Predicting Diagnosis of Liver Disease

Na Yun Cho

3/29/2021

Introduction

As alcohol consumption, exposure to contaminated food and air, and usage of drugs are rising, there has been an increase in diseases associated with the liver globally. As liver filters the blood that comes from the digestive tract, and detoxifies chemicals and metabolizes drugs, it is one of the most vital organs in our bodies. Therefore, the rise in diseases associated with liver is alarming and should be addressed with importance.

This dataset includes 416 patient records with liver-associated illness and 167 patient records without liver-associated illness. These patient records have been collected from North East of Andhra Pradesh, India. The predictor variables include age of the patient, gender of the patient, total bilirubin, direct bilirubin, alkaline phosphate, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin, albumin and globulin ratio, and the outcome of whether the patient has liver-associated illness or not. In this dataset, there are 414 liver patients and 165 non-liver patients.

In this project, I am trying to use machine learning classification methods to accurately predict whether these patients have liver-associated disease or not based on the predictor variables. To make the most accurate prediction, I will select the model that shows the best performance.
To tidy the data, I recoded the binary classes of the outcome variable as 'yes', which indicates the presence of liver disease and 'no', which indicates the absence of liver disease. I also left out the gender predictor as it is a categorial variable and recoded some other predictor names for easier identification. I also proceeded to drop all NA values in the data.

Exploratory Analysis

Feature plots (one-to-one relation between classes and covariates)
According to the feature plots **(Figure1),** it was seen that the outcome classes are not clearly separated. I was also able to identify that predictors such as albumin-globulin ratio, total bilirubin, direct bilirubin, total proteins, and albumin are heavily skewed to the right. The rest of the predictors showed a relatively normal distribution. Thus, since LDA and QDA would be used for model fitting, I transformed all the predictors on a logarithmic scale for consistency and better performance. After the transformation, the previously skewed predictors showed a more normal distribution.

Correlation plot
The correlation plot (**Figure2**) showed high correlation among some of the predictors. Noticeable correlation was shown between alamine aminotransferase and aspartate aminotransferase, direct bilirubin and total bilirubin, and albumin and albumin-globulin ratio. Intuitively, these correlations make sense since these intercorrelated predictors are related. I kept all these predictors at this point, and decided to check which predictors are more important later by checking their variable importance measures.

Models

To perform classification, logistic regression, linear discriminant analysis, quadratic discriminant analysis, naïve bayes, and k-nearest neighbors methods were used. 80% of the dataset was partitioned into training data and the remaining 20% consisted of test data.

Logistic Regression
First, logistic regression model was fit with all the predictors. The confusion matrix showed an accuracy of 0.7391, no information rate of 0.713, kappa of 0.2216, sensitivity of 0.9390, and specificity of 0.2424. The AUC was 0.756.

In order to identify which predictors play important roles in predicting the response and select these important predictors, I ran the vip function to get the variable importance measures of my predictors. As **Figure 3** shows, 'age', 'albumin', 'aspartate', 'alamine', 'alkaline phosphate', 'total proteins', 'total bilirubin', 'age ratio' predictors show their importance in decreasing order. All of these predictors have already been transformed with logarithmic scale. All of these important predictors were selected for model fitting except 'age ratio', which not only shows the smallest importance measure but also has shown intercorrelation with 'albumin'. Here, the suggestion that intercorrelation between predictors should be avoided as much as possible is being followed.

Then, another logistic regression model was fit with these selected 7 predictors. Its confusion matrix showed an accuracy of 0.7478, no information rate of 0.713, kappa of 0.2557, sensitivity of 0.9390, and specificity of 0.2727. The AUC was 0.757. This improved accuracy of the logistic regression model strengthened the validity of the predictor selection technique, and these selected predictors have been used for all other fitted models.

LDA
The assumption of LDA is that all the variables follow a multivariate normal distribution. Thus, each covariate has its own mean and shares a common variance-covariance matrix. But LDA can still be applied to predictors that do not have normal distributions. The model was fit with the cross-validation method, and the optimal LDA model has a AUC value of 0.752, sensitivity of 0.21, and specificity of 0.93.

QDA
The assumption of QDA is that all the variables follow a Gaussian distribution. Thus, each covariate has its specific mean and different variance-covariance matrix. The optimal QDA model has a AUC value of 0.7395, sensitivity of 0.7839, and specificity of 0.59.

