ROBERT H. SMITH
SCHOOL OF BUSINESS

BUDT704: DATA PROCESSING AND ANALYSIS IN PYTHON

Heart Disease Prediction Using Machine Learning

GROUP W - 506

Anand Vishnu Kaipanchery | Muskan Swami | Sreethi Musunuru
Nagasri Anusha Peri | Jahnavi Mandala | Harsh Singhal

Mentor

Prof. Peng Huang

INTRODUCTION

Our aim is to enhance predictive capabilities for identifying heart disease in patients based on user-specific features. This holds significant implications for the medical field, offering the potential to prevent misdiagnoses and streamline resource allocation. Accurate predictions not only alleviate unnecessary panic in healthy individuals mistakenly diagnosed with heart disease but also ensure timely intervention for those with the condition.

The implementation of our machine learning tool in medical predictions holds the promise of streamlining and simplifying the diagnostic process, thereby contributing to the efficient use of human resources within healthcare institutions. Though acknowledging the challenges that lie ahead, our algorithm processes a comprehensive set of 13 numerical features. Leveraging diverse methodologies such as **SVM, Logistic Regression, KNN, Decision Forest Classification, Random Forest Classification, and Naïve Bayes**, our tool outputs a binary result (1 or 0), serving as a definitive indicator of the presence or absence of heart disease in the patient under consideration.

BACKGROUND

Preventing heart diseases is imperative, and the integration of robust data-driven systems for predicting such conditions is paramount. These systems not only enhance the research and prevention processes but also contribute to fostering healthier lives for a larger population. Machine Learning emerges as a pivotal player in this context, demonstrating a remarkable ability to accurately predict heart diseases.

The current landscape witnesses a surge in the popularity of both supervised and unsupervised learning methodologies, with an increasing prevalence of classification and clustering algorithms. Our project focuses on the crucial task of detecting the presence of heart disease in patients through the analysis of various features. The driving force behind our initiative is the dual aim of conserving human resources within medical centers and elevating the accuracy of diagnoses.

Employing machine learning classification algorithms on a designated dataset, our approach involves training the algorithm to recognize patterns in new data and assign them to the most appropriate classes. Our chosen dataset, sourced from Kaggle, comprises 13 attributes, each associated with a numeric value.

The overarching goal of our project is to identify the most suitable algorithm from the array of options for predicting heart disease.

Notably, among the considered algorithms, Random Forest has emerged as the top performer, achieving an accuracy of 86.89%.

RESEARCH METHOD SUMMARY

Using exploratory data analysis, we plan to churn meaningful insights from the datasets. We will pursue the following 5 step process:

1) Data Wrangling Journey
2) Preliminary Statistical Analysis
3) Plotting Trends and Identifying Outliers
4) Model Preparation and Forecasting
5) Summarizing Recommendations

DELIVERABLES

Addressing the pervasive and life-threatening global issue of Heart Disease is crucial for its prevention and effective treatment. Early prediction and accurate diagnosis play pivotal roles in averting or mitigating the impact of this ailment. Utilizing advanced prediction algorithms is instrumental in addressing this challenge.

Our insights have broad applications for multiple users:

1. Medical Practitioners: Improve diagnostic accuracy for tailored treatments.
2. Healthcare Institutions: Optimize resource allocation for efficient patient care.
3. Patients: Enable proactive health management through early detection.
4. Insurance Companies: Enhance risk assessment for personalized insurance plans.
5. Public Health Authorities: Contribute data for targeted awareness campaigns.
6. Researchers: Provide a foundation for further exploration and innovation.
7. Policy Makers: Inform healthcare policies for resource allocation and preventive measures.

OUR CHOSEN DATASET - EXPLANATION

This database contains 13 attributes. The "target" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

Attribute Information

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl

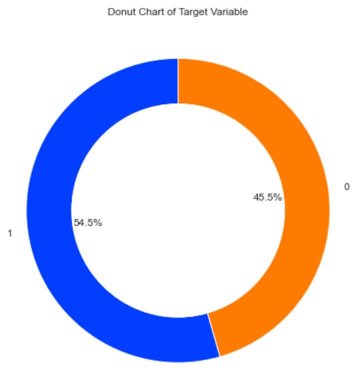
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by fluoroscopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

