23BCE9719_L55+L56_Heart Disease Prediction

TEAM MEMBERS:

SK.ZAAFIRA YUMN (23BCE9719)
M.NARENDRA KUMAR (23BCE20218)
NSK. SAI KARTHIK (23BCE8104)

Abstract

Heart disease is a leading cause of death worldwide, making early detection crucial. This project uses machine learning to predict heart disease based on patient data such as age, blood pressure, cholesterol, and ECG readings. Various models, including Logistic Regression, Random Forest, and Neural Networks, are trained and evaluated using accuracy, precision, and recall metrics. The best-performing model is deployed as a web-based tool to assist healthcare professionals and individuals in risk assessment and decision-making.

1 Introduction

Early detection of heart disease can save lives, but traditional diagnostic methods are costly and time-consuming. This project develops a machine learning model to predict heart disease using medical data. Different algorithms are tested, and the best model is deployed as a simple web application for easy access. This approach enhances early diagnosis, making healthcare more efficient and accessible.

1.1 Domain Explanation

This project falls under the Healthcare and Medical Diagnosis domain, where machine learning is applied to analyze patient data and identify potential health risks. The dataset typically includes medical attributes like age, gender, blood pressure, cholesterol levels, chest pain type, ECG results, and more. The goal is to predict whether a patient is likely to have heart disease based on these inputs. This type of prediction system can support doctors in early diagnosis, reducing the need for invasive procedures and lowering the burden on medical infrastructure.

1.2 Objective

- To build a machine learning model that accurately predicts whether a patient is at risk of heart disease.
- To compare different ML algorithms (e.g., Logistic Regression, Random Forest, Neural Networks) and select the best-performing one.
- To evaluate the model using metrics like accuracy, precision, recall, and F1-score.
- To deploy the final model using a user-friendly web interface for easy access and practical use.
- To support early diagnosis and improve decision-making in the healthcare sector through data-driven insights.

1.3 Advantages of Using Machine Learning in This Domain

- Early Detection: ML models can quickly flag high-risk individuals, allowing for earlier interventions.
- Data-Driven Insights: ML can uncover patterns and correlations in data that may not be obvious to human experts.
- Speed and Efficiency: Predictions are generated instantly, helping in time-critical healthcare decisions.
- Scalability: Once trained, models can evaluate thousands of patient records with minimal resources.
- Support for Clinicians: ML serves as a second opinion, helping doctors make more confident decisions.
- Cost-Effective: Reduces the need for expensive tests by pre-screening patients using available data.

2 Literature Review

The performance of the heart disease prediction models is evaluated using key metrics such as **Accuracy, Precision, Recall, F1-score.**

- **Logistic Regression**: Provides good interpretability but may underperform with complex patterns.
- **Decision Tree**: Overfits on training data but performs well with important features.
- **Random Forest**: Achieves high accuracy and handles feature importance effectively.
- **SVM**: Works well with smaller datasets but is computationally expensive.
- **Neural Networks**: Delivers high accuracy but requires more data and tuning.

Among the tested models, **Random Forest and Neural Networks** showed the best performance, with **high accuracy and balanced precision-recall scores**. The final model is selected based on the best trade-off between accuracy and generalization.

