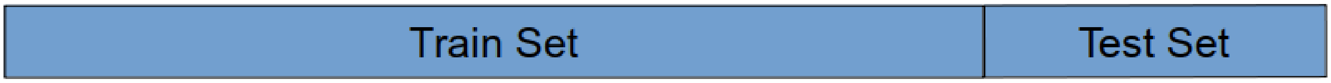
Data 301 test2

Assume a dataset is split in the following way;



1. (5 pts) Is it valid to train a random forest once on the Train set and verify it's performance on the Test set? Yes, but you are not tuning your hyperparameters, so you are likely not getting the best model.

2. (5 pts) Is it valid to perform hyper parameter tuning for a random forest using OOB scores and the Train set? Keep in mind you can estimate your models performance using the final tuned hyper parameters on the test set. Is there a better way to do this? Yes, the OOB score will tell you how well your hyperparameters worked for the model. But you will only be training on a portion of the train set. A better choice is to use k-fold cross validation for hyper parameter tuning.

3.(5 pts) Assume you have successfully discovered the best hyper parameters for your random forest model. What are the advantages and disadvantages of combining the train and test sets into one big dataset, and then retraining a random forest on that combined set? Advantages: more data to train on, assuming the hyper parameters are chosen so that you did not over/under fit the data then the newly trained model should perform as well or slightly better. Disadvantages: No test set to prove performance.

4.(5 pts) For question3 above, is there a way to estimate your models performance? If so how? Yes use OOB score

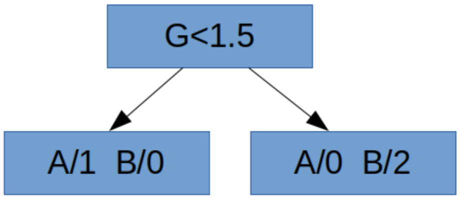
5. (5 pts) Suppose I train a random forest using 5-fold cross validation on the Train set. What do I learn, and what should I do next? Assuming I don't want to optimize hyper parameters. I can average the 5 folds to get an estimate of my models performance. Next: train a new model on the train set, get an estimate of it's performance on test set.

Train a decision tree on the following dataset. Use Gini impurity to determine split points.

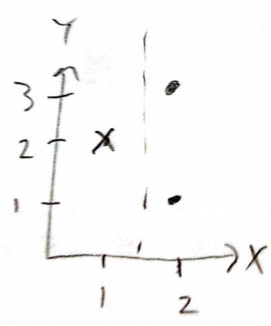
X	Y	target
1	2	A
2	1	B
2	3	B

6. (20 pts) Calculate the best column to split on, and the best split point.

Column X
split point is 1.5
for left node g=1-(1/1)**2-(0/1)**2=0
for right node g=1-(0/2)**2-(2/2)**2=0
g=(1/3)*0 +(2/3)*0=0
Use column X split point 1.5



7. (15 pts) Graph the above dataset and show the best split point.



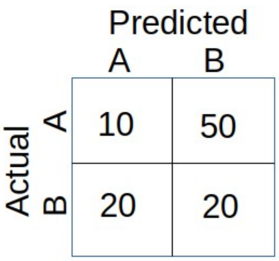
8. (10 pts) What will your trained decision tree predict for the following points?

X	Y	prediction
1	75	A
2	2	B
1	1	A
1	3	A
3	3	B

A dataset consists of:
40 instances of class A
60 instances of class B

Your algorithm predicts 30 class A's of which 10 are correct
70 class B's of which 20 are correct: **Note: should read 40 are correct!**

9. (10 pts)Fill in the confusion matrix

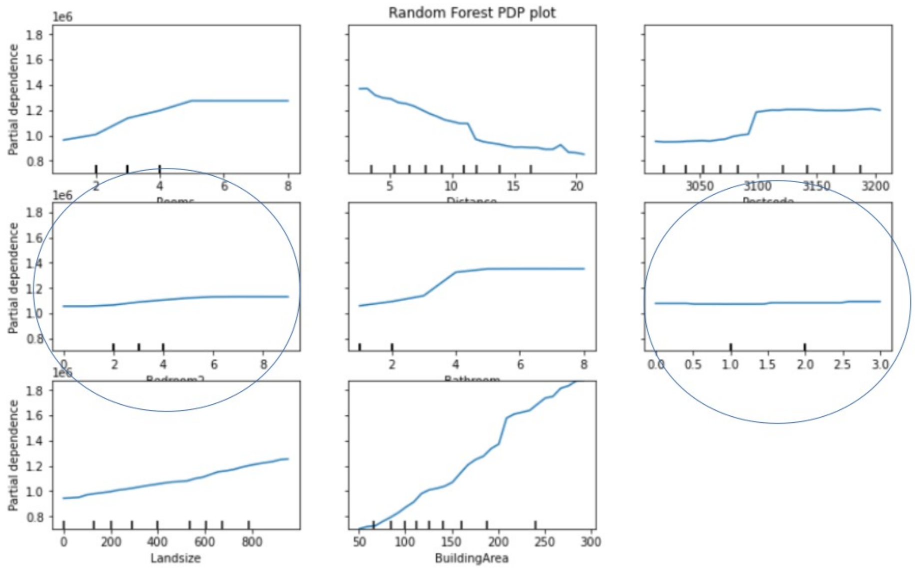


		Predicted	
		A	B
Actual	A	10	50
	B	20	20

10. (10 pts) What is the algorithms overall Precision

P=(10 +20)/((10+20) +(50+20))

11. (10 pts) For the following PDP plots. Circle the variable(s) that have the least predictive power. If more than one rank them with 1 being least predictive, 2 being next least predictive etc.


Random Forest PDP plot
1e6
Partial dependence
Rooms
Distance
Postcode
1e6
Partial dependence
Bedroom2
Bathroom
1e6
Partial dependence
Landsize
BuildingArea