MACHINE LEARNING

PROJECT REPORT

PROJECT TITLE: HEART FAILURE PREDICTION

SUPERVISED BY: DR SADAF HUSSAIN

SUBMITTED BY: MUHAMMAD ABDULLAH
FA-2020/BSCS/**014-(A)**

MUHAMMAD NABEEL AMJAD
FA-2020/BSCS/**221-(A)**

TOUSEEF ABBAS
FA-2020/BSCS/**525-(A)**

SUBMISSION DATE: FEB 4TH, 2024

DEPARTMENT OF COMPUTER SCIENCE

LAHORE GARRISON UNIVERSITY

Table of Contents

1. Abstract

This document presents a comprehensive exploration of heart disease prediction using machine learning algorithms. Our study investigates the performance of diverse algorithms, including KNN, Logistic Regression, SVM, Random Forest, Gradient Boosting, and Decision Tree. Through meticulous preprocessing, feature scaling, and outlier removal, we strive to enhance model accuracy.

Keywords: Machine Learning, Heart Disease, Random Forest, EDA

2. Introduction

With the rising importance of predictive healthcare, this study focuses on heart disease prediction using machine learning techniques. Utilizing datasets from renowned repositories, our investigation delves into the performance of various algorithms. The introduction outlines the significance of accurate heart disease prediction, emphasizing the potential impact on early diagnosis and intervention.

3. Literature Review

A comprehensive literature review highlights key studies in heart disease prediction using machine learning.

Sr.	Title Author	Year	Input features	Algorithm	Output	Dataset	Accuracy
1	Heart disease prediction using machine learning algorithms (Harshit Jindal1, Sarthak Agrawal1, Rishabh Khera1, Rachna Jain2 and Preeti Nagrath)	2021	Age, sex, CP, Trestbps, Chol, FBS, Restecg, Thalach, Exang, Oldpeak, Slope, Ca, Thal	KNN, Logistic Regression, Random Forest Classifiers,	Disorder	archive.ics.uci.edu	KNN: 88.5% Logistic Regression: 88.5%

2	Heart Failure Prediction using Machine Learning Techniques (Prasanta Kumar Sahoo, 2 Pravalika Jeripothula)	2021	Age, Gender ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, Exercise Angina, Oldpeak, ST_Slope,	SVM, Naïve Bayes, KNN, Logistic Regression, Decision tree	Class Label	archive.ics.uci.edu	SVM: 85.2% Naive Bayes: 75% Logistic Regression: 83% Decision Tree: 68% KNN: 81%
3	A Mini-Project Report On "Heart Disease Prediction" (Nirusha Manandhar, Sagun Lal Shrestha, Ruchi Tandukar)	2020	Male, age, education, currentSmooker, cigsPerday, BPMeds, prevalentStrokes, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartRate, glucose	Logistic Regression, Backward Elimination Method, Recursive Feature Elimination using Cross-Validation (RFECV)	HeartDisease	kaggle.com	RFECV: 85% Backward Elimination:83 %

4. Methodology

The purpose of designing this system is to do predictions if the patient has heart disease or not. Here is the methodology I have used in my project.

