My Name Biao Feng(myNetID: biaof2)

IE598 MLF F18

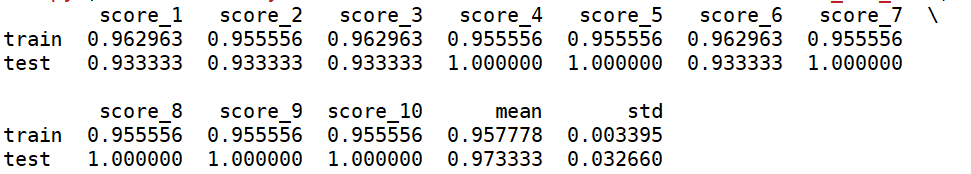
Module 6 Homework (Cross validation)

Using the Iris dataset, with 90% for training and 10% for test and the decision tree model that you submitted for Module 2:

**Part 1: Random test train splits**

Run in-sample and out-of-sample accuracy for 10 different samples by changing random\_state from 1 to 10 in sequence.

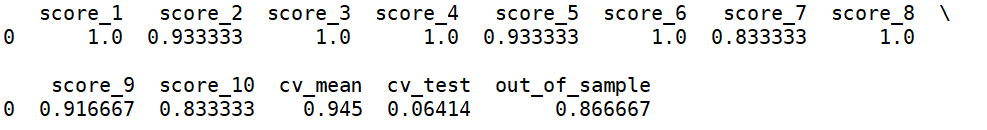
Display the individual scores, calculate the mean and standard deviation of the set. Report in a table format.



**Part 2: Cross validation**

Now rerun your model using cross\_val\_scores with k-fold CV (k=10).

Report the individual fold accuracy scores, the mean CV score and the standard deviation of the folds. Now run the out-of-sample accuracy score. Report in a table format.



**Part 3: Conclusions**

Write a short paragraph summarizing your findings. Which method of measuring accuracy provides the best estimate of how a model will do against unseen data? Which one is more efficient to run?

From the graph, it is apparently that the train\_test\_split is more useful for the unseen data since mean test scores are higher than out of sample score in cross validation.

Cross validation is more efficient to run since we need to write a for loop for train\_test\_split to set the random\_state.

**Part 4: Appendix**

Link to github repo