1. What is the estimated depth of a Decision Tree trained (unrestricted) on a one million instance training set?

Ans: **around 20**

1. Is the Gini impurity of a node usually lower or higher than that of its parent? Is it always lower/greater, or is it usually lower/greater?

Ans: **a node’s gini impurity is generally lower than its parent.**

1. Explain if its a good idea to reduce max depth if a Decision Tree is overfitting the training set?

Ans: **yes, by decreasing the max\_depth, the decision tree won’t run till its death and it will make some generalisation to fit the training dataset.**

1. Explain if its a good idea to try scaling the input features if a Decision Tree underfits the training set?

Ans: **decision tree is not affected by feature scaling as the magnitude of feature values doesn’t affect the splitting.**

1. How much time will it take to train another Decision Tree on a training set of 10 million instances if it takes an hour to train a Decision Tree on a training set with 1 million instances?

Ans: **more than 10 hours**

1. Will setting presort=True speed up training if your training set has 100,000 instances?

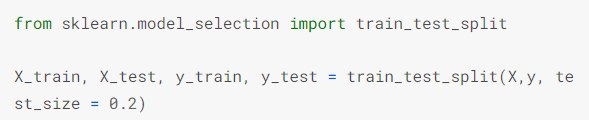
Ans: **it will add one more step before splitting the data so if you have a large dataset then it could slow down the process of splitting the dataset into nodes.**

7. Follow these steps to train and fine-tune a Decision Tree for the moons dataset:

1. To build a moons dataset, use make moons(n samples=10000, noise=0.4).



1. Divide the dataset into a training and a test collection with train test split().



1. To find good hyperparameters values for a DecisionTreeClassifier, use grid search with cross-validation (with the GridSearchCV class). Try different values for max leaf nodes.

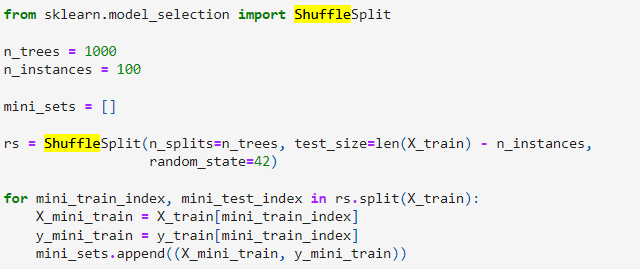


1. Use these hyperparameters to train the model on the entire training set, and then assess its output on the test set. You can achieve an accuracy of 85 to 87 percent.

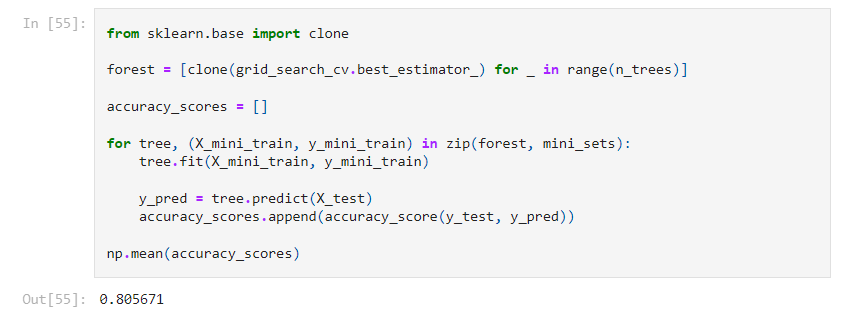


8. Follow these steps to grow a forest:

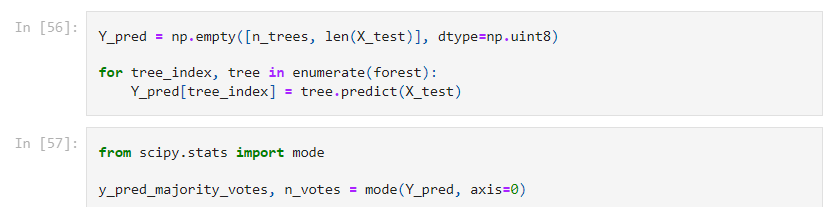
1. Using the same method as before, create 1,000 subsets of the training set, each containing 100 instances chosen at random. You can do this with Scikit-ShuffleSplit Learn's class.



1. Using the best hyperparameter values found in the previous exercise, train one Decision Tree on each subset. On the test collection, evaluate these 1,000 Decision Trees. These Decision Trees would likely perform worse than the first Decision Tree, achieving only around 80% accuracy, since they were trained on smaller sets.



1. Now the magic begins. Create 1,000 Decision Tree predictions for each test set case, and keep only the most common prediction (you can do this with SciPy's mode() function). Over the test collection, this method gives you majority-vote predictions.



1. On the test range, evaluate these predictions: you should achieve a slightly higher accuracy than the first model (approx 0.5 to 1.5 percent higher). You've successfully learned a Random Forest classifier!