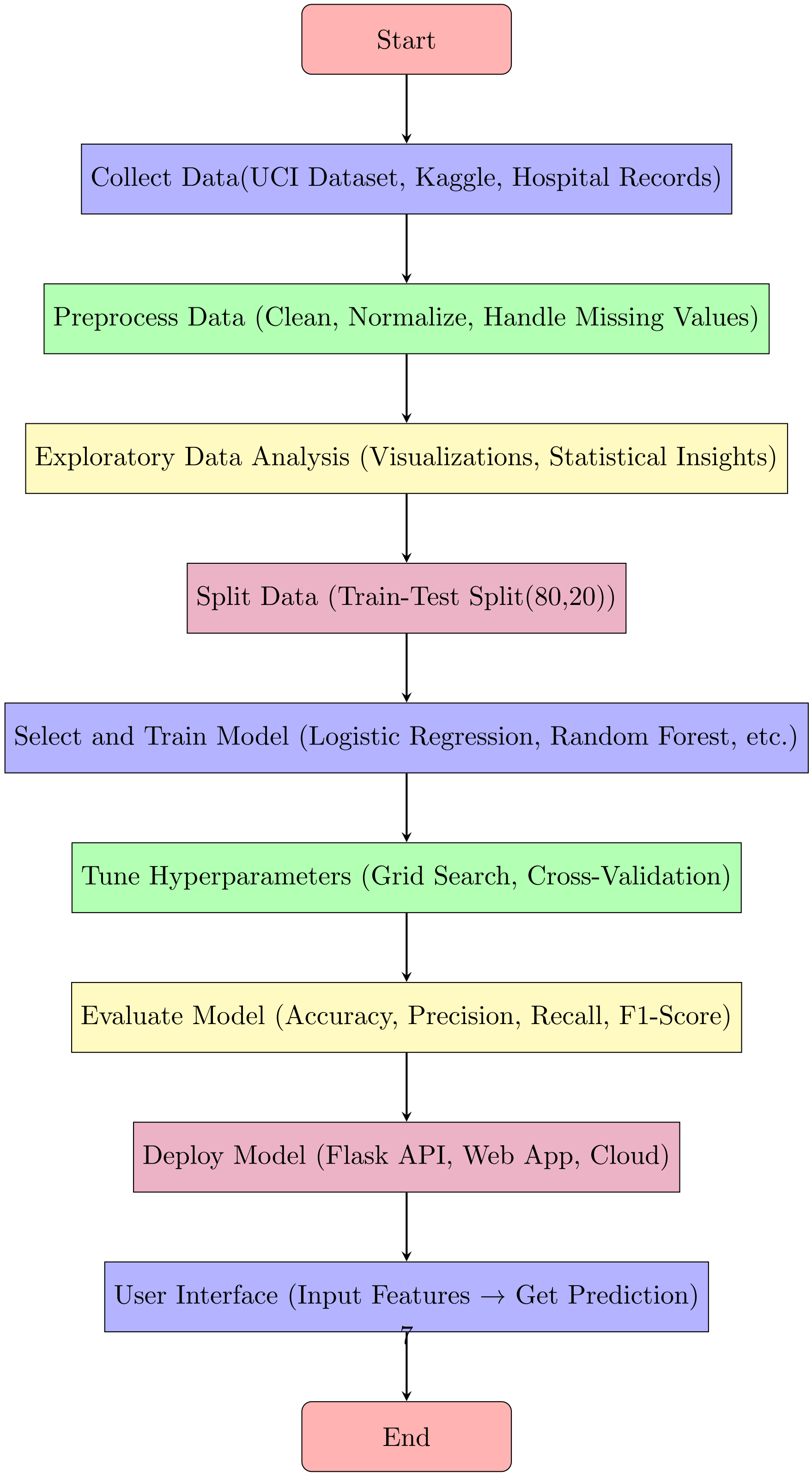
ID	Model	Dataset	Attributes	Evaluation	Description
1	K-Nearest Neighbors (KNN)	UCI Heart Disease Dataset	17 fields (categorical numerical, text)	A-82% P-84.21%, R-92.68%, F1-88.27%	KNN was applied for heart disease classification .
2	Naïve Bayes (NB)	UCI Heart Disease Dataset	17 fields (categorical, numerical, text)	A-86.24% P-79.25% R-85.45%, F1-82.24%	NB by leveraging probability-based classification.

