Decision Tree:

In both datasets, when the value of max\_depth is lower, the models underfit. As the max\_depth increases, the models overfit. When max\_depth increases beyond 10, as the models fit exactly like training data resulting to overfitted models. It looks like a good max\_depth for both datasets will be between 5 and 10 based on initial data exploration. We will now use GridSearchCV to find the most optimal max\_depth. While we are at it, let’s also find the optimal min\_samples\_leaf.

Looking at the learning curves of both datasets, we can see that our first dataset (with small sample) suffers from high variance (backed up by the spread of light green and light green area). As the training instances increase, there isn’t much improvement in cross validation score. The high variance can be reduced using less features and by increasing the training samples. In case of second dataset, as training instances increase, the cross-validation score increases. The predictions for second dataset also have a high variance. As mentioned earlier, we can decrease the variance by removing some features (making our tree less complex) or by using performing ensemble learning like boosting.