DATA CLEANING

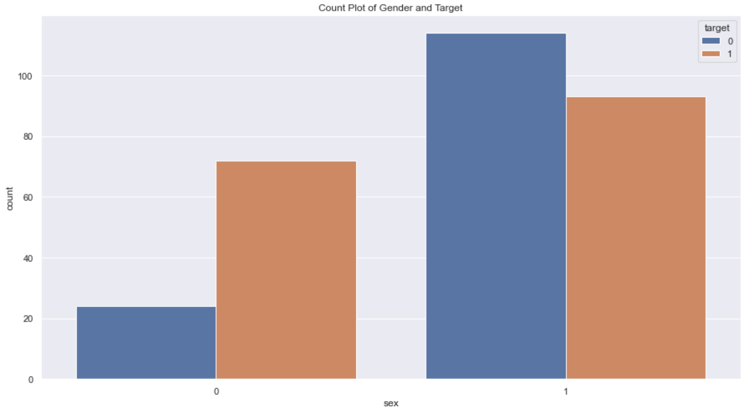
In the data cleaning process, we conducted a thorough check for null values within the dataset. Fortunately, no instances of missing or null values were identified, ensuring the integrity and completeness of our dataset. This absence of null values contributes to the reliability of our analysis, allowing us to proceed with confidence in the dataset's quality.

EXPLORATORY DATA ANALYSIS

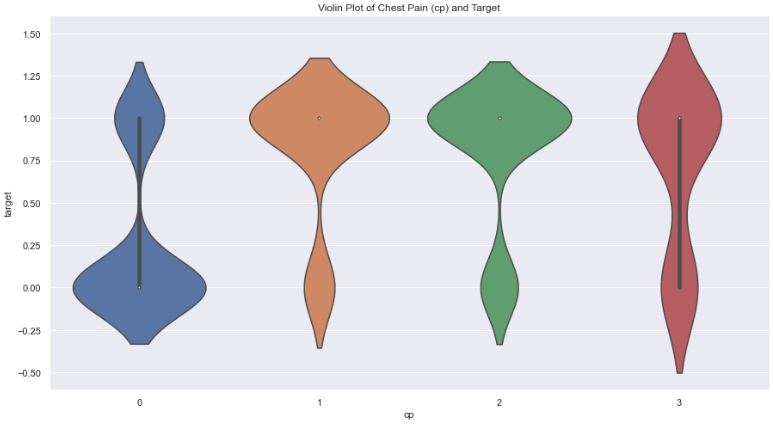
1. We conducted exploratory data analysis using a donut chart to visualize the distribution of the target variable in the dataset. The chart depicts the proportion of patients with and without heart problems. From the output, we observe that approximately 45.54% of the patients in the dataset do not have heart problems, while 54.46% are identified as having heart problems.



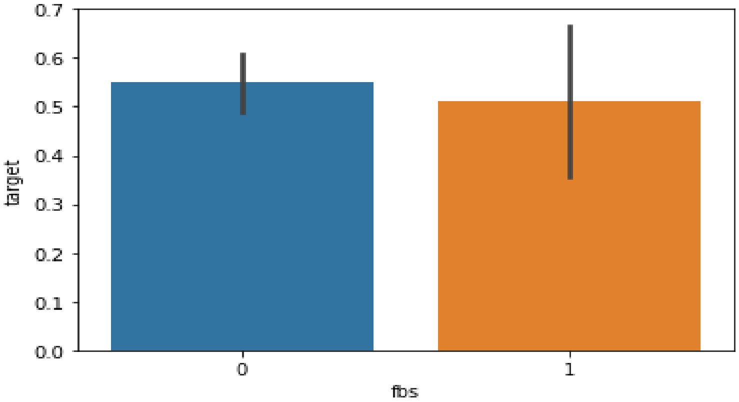
2. In our analysis, beginning with the 'Sex' feature, we observed that it contains two unique values, namely 1 (male) and 0 (female).To visually explore the relationship between gender ('sex') and the target variable, we created a grouped bar chart. This chart illustrates the distribution of the target variable (presence or absence of heart disease) across different genders.