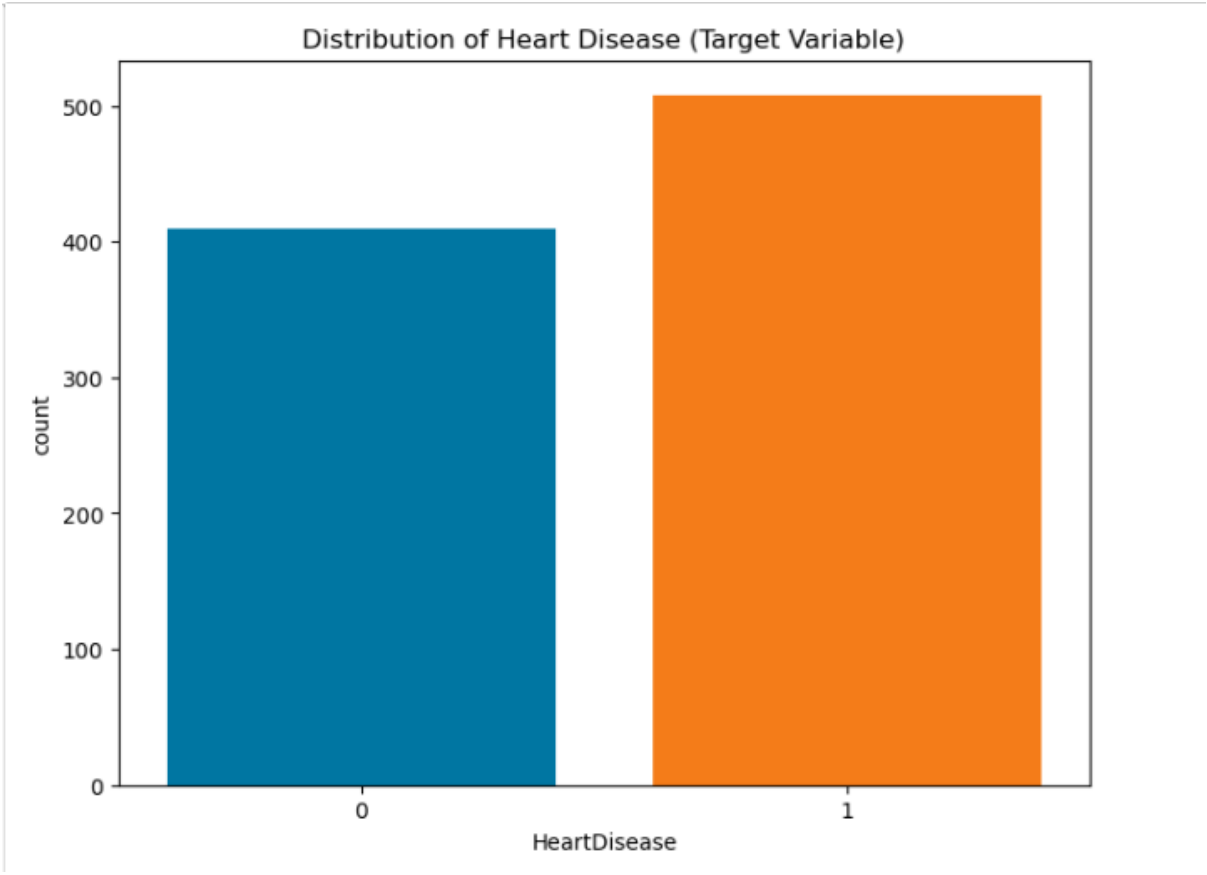
4.1 Dataset Description and Visualization

The dataset for this project is used from UCI repository. This dataset has 11 features. Below is the name of features and dataset and its description.

Features	Description	Datatype
Age	age of the patient [years]	Integer

Sex	gender of the patient [M: Male, F: Female]	Object
ChestPainType	chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]	Object
RestingBP	resting blood pressure [mm Hg]	Integer
Cholesterol	serum cholesterol [mm/dl]	Integer
FastingBS	fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]	Integer
RestingECG	resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]	Object
MaxHR	maximum heart rate achieved [Numeric value between 60 and 202]	Integer
ExerciseAngina	exercise-induced angina [Y: Yes, N: No]	Object
Oldpeak	oldpeak:(Depression induced by exercise relative to rest) = ST [Numeric value measured in depression]	Float
ST_Slope	the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]	Object
HeartDisease	output class [1: heart disease, 0: Normal]	Integer

Here is the visualization of our dataset in the Bar chart below. It shows how the data is distributed. Here the blue bar shows that how many people in our dataset don’t have heart disease and orange bar is showing the people having heart disease.



