Robert Genega

CMSC 471

Project 3 Report

Preprocessing:

I ignored the PassengerID, Name, ticket, fare, and cabin columns when parsing the training csv, as they should not have affected whether a passenger lived or died. The cabin number may have had an impact however fewer than half of the entries had cabins listed and it would've proved difficult to enumerate the different cabin names. Male and Female were changed to 1 and 0. Passenger ages were rounded to integers and enumerated as following:

0 – 6 = 0

7 – 13 = 1

14 – 20 = 2

21 – 27 = 3

28 – 34 = 4

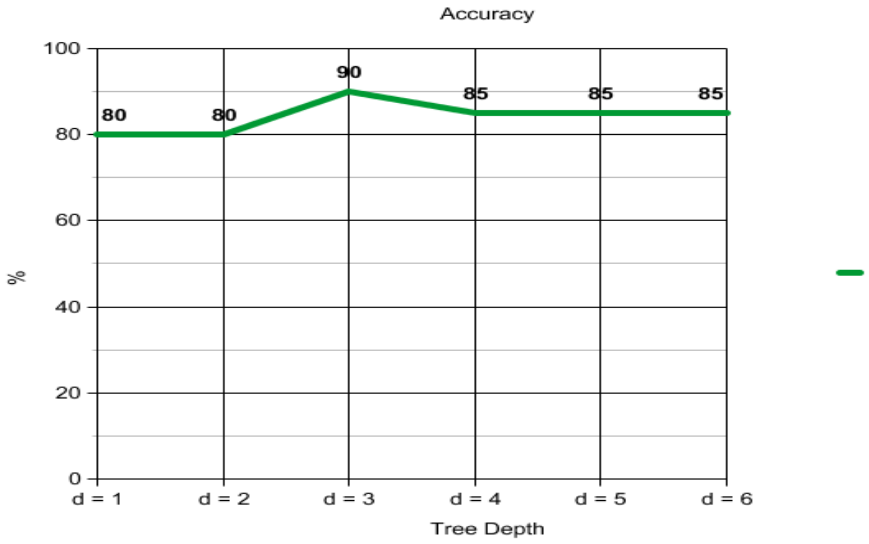
35 or greater = 5

Any entry with an age not listed was given the average age calculated (30) which was enumerated to 4.

Embarked was enumerated S = 1, C = 2, Q = 3. Pclass, SibSp, ParCh were left as the integers given in the dataset.

Parameters:

I modified my decision from part 1 to handle the titanic csv information, and I used the max decision tree depth as my adjustable parameter. Given that there were six functional columns in the dataset, I tested my decision tree from max depth = 1 to a max depth = 6.



Accuracy:

I used a train/test split of 871 / 20 to validate my decision tree. 20 entries were removed from the training set and added to my test set.

0 = did not survive 1 = survived

Tree depth = 1, test size = 20	Predicted 0	Predicted 1	
Actual 0	10	3	13
Actual 1	1	6	7
	11	9	Acc = .80

Tree depth = 2, test size = 20	Predicted 0	Predicted 1	
Actual 0	10	3	13
Actual 1	1	6	7
	11	9	Acc = .80

Tree depth = 3, test size = 20	Predicted 0	Predicted 1	
Actual 0	11	2	13
Actual 1	0	7	7
	11	9	Acc = .90

Tree depth = 4, test size = 20	Predicted 0	Predicted 1	
Actual 0	11	2	13
Actual 1	1	6	7
	12	8	Acc = .85

Tree depth = 5, test size = 20	Predicted 0	Predicted 1	
Actual 0	12	1	13
Actual 1	2	5	7
	14	6	Acc = .85

Tree depth = 6, test size = 20	Predicted 0	Predicted 1	
Actual 0	12	1	13
Actual 1	2	5	7
	14	6	Acc = .85

Results:

As shown from the graph and confusion matrices, the decision tree predicted more accurately at a tree depth of 3. This could mean that overfitting of the graph occurred after the first three levels of the tree or more simply that the first three features that were split had a stronger correlation to the passenger

survival. Given more time I would test with k-fold cross validation to get more accurate test results. I would also adjust the enumeration type for age to fit a normal distribution.