Indian Liver Patient Prediction

Zhe Chen (zc2556), Group 29, 03/29/2021

Introduction:

The liver is a multifunctional organ that plays an important role in metabolism and digestion, controlling the various synthesis of necessary proteins as well as cleaning poisonous substances through the blood. As one of the most common causes of death worldwide, liver disease is responsible for nearly 2 million deaths per year, contributing to 3.5% of all deaths globally. (Asrani, 2019) Among deaths of liver disease, half of the deaths are due to cirrhosis, the progression of early liver disorder, and the other half is due to viral infections. Therefore, liver disease is one substantial health burden for many countries, especially in central Asia, Africa, and Southeast Asia. (Sepanlou, 2020)

The burden of liver disease is heavy in India because of excessive alcohol consumption, contaminated food intake, and bad air quality. Since early diagnosis can guide early intervention of liver disease, helping decrease the burden of diagnosis, the goal of the study is to train a reliable model to make an accurate prediction for liver disease.

Data Collection:

Data was originally collected from North East of Andhra Pradesh, India, and downloaded from UCI Machine Learning Repository. The dataset includes 416 liver patients and 167 nonliver patients, as well as some medical records which will be predictors in later analysis. Dataset was relatively tidy and clean, and only a few missing values were excluded. Finally, 579 observations were included in the analysis, and for convenience of the analysis, variable names were recorded. (**Table 1**)

Dataset had 10 predictors and 1 outcome variable("status"). Most of the predictors were numeric type except for the "Gender". Since the outcome "status" was binary, classifier methods would be implemented to build the model later. Also, since the majority of the predictors were related metabolism of the liver, all predictors were expected to be included in the model building later.

Since the dataset was relatively small, a 0.8 ratio was selected to split the dataset into training and testing sets. The split was conducted by the "createDataPartition" function to ensure the equal distribution of outcome status.

Exploratory Analysis:

Data was visualized to check the structure. Since "Gender" was a binary predictor while all other predictors were continuous, continuous predictors were visualized first in the feature plot (library "caret"). From the feature plot, bell shape was observed for "tot_pro", "Albumin", "alb_glob" and "Age", but other predictors presented a large extent of skewness. (**Supplement, Exploratory Plot**) Although non-parametric methods generally don't require data to fit assumptions, since parametric methods might be implemented, Box-Cox transformation was implemented to preprocess the

data. After Box-Cox transformation, all continuous predictors presented a bell shape pattern. (**Figure 1**) Detailed information is provided in **Supplement**, **Transformation**. As for the only binary predictor "Gender", a proportional test was conducted to check whether liver disease distributed differently by gender. In the result, the p-value was 0.064, and liver disease distributed evenly was concluded. Details were provided in **Supplement, Exploratory Plot**.

Feature Selection:

Feature selection was performed through an exhaustive fitting of logistics regression. From the exhaustive fitting with full features, 6 features (Age, alk_phosph, ala_aminotrans, tot_pro, Albumin, and alb_glob) showed significance, which matched the result of the variable importance plot. (**Figure 2**) Although the refitted model, with selected features, didn't perform better than the full model in ROC and sensitivity, since the p-value of ANOVA test between the full logistics model and the "smaller" logistics model is significant (0.03), these 6 features were selected for the later model building. Detailed information is presented in **Supplement**, **Feature Selection**.

Methods:

5 models were built to fit this data. They were logistics regression, LDA, QDA, KNN and Naïve Bayes. Since the dataset has had been preprocessed with Box-Cox transformation, the assumptions of multinormal distribution of LDA and QDA were satisfied. Logistics regression does not require many assumptions, as long as observations are selected independently, and predictors are independent with each other. Similarly, KNN, as a non-parametric method, and Naïve Bayes both do not have many assumptions for the dataset to satisfy. However, since some predictions are extremely skewed, which might influence the classifiers to distinguish clearly, Box-Cox transformed data was applied to all 5 models.

