**Decision Tree:**

Decision Tree is a supervised-learning algorithm, the principle of this algorithm is to build binary tree to partition the data into the leaves of the tree. Each node of the tree is a chosen rule that maximise the “purity” of the leaf. There are two indexes to measure the purity, Gini index and Cross-entropy. In this project, we import the built-in Decision tree model from Scikit-learn library’ and then tune the parameters of this model to make better prediction for our data set. There are two main parameters: criterion and max\_depth. Criterion decide whether Gini index or Cross-entropy is used and Max\_depth decide the max depth of the tree. To set a proper depth for the tree is the core task of this model since too shallow will reduce the accuracy and too high will also cause problems such as overfitting.

After many attempts, we conclude that criterion ‘Gini index’ and ‘Cross-entropy’ are similar in performance while ‘Gini index’ do better in lower depth and ‘Cross-entropy’ perform well in higher depth. We can see the accuracy of different tree depth:

C:\Users\ZHUCH\AppData\Local\Temp\WeChat Files\4ff73e15c7e1c092e3c38c82d2bcc95.jpgC:\Users\ZHUCH\AppData\Local\Temp\WeChat Files\f472f31925ab0f73e1fe09ad561c363.jpg

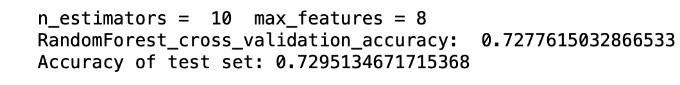
C:\Users\ZHUCH\AppData\Local\Temp\WeChat Files\8ecde38f459ed8fb462c911961143c6.jpgC:\Users\ZHUCH\AppData\Local\Temp\WeChat Files\a8b895166f7feec9668fde7428d609d.jpg

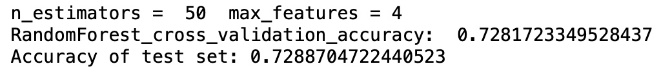
These four figures show the accuracy of four max depth: 2, 5, 16 and 20. The results show that the accuracy increase with the tree depth at first but turn to decrease when the max depth is too big and the best tree depth is about 6 with accuracy around 73.5%. The accuracy is calculated by 10-folds cross validation.

**Random Forest:**

Random Forest is ensemble algorithms, it utilizes the principle of decision tree but average the results from several different decision trees. The advantage of Random Forest is: It can deal with high dimensional dataset and always performs better than single decision tree. It is hard to overfit either. However, the disadvantage of Random Forest is it requires more computing resources and has poor performance on low dimensional dataset.

The parameters of Random Forest are similar to decision tree and the most importance parameters are n\_estimators and max\_features. n\_estimators represent the number of trees and max\_features is the maximum number of features that allowed to use in the tree.

 C:\Users\ZHUCH\AppData\Local\Temp\WeChat Files\ec640b182e33ad25a237ae6465839d2.jpg

In theory, the accuracy will increase with more tree and larger maximum number of features, that is the same with the results of our attempts shown in these figures.

Random Forest is the ensemble method of Decision tree and the performance of Random Forest model is usually better than that of Decision tree model.