4.2 Preprocessing

Preprocessing is most important part. By doing preprocessing we can have higher accuracy when a model is trained and tested on our dataset. In our project we have used 2 techniques for preprocessing which are Label encoding and Exploratory data analysis (EDA).

Label Encoding

In label encoding we usually convert the string values in our dataset into integer value. So we in our dataset we have some feature having strings values and we also observed that these string values can be converted into categorical numerical values. The features such as Sex, ChestPainType, RestingECG, ExerciseAngina, and St_Slope are converted to numerical values using label encoding. The label encoded dataset is shown below.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	1	1	140	289	0	1	172	0	0.0	2	0
1	49	0	2	160	180	0	1	156	0	1.0	1	1
2	37	1	1	130	283	0	2	98	0	0.0	2	0
3	48	0	0	138	214	0	1	108	1	1.5	1	1
4	54	1	2	150	195	0	1	122	0	0.0	2	0

Exploratory Data Analysis (EDA):

After doing label encoding now we have Explored our dataset.

Here are the steps we followed during EDA.

i. Features Scaling:

In feature scaling we check if the difference between the record values of our feature is high we must have to scale those features. But in this dataset we have looked that all the features are scaled.

ii. Checking Null Values:

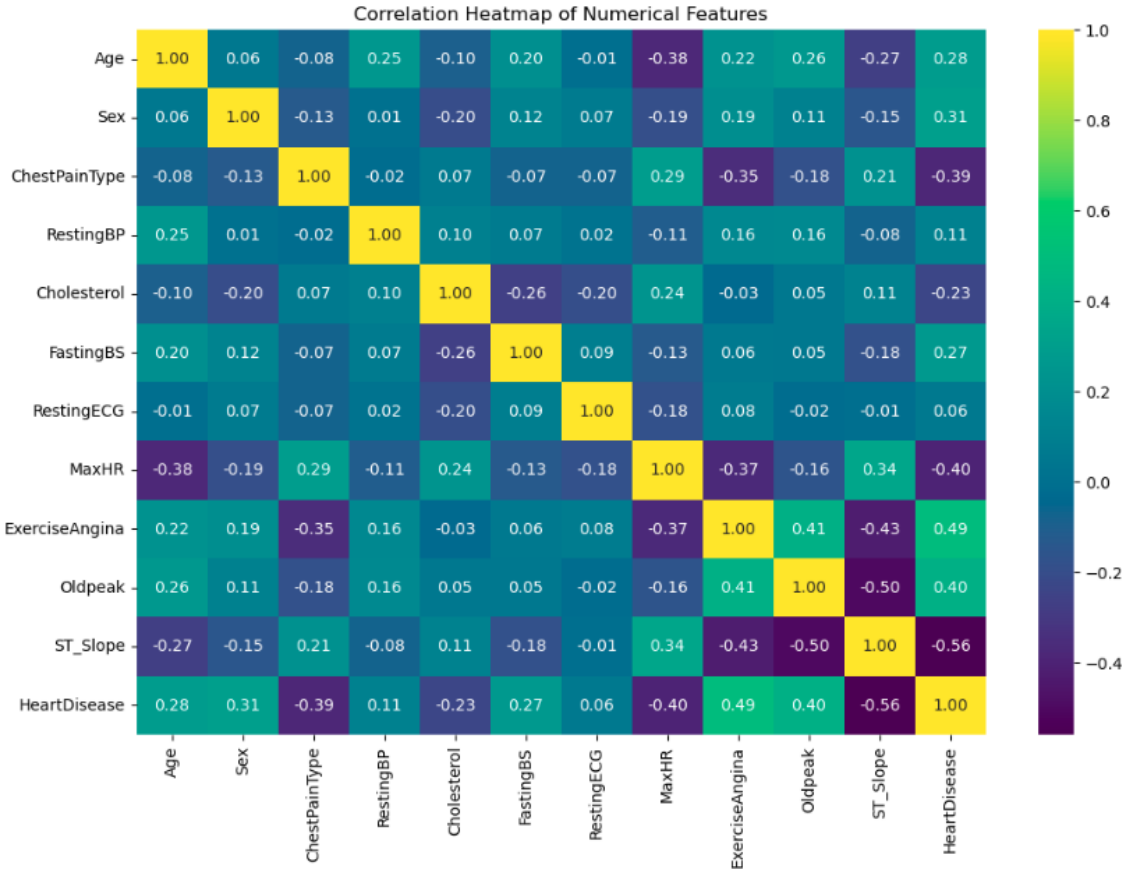
The records which have some missing values are called null values. We checked in our dataset if it has null values and we got the result with 0 null values.

