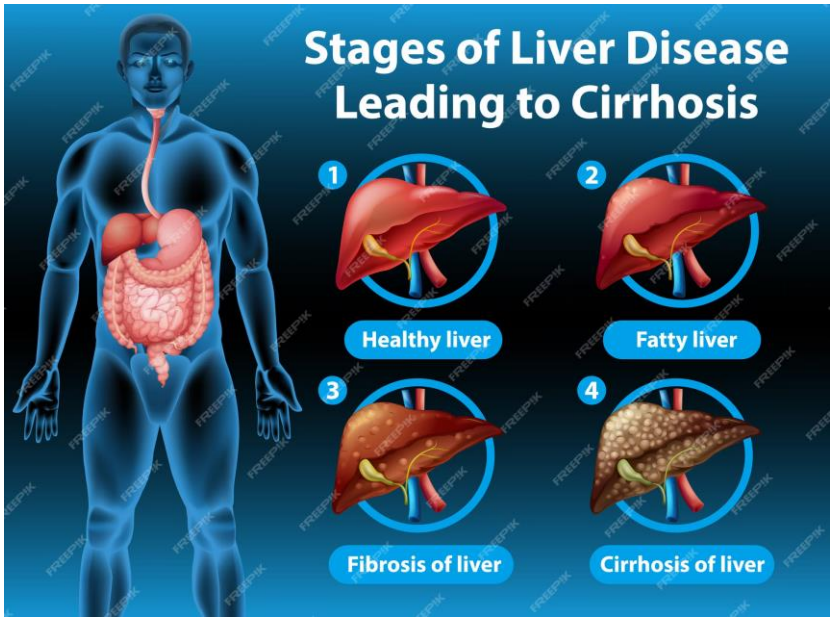
Good 75

MATH 4322
Group Project Final
Group 6

Students
Le Bui
Ethan Pradhan
Khoi Phan
Thien Ngo

Cirrhosis Prediction



December 26, 2024

1) Introduction

The data set contains a record of 418 patients with cirrhosis including their biological characteristics (Age and Sex), clinical signs (Ascites, Hepatomegaly, Spiders, and Edema), and blood tests (Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos, SGOT, Triglycerides, Platelets, and Prothrombin). Those 15 variables are related to liver function.

However, the cirrhosis stage is classified based on the histologic stage, which is an invasive method. Therefore, we are interested in finding out if late stages (3 and 4) can be predicted based on non-invasive methods since the identification of early and late-stage cases of cirrhosis may relate closely to a patient's survival outcome.

To accomplish this goal, we require some preprocessing before we can use it to fit our models. The dataset contained multiple missing values; therefore we had to use the na.omit() function to drop entries containing missing data. We also had to map the **Stage** variable to **Early** (1 and 2) or **Late** (3 and 4) stages, since in the dataset it was classified into stages 1, 2, 3, or 4 and we would like to use a logistic regression model. Finally, we excluded variables such as patient ID, number of days, status, and drug since those variables are unrelated to our goal.

After cleaning the data set, we randomly split out data into 80% of the training set, and 20% of the test set for ease of exploring the relationships between variables, constructing models, and performing cross-validation.

ID: unique identifier for each instance (person) **N_Days**: number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986 **Status**: status of the patient C (censored), CL (censored due to liver tx), or D (death) **Drug**: type of drug D-penicillamine or placebo **Age:** age in days **Sex:** M (Male) of F (Female) **Ascites:** presence of ascites N (No) or Y (Yes) **Hepatomegaly:** presence of hepatomegaly N (No) or Y (Yes) **Spiders:** presence of spiders N (No) or Y (Yes	**Edema:** presence of edema N (no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy) **Bilirubin**:serum bilirubin in [mg/dl] **Cholesterol**:serum cholesterol in [mg/dl] **Albumin**:albumin in [gm/dl] **Copper**:urine copper in [ug/day] **Alk_Phos**:alkaline phosphatase in [U/liter] **SGOT**: SGOT in [U/ml] **Tryglicerides**: triglicerides **Platelets:** platelets per cubic [ml/1000] **Prothrombin:** prothrombin time in seconds [s] **Stage:** histologic stage of disease (1, 2, 3, or 4)

The question we want to answer: Are non-invasive measurements like clinical examinations and blood tests useful for predicting the early and late **Stages** of cirrhosis? And if yes, which predictors relate closely to stages of cirrhosis?