All five models were trained by the "train" function from "caret" library with repeated cross-validation for 10 repeats and seeds were stabilized to "621". Evaluation standard was selected to "ROC" for all models. KNN's "k" value was selected automatically by the "train" function and Naïve Bayes' tuning number was selected to 2.8, according to the plot. (**Figure 3**)

Detailed fitting information can be found in **Supplement**.

Results:

5 models were evaluated in the training set as well as the testing set. For the training set, five models were resampled with the function "resamples" and the comparison was visualized in a boxplot. (**Figure 4**) In general, all five models had satisfactory performance. The mean ROC for the five models was above 0.65. Among these models, LDA, logistics regression, and Naïve Bayes had a little bit better performance than others, which closed to 0.75. As for the sensitivity, QDA had the best performance whose mean was around 0.7, following by Naïve Bayes and logistics regression.

Lastly, as for specificity, LDA, logistics regression and Naïve Bayes had the top 3 performance, which closed to 0.9. Boxplots of sensitivity and specificity can be found in **Supplement**, **Box Plot of Resamples**.

Then models were evaluated in the test dataset with a ROC plot. (**Figure 5**) Logistics regression and Naïve Bayes had an acceptable result (above 0.7) while the other three models showed AUC values closed to 0.5, meaning the performance of LDA, QDA, and KNN models did not surpass random guess. Thus, the logistics regression model and Naïve Bayes had the best flexibility among the five models.

Conclusion and Discussion:

From the comparison in the training set, it is hard to select the best model that has the best performance for all three criteria (ROC, sensitivity, and specificity). However, since sensitivity and specificity are balanced criteria, cannot be high for both, selecting the best model is situation-dependent. If the dataset collected patients with infectious liver disease, ROC and specificity become important since high specificity prevents the spread of infection. But for chronic disease, high sensitivity is important because early intervention is vital to stop the progression. If the model is used to predict all types of liver disease, the Naïve Bayes model might be selected since it performs relatively well in all three criteria.

As for the comparison in the testing set, we had some interesting observations. LDA, QDA, and KNN all had an unexpected performance (**Figure 5**) These three models' bad performance presented the sign of overfitting since they all had highly repetitive predicted values on the testing set. KNN is a non-parametric method; LDA is a more parametric and less non-parametric method; QDA is in between. It is quite risky to conclude that the three models' bad performance is due to their non-parametric properties, but it might because the test set has some pattern that these relative non-parametric methods do not work well. Also, a small dataset contributes to the overfitting of these relative non-parametric models. Naïve Bayes seems to be the most robust model among these five models.

Additionally, another criterion to consider is the confusion matrix of these models. Detailed information is available in **Supplement, Confusion Matrix**. The confusion matrix can provide more detailed information than the ROC curve. More importantly, the confusion matrix is more useful when a certain threshold is needed to be fixed.

Reference

1. Asrani, S. K., Devarbhavi, H., Eaton, J., & Kamath, P. S. (2019). Burden of liver diseases in the world. *Journal of hepatology*, *70*(1), 151–171. https://doi.org/10.1016/j.jhep.2018.09.014.

2. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

3. Sepanlou, SG; Safiri, S; Bisignano, C; Ikuta, KS; Merat, S; Saberifiroozi, M; Poustchi, H; ... Malekzadeh, R.etc.. (2020) The global, regional, and national burden of cirrhosis by cause in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Gastroenterology and Hepatology*, 5 (3) pp. 245-266. 10.1016/S2468-1253(19)30349-8.

