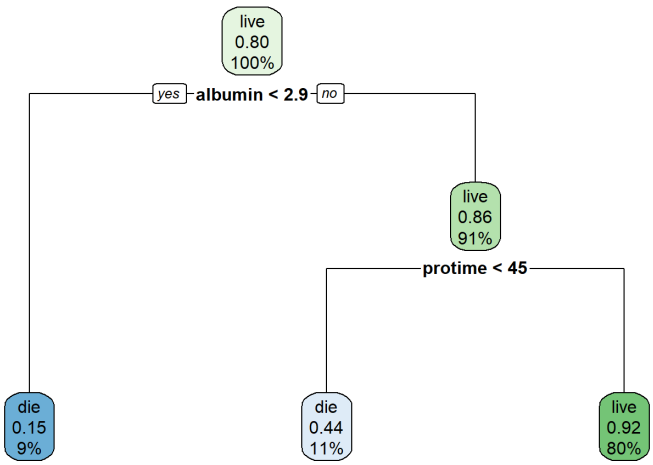
Sunniul Alam
SXA180118
CS4375.002
30 March 2022

HW3 Writeup

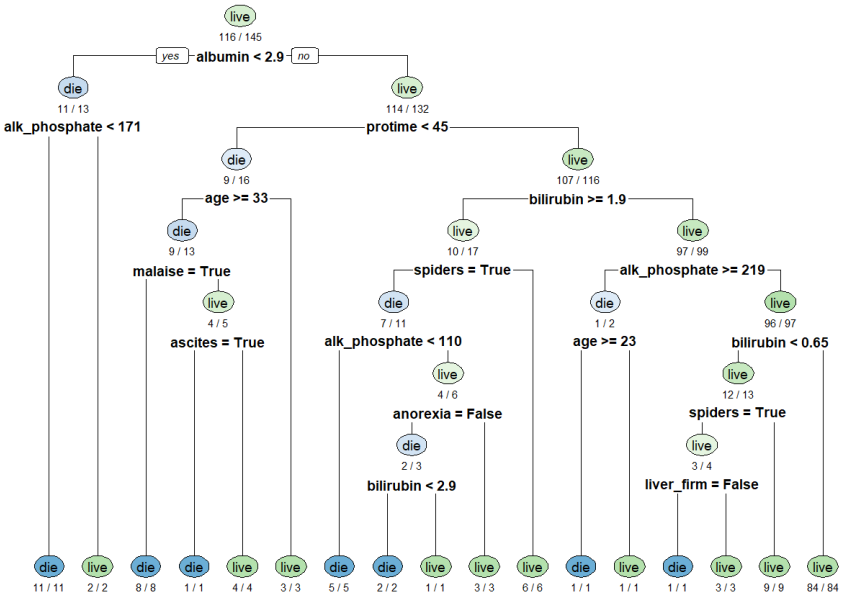
1. Improve the quality of "Hepatitis" dataset by handling duplicate objects and missing values. Using the default setting of rpart, create a decision tree to classify the objects of the improved dataset into two categories: die and live.
    a. Plot the tree



    b. Describe the characteristics of the objects that were classified as desired target.
    The desired target is for the class to be live. The objects that were classified as the desired target have an albumin level less than or equal to 2.9 and an protime level less than or equal to 45.
    c. Predict a class for each object in the improved Hepatitis data set using the constructed tree. Find TPR, FPR, TNR, FNR, and accuracy for this prediction.
    RScript
    d. Display the actual class label and the predicted class label for the first ten objects of the improved Hepatitis data set. Find the error rate for these ten objects.
    RScript
    e. After finding the training error, estimate generalization error for this tree using pessimistic and optimistic approach.
    RScript
2. Using rpart, create a fully grown decision tree to classify the objects in the improved "Hepatitis" dataset into two categories: die and live. We can control splitting the node

Sunniul Alam
SXA180118
CS4375.002
30 March 2022

using minsplit and cp. To have a fully grown decision tree, consider minsplit as 2 and cp as 0 for this tree. A.



a. Predict a class for each object in the improved Hepatitis data set using the constructed tree. Find TPR, FPR, TNR, FNR, and accuracy of this prediction.

   Check RScript

b. After finding the training error, estimate generalization error for this tree using pessimistic and optimistic approach.