2) Methods

DECISION TREE *(Le Bui, Khoi Phan):*

Reasons for decision tree in this project: more closely mirrors human decision-making than regression and certain other methods and able to do a binary or multi-class classification task. The decision tree can provide the most significant predictors in its output.

Advantages:

- Easy to explain, visualize, and interpret.
- More closely mirror human decision-making than regression and certain other methods.
- Easily handle both quantitative and qualitative predictors (and responses).

Disadvantages:

- Trees generally do not have the same level of predictive accuracy as some other regression and classification approaches.
- They struggle with prediction accuracy.
- They suffer from high variance - different subsets of the same data could yield greatly different results.

a) Model's formula:

The training data was obtained by randomly splitting the whole dataset into 80% of the training set and 20% of the test set.

We use all predictors, except "ID", "N_Days", "Status", and "Drug" in the model since those variables are not related.

- Stage ~ Age + Sex + Ascites + Hepatomegaly + Spiders + Edema + Bilirubin + Cholesterol + Albumin + Copper + Alk_Phos + SGOT + Triglycerides + Platelets + Prothrombin
- R codes: tree_model <- tree(Stage ~ . , data = train_data_unique)

b) The thought process of our considerations while fitting the model

The decision tree uses 80% of the data set for training purposes and the rest 20% for testing.

There are 15 predictors used as input, but only Hepatomegaly, Copper, SGOT, Platelets, Cholesterol, Bilirubin, Age, Albumin, and Prothrombin are used by the tree.

However, one of the disadvantages of decision trees is that they struggle with prediction accuracy and suffer from high variance since different subsets of the same data could yield greatly different results. The purpose of pruning in decision trees is to improve the model's ability to generalize on unseen data by preventing overfitting.

Therefore, we used cv.tree function to perform cross-validation and pruning to get the tree with the optimal size corresponding to the deviance that yields the best goodness of fit for the tree model. After performing such actions, we got a smaller pruned tree and used that to make predictions on the test dataset again.

The result from the R script shows that the original tree obtains a training error as low as 0.1 while the pruned tree obtains a training error of 0.1682. However, the test error of the original tree is 0.3035714 while the test error of the pruned tree is as low as 0.1785714.

c) Randomly subdivide your full data set in 80% for training, and 20% for testing.

Before implementing our 10-time subdivision loop, we acquired the test error from the pruned tree of 0.1786, which is much better than the unpruned tree with a test error of 0.3035714 (as mentioned above). However, when randomly splitting the tree (80% for training, 20% for testing) 10 times, we obtain the test result with high variance, and the average test error is 0.3054, which is a drawback of the decision tree:

```
> summary(test_error_tree)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2143  0.2545  0.3036  0.3054  0.3527  0.3929
```

LOGISTIC REGRESSION MODEL *(Thien Ngo, Ethan Pradhan)*:

Reason for logistic regression: This model focused on its simplicity in finding the response variables in terms of relationships between the disease stage and the patient's other predictors. The model is also computationally efficient and performs well with a small-size dataset.

Advantages:

- Simplicity interpretability.
- Good with small data sets.
- Many of the predictors in the dataset, such as Age, Bilirubin, Albumin, and Prothrombin, are continuous or binary variables, which logistic regression can handle efficiently.
- Specializes in binary classification.

Disadvantages:

- Cannot handle large data sets and extreme outliers in the data set.
- Limited in interpreting complex patterns.

a. Model Formula:

We attempted to fit the logistic regression model using all predictors, except patient ID, number of days, status, and drug as explained above.