NB
Naïve Bayes is an approximation to the Bayes classifier and assumes that the features are conditionally independent given the class. The best AUC value is 0.749.

KNN
From repeated 10-fold cross validation, the best tuning parameter was chosen to be 106, and the AUC value is 0.529. The predictors were centered and scaled. As a non-parametric model, it tends to require large sample size for better performance and it is more susceptible to overfitting.

Model Comparison

The final model was selected based on cross validation performance, which was assessed by comparing each model’s mean cross-validation AUC. As **Figure 4** shows, the logistic regression model has the highest mean cross-validation AUC (0.754) and LDA has the second highest mean cross-validation AUC (0.751). On the other hand, QDA shows the lowest mean cross-validation AUC (0.7395). The ordering of the models by the mean cross-validation AUC values from highest to lowest is logistic regression, LDA, Naïve Bayes, KNN, and QDA. Because the logistic regression model has the largest AUC, I chose this as the final model.

As shown in **Figure 5**, test set performance was also assessed by comparing each model’s AUC value. Whereas the test AUC values of the logistic regression model and Naïve Bayes are very similar to their mean cross-validation AUC values, the test AUC values of LDA, QDA, and KNN are a bit lower their cross-validation AUC values. The test AUC values are especially lower in KNN and QDA, which indicates overfitting. The ordering of the models by the test AUC values from highest to lowest is logistic regression, Naïve Bayes, LDA, KNN, and QDA.

Thus, the ordering of the models by the mean cross-validation AUC values from highest to lowest does not differ much from the ordering of the models by the test AUC values from highest to lowest.

Conclusion

I was able to find that according to the test AUC values, the logistic regression model has the highest accuracy, which is followed by the Naïve Bayes model, LDA, KNN, and QDA. KNN and QDA show relatively low and similar test AUC values, which are 0.56 and 0.5, respectively. On the other hand, the logistic regression model and the Naïve Bayes model show relatively high and similar test AUC values, which are 0.76 and 0.74. LDA shows lower accuracy than these two models, but still perform better than KNN and QDA.

From these results, I can see that parametric models such as logistic regression, Naïve Bayes, and LDA tend to perform better than KNN, which is a non-parametric model. This better performance of the parametric models can indicate that the assumptions used for fitting these models hold in my dataset.

The relatively good performance of the logistic regression model makes sense since my data meet a lot of the assumptions of this model, some of which are having an outcome that is a binary variable, not having extreme outliers, and not having severe multicollinearity among the predictors. The outcome classes in my data are not well-separated either, which prevents the regression parameter estimates from being unstable. If the outcome classes were well-separated, LDA could have shown better performance than the logistic regression model. In addition, I can also make an insight that there must be a sort of linear relationship between the logit of the outcome and each predictor. The relatively good performance of the Naïve Bayes model also makes sense since the majority of the predictors used for model fitting are independent in each class and this model can still fit well with data that is not too large.

On the other hand, KNN and QDA show bad performance, since they have AUC values around 0.5, which indicates that these models perform as well as guessing at random. This bad performance may be due to the fact that these models are too flexible for this dataset. Additionally, since non-parametric models tend to require large data for good performance, my dataset may not have been large enough.