   Check RScript

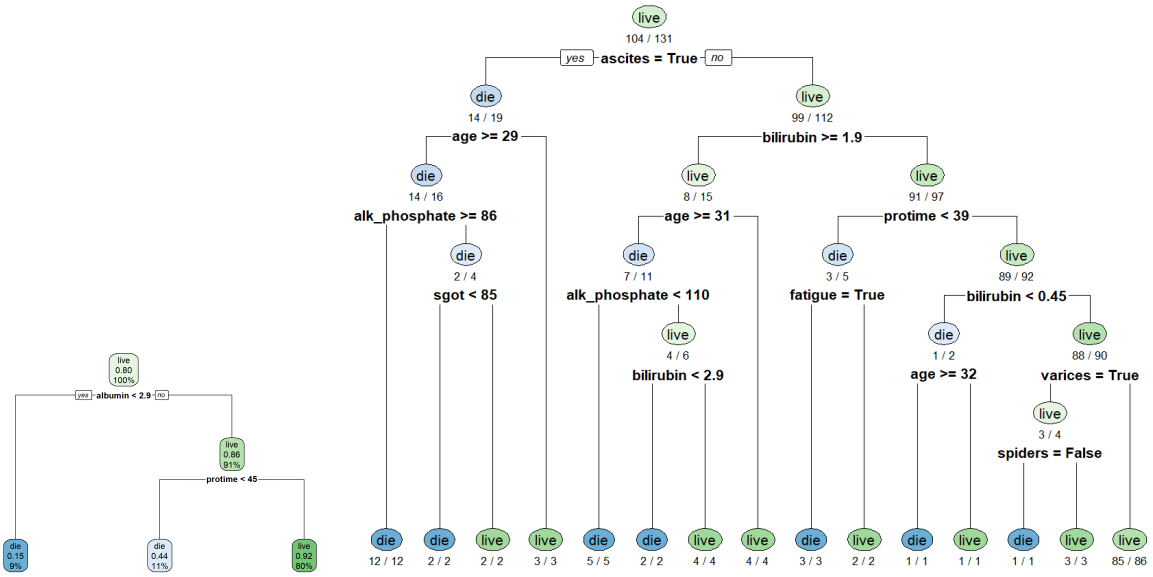
c. Compare the training error based on the fully grown tree (constructed tree in Q2) and the default tree (constructed tree in Q1). Which one has a higher training error rate? Why?

   When comparing the training errors on the fully grown tree and the default tree, it is evident that the default tree has the higher training error rate. We can conclude that this is because the splitting of the tree is more accurate. This is seen through the fact that the false positive and false negative rates for the fully grown tree are 0, while the false positive and false negative rates for the default tree are 0.045 and 0.519, respectively.

Sunniul Alam
SXA180118
CS4375.002
30 March 2022

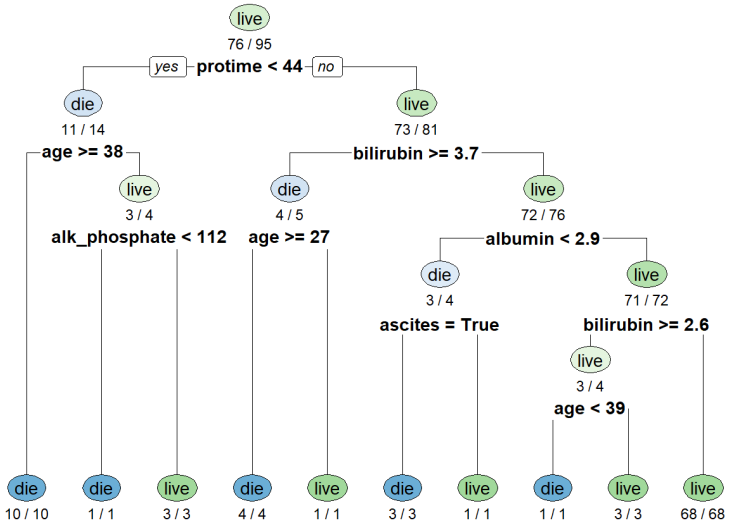
3. What is underfitting in machine learning? Construct a decision tree for the improved "Hepatitis" data set to illustrate under-fitting.

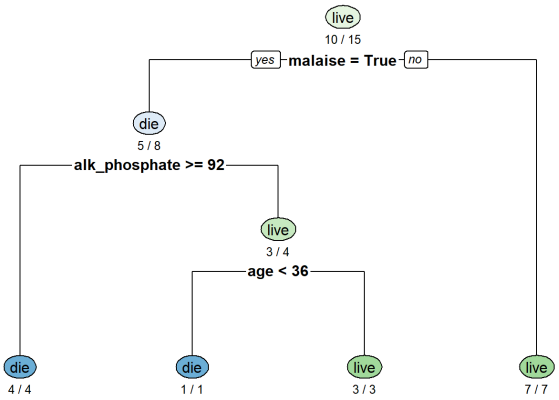
   Underfitting is when the relationship between the variables can not be seen and thus predictions on the test data will be incorrect.
   The first tree does not properly sort the training data enough to accurately determine which class is live and which is die. The other tree is more developed and, as seen below, the sorting is more specific.



