Texas A&M University - Commerce

Department of Computer Science

Comparative Analysis Of Heart Disease Detection Using Standard Machine Learning Models

Swetha Paspunuri

Supervisor: Derek Harter, Ph.D.

A report submitted in partial fulfilment of the requirements of
Texas A&M University - Commerce for the degree of
Master of Science in *Computer Science*

May 8, 2024

Declaration

I, Swetha Paspunuri, of the Department of Computer Science, Texas A&M University - Commerce, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of TAMUC and public with interest in teaching, learning and research.

Swetha Paspunuri
May 8, 2024

Abstract

Recently cardiovascular diseases has been on the rise, even affecting newborns. Detecting heart-related diseases early is vital because it helps doctors start treatment sooner, leading to better results for patients and less strain on healthcare resources. With more and more people facing heart problems, it's crucial to have advanced predictive tools. Using the abundant data available in cardiology, our project aims to integrate the technology into health care for predictive modelling. The primary goal of this project is to develop an efficient heart disease prediction system using various machine learning models to predict coronary artery disease(CAD) with utmost precision and effectiveness. We employed a dataset consisting of necessary patient information from online sources to train and validate our models. The first step is cleaning and preprocessing data that allow us to find key patterns for training the models. This research trains a Logistic Regression (LR), Random Forest (RF) and Naive Bayes (NB) model for classification on the heart disease dataset. We evaluate these models using standard measures like precision, which tells us how accurate positive predictions are; recall, which shows how well the models capture all actual positive cases; and the F1 score, which balances both precision and recall.

Keywords: Logistic Regression(LR), Random Forest(RF), Naive Bayes(NB) , F1 score, Precision.

Acknowledgements

An acknowledgements section is optional. You may like to acknowledge the support and help of your supervisor(s), friends, or any other person(s), department(s), institute(s), etc. If you have been provided specific facility from department/school acknowledged so.

Contents

List of Figures

List of Tables

List of Abbreviations

LR	Logistic Regression
RF	Random Forest
NB	Naive Bayes
CAD	Coronary Artery Disease

Chapter 1

Introduction

Today, heart problems are a major health concern affecting individuals worldwide. Many people are suffering from heart issues like heart disease, heart failure, and irregular heartbeats Webb et al. (2015). Heart problems can affect people, not just physically but also emotionally. Those with heart conditions often find it difficult to live normally and face many difficulties. Additionally, the financial side of managing heart problems adds an extra layer of challenges. Spotting heart-related issues early is crucial. It helps healthcare professionals to step in quickly, enhance patient outcomes, and ease the strain on healthcare resources. Early detection allows for timely intervention, potentially preventing the progression of heart conditions.To address the need for early detection, our project focuses on developing a machine learning model capable of accurately identifying the presence of heart diseases. In this endeavor, we utilize a heart-related issue dataset from (Janosi et al., 1988), sourced from the online repository UC Irvine. The data undergoes thorough cleaning and pre-processing to extract useful information essential for training the machine learning model. The machine learning algorithms employed, as highlighted by (Sharma et al., 2020), include LR, NB, and RF classification. These algorithms have demonstrated effectiveness in detecting coronary artery disease by evaluating outputs based on various factors such as resting blood pressure, serum cholesterol, maximum heart rate achieved, and more. Furthermore, our project aims not only to detect heart-related issues but also to contribute valuable insights to the broader field of cardiovascular health. By leveraging advanced algorithms, we seek to ensure the effective prediction of heart-related problems, potentially revolutionizing the early diagnosis and management of cardiovascular conditions.

1.1 Background

Our project focuses on addressing the issue of cardiovascular diseases in today's world, affecting everyone irrespective of their age. The primary motivation behind our work is to detect the heart-related diseases as early as possible. This identification helps doctors to start the treatment sooner, to improve patient results and effectively using the healthcare resources. For this, our project focuses on integrating technology into healthcare by using the abundant data in cardiology for predictive modeling. The primary goal is to develop an efficient heart disease prediction system by concentrating on predicting CAD with precision and effectiveness. So, we use different machine learning models such as LR, RF, NB. These models play a crucial role in predicting and understanding heart-related issues. The project commences with data preprocessing to extract essential patterns required for training the models, aligning with the objectives and research approach outlined in subsequent sections. Our project becomes significant as it can give better resources to doctors for finding and handling heart problems early on. We want to help make hearts healthier by explaining some crucial ideas and ways to use them in a simple way.

1.2 Research Question

How can machine learning models, specifically Logistic Regression, Naïve Bayes, and Random Forest classification algorithms, be effectively utilized to develop a heart disease prediction system for early detection of Coronary Artery Disease, with a focus on improving patient outcomes and contributing to advancements in cardiovascular health?