iii. Duplication of Rows:

we checked if there is any duplicate row in our dataset and we got the result with 0 duplicate rows.

iv. Correlation Matrix:

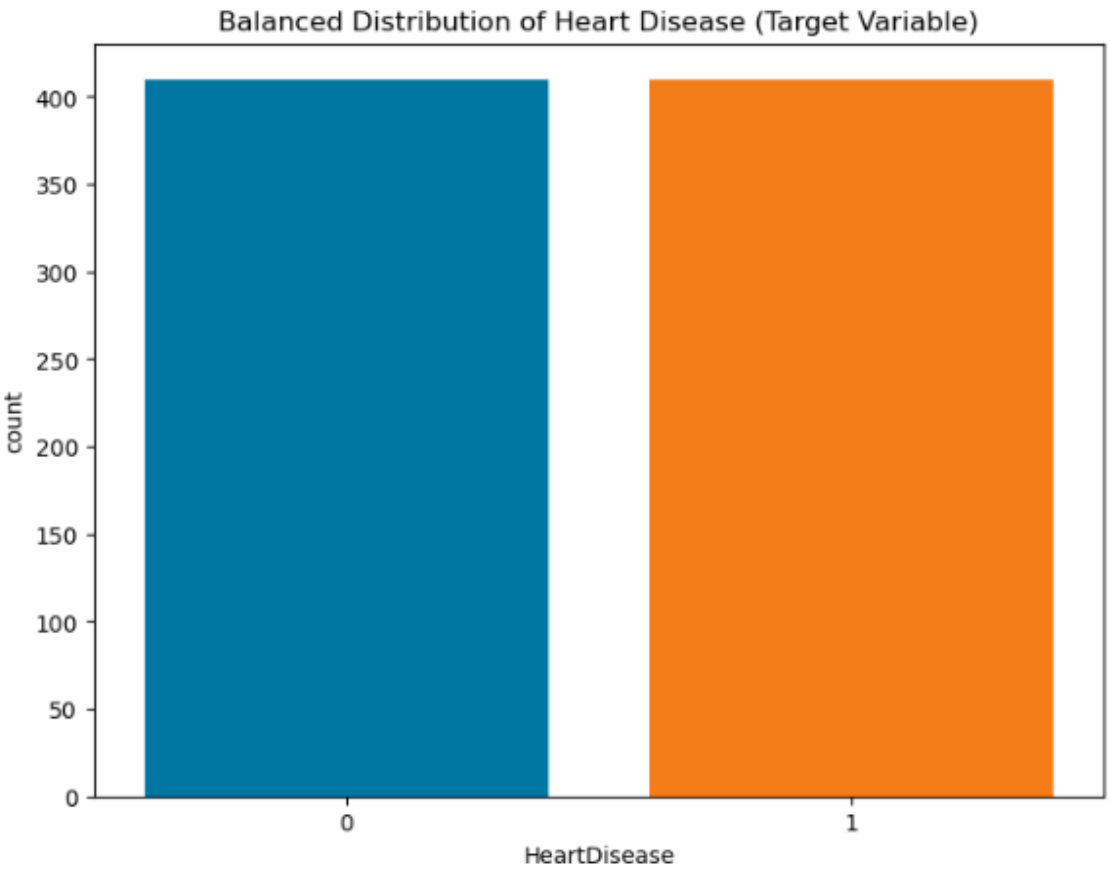
we did visualization of Correlation Matrix given below:



In this heat Map we can easily see that there is no feature which have very high Correlation with our target value.