Table and Figures

Table 1.Variable Description

Variable Name	Recoded Name	Description	Type
Age	Age	age of patients	Numeric
Gender	Gender	gender of patients	Factor
Total_Bilirubin	tot_bili	test of total bilirubin in blood	Numeric
Direct_Bilirubin	dir_bili	test of direct bilirubin in blood	Numeric
Alkaline_Phosphotase	alk_phosph	test of alkaline phosphatase in blood	Numeric
Alamine_Aminotransferase	ala_aminotrans	test of alanine aminotransferase (an enzyme) in blood	Numeric
Aspartate_Aminotransferase	asp_aminotrans	test of aspartate aminotransferase (an enzyme) in blood	Numeric
Total_Protein	tot_pro	test of total proteins in blood	Numeric
Albumin	Albumin	test of albumin in blood	Numeric
Albumin_and_Globulin_Ratio	alb_glob	the ratio of albumin and globulin	Numeric
Dataset	**status**	**disease status**	**Factor**

Figure 1. Feature Plots of Continuous Predictors after Box-Cox Transformation

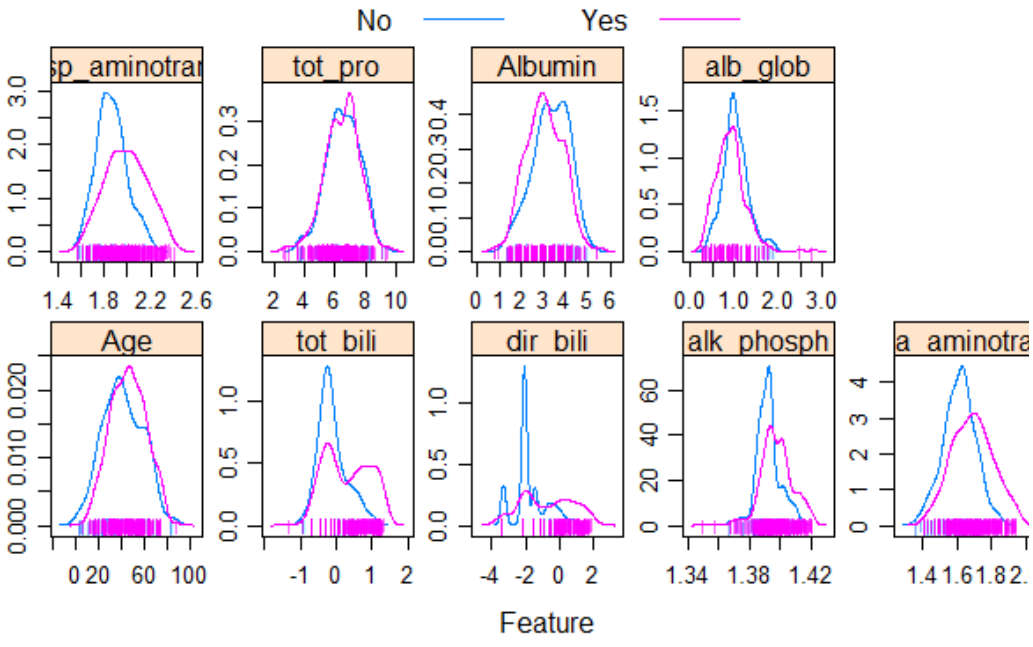


Figure 2. Variable Importance Plot

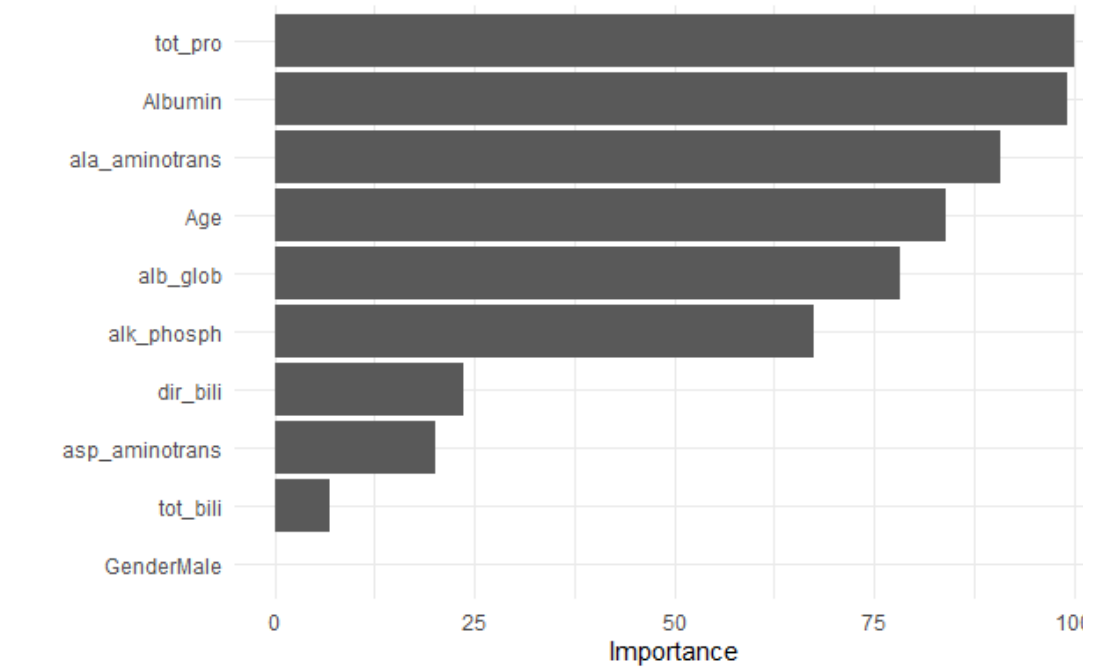


Figure 3. ROC vs Bandwidth Adjustment Plot

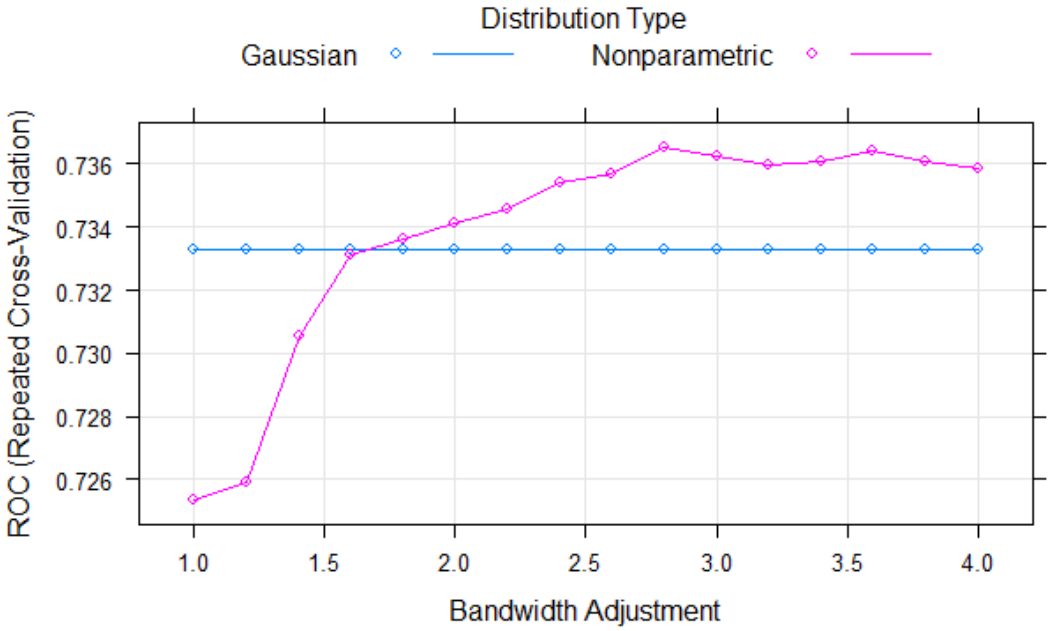


Figure 4. Boxplot of Models Comparison (ROC)