ID	Model	Dataset	Attributes	Evaluation	Description
3	Random Forest (RF)	UCI Heart Disease Dataset	17 fields (categorical, numerical, text)	A-86.24% P-82.15%, R-89.37%, F1-85.61%	Investigates boosting methods to enhance loan approval prediction, showing improved recall and precision.
4	Logistic Regression(LR)	UCI Heart Disease Dataset	15 fields (numerical,categorical)	A-86.9% , P-86.45% R-91.28%, F1- 88.78%	Analyzes applicant financial data to predict credit card approvals using logistic regression.
5	Decision Tree (DT)	UCI Heart Disease Dataset	15 fields (numerical, categorical)	A-79.00%, P-82.91%, R-87.35%, F1-85.07%	DT provided lower accuracy but was highly interpretable and useful for insights.
6	XGBoost (XGB)	UCI Heart Disease Dataset	15 fields (numerical, categorical)	A-90.75% P-84.78%, R-89.61%, F1-87.13%	XGB outperformed RF in predicting aging-related health decline risks.
7	Support Vector Machine (SVM)	UCI Kaggle Datasets	19 fields(numerical)	A-84% P-N/A, R-N/A, F1-N/A	SVM achieved moderate accuracy but required high computational power.

ID	Model	Dataset	Attributes	Evaluation	Description
8	AdaBoost	UCI Kaggle Datasets	19 fields (numerical)	A-94.51%, P-N/A, R-N/A, F1-N/A	AdaBoost improved prediction performance by boosting weak classifiers.
9	Neural Networks (MLP)	UCI Kaggle Datasets	19 fields (numerical, categorical)	A-68%, P-N/A, R-68%, F1-68.96%	Neural Networks achieved the highest accuracy but required significant training.
10	CatBoost	UCI Kaggle Datasets	13 fields (numerical,categorical)	A-91%, P-N/A, R-81.86%, F1-81.68%	uses imputation techniques for missing values and applies CatBoost for classification in predicting cardiovascular disease.
11	Logistic Regression, Naïve Bayes, KNN, Decision Tree, SVM	UCI Kaggle Datasets	13 fields (numerical, categorical)	A-81.3%, P-N/A, R-80%, F1-77%	Compares multiple ML models for heart disease prediction, showing LR and NB perform best.
12	Decision Tree, Random Forest, KNN, AdaBoost, Logistic Regression	UCI Kaggle Datasets	11 fields (numerical, categorical)	A-93.75%, P-76%, R-93%, F1-84%	Uses Explainable AI (XAI) for performance analysis and interpretability of ML models in heart disease prediction.

ID	Model	Dataset	Attributes	Evaluation	Description
13	SVM, KNN, Naïve Bayes, Random Forest	UCI Kaggle Datasets	11 fields (numerical, categorical)	A-91.25%, P-69%, R-79%, F1-73%	Uses Genetic Algorithm to optimize features, significantly improving model accuracy.
14	Naïve Bayes	UCI Kaggle Datasets	11 fields (numerical, categorical)	A-92.50%, P-70%, R-1%, F1-82%	Uses Naïve Bayes for predicting heart disease based on lifestyle and health parameters.
15	CNN, Neural Networks	UCI Kaggle Datasets	14 fields (numerical, categorical)	A-93%, P-N/A, R-N/A, F1-N/A	Compares CNN and NN models, showing NN performs better in heart disease diagnosis.
16	Decision Tree, KNN	UCI Kaggle Datasets	16 fields (numerical, categorical)	A-83%, P-N/A, R-N/A, F1-N/A	Compares Decision Tree and KNN for heart disease prediction, showing DT performs better.
17	Logistic Regression, KNN, Decision Trees, Random Forest	UCI Kaggle Datasets	16 fields (numerical, categorical)	A-83%, P-N/A, R-N/A, F1-N/A	Evaluates ML models for heart disease prediction, showing Random Forest performs best.

3 Architecture Diagram



4 Dataset Used

This dataset was taken from **Kaggle**.