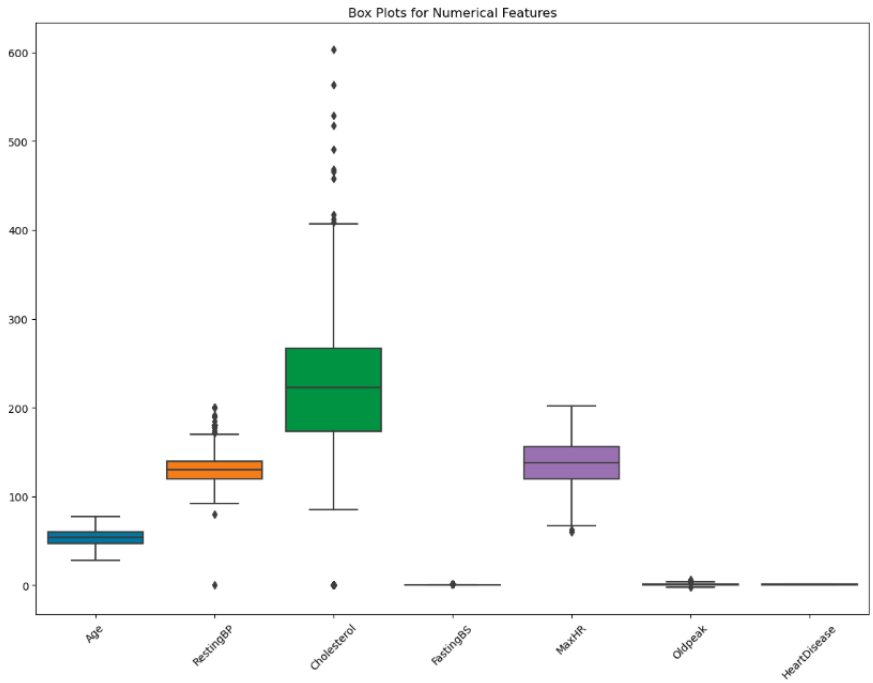
v. Balanced distribution of our dataset:

Balancing a dataset makes training a model easier because it helps prevent the model from becoming biased towards one class. Given below is the visualization when Balanced distribution is done.