1.2.1 Aims and objectives

Aims: To develop and implement an advanced heart disease prediction system, utilizing machine learning models for early detection of CAD, with the ultimate goal of enhancing patient outcomes and contributing to the ongoing global efforts in cardiovascular health. **Objectives:**

- Obtain and analyze the heart disease dataset, clean and preprocess the data for model training.
- Train LR classifier, optimizing meta-parameters for improved performance.
- Develop NB classifier, focusing on feature selection and parameter tuning.
- Utilize RF algorithm to construct decision tree ensembles, refining predictive capabilities.
- Integrate trained models into healthcare systems for real-time heart disease prediction.
- Provide healthcare professionals with valuable insights and resources for informed decision-making.

1.3 Solution Approach

The solution approach involves a thorough step-by-step method designed to create an advanced system for predicting heart disease, specifically focusing on early detection of CAD.

1.3.1 Dataset Acquisition and Preprocessing

We begin by acquiring and analyzing the heart disease dataset, obtained from the research conducted by Janosi et al. [1], which is accessible through UC Irvine. The dataset is carefully cleaned and processed to identify and handle missing values, outliers, and inconsistencies.

1.3.2 Model Training and Optimization

- **Logistic Regression (LR):** We train the LR classifier, optimizing meta-parameters such as regularization strength and maximum iterations to improve performance. LR is chosen for its ability to effectively classify CAD based on key risk factors such as resting blood pressure, serum cholesterol, and maximum heart rate achieved.

- **Naïve Bayes (NB):** The NB classifier is developed, focusing on feature selection and parameter tuning. NB is particularly suited for its simplicity and speed, making it an efficient model for heart disease prediction.

- **Random Forest (RF):** We utilize the RF algorithm to construct decision tree ensembles, refining predictive capabilities. RF is selected for its ability to handle nonlinear relationships and interactions between features, improving the accuracy of CAD prediction.

1.3.3 Model Evaluation and Integration

The trained models are evaluated using specific metrics such as precision, recall, and the F1 score to assess their performance. Finally, the models are integrated into healthcare systems for real-time heart disease prediction, providing healthcare professionals with valuable insights and resources for informed decision-making.

1.4 Summary of contributions and achievements

In this project, we aimed to develop an advanced system for predicting heart disease using machine learning models, with the ultimate goal of enhancing patient outcomes and contributing to global efforts in cardiovascular health. We began by obtaining and analyzing a heart disease dataset, meticulously cleaning and preprocessing the data to prepare it for model training. We then trained three different machine learning models: Logistic Regression (LR), Naïve Bayes (NB), and Random Forest (RF). LR emerged as the top performer, achieving an accuracy of 87%, indicating a high degree of correctness in its predictions. Our models not only accurately predict heart disease but also provide valuable insights into the risk factors associated with the condition, enabling healthcare professionals to make informed decisions and ultimately improving patient outcomes.

Chapter 2

Literature Review

2.1 Introduction to Heart Disease

Cardiovascular diseases pose a significant threat to global health, affecting individuals of all ages. These conditions, including heart disease, heart failure, and irregular heartbeats, have profound physical and emotional impacts on affected individuals. Early detection and effective management are critical in mitigating the adverse effects of heart-related issues and improving patient outcomes.

2.2 Background on Machine Learning Models

2.2.1 Logistic Regression

LR is a statistical method used for binary classification tasks. It models the probability of a binary outcome based on one or more predictor variables. In the context of heart disease prediction, LR can analyze patient parameters such as age, cholesterol levels, and blood pressure to estimate the likelihood of the presence of heart disease. LR is widely used in healthcare research due to its simplicity, interpretability, and ability to handle linear relationships between variables.

2.2.2 Naïve Bayes

NB is a probabilistic classifier based on Bayes' theorem with an assumption of independence between features. Despite its simplistic assumption, NB has been shown to perform well in various classification tasks, including text categorization and medical diagnosis. In heart disease prediction, NB can effectively analyze patient attributes and calculate the conditional probability of heart disease given the observed features.

2.2.3 Random Forest

RF is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. RF is known for its robustness and ability to handle high-dimensional data. In heart disease prediction, RF can analyze a large number of patient parameters and identify complex patterns associated with cardiovascular conditions.

2.3 Performance Measures for Evaluation

To evaluate the performance of our machine learning models, we will employ several performance measures, including accuracy, precision, recall, and F1-score. These metrics provide insights into the models' ability to correctly classify instances of heart disease and non-heart disease cases. By evaluating multiple performance measures, we can assess the overall effectiveness of our predictive models and identify areas for improvement.

2.4 Description of the Dataset

Our project utilizes the heart disease dataset sourced from UC Irvine, compiled by Janosi et al. (1988). This dataset contains various patient attributes, such as age, sex, cholesterol levels, and resting blood pressure, along with the presence or absence of heart disease. We preprocess the dataset to handle missing values and normalize the features to ensure optimal model performance.