- fit.bc_all <- glm(Stage ~ ., family = "binomial", data = train_data_unique)

There is an issue occurring when fitting the logistic regression model using all predictors in the cleaned data. As explored from exploratory data analysis, the Ascites variable has the perfect separation issue. Therefore, we exclude Ascites in our logistic regression model and use all of the other variables.

Fitting logistic regression model using all predictors but Ascites:

- fit.bc <- glm(Stage ~ . -Ascites, family = "binomial", data = train_data_unique)

summary(fit.bc)

The preliminary model formula for logistic regression:

- Stage ~ Age + Sex + Hepatomegaly + Spiders + Edema + Bilirubin + Cholesterol + Albumin + Copper + Alk_Phos + SGOT +Triglycerides + Platelets + Prothrombin

$$P(X) = \frac{e^{(\hat{\beta}_0+\hat{\beta}_1\times Hepatomegaly+\hat{\beta}_2\times Copper+\hat{\beta}_3\times Spiders)}}{1+e^{(\hat{\beta}_0+\hat{\beta}_1\times Hepatomegaly+\hat{\beta}_2\times Copper+\hat{\beta}_3\times Spiders)}}$$

After using the stepwise function, we discovered that Hepatomegaly, Copper, and Spiders are significant predictors. Therefore, we exclusively use those predictors for model formula interpretation. P(X) means the probability of Stage, where an observation has a probability of 0.5 or less is classified as Early stage. Otherwise, it is classified as a Late state.

b. Thought Process:

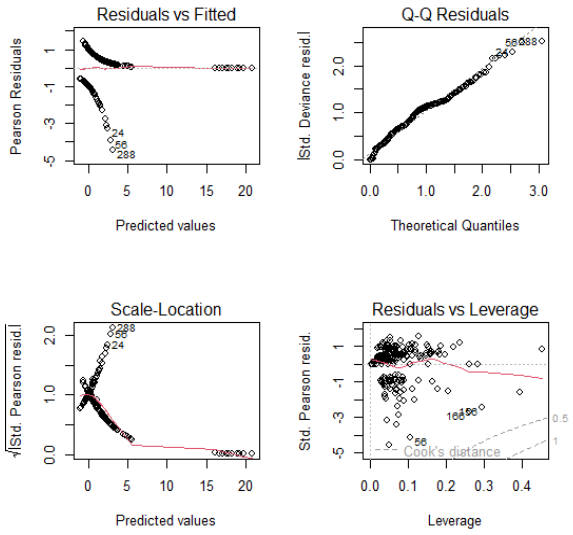
At first, we excluded variables such as patient ID, number of days, status, and drug since those variables are irrelevant to our goal and the patient's medical history.

Secondly, the Ascites variable has the perfect separation issue, so we exclude it.

```
> fit.bc_all <- glm(Stage ~ .,
+                   family = "binomial", data = train_data_unique)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Thirdly, a logistic model was trained with all the predictors, excluding the Ascities variable due to the perfect separation issue. After fitting the model with our training set, we received a model like this:

```
Call:
glm(formula = Stage ~ . - Ascites, family = "binomial", data = train_data_unique)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.621e+00  3.083e+00  -0.526 0.599041
Age           -1.890e-02  1.919e-02  -0.985 0.324742
SexM          -4.135e-01  6.256e-01  -0.661 0.508570
HepatomegalyY  1.599e+00  4.256e-01   3.757 0.000172 ***
SpidersY       7.219e-01  5.185e-01   1.392 0.163851
EdemaS         1.077e+00  8.673e-01   1.241 0.214450
EdemaY         1.657e+01  1.008e+03   0.016 0.986884
Bilirubin     -8.840e-02  6.753e-02  -1.309 0.190525
Cholesterol    4.582e-04  1.106e-03   0.414 0.678603
Albumin        2.225e-01  5.222e-01   0.426 0.670006
Copper         9.299e-03  4.273e-03   2.176 0.029535 *
Alk_Phos      -8.225e-05  8.844e-05  -0.930 0.352344
SGOT          -1.115e-03  3.788e-03  -0.294 0.768484
Trygliceride   3.784e-03  3.836e-03   0.986 0.323973
Platelets     -2.162e-03  2.048e-03  -1.055 0.291288
Prothrombin    1.584e-01  1.892e-01   0.837 0.402427
```



The diagnostic plot shows that the assumption of linearity and homoscedasticity is violated, but normality is pretty clear.

Looking at the previous model, we were able to define the relationship between the predictors and the response variable as either positive or negative and found the most significant predictors to be Hepatomegaly, Copper, and perhaps Spiders (though it has a p-value > 0.05). To confirm this, we ran forward, backwise, and bidirectional stepwise regression and found the best predictors to once again be Hepatomegaly, Copper, and Spiders, but with the addition of Edema. Although this model proved to result in the best AIC, we found the p-value of Edema to be too high for our comfort especially since EdemaY is close to 1, therefore we excluded it in the final model and ended with a summary like this:

```
> summary(step.logistic)

Call:
glm(formula = Stage ~ Hepatomegaly + Spiders + Edema + Copper,
    family = "binomial", data = train_data_unique)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.204e-01  2.978e-01  -1.076  0.28187
HepatomegalyY  1.425e+00  3.791e-01   3.760  0.00017 ***
SpidersY       8.346e-01  4.971e-01   1.679  0.09314 .
EdemaS         7.976e-01  8.163e-01   0.977  0.32851
EdemaY         1.591e+01  1.038e+03   0.015  0.98777
Copper         6.353e-03  3.024e-03   2.101  0.03566 *
```

After all, we fit the three predictors:

```
> step.model = glm(Stage ~ Hepatomegaly + Copper + Spiders, family='binomial', data = train_data_unique)
> summary(step.model)