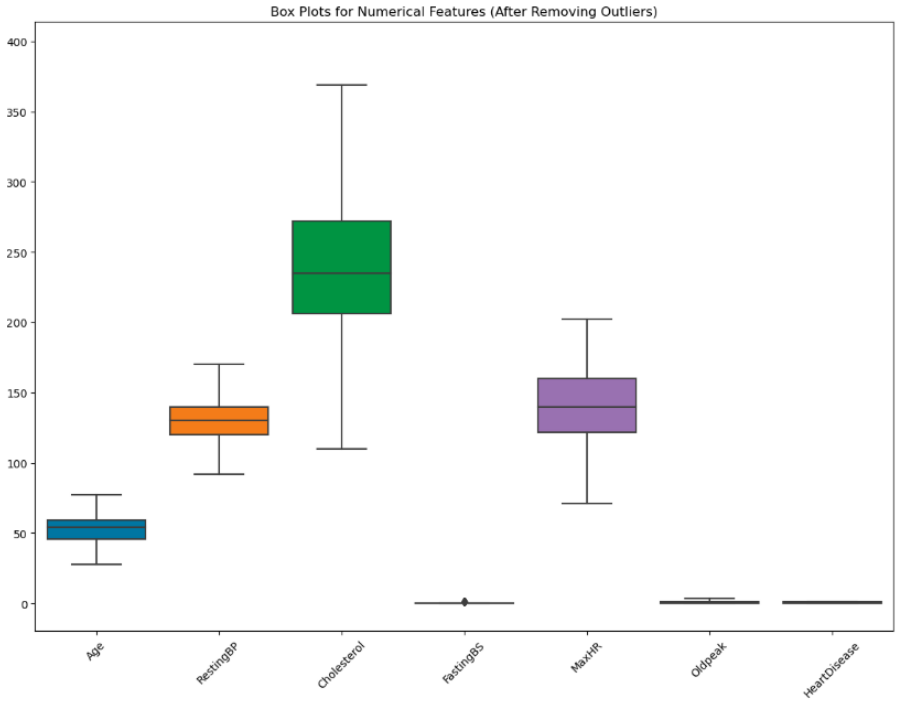


vi. Removing Outliers:

Outliers are data points that significantly deviate from the rest of the data. So we visualized our features with Box plot and experienced outliers in some of our features. The figure shows features with outliers.



We used modified Interquartile Range (IQR) method for outlier removal.



4.3 Algorithm Selection and Usage

These are the Algorithms We Selected:

Logistic regression: I do Binary or multiclass classification by modeling the probability of an event using a logistic function.

Support vector Machine (SVM): It does Classification or regression by finding a hyperplane that best separates data points with the largest margin.

K Nearest Neighbor (KNN): It does Classification or regression by assigning a data point the majority class or average value of its k nearest neighbors.

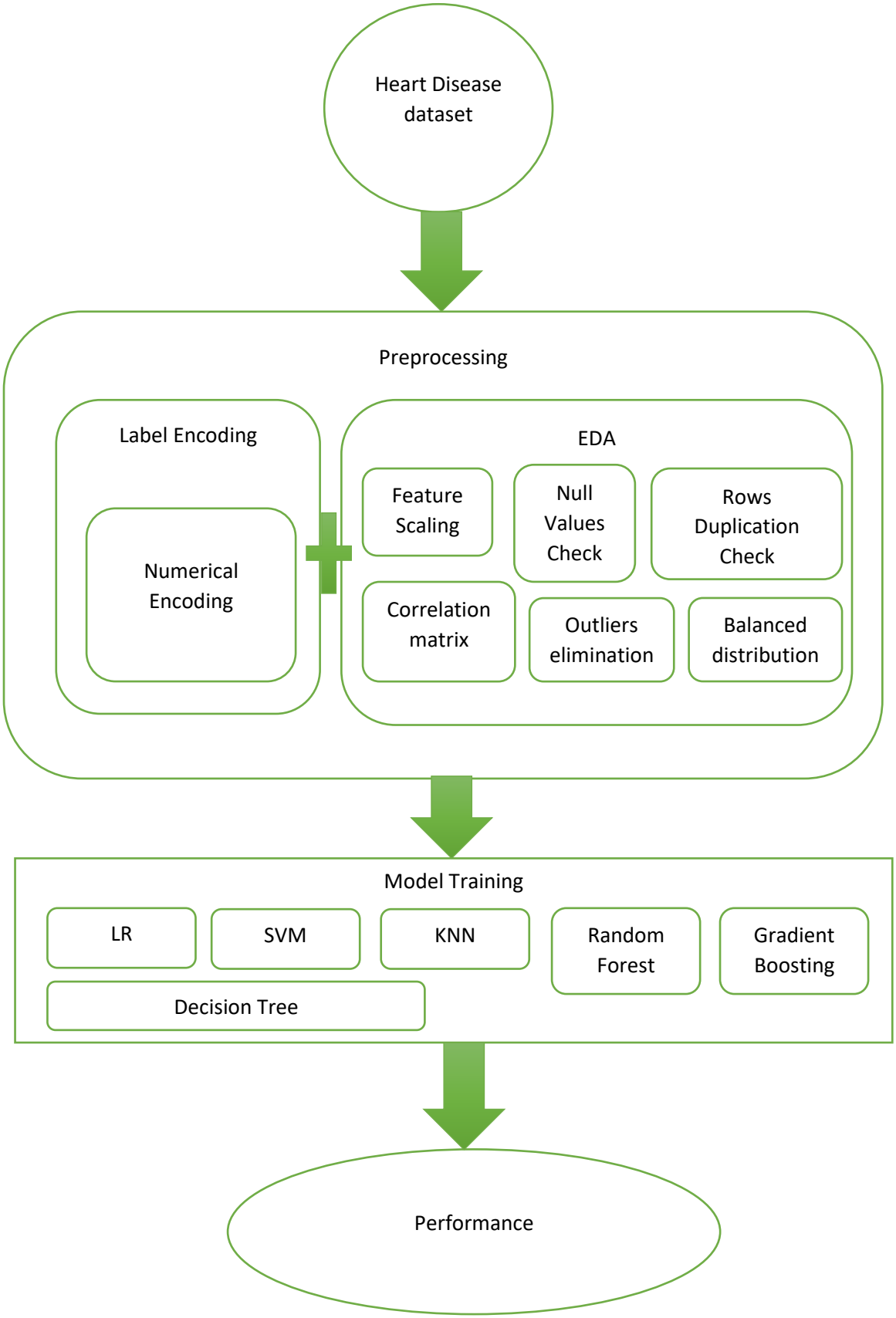
Random Forest: It Ensemble learning method using multiple decision trees to improve accuracy and control overfitting.