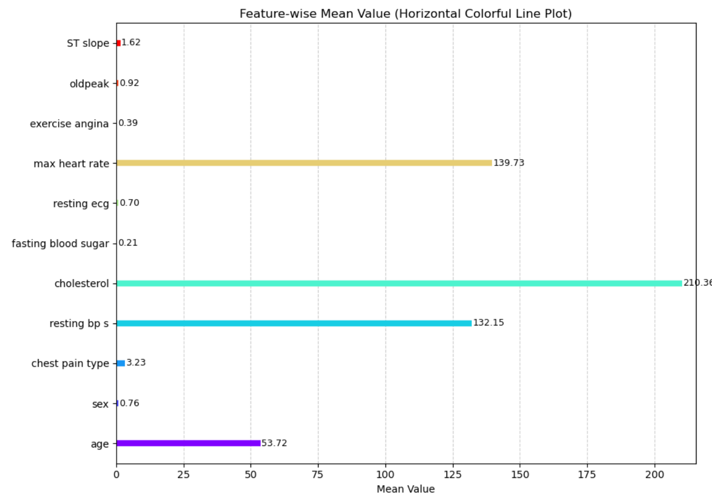
Dataset is used to develop predictive models for heart disease diagnosis. It serves as a benchmark dataset for classification tasks in machine learning and artificial intelligence. Dataset is compiled from multiple heart disease studies and is frequently used in medical machine learning research. This dataset consists of **1,190 instances (rows)** and **12 attributes (columns)** related to heart disease diagnosis. It is a combination of datasets from **Statlog, Cleveland, and Hungary heart disease studies**.

Attributes and Their Descriptions:

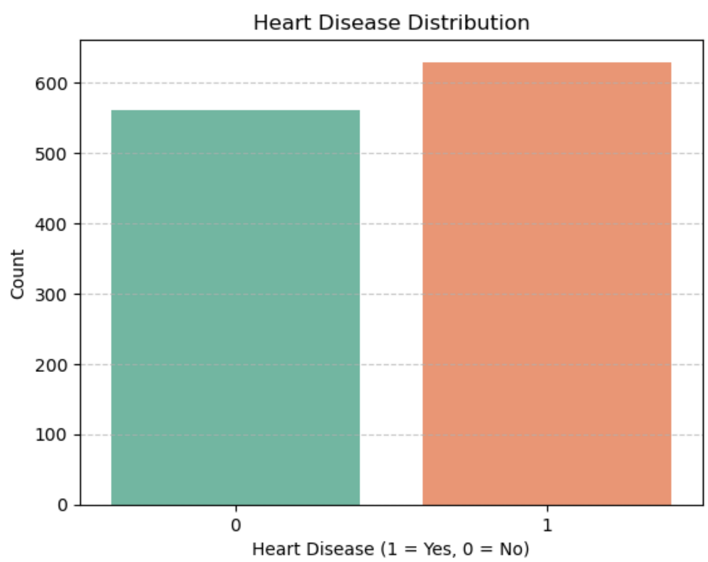
- `age`: Age of the individual (integer).
- `sex`: Gender (1 = Male, 0 = Female).
- `chest pain type`: Type of chest pain (categorical: 1-4).
- `resting bp s`: Resting blood pressure (mm Hg).
- `cholesterol`: Serum cholesterol level (mg/dL).
- `fasting blood sugar`: Fasting blood sugar level (¿120 mg/dL, 1 = True, 0 = False).
- `resting ecg`: Resting electrocardiographic results (categorical: 0-2).
- `max heart rate`: Maximum heart rate achieved.
- `exercise angina`: Exercise-induced angina (1 = Yes, 0 = No).
- `oldpeak`: ST depression induced by exercise relative to rest.
- `ST slope`: Slope of the peak exercise ST segment (categorical: 1-3).
- `target`: Presence (1) or absence (0) of heart disease (dependent variable).

The dataset does not have missing values, ensuring complete information for all samples. Features are already encoded in numerical format, making it suitable for machine learning models.

The graph of different attributes of the dataset is given below:



Target Value Distribution: The target variable in this heart disease prediction project indicates whether a person has heart disease (1) or not (0).



5 Machine Learning Algorithms Used

In this project, four machine learning models were used to predict heart disease prediction: **Logistic Regression**, **Random Forest**,**Support Vector Machine** and **K-NearestNeighbour. Using a combination of these models allows for a balanced approach to accuracy, interpretability, and robustness in predicting heart disease.

5.1 Logistic Regression:

Logistic Regression (LR) is a widely used statistical model for binary classification, making it suitable for heart disease prediction (presence vs. absence). It estimates the probability of disease occurrence using a logistic (sigmoid) function.