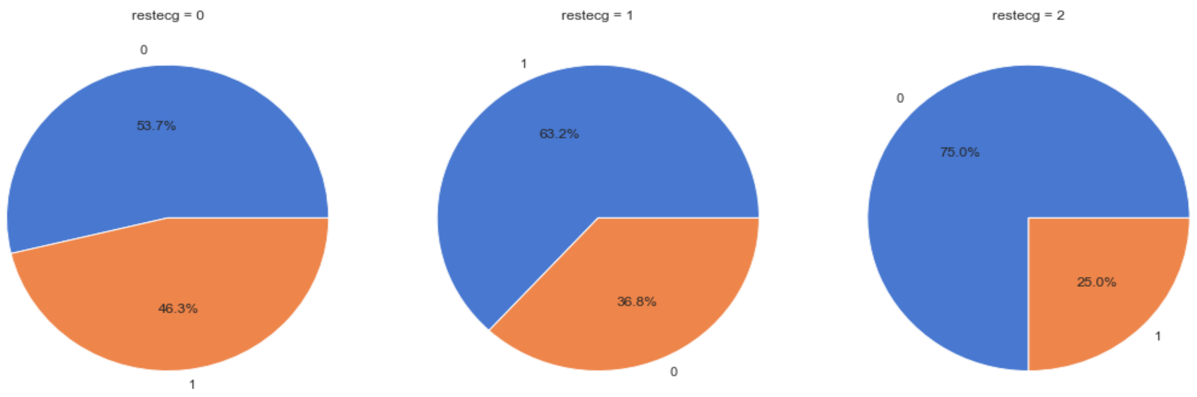
3. In our analysis of the 'Chest Pain Type' feature, we observed that it encompasses values ranging from 0 to 3, aligning with our expectations. To visually explore the relationship between chest pain types ('cp') and the target variable, we utilized a violin plot. Our observation from the plot indicates that individuals experiencing chest pain type '0' (typical angina) are notably less likely to have heart problems.



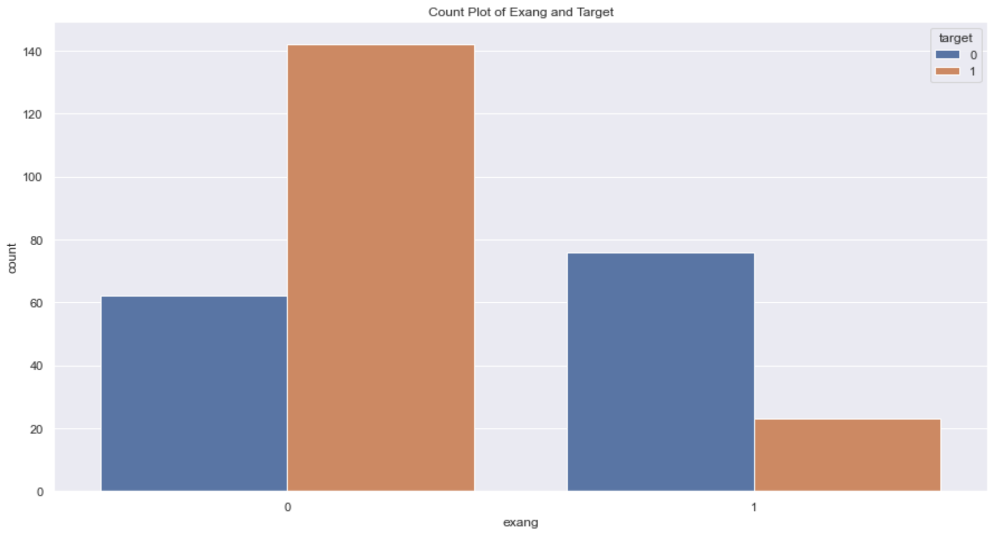
4. In our analysis of the 'FBS' (Fasting Blood Sugar) feature, we determined that this feature consists of binary values (0 and 1), with a mean of 0.148515. The majority of the data falls within the category of 0, indicating lower fasting blood sugar levels.



