Predicting Coffee Country of Origin

Derek Dattero

Objective:

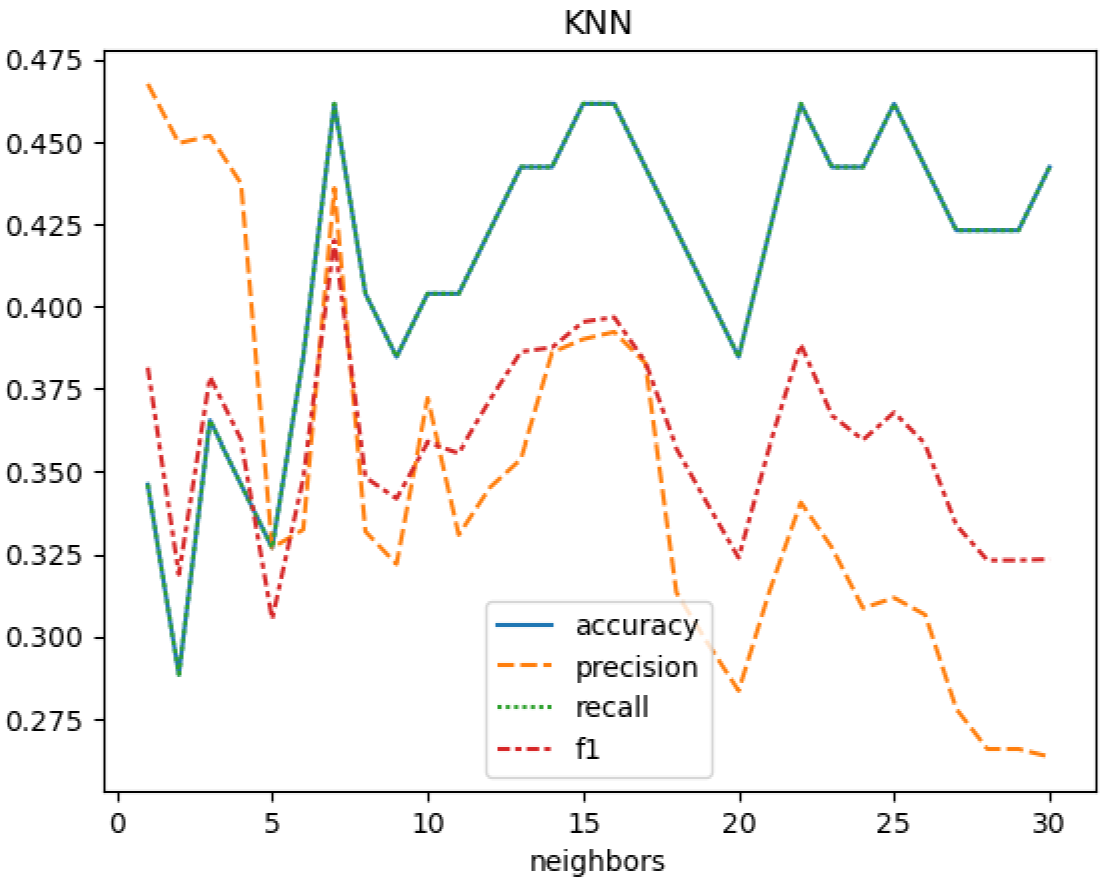
The goal of this project is to determine how well a classification model can predict the country of origin of a batch of coffee. Two applications of this model could be determining where an unknown batch of coffee originates as well as helping customers select coffee beans from countries based on their preferences.

Data:

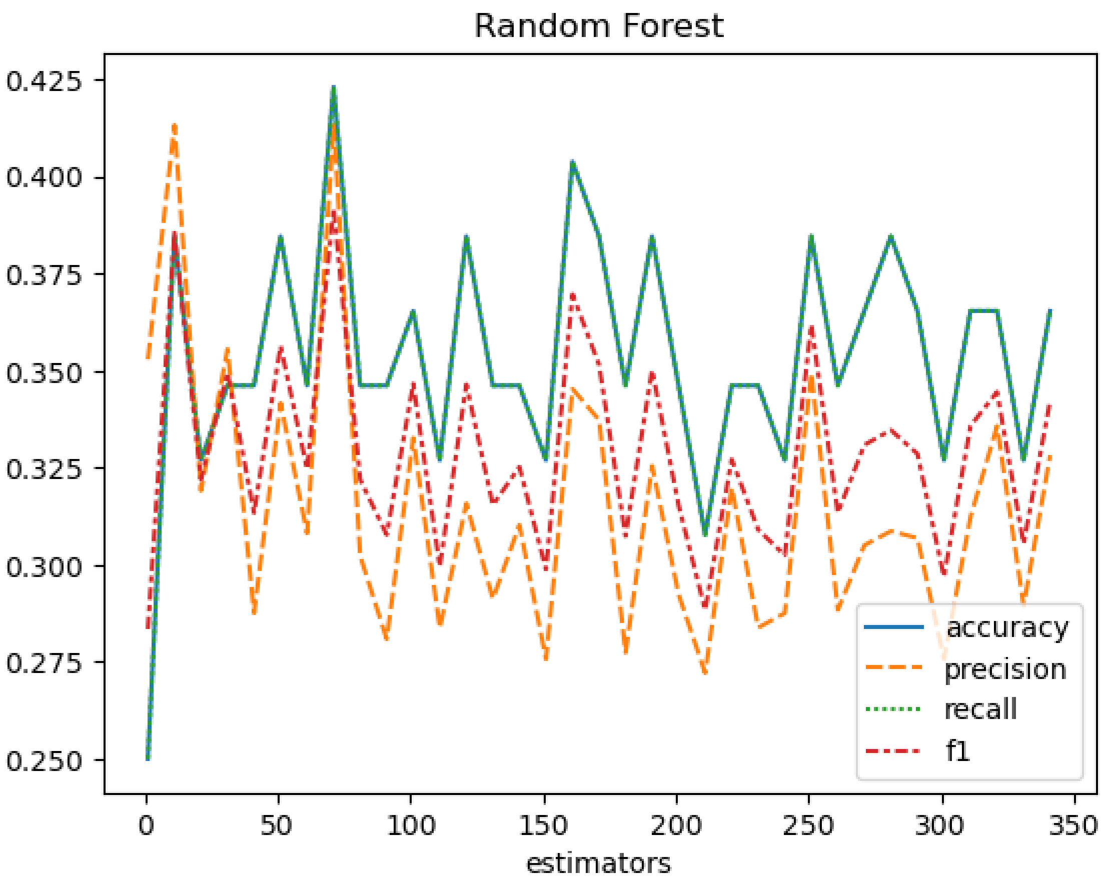
The data set used for this project was found on Kaggle. Three main groups of data were kept in order to train and test the models. The first is the country of origin, which is our target. The next two are the features used to classify the coffee: scores and moisture percentage. The scores include aroma, flavor, aftertaste, acidity, body, and balance. Sweetness and clean cup were included in the original data set, but were both the same value for all coffees. These attributes were scored on a 1-10 scale, and were scaled to 0-1 for training the models. Additionally, the moisture percentage was scaled to the same range. Lastly, an overall score for each coffee was included in the original data set, but is likely a composite score since it did not improve any of the models performance, so it was dropped.

Models:

Three types of classification models were tested: K Nearest Neighbors, Random Forest, and Support Vector. The model that performed the best was the K Nearest Neighbors model. When setting the number of neighbors to 7, the model reached 0.46 accuracy, 0.43 precision, 0.46 recall, and an F1 score of 0.42. This optimal number of neighbors was found by plotting the four metrics for models with the neighbors hyperparameter of 1 through 30.



The next most performative model was the Random Forest Classifier. With the number of estimators set to 71, the model achieved an accuracy of 0.42, a precision 0.42, recall of 0.42, and an F1 score of 0.39. The estimators hyperparameter was found in a similar way to the neighbors of the last model. The number of estimators checked ranged from 1 to 341, with 10 step increments.



Lastly, the least performative model was the Support Vector Classifier. GridSearchCV was used to tune the C and kernel hyperparameters. The values of C that were tested are 1, 5, 10, 50, and 100. The kernels tested were rbf, poly and sigmoid. The GridSearchCV found a C of 1 and the rbf kernel to be optimal. This model resulted in an accuracy of 0.35, a precision of 0.11, a recall of 0.35, and an F1 score of 0.18.

Summary:

For this application, the K Nearest Neighbors classifier performed the best at predicting the country of origin for a batch of coffee. Additionally, it is also the simplest model to interpret, as its predictions are based on the surrounding data points.

Ultimately however, none of the models were able to predict the country of origin more than half of the time. From looking at the decision making of the model it seems that it is hard for the model to use these attributes to determine the country of origin, as coffee from different countries overlap in score and moisture percentage. Perhaps other information about the coffee, such as flavor profile or color of the roasted beans, could help the model to be more performative.