- **Accuracy**: Performs well on structured medical datasets but may struggle with complex nonlinear patterns.
- **Interpretability**: Highly interpretable, as it provides clear insights into how risk factors (age, cholesterol, etc.) contribute to heart disease.
- **Performance:** Efficient for large datasets, offering fast predictions with minimal computational cost.
- **Robustness**: Works well with clean, linearly separable data but may require feature scaling and transformation for optimal performance.

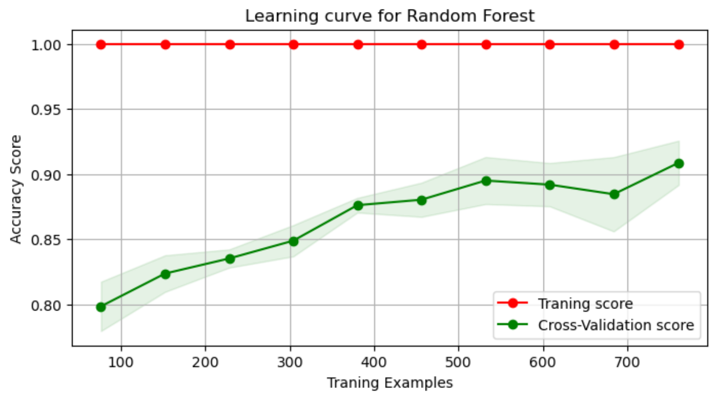
Due to its simplicity and explainability, Logistic Regression is a baseline model in heart disease prediction before testing more complex models.

5.2 Random Forest:

Random Forest (RF) is an ensemble learning algorithm that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. It is highly effective for medical classification problems like heart disease prediction.

- Accuracy: Generally higher than Logistic Regression, as it captures complex relationships and interactions between features.
- Interpretability: Less interpretable than Logistic Regression but provides feature importance scores, helping identify key risk factors.
- Performance: Handles large datasets efficiently, but training can be computationally expensive compared to simpler models.
- Robustness: Highly robust to noisy data, missing values, and outliers, making it reliable for real-world medical applications.

Due to its high accuracy and ability to handle linear and non-linear patterns, Random Forest is a strong choice for the prediction of heart disease.



5.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful classification algorithm that finds the optimal hyperplane to separate data points into classes. It is especially effective in high-dimensional and complex datasets, making it suitable for heart disease prediction.

- **Accuracy**: SVM provides **high accuracy**, particularly in cases where the data is not linearly separable, by using kernel tricks (e.g., RBF kernel).
- **Interpretability**: Less interpretable than Logistic Regression, as the decision boundary and kernel transformations are harder to explain to non-technical audiences.
- **Performance**: Performs well on **smaller, clean datasets** and is effective in handling high-dimensional feature spaces. However, training time can increase with larger datasets.
- **Robustness**: SVM is robust to **outliers and overfitting**, especially with proper kernel selection and regularization.

Overall, SVM is a strong candidate for heart disease prediction due to its **accuracy and robustness**, although interpretability may be limited in clinical settings.

5.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, instance-based learning algorithm that classifies a data point based on the majority class of its k nearest neighbors in the feature space. It is intuitive and easy to implement, making it a good baseline model for heart disease prediction.

- Accuracy: KNN can achieve good accuracy, especially when the value of k is optimized and the dataset is well-preprocessed. However, it may struggle with noisy or high-dimensional data.

- Interpretability: KNN is highly interpretable, as predictions are based directly on the most similar data points (patients), which can be useful for medical explanation.