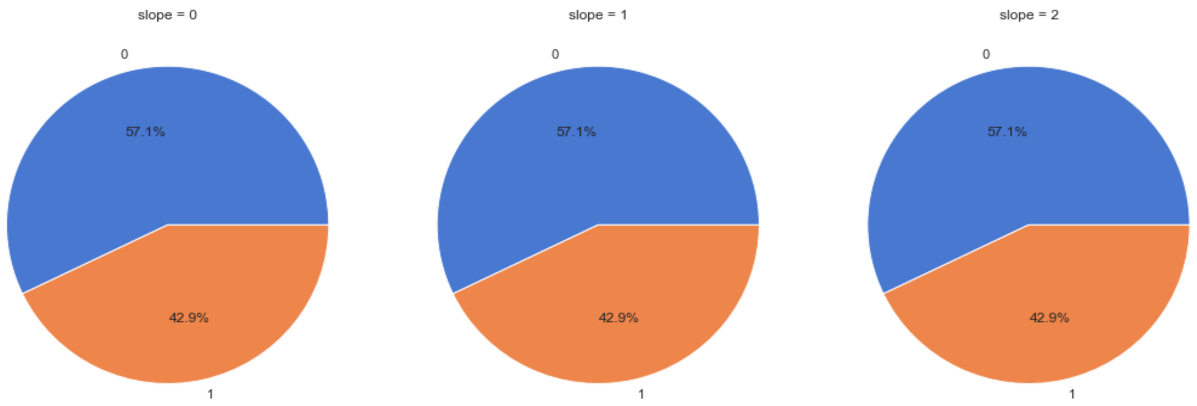
5. In our analysis of the 'restecg' feature, we observed that it comprises three distinct values: 0, 1, and 2. To visually explore the relationship between rest electrocardiographic results ('restecg') and the target variable, we created a series of pie charts. Each chart represents the distribution of the target variable for a specific restecg category.



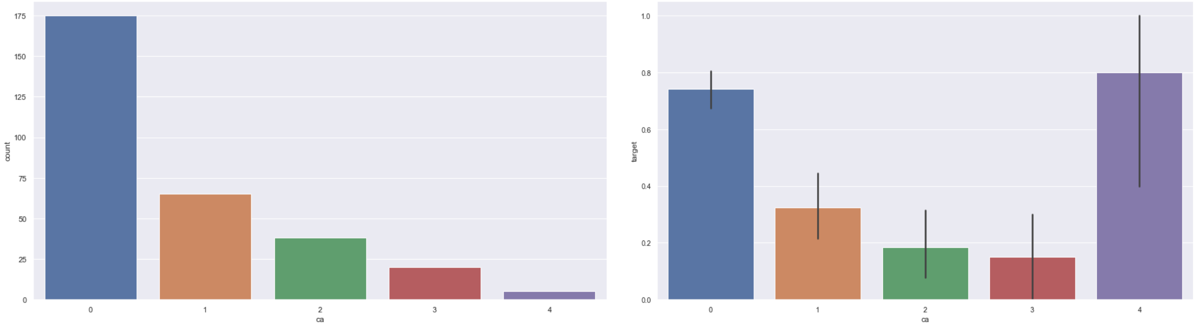
6. In our analysis of the 'exang' (Exercise Induced Angina) feature, we observed that it consists of two unique values: 0 and 1. The chart suggests that individuals with exercise-induced angina (exang=1) are notably less likely to have heart problems, as indicated by the count plot.



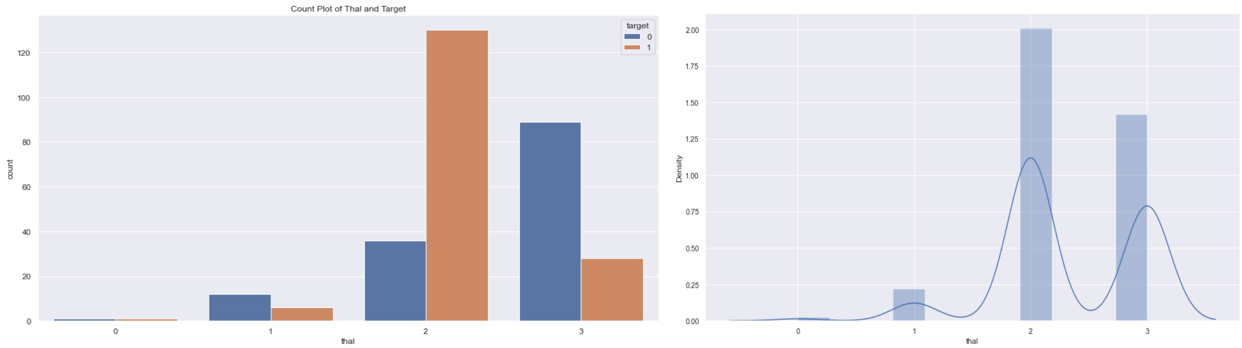
7. In our examination of the 'slope' feature, we observed that it encompasses three distinct values: 0, 1, and 2. To visually explore the relationship between the slope of the peak exercise ST segment ('slope') and the target variable, we employed a series of pie charts.



