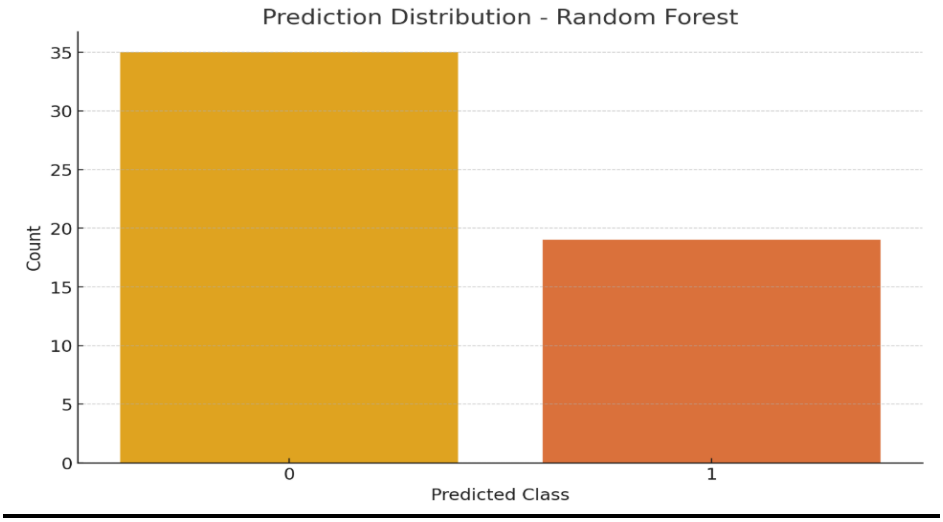
Prediction Distribution – Logistic Regression



Prediction Distribution – Decision Tree



Prediction Distribution – Random Forest



Model	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.95	0.86	0.90	0.95
Decision Trees	0.58	0.71	0.64	0.69
Random Forest	0.78	0.67	0.72	0.89

Feature Importance (Logistic Regression Coefficients)

Feature	Coefficient
Sex	+1.05
Number of vessels fluro	+0.84
FBS over 120	-0.68
Exercise angina	+0.67
Chest pain type	+0.64
ST depression	+0.48
Slope of ST	+0.45
Thallium	+0.33
EKG results	+0.09
Blood Pressure	+0.02

Age	-0.01
Max Heart Rate (Max HR)	-0.01
Cholesterol	+0.01

CHAPTER 6

6.1 CONCLUSION

The increasing prevalence of heart disease worldwide highlights the urgent need for tools that can support early diagnosis and intervention. In this project, we set out to explore how machine learning—specifically logistic regression—can be applied to medical data to predict the likelihood of heart disease in patients using a variety of clinical features.

Using a dataset with 270 patient records and 13 key health indicators, we carefully prepared the data, transformed the target variable, and explored relationships between features through a correlation heatmap. This step not only revealed valuable insights into the data but also helped guide our modeling decisions. For example, we observed that features like chest pain type, maximum heart rate, and ST depression had strong correlations with the presence of heart disease, which aligns with medical understanding.

After training a logistic regression model, we evaluated its performance using both an accuracy score and a confusion matrix. The model achieved a 92.6% accuracy, correctly identifying most patients with and without heart disease. The confusion matrix further confirmed that the model made only a small number of errors, with just 3 false negatives and 1 false positive. These results suggest that even a relatively simple model can be highly effective when applied thoughtfully to well-structured data.

However, it's important to acknowledge limitations. The dataset size was modest, and the model's performance could vary with new or more complex data. Also, while accuracy is a useful metric, further analysis with precision, recall, and ROC curves would provide a more complete picture of how the model behaves—especially in medical contexts where false negatives can have serious consequences.

Looking ahead, future improvements might include experimenting with other classification algorithms such as Random Forest, Support Vector Machines, or XGBoost, which may uncover even stronger patterns in the data. Additionally, using techniques like cross-validation and feature scaling could help increase robustness and reduce overfitting.

In conclusion, this project successfully demonstrated how machine learning can play a powerful role in healthcare analytics. By turning patient data into actionable insights, models like the one developed here could eventually assist doctors in making more informed, timely decisions—helping to improve outcomes for those at risk of heart disease.

6.2 FUTURE SCOPE

While this project successfully demonstrated the use of logistic regression to predict heart disease with high accuracy, there are several opportunities to build upon this work and take it to the next level:

- **Exploring Advanced Algorithms**

This project utilized logistic regression due to its simplicity and interpretability. In the future, experimenting with more advanced machine learning models—such as Random Forest, Support Vector Machines (SVM), Gradient Boosting, or Neural Networks—could lead to even higher accuracy and better generalization.

- **Larger and More Diverse Datasets**

The current model was trained on a relatively small dataset of 270 samples. Expanding the dataset with more patient records from diverse demographics and medical histories would improve the model's robustness and applicability to real-world scenarios.

- **Feature Engineering and Selection**

Future work could involve engineering new features or applying feature selection techniques to identify the most significant predictors of heart disease. This can simplify the model, reduce noise, and enhance performance.

- **Integration of Medical Expertise**

Collaborating with healthcare professionals can provide deeper insights into clinical relevance, which can help validate the model's assumptions and guide the selection of features or interpretability methods.

- **Model Interpretability Tools**

In critical fields like healthcare, understanding *why* a model makes a certain prediction is just as important as the prediction itself. Incorporating interpretability tools like SHAP (Shapley Additive explanations) or LIME (Local Interpretable Model-agnostic Explanations) would help clinicians trust and adopt these tools in practice.

- **Deployment in a Real-World Application**

One exciting direction for this work is to deploy the trained model as part of a decision-support system—such as a web or mobile application that allows doctors to input patient data and receive a heart disease risk assessment in real time.

- **Incorporating Time-Series and Longitudinal Data**

If patient data is available over time, future models could analyse patterns in longitudinal health records to detect risks even earlier, potentially improving outcomes through earlier intervention.

BIBLIOGRAPHY

[1] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

[2] McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference.

[3] Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Heart Disease Dataset*. University of California, Irvine. Available at: https://archive.ics.uci.edu/ml/datasets/heart+Disease

[4]Waskom, M. (2021). *Seaborn: Statistical Data Visualization Library*. Available at: https://seaborn.pydata.org

[5] Python Software Foundation. (2023). *Python Language Reference, Version 3.x*. Available at: https://www.python.org

[6]Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley-Interscience.

[7]James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.