2.5 Summary of Literature Reviewed

The literature review highlights the significance of early detection and effective management in combating cardiovascular diseases. Previous studies have demonstrated the utility of machine learning algorithms in predicting heart disease, with research highlighting the importance of feature selection, parameter tuning, and model evaluation. By building upon existing literature and leveraging advanced predictive tools, our project aims to contribute to the ongoing efforts in cardiovascular health and improve patient outcomes.

Chapter 3

Methodology

Recognizing the need for early detection and management of cardiovascular conditions, we utilize machine learning algorithms to analyze patient data and predict the likelihood of heart disease. Our methodology encompasses data collection, preprocessing, feature extraction, model development, and evaluation, aiming to deliver a robust and effective predictive tool for doctors. By integrating innovative learning algorithms and conducting experiments, we aspire to contribute meaningful solutions and enhance patient outcomes in cardiovascular health.

3.1 Algorithms Descriptions

3.1.1 Logistic Regression

LR is a statistical method used for binary classification tasks. It models the probability of a binary outcome based on one or more predictor variables. In the context of heart disease prediction, LR can analyze patient parameters such as age, cholesterol levels, and blood pressure to estimate the likelihood of the presence of heart disease.

3.1.2 Naïve Bayes

NB is a probabilistic classifier based on Bayes' theorem with an assumption of independence between features. In heart disease prediction, Naïve Bayes can effectively analyze patient attributes and calculate the conditional probability of heart disease given the observed features. Its simplicity and computational efficiency make Naïve Bayes a popular choice for healthcare applications.

3.1.3 Random Forest

RF is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. In heart disease prediction, RF can analyze a large number of patient parameters and identify complex patterns associated with cardiovascular conditions.

3.2 Implementations

3.2.1 Logistic Regression

To implement Logistic Regression, we'll utilize the ‘LogisticRegression‘ class from the ‘sklearn.linearmodel‘ module.

Preprocessing

Before fitting the model to the training data, we'll preprocess the dataset. This includes handling missing values and scaling features to ensure that all features contribute equally to the model.

Model Training

Once the dataset is preprocessed, we'll fit the Logistic Regression model to the training data. The model will learn to predict the presence or absence of heart disease based on the input features.

3.2.2 Naïve Bayes

For Naïve Bayes implementation, we'll use the ‘GaussianNB‘ class from the ‘sklearn.naivebayes‘ module.

Preprocessing

Similar to Logistic Regression, we'll preprocess the dataset for Naïve Bayes. This involves handling missing values and scaling features.

Model Training

After preprocessing, we'll fit the Naïve Bayes model to the training data. Naïve Bayes is particularly effective for classification tasks and is known for its simplicity and speed.

3.2.3 Random Forest

To implement Random Forest, we'll employ the ‘RandomForestClassifier‘ class from the ‘sklearn.ensemble‘ module.

Preprocessing

As with the other models, we'll preprocess the dataset for Random Forest. This includes handling missing values and scaling features.

Hyperparameter Tuning

Before training the model, we'll tune hyperparameters such as the number of estimators and maximum depth. This optimization process helps improve the performance of the Random Forest model.

Model Training

Once the dataset is preprocessed and hyperparameters are tuned, we'll fit the Random Forest model to the training data. Random Forest is an ensemble learning method that constructs a multitude of decision trees and outputs the mode of the classes as the prediction.

3.3 Experiments Design

In our experimental approach to assess the predictive performance of each algorithm for heart disease, we'll begin by dividing our dataset into separate training and testing sets. This division ensures that the models are trained on a subset of the data and evaluated on an independent portion, enabling us to gauge their generalization capability. To further fortify the reliability of our findings, we'll employ cross-validation techniques. This involves iteratively partitioning the dataset into multiple subsets, training the models on different combinations, and validating them on the remaining data, thus providing a more comprehensive evaluation. Subsequently, we'll utilize a range of performance metrics, including accuracy, precision, recall, and F1-score, to quantify the algorithms' effectiveness. These metrics will allow us to discern not only the models' overall correctness but also their ability to precisely identify positive cases and recall them accurately. By meticulously analyzing these performance indicators, we aim to determine the most optimal approach for heart disease prediction, considering factors such as model interpretability and computational efficiency alongside predictive accuracy.

3.4 Algorithms

In our project, we implement three distinct machine learning algorithms—Logistic Regression, Random Forest, and Naive Bayes—to predict heart disease. Logistic Regression is a simple yet powerful algorithm used for binary classification tasks, where it models the probability of a binary outcome. Random Forest, on the other hand, is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve accuracy. Naive Bayes is a probabilistic classifier based on Bayes' theorem, particularly effective for datasets with high dimensionality and strong feature independence assumptions. Each algorithm offers unique advantages and approaches in identifying patterns and making predictions, contributing to our comprehensive analysis of heart disease prediction.