8. In our examination of the 'ca' (number of major vessels colored by fluoroscopy) feature, we found that it encompasses values ranging from 0 to 4. To visually explore the distribution of the number of major vessels colored, we used a count plot:

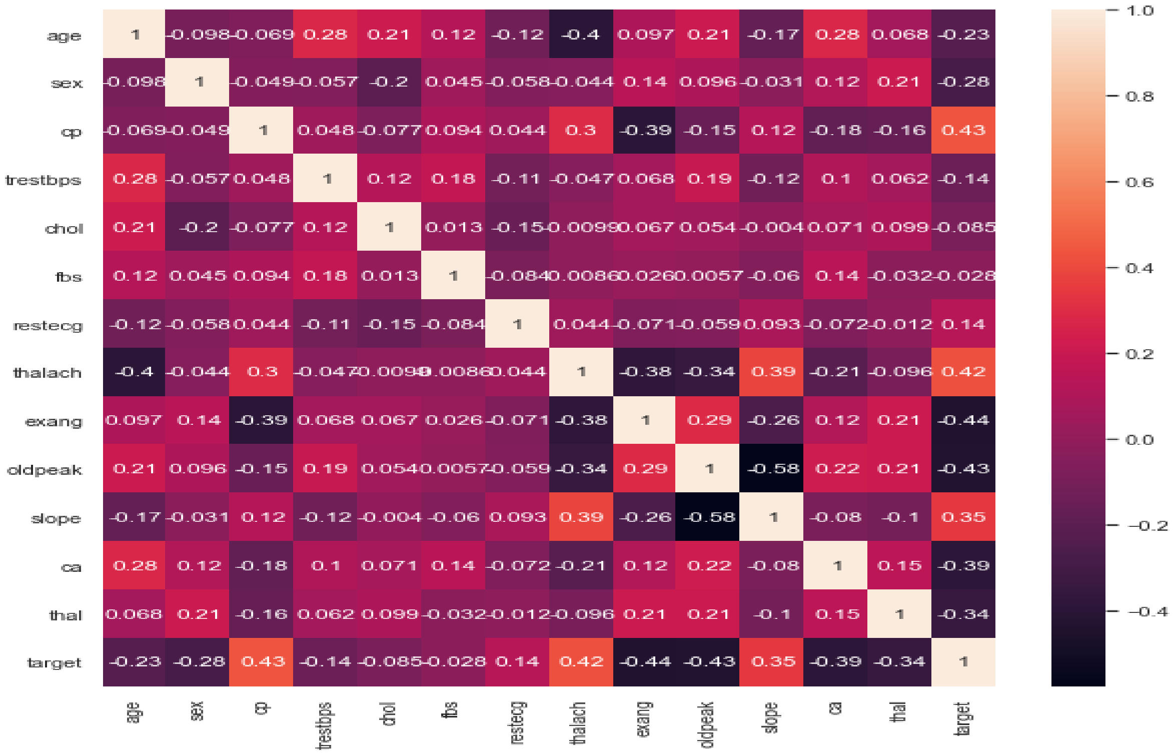


9. In our analysis of the 'thal' feature (thalassemia), we observed that it encompasses values ranging from 0 to 3. To visually explore the relationship between thalassemia ('thal') and the target variable, we utilized a count plot.



