My Name (myNetID)

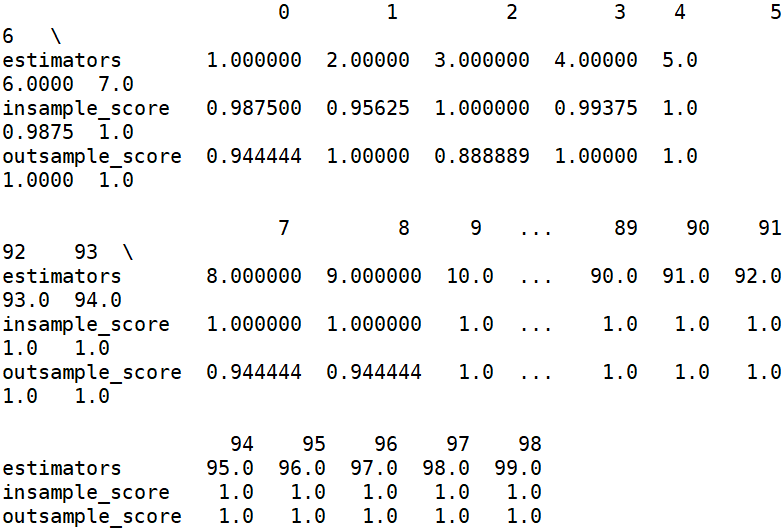
IE598 MLF F18

Module 7 Homework (Random Forest)

Using the Wine dataset, described in Raschka chapter 4 and 10 fold cross validation;

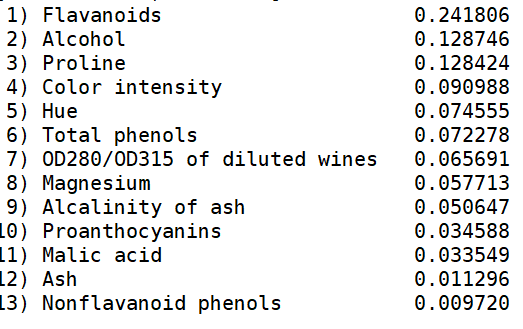
**Part 1: Random forest estimators**

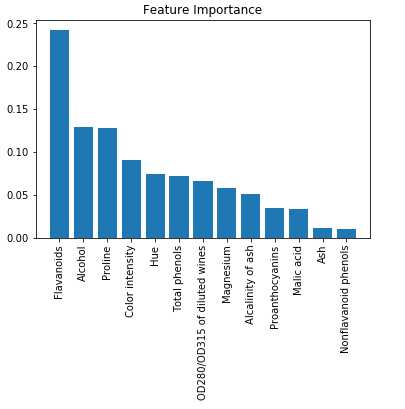
Fit a random forest model, try several different values for N\_estimators, report in-sample accuracies.



**Part 2: Random forest feature importance**

Display the individual feature importance of your best model in Part 1 above using the code presented in Chapter 4 on page 136. {importances=forest.feature\_importances\_ }





**Part 3: Conclusions**

Write a short paragraph summarizing your findings. What is the relationsjhip between n\_estimators, in-sample CV accuracy and computation time? What is the optimal number of estimators for your forest? Which features contribute the most importance in your model according to scikit-learn function? What is feature importance and how is it calculated? (If you are not sure, refer to the Scikit-Learn.org documentation.)

Generally, the larger n\_estimators we have, the higher in-sample CV accuracy we get but we need to spend more time.



The optimal number of estimators is 16.

Flavanoids is the most important feature I get.

Feature importance means that the correlation of features and target.

From the textbook, feature importance is calculated by average impurity computed from all the decision trees in the forest.

**Part 4: Appendix**

Link to github repo