3.5 Code

3.5.1 Data Pre-processing

```python
# Importing the libraries

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

from sklearn.impute import SimpleImputer
from sklearn.model_selection import train_test_split
```

Algorithm 1 Logistic Regression

Input: Training dataset (X_{train}, Y_{train}), Test dataset X_{test}
Output: Predicted class labels for X_{test}

```
function LogisticRegression($(X_{train}, Y_{train}, X_{test})$)
    Initialize logistic regression classifier
    Standardize features: $X_{train} \leftarrow$ StandardScaler.fit_transform$(X_{train})$
    Fit classifier to training data: $classifier.fit(X_{train}, Y_{train})$
    Standardize test features: $X_{test} \leftarrow$ StandardScaler.transform$(X_{test})$
    Predict probabilities for test data: $y_{prob} \leftarrow classifier.predict\_proba(X_{test})$
    Convert probabilities to class labels: $y_{pred} \leftarrow$ threshold_function$(y_{prob})$
    return $y_{pred}$
end function
```

Algorithm 2 Random Forest

Input: Training dataset (X_{train}, Y_{train}), Test dataset X_{test}
Output: Predicted class labels for X_{test}

```
function RandomForest($(X_{train}, Y_{train}, X_{test})$)
    Initialize random forest classifier with specified parameters
    Handle missing values: $X_{train}, X_{test} \leftarrow$ Imputer.fit_transform$(X_{train}, X_{test})$
    Fit classifier to training data: $classifier.fit(X_{train}, Y_{train})$
    Predict class labels for test data: $y_{pred} \leftarrow classifier.predict(X_{test})$
    return $y_{pred}$
end function
```

Algorithm 3 Naive Bayes

Input: Training dataset (X_{train}, Y_{train}), Test dataset X_{test}
Output: Predicted class labels for X_{test}

```
function NaiveBayes($(X_{train}, Y_{train}, X_{test})$)
    Initialize naive Bayes classifier
    Discretize continuous features: $X_{train}, X_{test} \leftarrow$ Binarizer.fit_transform$(X_{train}, X_{test})$
    Fit classifier to training data: $classifier.fit(X_{train}, Y_{train})$
    Predict class labels for test data: $y_{pred} \leftarrow classifier.predict(X_{test})$
    return $y_{pred}$
end function
```