4. Randomly select 2/3 of the objects in the improved Hepatitis data set as the training dataset and 1/3 of the objects as the test dataset. Create a decision tree using the training set. Then, randomly select 15 objects of the improved "Hepatitis" data set as the training dataset and consider the remaining objects as the test dataset. Create another decision tree using this training set. Find the training error and testing error for both trees. Which one has the higher the training error and testing error? Why?

Sunniul Alam
SXA180118
CS4375.002
30 March 2022

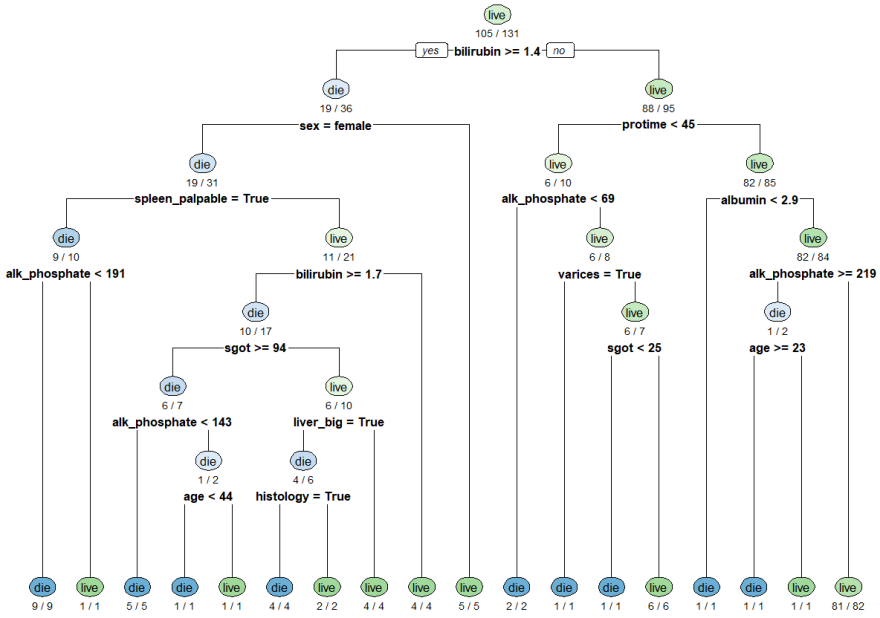




They both have the same training and testing error. This is because they are both zero.

5. After shuffling the improved Hepatitis data set, Partition the dataset into 10 disjoint subsets. Consider the first partition (fold) as the testing set and the remaining partitions (folds) as the training set. Find the size of the training set and test set. Using rpart, create a decision tree based on the training set. (consider minimum split as 2). Calculate the training error and the testing error . Repeat this process for every other folds. Find the average of testing error.

Sunniul Alam
SXA180118
CS4375.002
30 March 2022



6. -Consider the following training dataset for a binary classification problem
   a. What is the entropy of this collection of training examples with respect to the True class

      Entropy= - P(T) * $\log_2$(P(T)) - P(F) * $\log_2$((P(F))

      Entropy= - 4/9 * $\log_2$(4/9) - 5/9 * $\log_2$(5/9)
      Entropy= 0.5199666 + 0.471109= 0.99107606
   b. What are the information gains of A1 and A2 relative to the training dataset For A3, which is a continuous attribute, compute the information gain for every possible split.

a_1	T	F
T	5	0
F	0	4

a_1:
Entropy= P(T)(-P(T,T) * $\log_2$(P(T,T)) - P(T,F) * $\log_2$(P(T,F)) + P(F)(-P(F,T) * $\log_2$(P(F,T)) - P(F,F) * $\log_2$(P(F,F))
Entropy= 5/9(-5/5 * $\log_2$(5/5) - 0 * $\log_2$(0)) + 4/9(-0 * $\log_2$(0) - 4/4 * $\log_2$(4/4))
Entropy= 4/9(0 + 0) + 5/9(0 +0)

Sunniul Alam
SXA180118
CS4375.002
30 March 2022

Entropy= 0+ 0= 0

Gain= Entropy of Training Examples - Entropy of A1= 0.9911 -0= 0.9911

a_2	T	F
M	3	1
F	2	3

a_2:

Entropy= $P(M)(-P(M,T) * \log_2(P(M,T)) - P(M,F) * \log_2(P(M,F)) + P(F)(-P(F,T) * \log_2(P(F,T)) - P(F,F) * \log_2(P(F,F))$

Entropy= $4/9(-3/4 * \log_2(3/4) - 1/4 * \log_2(1/4)) + 5/9(-2/5 * \log_2(2/5) - 3/5 * \log_2(3/5))$

Entropy= 4/9(0.3113 + 0.5) + 5/9(0.5288 +0.4422)

Entropy= 0.3606 + 0.5394= 0.9000

Gain= Entropy of Training Examples - Entropy of A2= 0.9911 - 0.9000= 0.0911

c. What is the best split (among A1,A2, and A3) according to the information gain?

