Statlog (Heart) Data Set

Introduction:

The Statlog (Heart) Data Set is a dataset containing information on individuals with suspected heart disease. The dataset was collected by the Cleveland Clinic Foundation, and it consists of 303 instances and 14 attributes, including demographic, clinical, and laboratory features. The goal of the dataset is to predict whether an individual has heart disease or not, which is a binary classification problem.

The dataset was created in the late 1980s and early 1990s and was made publicly available as part of the Statlog project, which aimed to evaluate the performance of various machine learning algorithms on real-world datasets. The dataset has been widely used in research and education in machine learning and statistics.

The 14 features in the dataset include age, sex, chest pain type, resting blood pressure, serum cholesterol level, fasting blood sugar level, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, thallium stress test result, and the target variable which indicates whether the individual has heart disease or not.

The Statlog (Heart) Data Set is a valuable resource for researchers and educators in machine learning and statistics, as it provides a real-world problem with a well-defined target variable and a set of features that are relevant to the problem. Additionally, the dataset has been widely used as a benchmark dataset for evaluating the performance of various classification algorithms, which has led to the development of new and improved methods for solving binary classification problems.

Research problem:

Do different statistical modeling techniques, such as logistic regression, Naive Bayes, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA), perform in predicting or classifying a specific outcome in the 'heart.dat' dataset, and what are the corresponding Area Under the Curve (AUC) scores for these models?" And compare the multiple classification models and the calculation of AUC scores to evaluate their performance in classifying the outcome.

Data set contains:

The dataset contains 270 instances and 14 attributes. The attributes are:

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. chest pain type: chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
4. resting blood pressure: resting blood pressure (in mm Hg on admission to the hospital)
5. serum cholesterol: serum cholesterol in mg/dl
6. fasting blood sugar: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. resting electrocardiographic results: resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or definite left ventricular hypertrophy)
8. maximum heart rate achieved: maximum heart rate achieved during exercise
9. exercise induced angina: exercise induced angina (1 = yes; 0 = no)
10. oldpeak: ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
12. number of major vessels: number of major vessels (0-3) colored by fluoroscopy
13. thal: thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
14. target: diagnosis of heart disease (1 = presence; 2= absence)

Models I have Developed:

1.Logistic regression

AIC : 207.5817

BIC: 257.9076

Confusion matrix:

a

C_logis 1 2

0 17 97

1 133 22

Accuracy: 0.1449814

2.Naive bayes:

Confusion matrix:

C_NB 1 2

1 133 21

2 17 98

Accuracy: 0.8587361

3.LDA

Confusion matrix:

C_LDA 1 2

1 133 24

2 17 95

Accuracy: 0.8475836

4. QDA Confusion matrix:

C_QDA 1 2

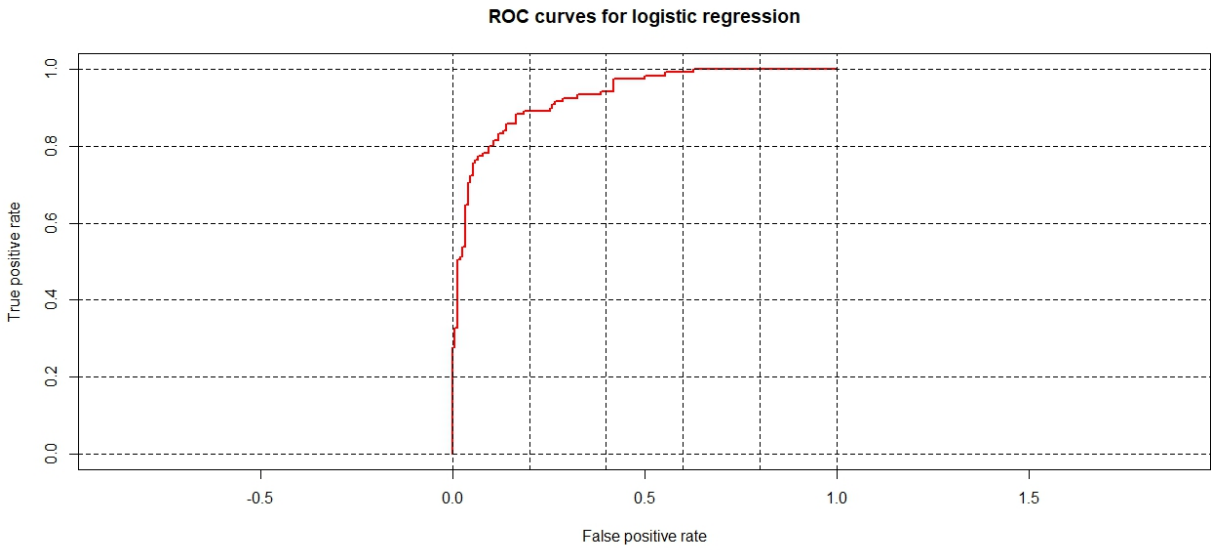
1 133 18

2 17 101

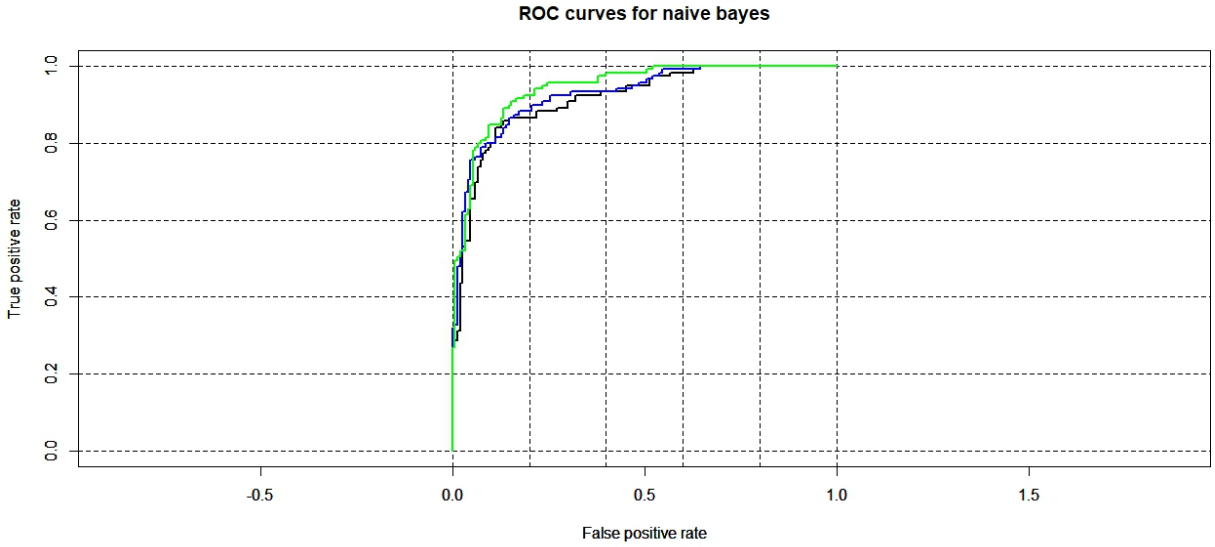
Accuracy: 0.8698885

5.ROC and AUC Scores

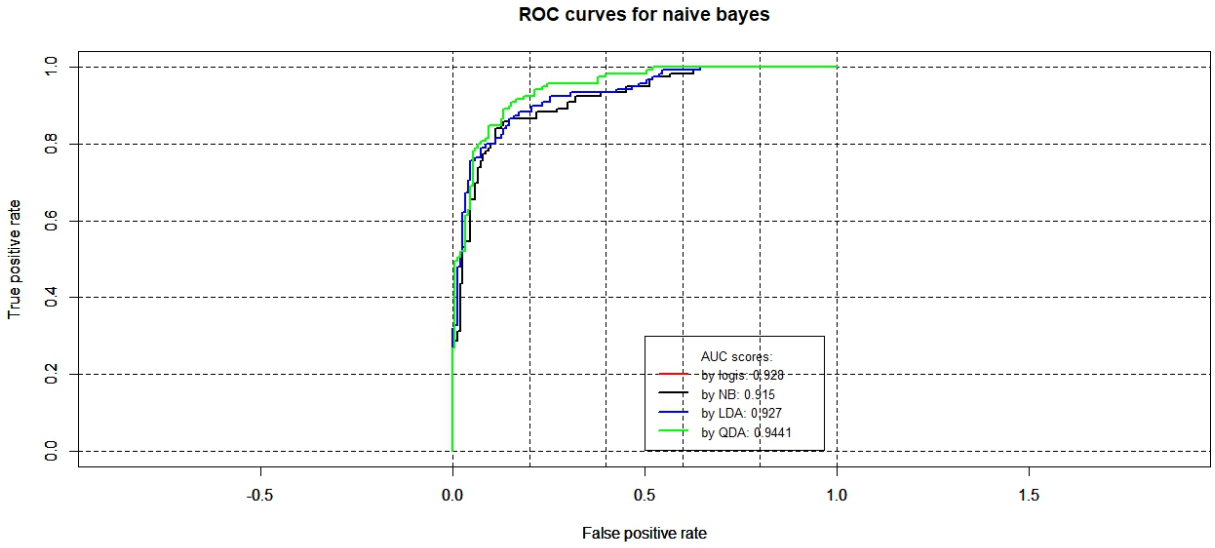
ROC for logistic:



ROC for NB(BLACK), LDA (Blue)and QDA(Green)



AUC:



Result:

Based above accuracy scores and AUC scores we can conclude that QDA is the best model for the statlog(heart) data set.Compare to the other models QDA is more accurate.

. QDA Confusion matrix:

C_QDA 1 2

1 133 18

2 17 101

Accuracy: 0.8698885

AUC score:0.9441