Figures and Tables

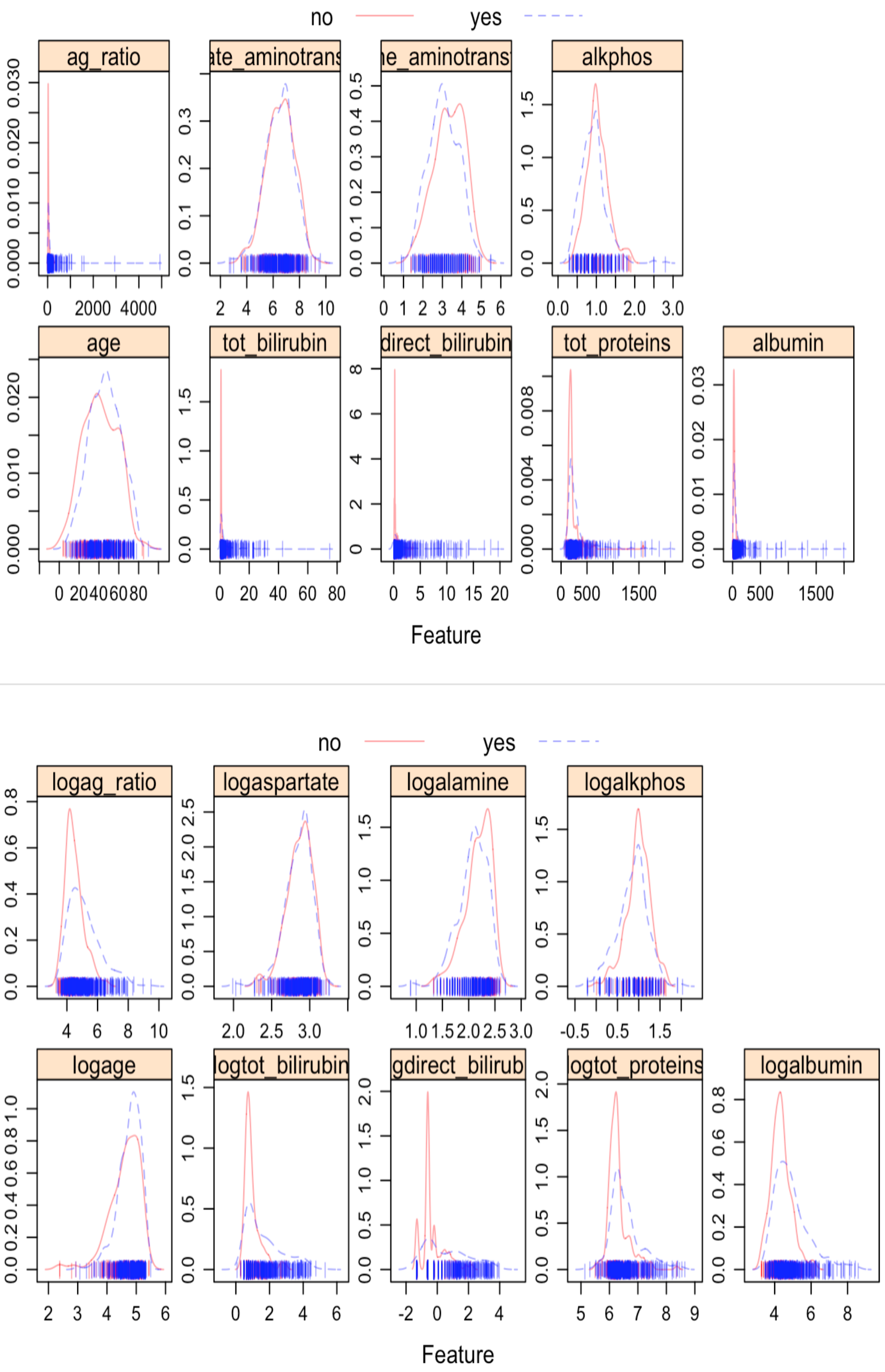

Figure 1 Feature plots before log-transformation(above) and after-log transformation(below) on predictor variables

