Predicting Hepatitis C Virus Patient Using Multiple Classification Methods

P8106 Data Science 2 Final

Jibei Zheng, Renjie Wei, Shihui Zhu

2022-05-09

Contents

1 Introduction

1.1 Data Source

The original purpose of the research where the data set was built was to replace liver biopsy for disease staging. In the study, multiple serum markers in this dataset are under evaluation with multi-parametric panels yielding the most promising results[1,2].

1.2 Motivation

According to the Centers for Disease Control and Prevention (CDC): Hepatitis C is a liver infection caused by the hepatitis C virus (HCV). Hepatitis C is spread through contact with blood from an infected person. Today, most people become infected with the hepatitis C virus by sharing needles or other equipment used to prepare and inject drugs. For some people, hepatitis C is a short-term illness, but for more than half of people who become infected with the hepatitis C virus, it becomes a long-term, chronic infection. Chronic hepatitis C can result in serious, even life-threatening health problems like cirrhosis and liver cancer. People with chronic hepatitis C can often have no symptoms and don't feel sick. When symptoms appear, they often are a sign of advanced liver disease. There is no vaccine for hepatitis C. The best way to prevent hepatitis C is by avoiding behaviors that can spread the disease, especially injecting drugs. Getting tested for hepatitis C is important, because treatments can cure most people with hepatitis C in 8 to 12 weeks[3].

Creating a predictive model that could perform early detection of Hepatitis C and other liver diseases would allow people to quickly and easily determine their risk/get treatment.

1.3 Data Description

The data contains 615 observations and 13 attributes of blood donors and Hepatitis C patients laboratory (10 laboratory results) and demographic values (age and gender), as well as a subject Category indicator. All attributes except the outcome indicator Category (blood donors vs. Hepatitis C, including its progress-Hepatitis C, Fibrosis, Cirrhosis) and Sex are numerical. Package `tydiverse` were used to clean data and transform data types for analysis convenience, package `caret` were used to partitioning data to training and testing set, 70% of the data to be train data and the 30% rest to be test data. I investigated for abnormal values in laboratory values, as well as the missing data. The descriptive analysis is shown in Table 1.

Table 1: **Summary of Dataset**

Variable	**N**	**Overall**, N = 615	**0=Blood Donor**, N = 533	**0s=suspect Blood Donor**, N = 7	**1=Hepatitis**, N = 24	**2=Fibrosis**, N = 21	**3=Cirrhosis**, N = 30	**p-value**
Age	615	47 (39, 54)	47 (39, 53)	55 (48, 65)	37 (32, 47)	51 (48, 57)	56 (46, 59)	<0.001
Sex	615							0.10
f		238 (39%)	215 (40%)	1 (14%)	4 (17%)	8 (38%)	10 (33%)	
m		377 (61%)	318 (60%)	6 (86%)	20 (83%)	13 (62%)	20 (67%)	
ALB	614	42.0 (38.8, 45.2)	42.2 (39.2, 45.4)	21.6 (19.8, 23.7)	43.5 (41.8, 46.2)	41.0 (39.0, 45.0)	33.0 (29.0, 36.0)	<0.001
Missing/NA		1	0	0	0	0	1	
ALP	597	66 (52, 80)	67 (55, 80)	106 (77, 120)	35 (31, 40)	40 (33, 43)	80 (49, 104)	<0.001
Missing/NA		18	0	0	3	9	6	
ALT	614	23 (16, 33)	23 (17, 32)	49 (21, 144)	15 (10, 40)	34 (10, 114)	6 (4, 25)	<0.001
Missing/NA		1	0	0	1	0	0	
AST	615	26 (22, 33)	25 (21, 30)	47 (34, 113)	47 (38, 82)	70 (43, 106)	93 (60, 120)	<0.001
BIL	615	7 (5, 11)	7 (5, 10)	5 (2, 6)	13 (8, 16)	13 (10, 15)	34 (14, 56)	<0.001
CHE	615	8.26 (6.94, 9.59)	8.35 (7.10, 9.62)	5.33 (4.33, 10.07)	9.51 (7.40, 10.15)	8.59 (7.28, 9.45)	3.42 (1.80, 5.72)	<0.001
CHOL	605	5.30 (4.61, 6.06)	5.40 (4.70, 6.17)	4.30 (3.10, 4.98)	5.06 (4.12, 5.80)	4.58 (4.20, 4.92)	3.87 (3.58, 4.59)	<0.001
Missing/NA		10	7	0	0	1	2	
CREA	615	77 (67, 88)	78 (69, 89)	52 (30, 70)	72 (62, 81)	71 (65, 79)	68 (61, 102)	0.003
GGT	615	23 (16, 40)	21 (15, 32)	83 (55, 257)	46 (34, 91)	72 (53, 95)	96 (50, 141)	<0.001
PROT	614	72.2 (69.3, 75.4)	72.2 (69.4, 75.2)	47.8 (47.4, 55.9)	73.7 (71.0, 77.1)	76.1 (72.3, 80.9)	70.0 (65.3, 77.0)	<0.001
Missing/NA		1	0	0	0	0	1	

2 Exploratory analysis/visualization

TODO

3 Modeling

3.1 Data Preprocessing

The data is already prepared in .csv table format. However, the dataset has 31 missing values and we consider them to be missing-at-random (MAR). We implemented 5-nearest neighbor imputation on the original data to accommodate those missing values, then applied Box Cox transformation and standardization (center and scale) to re-scale the numerical covariates so they are approximately normally distributed. This normalization helps us reduce the influence from outliers and achieve better classification result.

3.2 Predictors

In the modeling part, all variables were included and there wasn't variable selection procedure prior to modeling process. The target response variable is Category, which was recoded as 0 and 1, representing healthy blood donors and kinds of liver diseases patients.