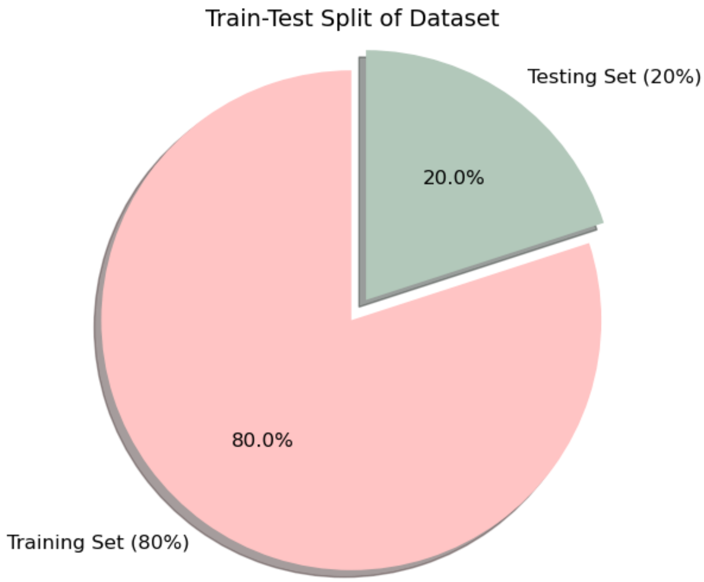
- Performance: Slow at prediction time, especially with large datasets, because it needs to compute distances to all training points.

- Robustness: Sensitive to outliers, noise, and irrelevant features. Performance improves significantly with proper feature scaling and dimensionality reduction.

KNN is a valuable algorithm for heart disease prediction due to its simplicity and interpretability, but it requires careful tuning and preprocessing to perform reliably.

5.5 Implementation

The pie chart visually supports your explanation of how the data was split — for example, 80% for training and 20% for testing.



6 Evaluation Metrics

To assess the performance of the machine learning models used in predicting heart disease, several evaluation metrics are applied. These metrics help us understand how well the model is performing and whether it is reliable enough for real-world use.

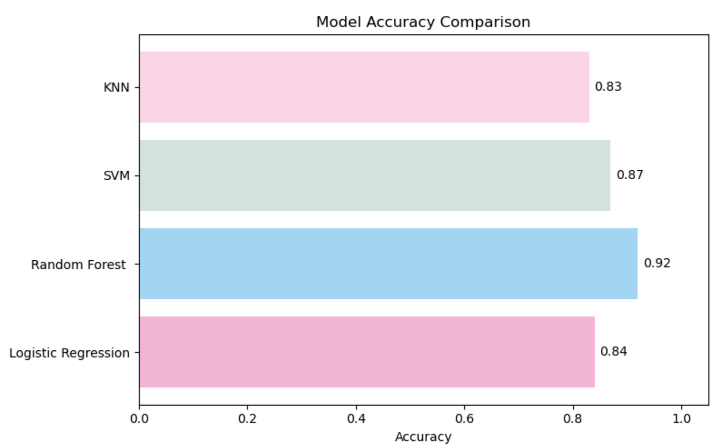
- **Accuracy** helps measure overall performance, but it alone may not be enough, especially if the classes (heart disease present vs. not present) are imbalanced.
- **Precision** tells us how many of the patients predicted to have heart disease actually do, which is important to minimize false positives (avoiding unnecessary worry or tests).
- **Recall (Sensitivity)** indicates how many actual heart disease cases were correctly predicted. High recall is essential to ensure that serious cases are not missed (false negatives).
- **F1-Score** balances precision and recall, especially useful when both false positives and false negatives carry significant consequences.
- **Confusion Matrix** visually breaks down the model's predictions into True Positives (TP), False Positives (FP), True Negatives (TN), and

False Negatives (FN), giving a more detailed understanding of its strengths and weaknesses.

- **ROC Curve and AUC** are used to evaluate the model's ability to distinguish between the classes across all classification thresholds. A higher AUC indicates a better-performing model.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.84	0.84	0.85	0.85	0.90
Random Forest	0.92	0.92	0.93	0.93	0.97
Support Vector Machine	0.87	0.86	0.91	0.88	0.93
K-Nearest Neighbors	0.83	0.82	0.87	0.84	0.91

