Crop Recommendations

Dataset:

The dataset I have chosen is from kaggle ([https://www.kaggle.com/datasets](https://www.kaggle.com/datasets/madhuraatmarambhagat/crop-recommendation-dataset?resource=download)

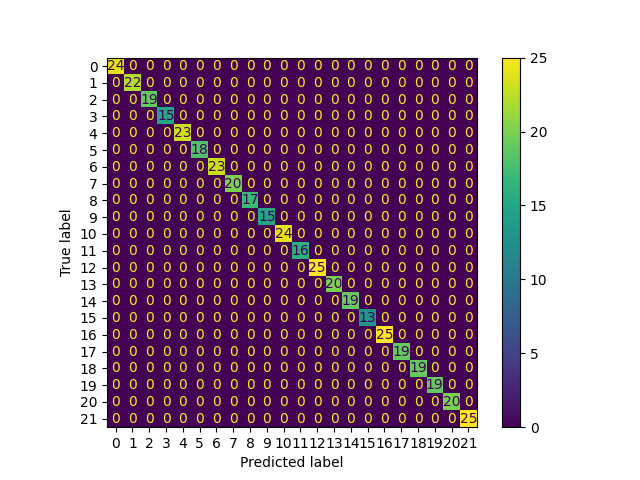
[/madhuraatmarambhagat/crop-recommendation-dataset?resource=download](https://www.kaggle.com/datasets/madhuraatmarambhagat/crop-recommendation-dataset?resource=download)) and has information about soil. The goal of the dataset is to be able to recommend the best kind of plant to grow in the soil. It has 6 columns describing the soil: the nitrogen content (in mg/kg), the phosphorus content (in mg/kg), the potassium content (in mg/kg), the temperature (in celsius), the humidity (in % of air), the ph level, and rainfall (in mm). The labels for the plants are rice, maize, chickpea, kidney beans, pigeon peas, moth beans, mung beans, blackgram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, jute, and coffee. The dataset does not list where these data points are taken from.

Splitting up the data:

For the features, we take just the first 6 columns of each entry. For the labels, we just use the last column. Train test split is then utilized to create train and test lists. The training set is 80% of the original dataset, while the test is only 20%.

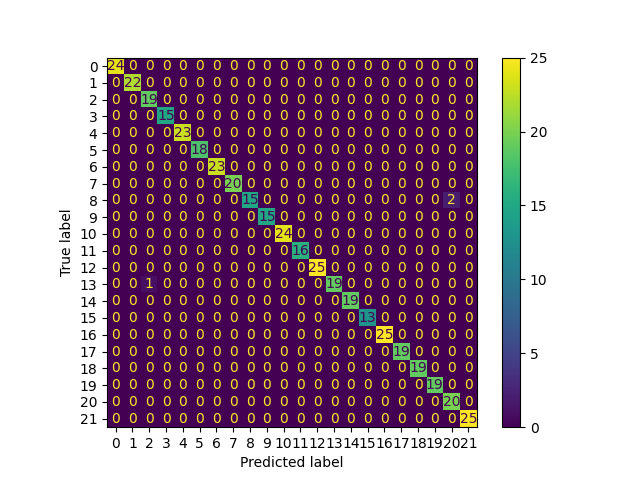
Random forest classifier:

After splitting up our data, we make a random forest classifier(rfc). In order to find the best max depth of our rfc, we will pass the rfc into a grid search and have the max\_depth parameter go over a linspace from 5 to 20. After running and printing out the best max\_depth, we find that it is 17. Using our newly found best max depth, the score on the test set is 99.772, and the OOB score is 99.545.



Looking at the confusion matrix, we see that the random forest got nothing wrong in the test set, which is incredibly good.

Bagging classifier with SVC:

Next we make a bagging classifier and pass in a support vector classifier with a linear kernel. To find the best n\_estimators for our classifier we have to make another grid search. This time our linspace goes from 10 to 50. After running and printing out the best number of estimators, we find it to be 35. Creating a new bagging classifier with svc and using 35 as our n\_estimators, our score over the test set is 98.579, and our OOB score is 99.090.

Looking at the confusion matrix, we see that the bagging classifier got a total of 3 plants wrong, which still is super low.

Compare outcomes:

It is clear that both methods are pretty good at classifying the dataset correctly. The random forest classifier has a better score and OOB score, but not by much. Each of the method’s scores on the test set was consistently 1 or 2 percent worse than the OOB score.