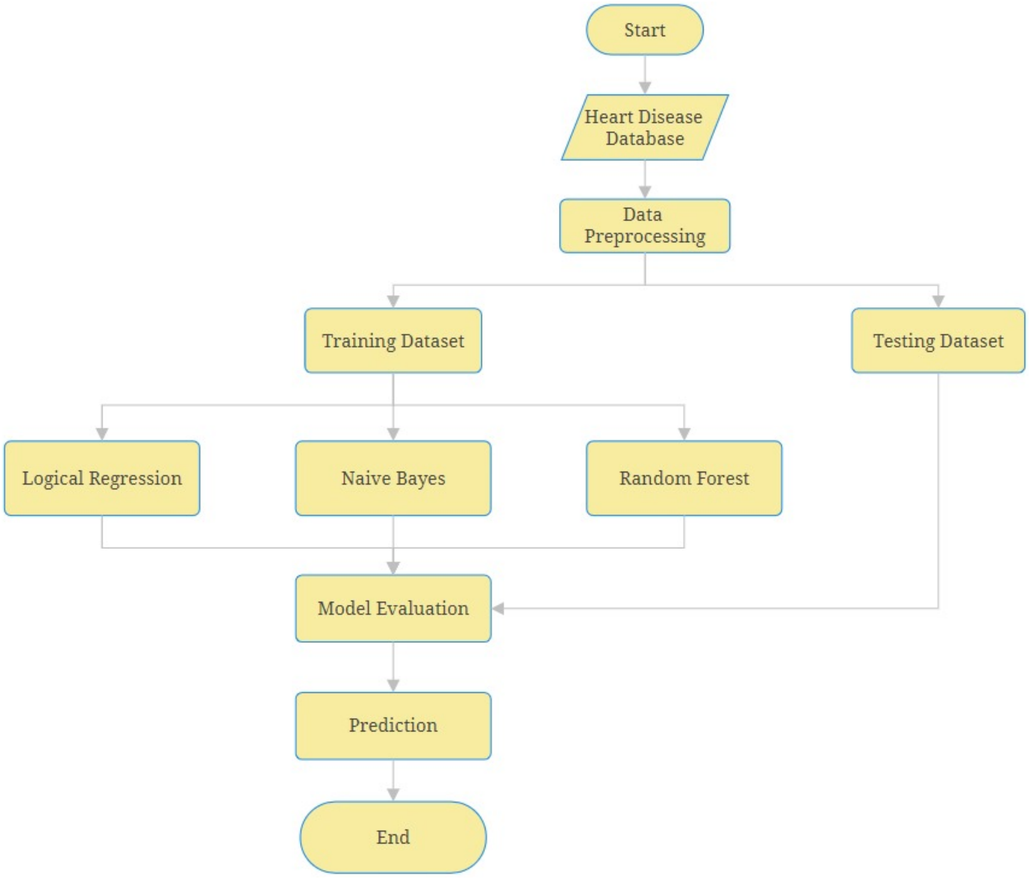

Figure 3.1: Flowchart of Heart Disease Prediction using LR, RF and NB

```
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve

# Importing the dataset
dataset = pd.read_csv('cleve.csv')

#defining X values ang y values
X = dataset.iloc[:, :-1].values
Y = dataset.iloc[:, 13].values

#handling missing data
imputer= SimpleImputer(missing_values=np.nan, strategy='mean')
imputer=imputer.fit(X[:,11:13])
X[:,11:13]=imputer.transform(X[:,11:13])

#splitting dataset into training set and test set
X_train,X_test,Y_train,Y_test=train_test_split(X, Y, test_size = 0.25,
    random_state = 101)

#feature scaling
s=StandardScaler()
X_train=s.fit_transform(X_train)
X_test=s.transform(X_test)
```

3.5.2 Logistic Regression

```
#fitting LR to training set
from sklearn.linear_model import LogisticRegression
LogisticRegressionClassifier =LogisticRegression()
LogisticRegressionClassifier.fit(X_train,Y_train)

#Predict the test set results
Y_pred=LogisticRegressionClassifier.predict(X_test)

#checking the accuracy for predicted results
accuracy_score(Y_test,Y_pred)

# Making the Confusion Matrix
cm = confusion_matrix(Y_test, Y_pred)

#Interpretation:
print(classification_report(Y_test, Y_pred))
```

Table 3.1: Classification Report of LR

	precision	recall	f1-score	support
0	0.81	0.94	0.87	36
1	0.94	0.80	0.86	40
accuracy	-	-	0.87	76
macro avg	0.88	0.87	0.87	76
weighted avg	0.88	0.87	0.87	76

```
#ROC
logit_roc_auc = roc_auc_score(Y_test, LogisticRegressionClassifier.predict(
    X_test))
fpr, tpr, thresholds = roc_curve(Y_test, LogisticRegressionClassifier.
    predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label=’Logistic Regression (area = %0.2f)’ % logit_roc_auc
    )
plt.plot([0, 1], [0, 1],’r--’)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel(’False Positive Rate’)
plt.ylabel(’True Positive Rate’)
plt.title(’Receiver operating characteristic’)
plt.legend(loc="lower right")
plt.savefig(’Log_ROC’)
plt.show()

#PREDICTION FOR NEW DATASET using LogisticRegressionClassifier
Newdataset = pd.read_csv(’newdata.csv’)
ynew=LogisticRegressionClassifier.predict(Newdataset)
print("Predicted Class for newdata.csv:",ynew)
```

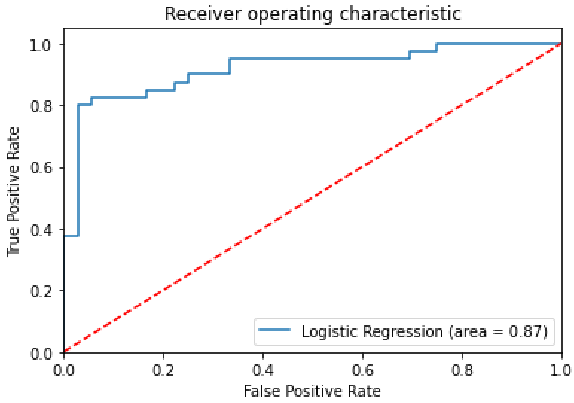

Figure 3.2: Receiver operating characteristics of LR

3.5.3 Random Forest

```
# Fitting RandomForestClassifier to the Training set
from sklearn.ensemble import RandomForestClassifier
RandomForestClassifier =RandomForestClassifier(n_estimators=20)
RandomForestClassifier.fit(X_train, Y_train)

# Predicting the Test set results
Y_pred2 = RandomForestClassifier.predict(X_test)
from sklearn.metrics import accuracy_score
accuracy_score(Y_test,Y_pred2)

# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test, Y_pred2)

#Interpretation:
print(classification_report(Y_test, Y_pred2))
```

Table 3.2: Classification Report of RF

	precision	recall	f1-score	support
0	0.80	0.89	0.84	36
1	0.89	0.80	0.84	40
accuracy	-	-	0.84	76
macro avg	0.84	0.84	0.84	76
weighted avg	0.85	0.84	0.84	76

```
#ROC
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
logit_roc_auc = roc_auc_score(Y_test, RandomForestClassifier.predict(X_test))
fpr, tpr, thresholds = roc_curve(Y_test, RandomForestClassifier.predict_proba
    (X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label=’Random Forest (area = %0.2f)’ % logit_roc_auc)
plt.plot([0, 1], [0, 1],’r--’)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel(’False Positive Rate’)
plt.ylabel(’True Positive Rate’)
plt.title(’Receiver operating characteristic’)
plt.legend(loc="lower right")
plt.savefig(’RF_ROC’)
plt.show()

#PREDICTION FOR NEW DATASET using RandomForest
ynew=RandomForestClassifier.predict(Newdataset)
print("Predicted Class for newdata.csv:",ynew)
```