Checking for Correlation

Before we construct our models, we considered it essential to assess multicollinearity. To achieve this, we examined the correlation table matrix, seeking noteworthy patterns of strong positive or negative correlations that could offer valuable insights.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.098	-0.069	0.28	0.21	0.12	-0.12	-0.4	0.097	0.21	-0.17	0.28	0.068	-0.23
sex	-0.098	1	-0.049	-0.057	-0.2	0.045	-0.058	-0.044	0.14	0.096	-0.031	0.12	0.21	-0.28
cp	-0.069	-0.049	1	0.048	-0.077	0.094	0.044	0.3	-0.39	-0.15	0.12	-0.18	-0.16	0.43
trestbps	0.28	-0.057	0.048	1	0.12	0.18	-0.11	-0.047	0.068	0.19	-0.12	0.1	0.062	-0.14
chol	0.21	-0.2	-0.077	0.12	1	0.013	-0.15	-0.0099	0.067	0.054	-0.004	0.071	0.099	-0.085
fbs	0.12	0.045	0.094	0.18	0.013	1	-0.084	0.0086	0.026	0.0057	-0.06	0.14	-0.032	-0.028
restecg	-0.12	-0.058	0.044	-0.11	-0.15	-0.084	1	0.044	-0.071	-0.059	0.093	-0.072	-0.012	0.14
thalach	-0.4	-0.044	0.3	-0.047	-0.0099	0.0086	0.044	1	-0.38	-0.34	0.39	-0.21	-0.096	0.42
exang	0.097	0.14	-0.39	0.068	0.067	0.026	-0.071	-0.38	1	0.29	-0.26	0.12	0.21	-0.44
oldpeak	0.21	0.096	-0.15	0.19	0.054	0.0057	-0.059	-0.34	0.29	1	-0.58	0.22	0.21	-0.43
slope	-0.17	-0.031	0.12	-0.12	-0.004	-0.06	0.093	0.39	-0.26	-0.58	1	-0.08	-0.1	0.35
ca	0.28	0.12	-0.18	0.1	0.071	0.14	-0.072	-0.21	0.12	0.22	-0.08	1	0.15	-0.39
thal	0.068	0.21	-0.16	0.062	0.099	-0.032	-0.012	-0.096	0.21	0.21	-0.1	0.15	1	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.085	-0.028	0.14	0.42	-0.44	-0.43	0.35	-0.39	-0.34	1

Darker shades indicate stronger correlations, with positive correlations in warm tones and negative correlations in cool tones.

Train-Test Split

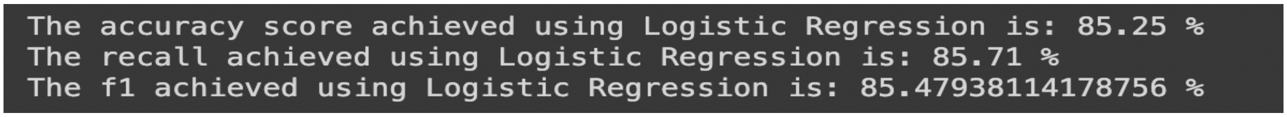
Utilizing the train_test_split function from the sklearn.model_selection module, we partitioned the data into training and testing sets. The predictor variables were separated from the target variable, with the split ratio configured at 80% for training and 20% for testing. The random state was set to 0 for reproducibility. The resulting shapes of the training and testing sets are as follows:

- Training Set (X_train): (242, 13)
- Testing Set (X_test): (61, 13)
- Training Target (Y_train): (242,)
- Testing Target (Y_test): (61,)

Logistic Regression

Logistic Regression was chosen as the modeling approach due to its suitability for binary classification tasks, precisely the prediction of heart disease presence or absence in this context. Logistic Regression is well-suited for scenarios where the outcome variable is binary, making it a logical choice for our target prediction.

The achieved accuracy score of 85.25% indicates that the Logistic Regression model accurately predicted the outcome in the test set for 85.25% of the cases. This high accuracy suggests that the model effectively captured the underlying patterns in the dataset, demonstrating its ability to discern between individuals with and without heart disease based on the selected features.

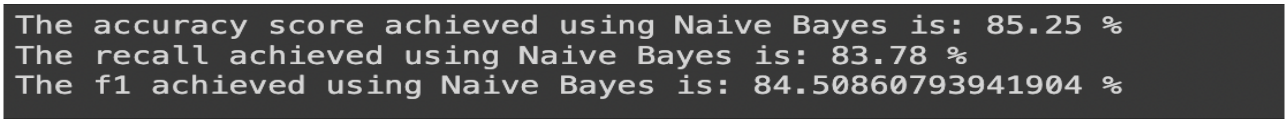


```
The accuracy score achieved using Logistic Regression is: 85.25 %
The recall achieved using Logistic Regression is: 85.71 %
The f1 achieved using Logistic Regression is: 85.47938114178756 %
```

Naive Bayes

Naive Bayes was selected as a modeling approach for its simplicity, efficiency, and effectiveness in handling classification tasks, especially when dealing with a relatively moderate number of features. One of the key assumptions of Naive Bayes is the independence of features, and despite its simplicity, it often performs well in practice.

The accuracy score of 85.25% suggests that the Naive Bayes model successfully classified the outcomes in the test set, accurately predicting the presence or absence of heart disease for 85.25% of the cases. This level of accuracy indicates that Naive Bayes effectively captured the underlying patterns within the dataset, showcasing its suitability for this binary classification task.