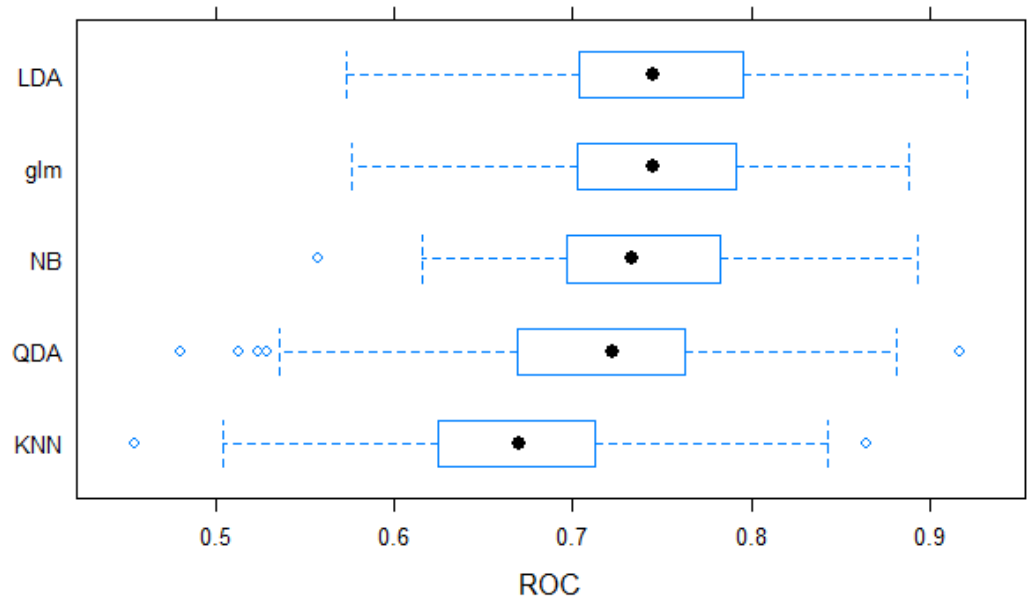
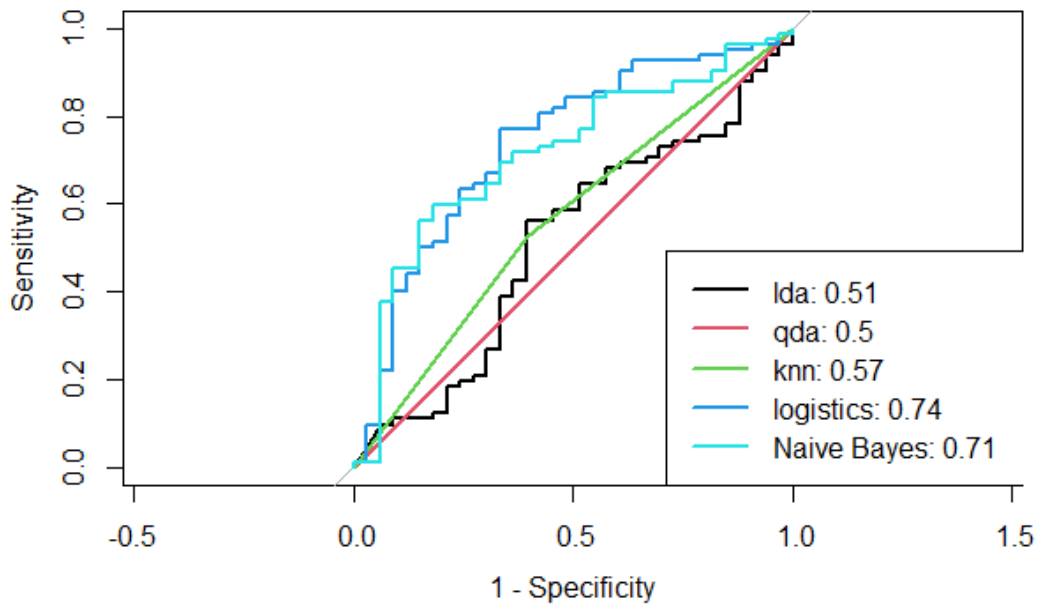


Figure 5. ROC Plot for 5 Models in Test Set


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
-0.5
0.0
0.5
1.0
1.5
1 - Specificity
lda: 0.51
qda: 0.5
knn: 0.57
logistics: 0.74
Naive Bayes: 0.71