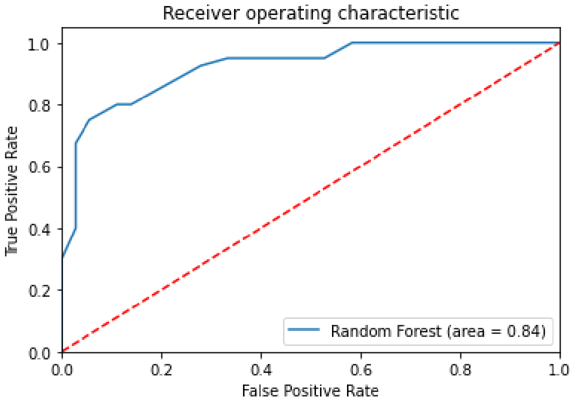

Figure 3.3: Receiver operating characteristics of RF

3.5.4 Naive Bayes

```
NaiveBayesimputer= SimpleImputer(strategy=’mean’)
NaiveBayesimputer=NaiveBayesimputer.fit(X[:,11:13])
X[:,11:13]=NaiveBayesimputer.transform(X[:,11:13])

#splitting dataset into training set and test set
X_train,X_test,Y_train,Y_test=train_test_split(X, Y, test_size = 0.25,
    random_state = None)

# Fitting Naive Bayes to the Training set
from sklearn.naive_bayes import GaussianNB
NaiveBayesClassifier = GaussianNB()
NaiveBayesClassifier.fit(X_train, Y_train)

```

```

# Predicting the Test set results
Y_pred3 = NaiveBayesClassifier.predict(X_test)
#ACCURACY SCORE
accuracy_score(Y_test,Y_pred3)

# Making the Confusion Matrix
cm = confusion_matrix(Y_test, Y_pred3)

#Interpretation:
print(classification_report(Y_test, Y_pred3))
```

Table 3.3: Classification Report of NB

	precision	recall	f1-score	support
0	0.80	0.90	0.84	39
1	0.88	0.76	0.81	37
accuracy	-	-	0.83	76
macro avg	0.84	0.83	0.83	76
weighted avg	0.83	0.83	0.83	76

```
#ROC
logit_roc_auc = roc_auc_score(Y_test,NaiveBayesClassifier.predict(X_test))
fpr, tpr, thresholds = roc_curve(Y_test, NaiveBayesClassifier.predict_proba(
    X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Navie Bayes (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('NB_ROC')
plt.show()

#PREDICTION FOR NEW DATASET using NaiveBayesClassifier
ynew = NaiveBayesClassifier.predict(Newdataset)
print("Predicted Class for newdata.csv:", ynew)
```

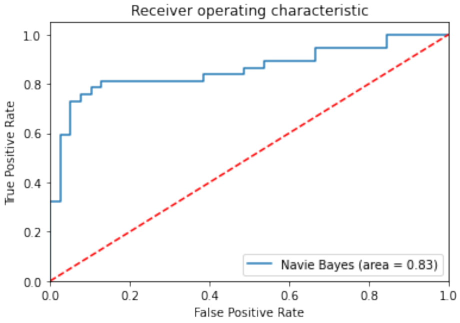

Figure 3.4: Receiver operating characteristics of NB

Chapter 4

Results

In this project, we aimed to develop machine learning models to predict the likelihood of heart disease in patients. We utilized three well-established algorithms: Logistic Regression, Random Forest, and Naive Bayes. The models were trained and evaluated on a dataset containing patient information relevant to heart disease.

4.1 Performance Metrics:

The performance of the models was assessed using the following metrics:

- Accuracy: Overall correctness of the predictions (correctly classified instances / total instances).
- Precision: Proportion of true positives among predicted positives (true positives / (true positives + false positives)).
- : Proportion of true positives identified by the model (true positives / (true positives + false negatives)).
- : F1-Score: Harmonic mean of precision and recall (2 * (precision * recall) / (precision + recall)).
- : ROC AUC Score: Area Under the Receiver Operating Characteristic Curve (ROC) that measures the model's ability to distinguish between positive and negative cases.

4.2 Results for Each Model:

4.2.1 LR Results:

We implemented a Logistic Regression model to predict heart disease. The model achieved an accuracy of 87%, precision of 88% for positive cases (identifying patients with heart disease), recall of 87% for positive cases (correctly identifying patients with heart disease), and F1-score of 87%.