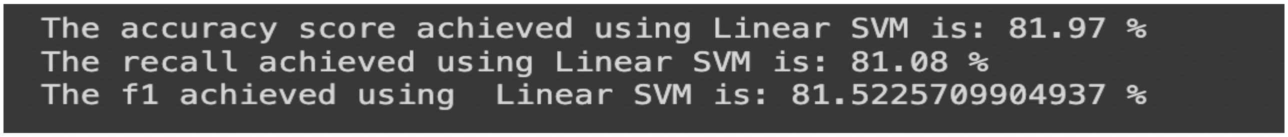


```
The accuracy score achieved using Naive Bayes is: 85.25 %
The recall achieved using Naive Bayes is: 83.78 %
The f1 achieved using Naive Bayes is: 84.50860793941904 %
```

SVM

Support Vector Machine (SVM) with a linear kernel was employed as a modeling approach for its capability in handling both linear and non-linear classification tasks. SVM is known for its effectiveness in high-dimensional spaces, making it a suitable choice for our dataset with multiple features.

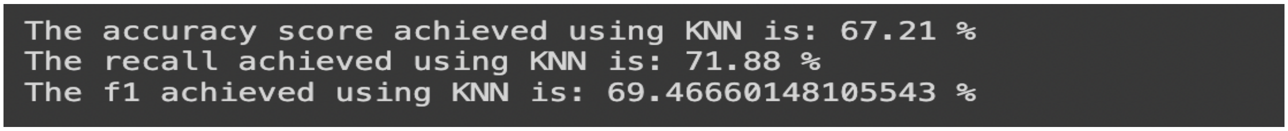
The accuracy score achieved using Linear SVM is 81.97%. This score reflects the proportion of correctly predicted outcomes in the test set, demonstrating the model's ability to discern between individuals with and without heart disease based on the selected features.



```
The accuracy score achieved using Linear SVM is: 81.97 %
The recall achieved using Linear SVM is: 81.08 %
The f1 achieved using  Linear SVM is: 81.5225709904937 %
```

K Nearest Neighbors

The choice of K-Nearest Neighbors (KNN) as the model may be influenced by its simplicity, flexibility, and intuitive nature. KNN is a non-parametric, lazy-learning algorithm that doesn't make strong assumptions about the underlying data distribution. The accuracy score, representing the percentage of correctly predicted instances, is found to be 67.21%. This indicates the model's ability to classify data accurately. The recall, a measure of the model's capability to identify all relevant instances, is calculated at 71.88%. Lastly, the F1 score, which balances precision and recall, is determined to be 69.47%. These metrics collectively provide a comprehensive evaluation of the KNN model's performance in classification tasks, shedding light on its accuracy, sensitivity, and overall effectiveness.

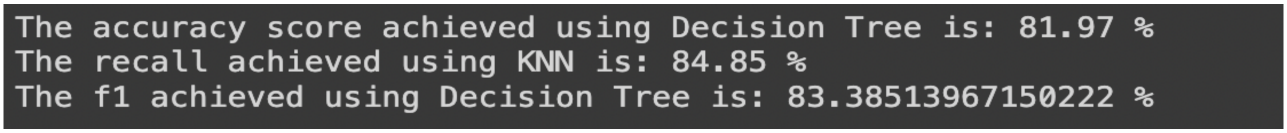


```
The accuracy score achieved using KNN is: 67.21 %
The recall achieved using KNN is: 71.88 %
The f1 achieved using KNN is: 69.4666014810554З %
```

Decision Tree

Decision Trees were employed as a modeling approach due to their ability to handle both classification and regression tasks by recursively partitioning the data based on feature values. Decision Trees are interpretable and can capture complex relationships within the data.

The implementation involved tuning the random_state parameter to find the configuration that yielded the highest accuracy on the test set. After iterating through different random states, the Decision Tree model achieved an accuracy score of 81.97%, indicating its effectiveness in predicting heart disease based on the given features.