Gradient Boosting: It Ensemble learning technique building decision trees sequentially, each correcting the errors of the previous one.

Decision Tree: It does Classification or regression by recursively partitioning data based on feature splits.

4.4 Model

Architectural Model of a proposed Solution:



5 Results

The result is shown and discussed.

5.1 Discussion

we applied the algorithms which we selected above and the highest accuracy we got was 88% when we used Random Forest. Similarly we got the lowest accuracy of 68% when we applied SVM.

5.2 Performance Matrices

Algorithm	Precision	Recall	F1 Score
Logistic Regression	0.897959	0.822430	0.858537
SVM	0.757895	0.672897	0.712871
KNN	0.770833	0.691589	0.729064
Random Forest	0.897196	0.897196	0.897196
Gradient Boosting	0.920000	0.859813	0.888889
Decision Tree	0.860215	0.747664	0.800000

5.3 Comparison

Algorithm	Accuracy %
Logistic Regression	84
SVM	68
KNN	70
Random Forest	88
Gradient Boosting	87
Decision Tree	78

Random Forest outperformed other algorithms with the highest accuracy of 88%, showcasing its effectiveness in capturing complex patterns. Gradient Boosting closely followed with an accuracy of 87%, demonstrating strong predictive capabilities. SVM lagged behind with 68% accuracy, indicating a need for optimization or consideration of alternative models.

6 Conclusion and Future work

The project concludes with the successful implementation of machine learning algorithms, particularly Random Forest and Gradient Boosting, for heart disease prediction, achieving accuracy rates of 88% and 87%, respectively. The preprocessing steps, including label encoding, exploratory data analysis, and outlier removal, significantly contributed to model performance.
Future work may involve exploring additional features, optimizing hyperparameters for algorithms with lower accuracy, and incorporating advanced techniques such as deep learning. Continuous refinement of the model and integration of more diverse datasets could enhance the

overall predictive power and robustness of the heart disease prediction system. Additionally, real-time deployment and monitoring mechanisms can be considered for practical applications in healthcare.

References

1. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
2. https://www.academia.edu/42249626/Mini_Project_Report_On_Heart_Disease_Prediction_
3. https://www.academia.edu/98198442/Heart_Failure_Prediction_using_Different_Machine_Learning_Techniques
4. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3759562