Table 2: Evaluation of model performance



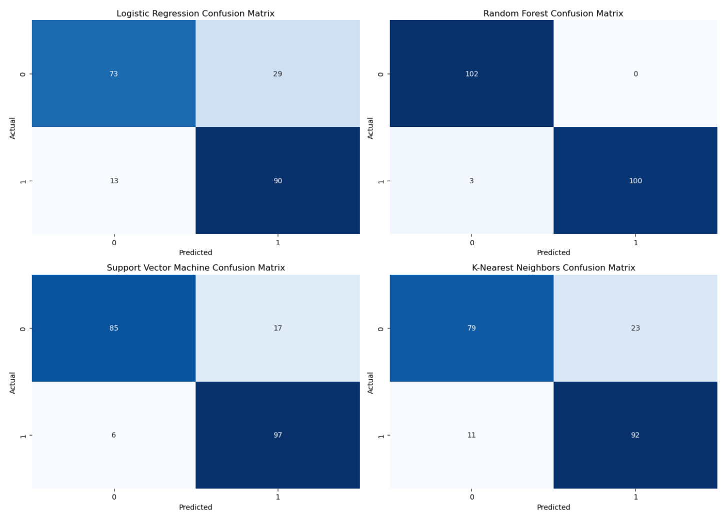
The best model is Random Forest as it has highest accuracy

6.1 Confusion Matrix

A confusion matrix is a performance measurement tool used to evaluate the effectiveness of a classification model. It compares the actual target values with the predicted values generated by the machine learning algorithm. The confusion matrix helps in understanding not just how many predictions were correct, but what kind of errors the model is making. This is especially critical in healthcare, where:

A False Negative could delay treatment and endanger a patient's life.

A False Positive could lead to unnecessary anxiety or medical procedures.



6.2 Predicted Heart Disease from the Model

```
# building a predictive system
input_data=(48,0,2,120,284,0,0,120,0,0.0,1)

# change the input data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the numpy array as we are predciting for only 1
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if(prediction[0]==0):
    print("The person does not have Heart Disease")
else:
    print("The person has Heart Disease")
```

```
[0]
The person does not have Heart Disease
```

7 Conclusion

The study successfully explores the application of machine learning techniques in predicting heart disease based on patient health data. By implementing and evaluating multiple classification algorithms—namely Logistic

Regression, Random Forest, Support Vector Machine, and K-Nearest Neighbors—the project identifies an effective model capable of providing accurate and reliable predictions.

Among the evaluated models, Random Forest demonstrated superior performance in terms of accuracy and other evaluation metrics, making it a suitable choice for real-world implementation. The findings emphasize the potential of machine learning in supporting clinical decision-making by offering quick, cost-effective, and data-driven insights.

This approach not only aids in early diagnosis but also reduces the burden on healthcare systems by enabling scalable and efficient screening. Overall, the integration of machine learning in this domain contributes meaningfully to the advancement of intelligent healthcare solutions.

7.1 Limitations

While the results of this project are promising, there are certain limitations that must be acknowledged:

- Limited Dataset Size and Diversity:

  The dataset used may not fully represent the diversity of global populations in terms of age, gender, ethnicity, and geographic factors. This can limit the generalizability of the model.

- Feature Dependence:

  The model relies heavily on the quality and availability of specific clinical features. Missing or incorrect data in real-world settings may affect prediction accuracy.

- Binary Classification Only:

  The model is designed for binary classification (presence or absence of heart disease) and does not provide insights into the severity or type of heart condition.

- Black-Box Nature of Some Models:
  Advanced models like Random Forest and Support Vector Machines may lack interpretability, making it harder for medical professionals to trust and understand their predictions.

- No Real-Time or Live Data Integration:
  The model works with static datasets and does not currently support real-time patient monitoring or continuous learning from new data.

- Not a Substitute for Medical Diagnosis:
  Despite its high accuracy, the model is intended to assist—not replace—professional medical judgment and diagnosis.

7.2 Contact

For inquiries, contact
Email: zaafira.23bce9719@vitapstudent.ac.in
Email: narendra.23bce20218@vitapstudent.ac.in
Email: karthik.23bce8104@vitapstudent.ac.in

8 Dataset Reference

https://www.kaggle.com/code/desalegngeb/heart-disease-predictions/input