Call:
glm(formula = Stage ~ Hepatomegaly + Copper + Spiders, family = "binomial",
    data = train_data_unique)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.288013   0.295031  -0.976 0.328960
HepatomegalyY  1.456640   0.375599   3.878 0.000105 ***
Copper         0.006864   0.003031   2.265 0.023542 *
SpidersY       0.950000   0.490317   1.938 0.052682 .
```

The model formula is:

$$P(\text{Stage}|\text{Hepatomegaly}, \text{Copper}, \text{Spiders}) = \frac{\exp(-0.288013 + 1.456640 \cdot \text{HepatomegalyY} + 0.006864 \cdot \text{Copper} + 0.950000 \cdot \text{SpidersY})}{1 + \exp(-0.288013 + 1.456640 \cdot \text{HepatomegalyY} + 0.006864 \cdot \text{Copper} + 0.950000 \cdot \text{SpidersY})}$$

The test error obtained from this subset is 0.2321429.

c. Randomly subdivide your full data set in 80% for training, and 20% for testing.

Our implementation for this task uses a for loop, which is attached to the R script. For each iteration, we randomly sampled 80% of our dataset to be used for training purposes and used the rest of the dataset to validate our model and store the test error rates. Also, in the loop, we construct the logistic model using Hepatomegaly, Spiders, and Copper, which is obtained from stepwise results.

The error rates range from 14.29% to 32.14%, with a mean error rate of 24.5%. With this result, we can use this model for predicting the cirrhosis stage.

3) Results of our methods: (Le Bui, Khoi Phan)

From logistic regression: the mean test error is 0.2446, the lowest is 0.1429 and the highest is 0.3214.

From the decision tree: the mean test error (of 10 different pruned trees) is 0.3054, the lowest is 0.2143 and the highest is 0.3929.

Therefore, our model is somewhat good to use for cirrhosis stage classification. We will discuss the most significant predictors below.

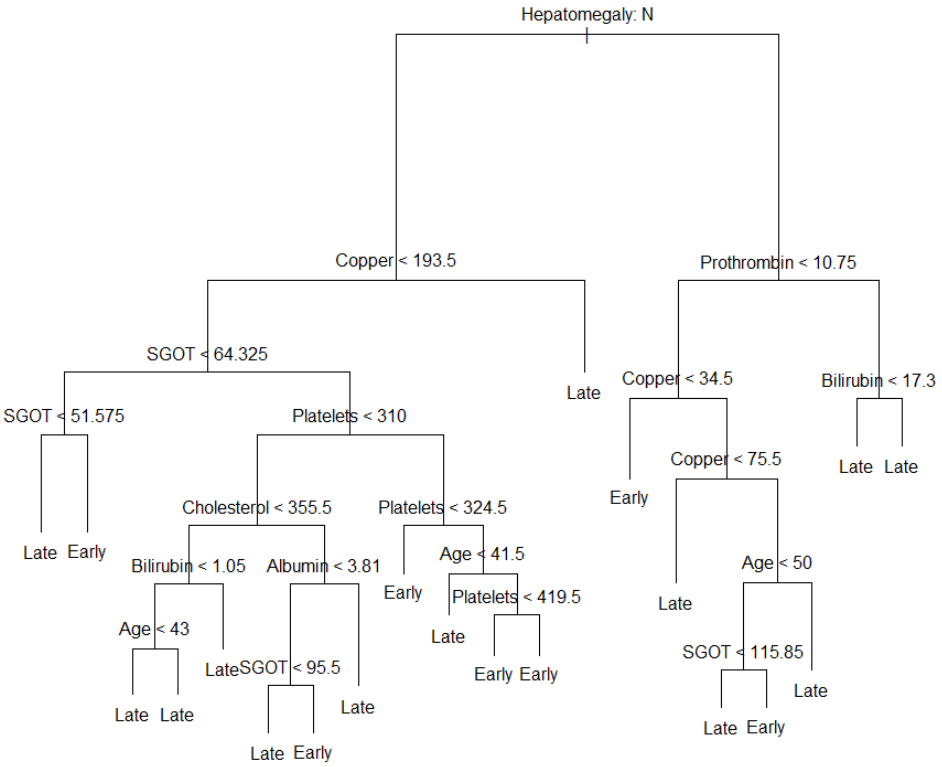
a) Results output, important model summaries, and images that resulted from model fitting:

The significance table that one gets from the summary() output, with all coefficient estimates and significance values of the **logistic regression** part is attached below.

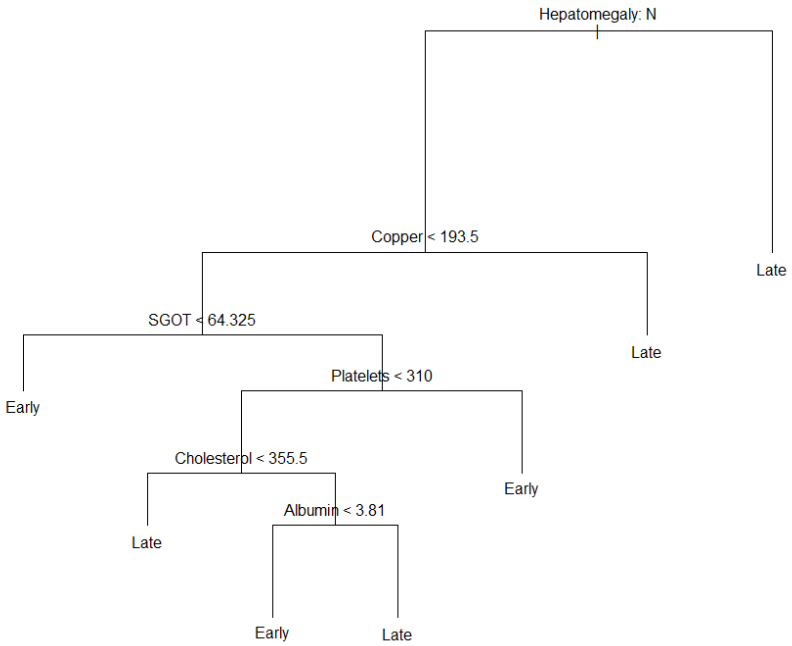
The step model is obtained from the stepwise function which was performed in previous parts, and we aim to choose the most significant predictors: Hepatomegaly, Copper, and Spiders.