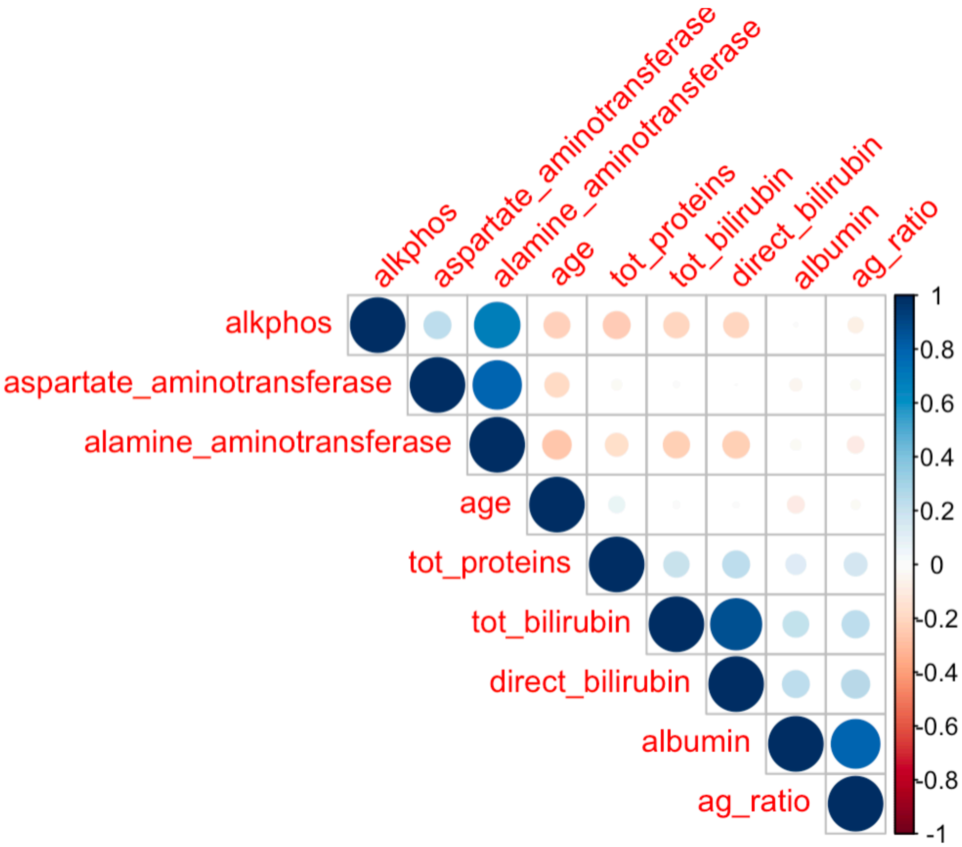

Figure 2 Correlation plot of the predictor variables

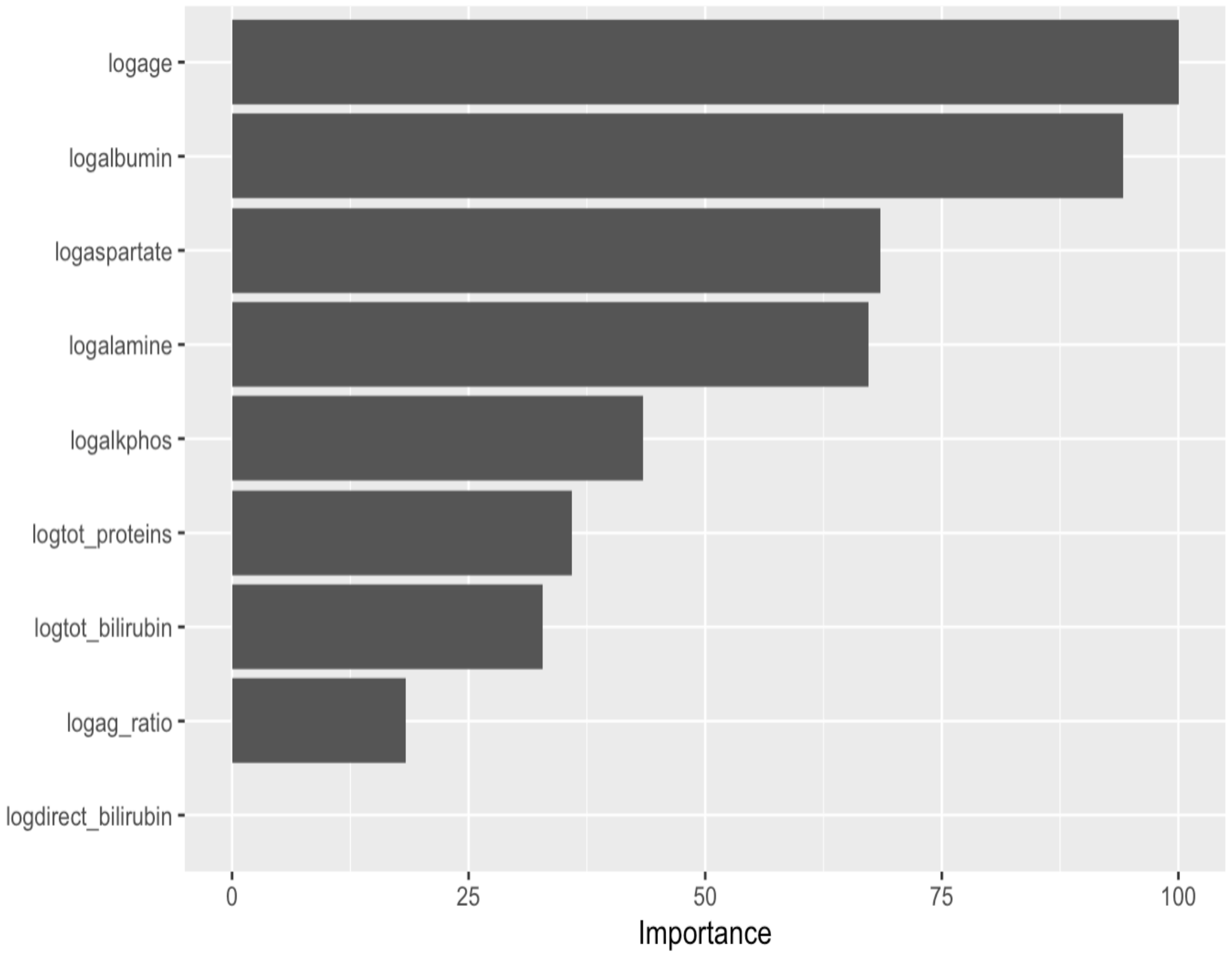