Specifically, here shows the predictors in the models: (1) `Age`: age of the patient in years; (2) `Sex`: sex of the patient; (3) `ALB`: amount of albumin in patient's blood; (4) `ALP`: amount of alkaline phosphatase in patient's blood; (5) `ALT`: amount of alanine transaminase in patient's blood; (6) `AST`: amount of aspartate aminotransferase in patient's blood; (7) `BIL`: amount of bilirubin in patient's blood; (8) `CHE`: amount of cholinesterase in patient's blood; (9) `CHOL`: amount of cholesterol in patient's blood; (10) `CREA`: amount of creatine in patient's blood; (11) `GGT`: amount of gamma-glutamyl transferase in patient's blood; (12) `PROT`: amount of protein in patient's blood;

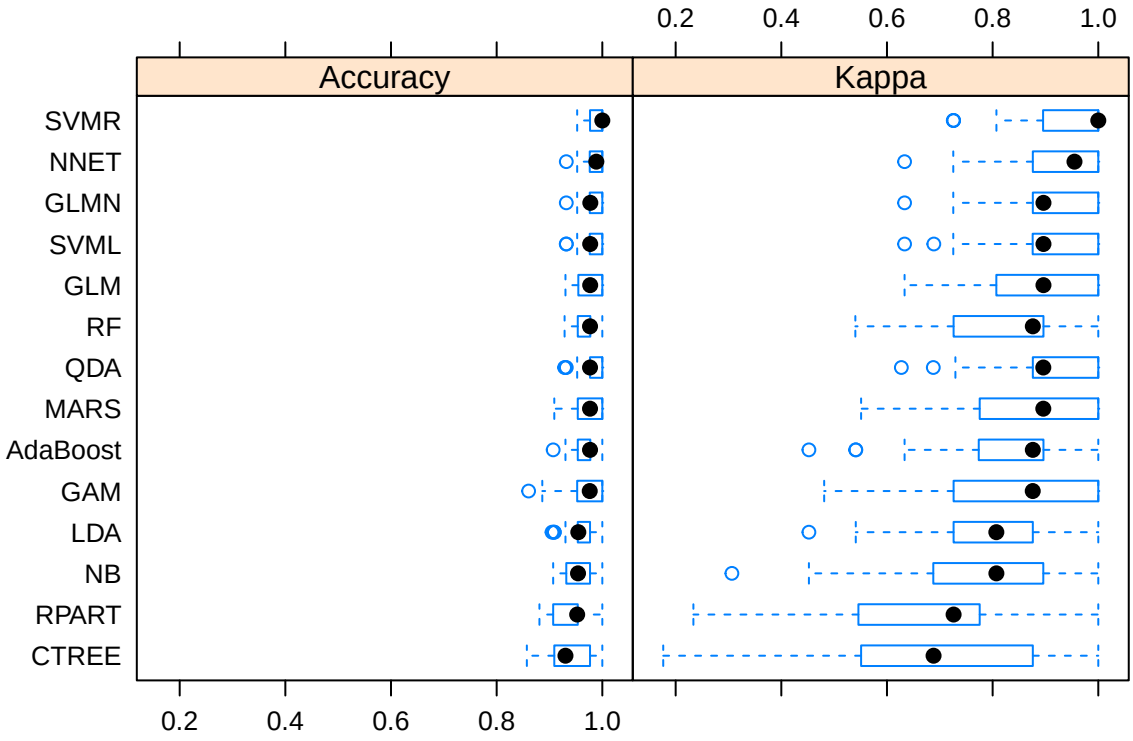
3.2 Used Techniques

models ...

Since the traditional approach in HCV diagnostic pathways is related to decision tree, two tree methods were used: the conditional inference trees (CTREE) and regression tree (RPART). The tree methods can be displayed graphically and more easily understood by physicians. Taking the diagnosis of HCV as a classification problem, other machine learning models were also trained for the purpose our project. We used the generalized additive model model (GAM), the generalized linear regression models, GLMNET (with penalization) and GLM (without penalization), linear and quadratic discriminant analysis models (LDA, QDA), as well as naive bayes (NB). **TODO We also implemented.....** All the models were trained using the package *caret*. The regression models (GLM, LDA, etc. al.) can accept mixture of variables which is suitable for our case. NB is useful when predictor number is large. GAM model can include any quadratically penalized GLM and a variety of other models, which induces great flexibility. Linear regression model also assumed the independence of the predictors.

3.3 Tuning parameters

3.4 Training Performance



3.5 Test performance

Models	Testing Error Rate
GLMNET	0.03261
GLM	0.03261
MARS	0.03804

Models	Testing Error Rate
GAM	0.02717
LDA	0.03804
NB	0.04891
QDA	0.03804
CTREE	0.06522
RPART	0.06522
RF	0.02717
AdaBoost	0.02717
SVML	0.02174
SVMR	0.02174
NNET	0.02717

3.6 Variable Importance

4 Limitations

5 Conclusion

6 Bibliography

1. Hoffmann, G., Bietenbeck, A., Lichtinghagen, R., & Klawonn, F. (2018). Using machine learning techniques to generate laboratory diagnostic pathways—a case study. Journal Of Laboratory And Precision Medicine, 3(6).

2. Lichtinghagen, R., Pietsch, D., Bantel, H., Manns, M. P., Brand, K., & Bahr, M. J. (2013). The Enhanced Liver Fibrosis (ELF) score: normal values, influence factors and proposed cut-off values. Journal of hepatology, 59(2), 236–242. https://doi.org/10.1016/j.jhep.2013.03.016

3. https://www.cdc.gov/hepatitis/hcv/index.htm

7 Appendix