```
> # We choose three best predictors with significant of p-values:
> step.model = glm(Stage ~ Hepatomegaly + Copper + Spiders, family='binomial', data = train_data_unique)
> summary(step.model)

Call:
glm(formula = Stage ~ Hepatomegaly + Copper + Spiders, family = "binomial",
    data = train_data_unique)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.288013   0.295031  -0.976 0.328960
HepatomegalyY  1.456640   0.375599   3.878 0.000105 ***
Copper         0.006864   0.003031   2.265 0.023542 *
SpidersY       0.950000   0.490317   1.938 0.052682 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 251.73  on 219  degrees of freedom
Residual deviance: 211.96  on 216  degrees of freedom
AIC: 219.96

Number of Fisher Scoring iterations: 5
```

With the **decision tree**, we provide pictures of the unpruned tree and the pruned tree.

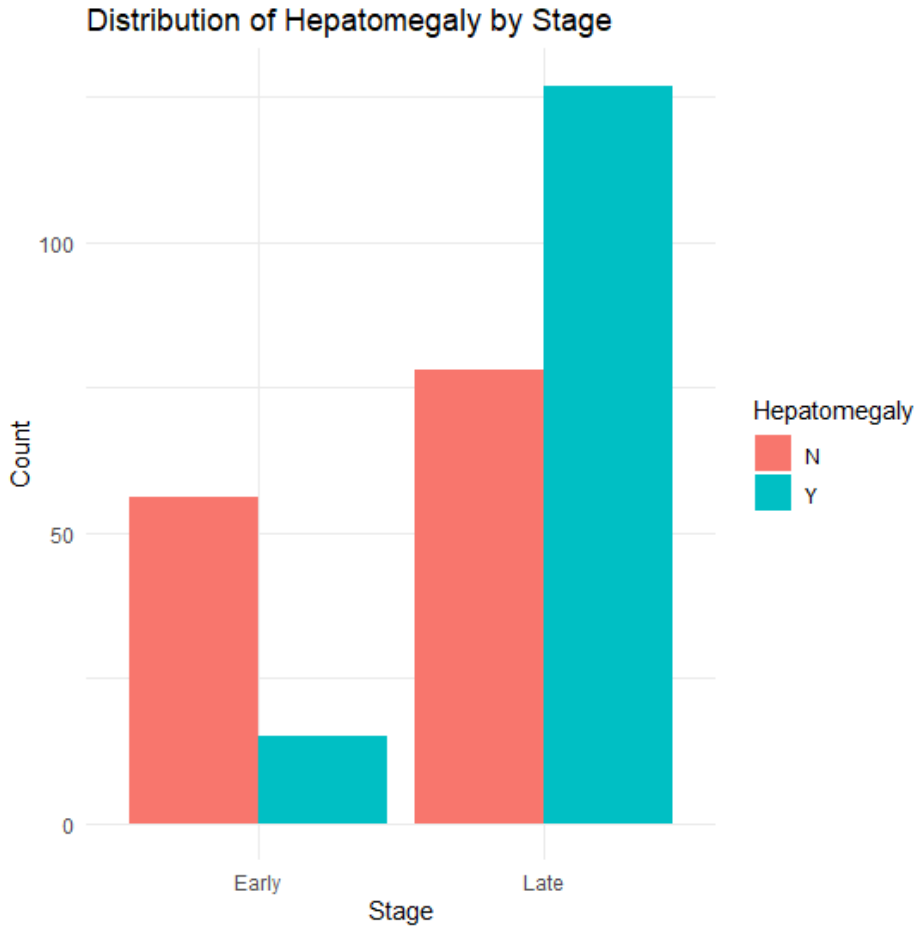


The unpruned tree