Figure 3 Variable Importance Plot

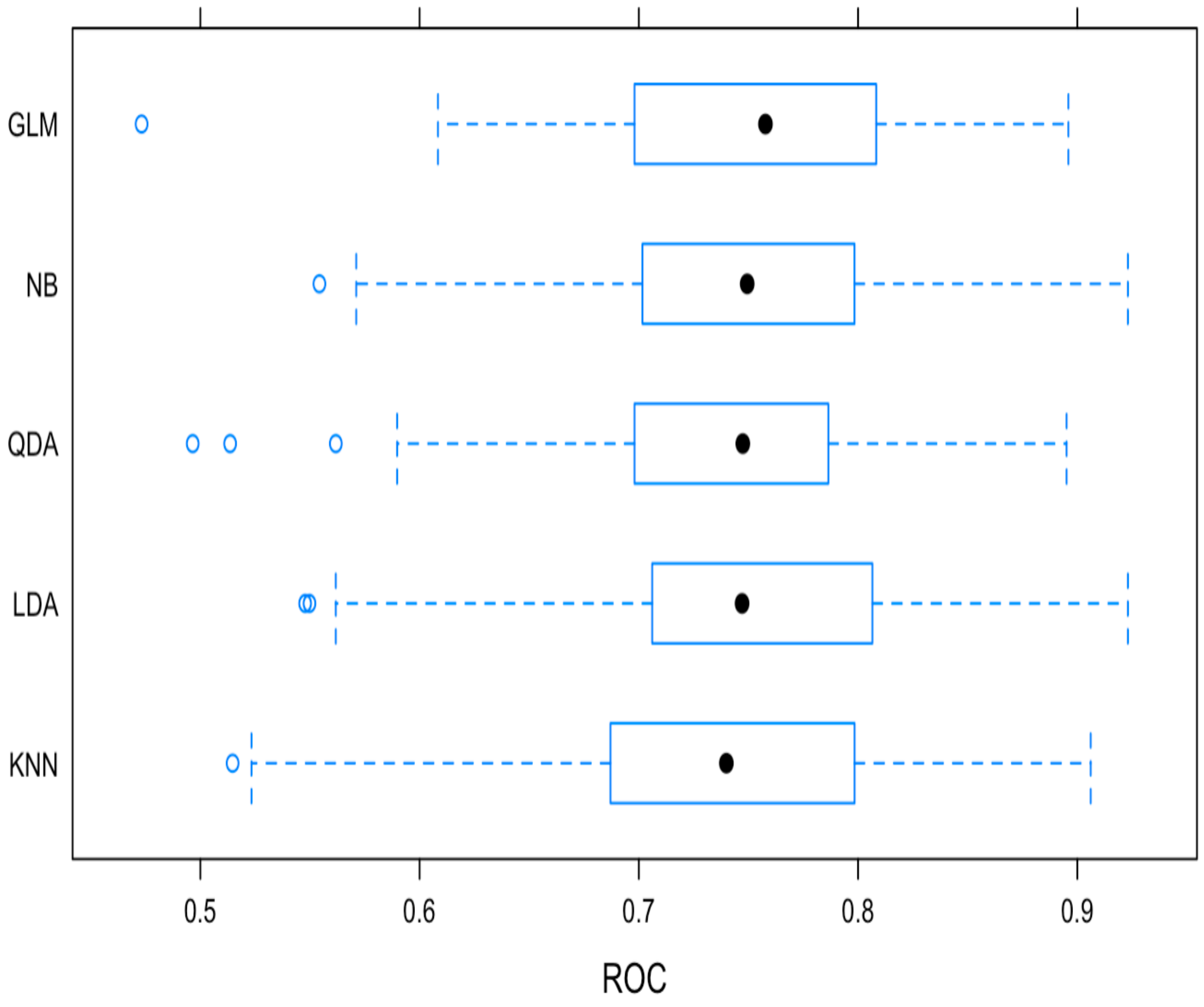

Figure 4 Boxplot of mean cross-validation