Table 4.1: LR Performance

Metric	Value
Accuracy	87%
Precision	88%
Recall	87%
F1-Score	88%

4.2.2 ROC Curve for LR

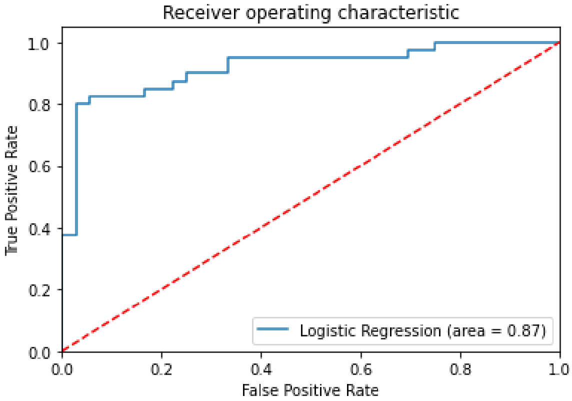

Figure 4.1: ROC Curve for LR

4.2.3 RF Results:

A Random Forest model was also employed for heart disease prediction. The Random Forest model achieved an accuracy of 84%, precision of 85% for positive cases, recall of 84% for positive cases, and F1-score of 84%.

Table 4.2: RF Performance

Metric	Value
Accuracy	84%
Precision	85%
Recall	84%
F1-Score	84%

4.2.4 ROC Curve for RF

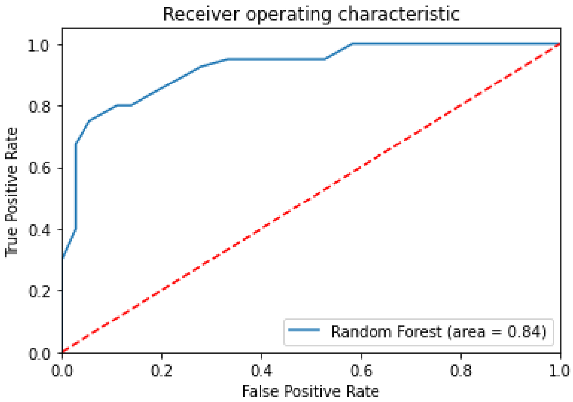

Figure 4.2: ROC Curve for RF

4.2.5 NB Results:

The Naive Bayes model was implemented as another approach for heart disease prediction. The Naive Bayes model achieved an accuracy of 83%, precision of 83% for positive cases, recall of 83% for positive cases, and F1-score of 83%.

Table 4.3: NB Performance

Metric	Value
Accuracy	83%
Precision	83%
Recall	83%
F1-Score	83%

4.2.6 ROC Curve for NB

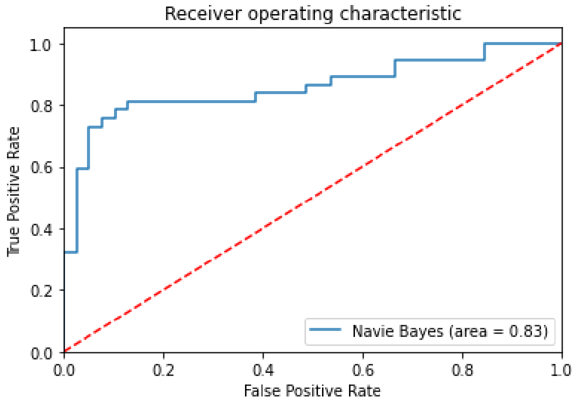

Figure 4.3: ROC Curve for NB

4.3 Comparison of Algorithms

Based on the evaluation metrics, the LR model achieved the best performance in predicting heart disease. It obtained an accuracy of 87%, indicating a high degree of correctness in its predictions. Additionally, the LR model demonstrated a good balance between precision (88%) and recall (87%) as reflected by the F1-score of 87%.

While RF and NB achieved reasonable performance (around 83-84% accuracy), LR outperformed them in terms of all chosen metrics. This could be due to the specific characteristics of the dataset or the inherent strengths of LR in handling linear relationships between features.

4.4 Why Did Logistic Regression Perform Well?

Logistic Regression demonstrated superior performance compared to Random Forest and Naive Bayes in predicting heart disease. Several factors contributed to its success:

- Linear Relationship Handling: Logistic Regression is particularly effective when the relationship between features and the target variable is linear. In our dataset, features might have linear relationships with the likelihood of heart disease, which Logistic Regression can exploit effectively.

- Interpretability Logistic Regression provides interpretable results, allowing us to understand the impact of each feature on the prediction. This transparency is essential in a medical context, where understanding the factors contributing to a prediction is crucial.

- Efficient with High-Dimensional Data Logistic Regression is efficient when dealing with high-dimensional data, making it suitable for datasets with a large number of features.

4.5 Interpretation of Results

By achieving an accuracy of 87%, Logistic Regression has demonstrated its potential as a valuable tool for early detection and risk assessment of heart disease. The balance between precision, recall, and accuracy indicates the model's effectiveness in correctly identifying patients with heart disease while minimizing false positives and false negatives.

4.6 Summary

This project explored the application of machine learning algorithms for heart disease prediction. The results demonstrate that the LR model achieved promising performance in predicting heart disease based on patient data. This approach has the potential to be a valuable tool for early detection and risk assessment of heart disease, ultimately contributing to improved patient outcomes.