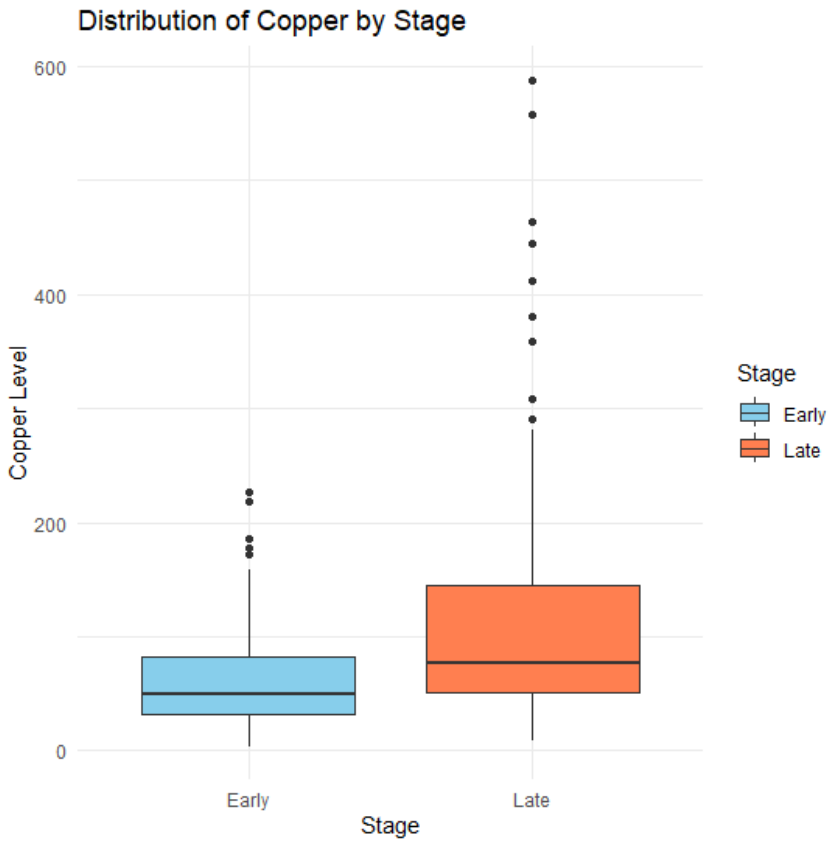


Using cv.tree() to find the optimal size of the pruned tree. A size of 7 appears to be optimal.

b) Interpretation of results and conclusions:



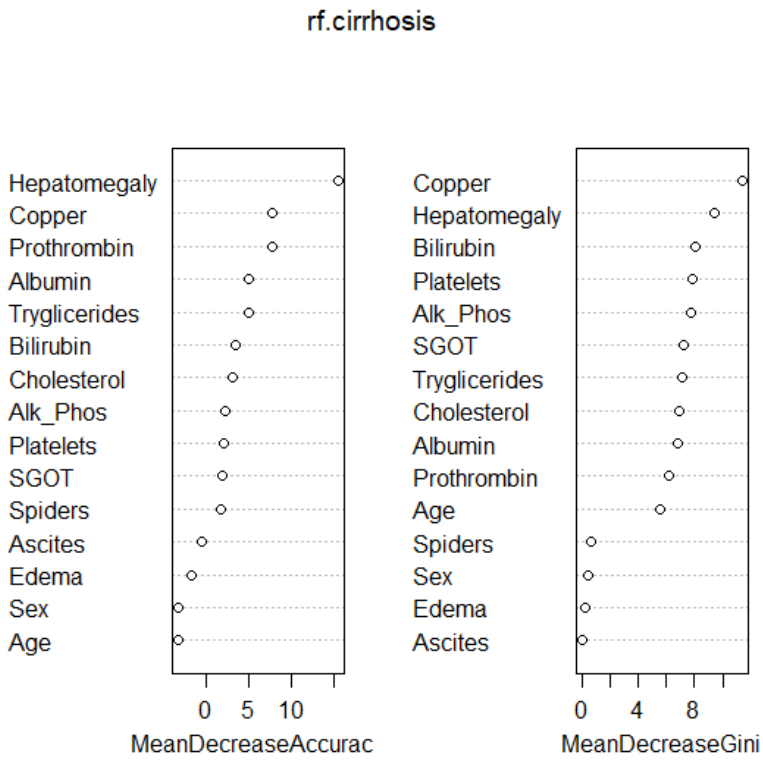
From the Exploratory data analysis (EDA) part, we see that patients with hepatomegaly tend to have late-stage cirrhosis.



Copper level tends to elevate in patients with late-stage cirrhosis. Some patients have an extremely high copper level.

From the stepwise function in part 2, Hepatomegaly, Copper, and Spiders are the most significant variables.

The pictures of the original tree and the pruned tree indicate that Hepatomegaly is the most important predictor. We are also interested in looking for other significant predictors using Random Forest and VarImplot.



Hepatomegaly and Copper are two of the most important predictors. This result is identical to the logistic regression model.