AUC values of each fitted model

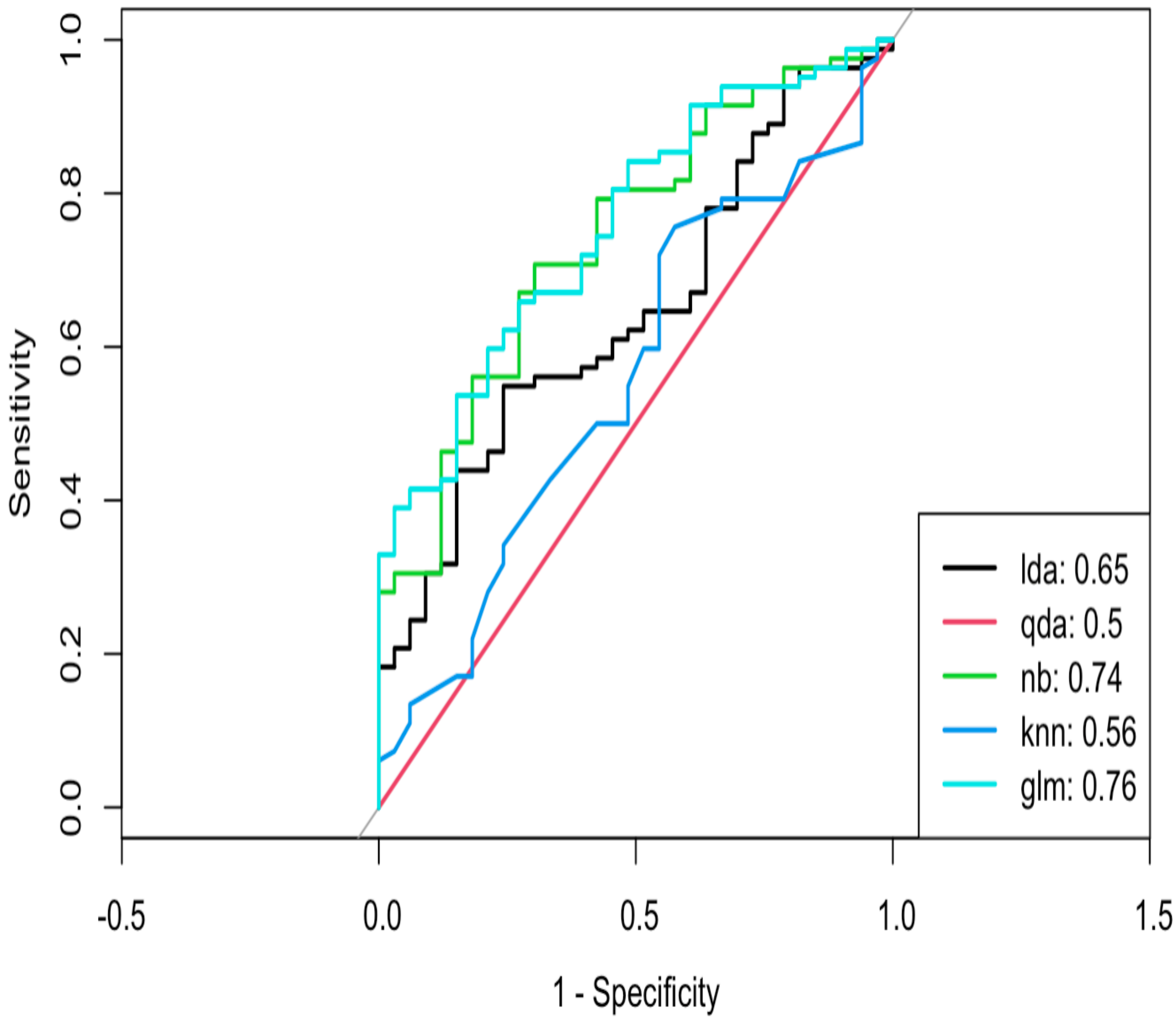

Figure 5 Plot of test AUC values of each fitted model