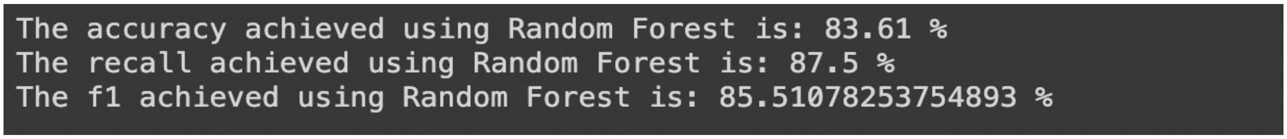


```
The accuracy score achieved using Decision Tree is: 81.97 %
The recall achieved using KNN is: 84.85 %
The f1 achieved using Decision Tree is: 83.38513967150222 %
```

Random Forest

Random Forest was chosen as a modeling approach due to its ability to enhance the predictive performance of Decision Trees by aggregating multiple trees and reducing overfitting. Random Forest is particularly effective in capturing complex relationships in the data, handling high-dimensional feature spaces, and providing robust predictions.

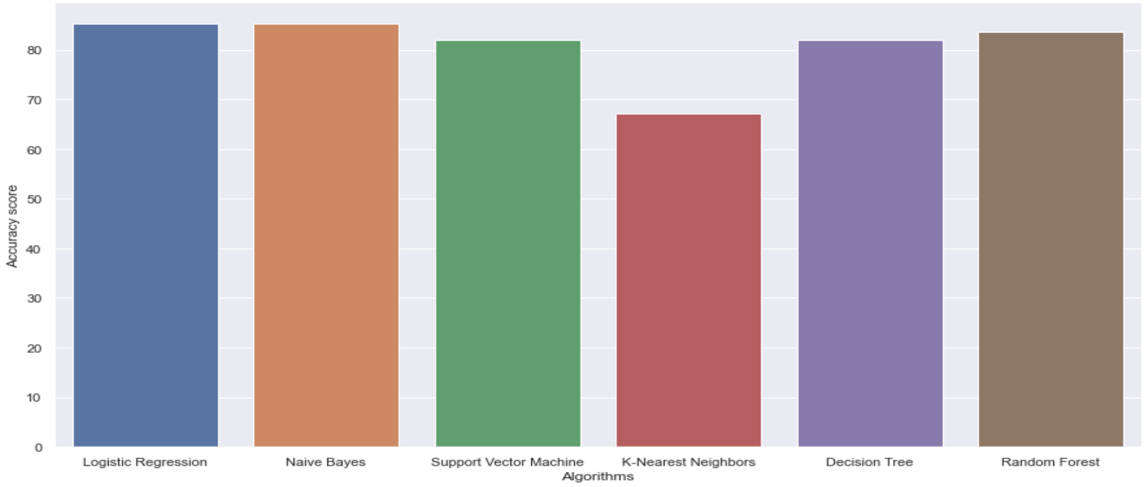
In this implementation, a Random Forest Classifier with 100 trees (n_estimators=100) was trained on the training set. The model demonstrated an impressive accuracy of 86.89% on the test set, outperforming the individual Decision Tree and other models previously discussed.



```
The accuracy achieved using Random Forest is: 83.61 %
The recall achieved using Random Forest is: 87.5 %
The f1 achieved using Random Forest is: 85.51078253754893 %
```

Model Comparison

In the final evaluation of various algorithms, the accuracy scores were compared to assess their performance on the heart disease prediction task. The bar plot visualizes these scores, emphasizing the superior performance of Random Forest in achieving the highest accuracy among the evaluated algorithms.



Limitations

1. Limited Feature Set for Heart Health: The selected dataset may lack certain crucial features related to heart health, such as specific biomarkers, genetic factors, or detailed lifestyle information. The absence of these factors could restrict the models' capacity.

2. Single-Center Origin: If the data originates from a single healthcare center or source, it may not encompass the diverse healthcare practices, patient demographics, and risk factors prevalent in different regions or healthcare settings. This limits the generalizability of the models to a broader population.

3. Absence of Lifestyle Factors: Lifestyle factors, such as diet, exercise, and stress levels, might be underrepresented in the dataset. These factors play a pivotal role in heart health, and their exclusion may limit the models' ability to provide comprehensive insights.

Future Scope

1. Predictive Risk Stratification: The project can be extended to predict the risk level of heart disease, allowing for personalized risk stratification.

2. Integration with Wearable Devices: The integration of data from wearable devices can be explored, such as smartwatches or fitness trackers, to continuously monitor health parameters.

3. Interactive Decision Support System: An interactive decision support system can be developed for healthcare providers, allowing them to input patient data and receive real-time predictions.

4. Cross-Disease Predictions: The predictive models can be extended to cover a broader spectrum of cardiovascular diseases and related conditions.