```
> summary(test_error_tree)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2143  0.2545  0.3036  0.3054  0.3527  0.3929
> summary(test_error_logistic)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1429  0.2054  0.2411  0.2446  0.2812  0.3214
```

As for the models' performance obtained from repeating the random subdivision of 80% training and 20% testing data 10 times, they are quite decent with a mean test error rate of 0.24 for the logistic regression model, and 0.31 for the decision tree model. These results suggest that our preprocessed data are good and these models are suitable for predicting the cirrhosis stage.

4) Conclusion: (Le Bui)

The decision tree and logistic regression models show that Hepatomegaly and Copper are two of the most important predictors.

The test errors obtained from both models are less than 0.40 and can be as low as 0.14. The average test errors are 0.31 and 0.25 for the decision tree and logistic regression models respectively. Overall, the test error from the logistic regression model tends to be better than the decision tree. Therefore, the logistic regression model is best to use in our project.

The decision trees are prone to high variance and low bias. Each subdivision of the data set can lead to significant differences in results.

Also, pruning the tree can help reduce the overfitting issue. Previous parts show that the original tree obtains a training error as low as 0.1 while the pruned tree obtains a training error of 0.1682. However, the test error of the original tree is 0.3035714 while the test error of the pruned tree is as low as 0.1785714. In the pruned tree, the difference between the training and test errors is quite small in comparison with the unpruned tree.

Our result indicates that our models can be applied for some real-world applications which is explained in part 6.

5) Bibliography:

Data set source:
https://archive.ics.uci.edu/dataset/878/cirrhosis+patient+survival+prediction+dataset-1

Materials referenced:
”An Introduction to Statistical Learning (with applications in R)” by James, Witten et al., Second Edition, ISBN: 978-1071614174

6) Professional product: (Le Bui)

Cirrhosis poses a profound challenge to patient health and imposes substantial economic burdens due to its treatment costs. Non-invasive diagnostic techniques serve as invaluable tools for the early detection and screening of individuals exhibiting clinical indicators, such as hepatomegaly, or presenting with abnormal blood test results. These methods are characterized by their efficiency, cost-effectiveness, and simplicity, making them accessible for widespread implementation. Early identification of cirrhosis is particularly crucial for optimizing public health strategies and mitigating disease progression.

Hepatomegaly, a hallmark clinical sign, can be readily assessed through physical examination or ultrasound imaging. Both approaches are straightforward, affordable, and widely available, enhancing their utility in routine medical practice.

7) R script:

The R script is uploaded in the file .R