Object	A3	Target Class
8	2.0	F
9	2.0	T
4	3.0	F
7	3.0	T
3	5.0	F
5	6.0	T
2	7.0	T
6	7.0	T
1	8.0	F

Split at 2.5

Entropy= 2/9 (-1/2 log2 1/2 -1/2 log2 1/2) + 7/9(-4/7 log2 4/7 -3/7 log2 3/7)

= 0.2222 + 0.5714= 0.7936

Info Gain= 0.9911-0.7936= 0.1975

Split at 4.0

Sunniul Alam
SXA180118
CS4375.002
30 March 2022

Entropy= 4/9 (-2/4 log2 2/4 - 2/4 log2 2/4) + 5/9(-3/5 log2 3/5 -2/5 log2 2/5)
= 0.4444 + 0.5394= 0.9838
Info Gain= 0.9911-0.9838= 0.0073
Split at 5.5
Entropy= 5/9 (-2/5 log2 2/5 - 3/5 log2 3/5) + 4/9(-3/4 log2 3/4 -1/4 log2 1/4)
= 0.5394+ 0.3506= 0.8900
Info gain = 0.9911 - 0.8900 = 0.1011
Split at 6.5
Entropy= 6/9 (-3/6 log2 3/6 - 3/6 log2 3/6) + 3/9(-2/3 log2 2/3 -1/3 log2 1/3)
= 0.6667 + 0.3061= 0.9728
Info gain = 0.9911 - 0.9728 = 0.0183
Split at 7.5
Entropy= 8/9(-5/8 log2 5/8 - 3/8 log2 3/8) + 1/9(-0/1log2 0/1 - 1/1log2 1/1)
= 0.8484 + 0= 0.8484
Info gain = 0.9911 - 0.8484 = 0.1427

The best split occurs is A1.

d. What is the best split (between A1 and A2) according to the classification error rate?

The Classification error for A1 is 0/9, or 0.

The Classification error for A2 is 3/9, or 0.3333.

The best split would be A1.

e. What is the best split (between A1 and A2) according to the Gini index?

5/9[1 - (5/5)^2 - (0/5)^2] + 4/9 [1 - (4/4)^2 - (0/4)^2]= 0
5/9[1 - (3/5)^2 - (2/5)^2] + 4/9 [1 - (1/4)^2 - (3/4)^2] = 0.43333
Since A1 is smaller, it produces a better split

7. Consider the decision tree shown in the following Figure.

- Compute the generalization error rate of the tree using the optimistic approach.

  The generalization error using pessimistic is 3/10.
- Compute the generalization error rate of the tree using the pessimistic approach. (For simplicity, use the strategy of adding a factor of 0.5 to each leaf node.)

  The generalization error using pessimistic is 5/10.
- Compute the generalization error rate of the tree using the validation set shown above. This approach is known as reduced error pruning.

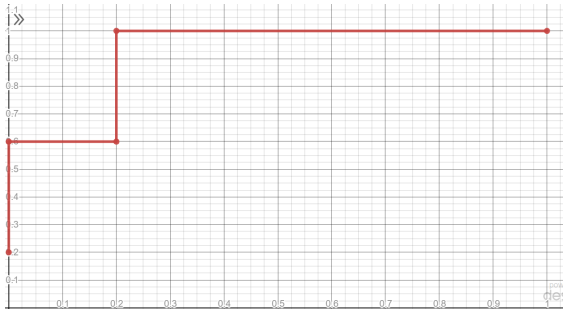
  The generalization error using pessimistic is 8/10.

Sunniul Alam
SXA180118
CS4375.002
30 March 2022

8. You are asked to evaluate the performance of two classification models, M1 and M2. The test set you have chosen contains 26 binary attributes, labeled as A through Z.

   a. Plot the ROC curve for both M1 and M2. (You should plot them on the same graph.) Which model do you think is better? Explain your reasons.

   M1

   M2

   The model for M1 is better since its AUC is higher.

   b. For model M1, suppose you choose the cutoff threshold to be t = 0.5. In other words, any test instances whose posterior probability is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure for the model at this threshold value.

M1	+	-
>0.5	3	2
<=0.5	1	4

Precision= ¾=0.75
Recall= ⅗=0.6
F-measure= .9/1.35= 0.6667