Chapter 5

Discussion and Analysis

The Discussion and Analysis chapter evaluates and interprets the results obtained from our heart disease prediction project. We analyze the performance of the machine learning models, including Logistic Regression, Random Forest, and Naive Bayes, in predicting heart disease based on patient data.

5.1 Significance of the findings

The findings of this project hold significant implications for the field of cardiovascular health. By successfully training and evaluating machine learning models for heart disease prediction, we have demonstrated the potential of these models as valuable tools for early detection and risk assessment. The high accuracy and balanced precision-recall trade-off achieved by the Logistic Regression model, in particular, highlight its effectiveness in identifying patients with heart disease. These findings enhance our understanding of the potential applications of machine learning in healthcare and contribute to ongoing efforts to improve patient outcomes in cardiovascular diseases.

5.2 Limitations

But, there are some things we need to think about. We only used one dataset, which might not cover all types of patients. Also, how good our models are might depend on how good the data is and what we look at. We need more research to check if our findings work in different hospitals and with different patients. Plus, we need to make sure our models make sense for doctors to use and don't give them wrong information.

5.3 Summary

In summary, the Discussion and Analysis chapter provides a comprehensive evaluation of the results obtained from our heart disease prediction project. The findings provide the potential of machine learning models, particularly Logistic Regression, in predicting heart disease and improving patient outcomes. While the results are promising, it is essential to consider the limitations and potential implications for future research and clinical practice.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In conclusion, this project aimed to develop machine learning models for predicting heart disease, with a primary focus on early detection of Coronary Artery Disease (CAD). Through the implementation of Logistic Regression, Random Forest, and Naive Bayes algorithms, we successfully trained models on a dataset containing relevant patient information. Our findings indicate that the Logistic Regression model achieved the highest accuracy and demonstrated a good balance between precision and recall. This suggests that machine learning algorithms, particularly Logistic Regression, hold promise as effective tools for predicting heart disease and contributing to improved patient outcomes. Overall, this project's central contributions lie in the successful implementation and evaluation of machine learning models for heart disease prediction, paving the way for future advancements in cardiovascular health.

6.2 Future work

While this project has made significant advancements in predicting heart disease, there are still opportunities for further exploration and improvement. Moving forward, it would be beneficial to conduct further research to refine the machine learning models and enhance their predictive capabilities. Additionally, exploring the integration of additional features or datasets could provide valuable insights into improving the accuracy and robustness of the models. Additionally, it would be helpful to study how well the prediction system works over a long time in real-life healthcare settings. In the future, we should keep improving and testing the prediction system to make sure it works well and can be trusted by doctors.

Chapter 7

Reflection

Undertaking this project has been a significant learning experience for me, extending far beyond the gaining of technical skills. While I did gain proficiency in using various programming languages and tools like LaTeX, the most valuable takeaway from this project was the development of problem-solving skills and research methodology. Through the process of identifying and solving a complex problem such as predicting heart disease, I learned the importance of thorough research inquiry and strategic planning.Figuring out how to predict heart disease was challenging, especially dealing with the complicated dataset and getting the data ready for analysis. Even though I faced some tough moments, like spending a lot of time cleaning the data, I learned a lot in the process. If I were to approach a similar problem in the future, I would focus more on data collection and pre-processing to streamline the model training process. Reflecting on the initial aims and objectives of the project, I realized the need for greater flexibility and adaptability in project planning. While my initial goals were clear and well-defined, the process of research and experimentation led to new insights and adjustments in approach. Overall, this project has not only enhanced my technical skills but also deepened my understanding of the research process and its implications for future work.

References

Janosi, A., Steinbrunn, W., Pfisterer, M. and Detrano, R. (1988), 'Heart disease', UCI Machine Learning Repository. DOI: `https://doi.org/10.24432/C52P4X`.

Sharma, V., Yadav, S. and Gupta, M. (2020), Heart disease prediction using machine learning techniques, *in* '2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)', pp. 177–181.

Webb, G., Mulder, B. J., Aboulhosn, J., Daniels, C. J., Elizari, M. A., Hong, G., Horlick, E., Landzberg, M. J., Marelli, A. J., O'Donnell, C. P. et al. (2015), 'The care of adults with congenital heart disease across the globe: current assessment and future perspective: a position statement from the international society for adult congenital heart disease (isachd)', *International journal of cardiology* **195**, 326–333.

Appendix A

An Appendix Chapter (Optional)

Some lengthy tables, codes, raw data, length proofs, etc. which are **very important but not essential part** of the project report goes into an Appendix. An appendix is something a reader would consult if he/she needs extra information and a more comprehensive understating of the report. Also, note that you should use one appendix for one idea.

An appendix is optional. If you feel you do not need to include an appendix in your report, avoid including it. Sometime including irrelevant and unnecessary materials in the Appendices may unreasonably increase the total number of pages in your report and distract the reader.

Appendix B

An Appendix Chapter (Optional)

...