Question 4 : Random Forest Classifier

Parameters:

We have performed a grid search on two parameters :

1. The ***max_depth*** parameter in the range **[2,20)** with a step-size of **2**: Analyzing the results from decision trees, we kept our parameter search for max_depth in this range because the performance drops and stays low after 20.
2. The number of trees is searched in the range **[3,200)** at regular interval of **20**

Accuracy Scores for different parameter settings

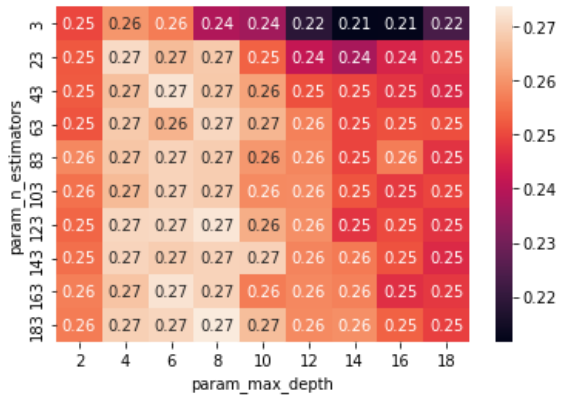
a. The best accuracy of 27.4% is obtained for the abalone-raw dataset at a maximum depth of 8 and 183 decision trees

b. The best accuracy of 54.4% is obtained for the wine-lda at a maximum depth of 4 and 123 decision trees

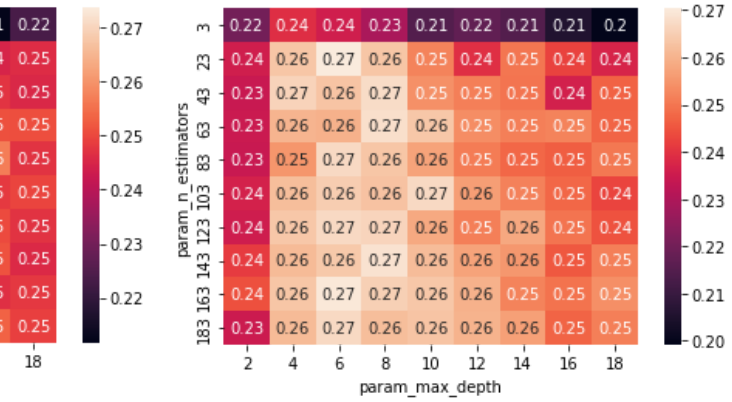
Analysis : For both datasets, lower values of max_depth seem to perform well.

Heat Plots:

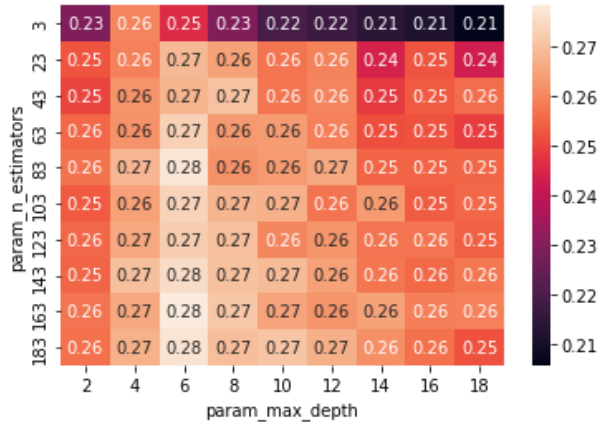
Abalone-Raw



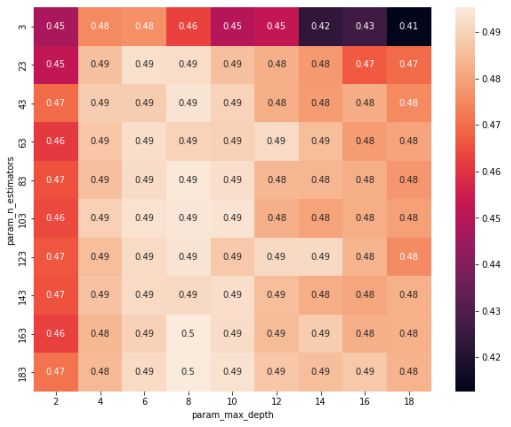
Abalone-PCA



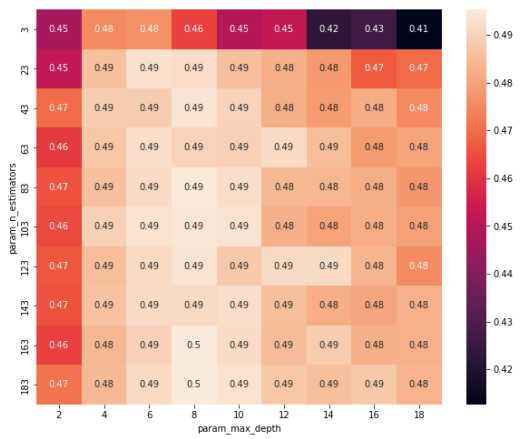
Abalone-LDA



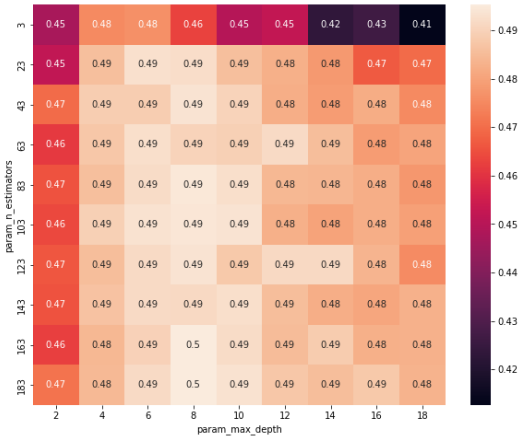
Wine-Raw



Wine-PCA



Wine- LDA



Observations:

- At constant values of maximum depth, the accuracy is observed to increase as we increase the number of trees for most maximum depths and then increase or plateau suggesting that either no new information is being captured by the addition of more trees for all the 6 datasets

- At constant values of the number of trees, the accuracy is first observed to increase as we increase the depth and then decrease suggesting that overfitting may be occurring for all the 6 datasets