Relating Coffee Species to Brew Taste and Region with Machine Learning By Sid Shah

Abstract

Buying coffee beans could be a confusing task, due to the large variety, each with their own region and flavor profiles. For many consumers this could lead to overbuying, or making unwanted purchases. If coffee flavor could be correlated to species and even country of origin, coffee purchasing would be much simpler, as consumers could use their own preferences to determine what to buy. This study aimed to find that correlation, using Data from the Coffee Quality Institute, taken from kaggle, and make predictions on the species based on the taste characteristics of the beans. Due to limitations with the data, and keeping in mind that the models tested weren't tuned to perform their best, the Random Forest model performed best, having one of the highest accuracies with the most favorable set advantages and disadvantages.

Introduction

About 2.25 billion cups of coffee are consumed in the world daily, some by those who go to coffee shops, some by people who make it at home. (Lee, 2023) Coffee beans constantly have variety, and each bag is not exactly the same as any other. With all the factors involved in the cultivation, roasting, and bagging of coffee beans, was there a way to use how it tastes and where the coffee originated to determine its species. One step further, could the taste of coffee help determine region, and in turn, help consumers decide which beans to buy? Because of the wide variety, it's easy to buy a bag that doesn't taste that great, leaving consumers less than satisfied for the weeks they have the bag for. If certain features could help decide how the coffee would probably taste, decisions at the grocery store would be made easier. The results of this study showed that coffee species could very accurately be predicted by flavor and regional features, but due to the small quantity of the data, further conclusions could not be reached.

Methodology

The data was found on Kaggle, which itself was retrieved from the Coffee Quality Institute website (B., n.d.). The data was collected by trained coffee reviewers who rated the coffee, and was retrieved from their website in January, 2018. The features included ratings for certain flavors found in the coffee, as described below:

Aroma: Refers to the scent or fragrance of the coffee.

Flavor: The flavor of coffee is evaluated based on the taste, including any sweetness, bitterness, acidity, and other flavor notes.

Aftertaste: Refers to the lingering taste that remains in the mouth after swallowing the coffee.
Acidity: Acidity in coffee refers to the brightness or liveliness of the taste.
Body: The body of coffee refers to the thickness or viscosity of the coffee in the mouth.
Balance: Balance refers to how well the different flavor components of the coffee work together.
Uniformity: Uniformity refers to the consistency of the coffee from cup to cup.
Clean Cup: A clean cup refers to a coffee that is free of any off-flavors or defects, such as sourness, mustiness, or staleness.
Sweetness: It can be described as caramel-like, fruity, or floral, and is a desirable quality in coffee. [Kaggle]

There was also farm data, including the region, the farm, the lot number, and country of origin, and coffee species. Any entries with missing values were removed from the dataset.

There are two main types of species— Arabica and Robusta. Arabica has been historically associated with better quality, even specialty coffees, while Robusta has been associated with cheaper, instant coffees. This distinction does have some ground due to the difference in flavor profile.

The taste of coffee however, is controlled by many factors. Even the altitude at which it was grown and the shade it received plays apart in taste. (Hanum & Karim, 2023). Location, due to soil composition and climate also plays a major role, as coffees from different countries have different flavor profiles (*Coffee Regions and their Different Flavour Profiles*, n.d.)

However, this dataset was mostly filled with entries of Arabica beans. To prepare the data for use, the data was balanced, to account for the discrepancy between Arabica and Robusta bean data.

Experiment

To better understand the data that was worked with, some of the following plots and graphs were generated.

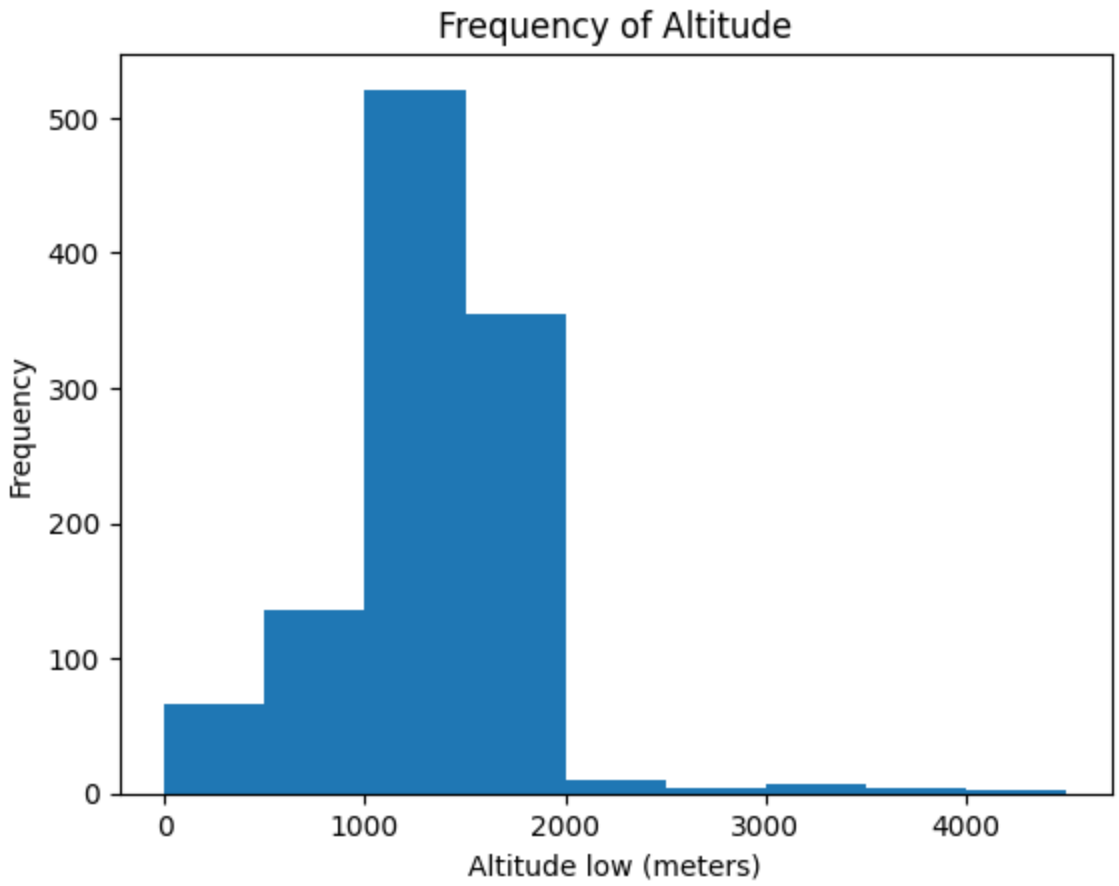

Figure 1: Frequency of growth altitude of various batches of coffee beans

Figure 1 is a histogram of the altitude of which each batch of coffee beans was grown at, and shows that regardless of species or region, most coffee grows between 1000 and 2000 meters above sea level. The distribution is right skewed and unimodal, with a mean of about 1500 meters, and a mode of 1000. In other words, most farms have a low point between 1000 to 2000 meters. Coffee grown at higher altitudes, up to 4499 meters, is much less frequently present.

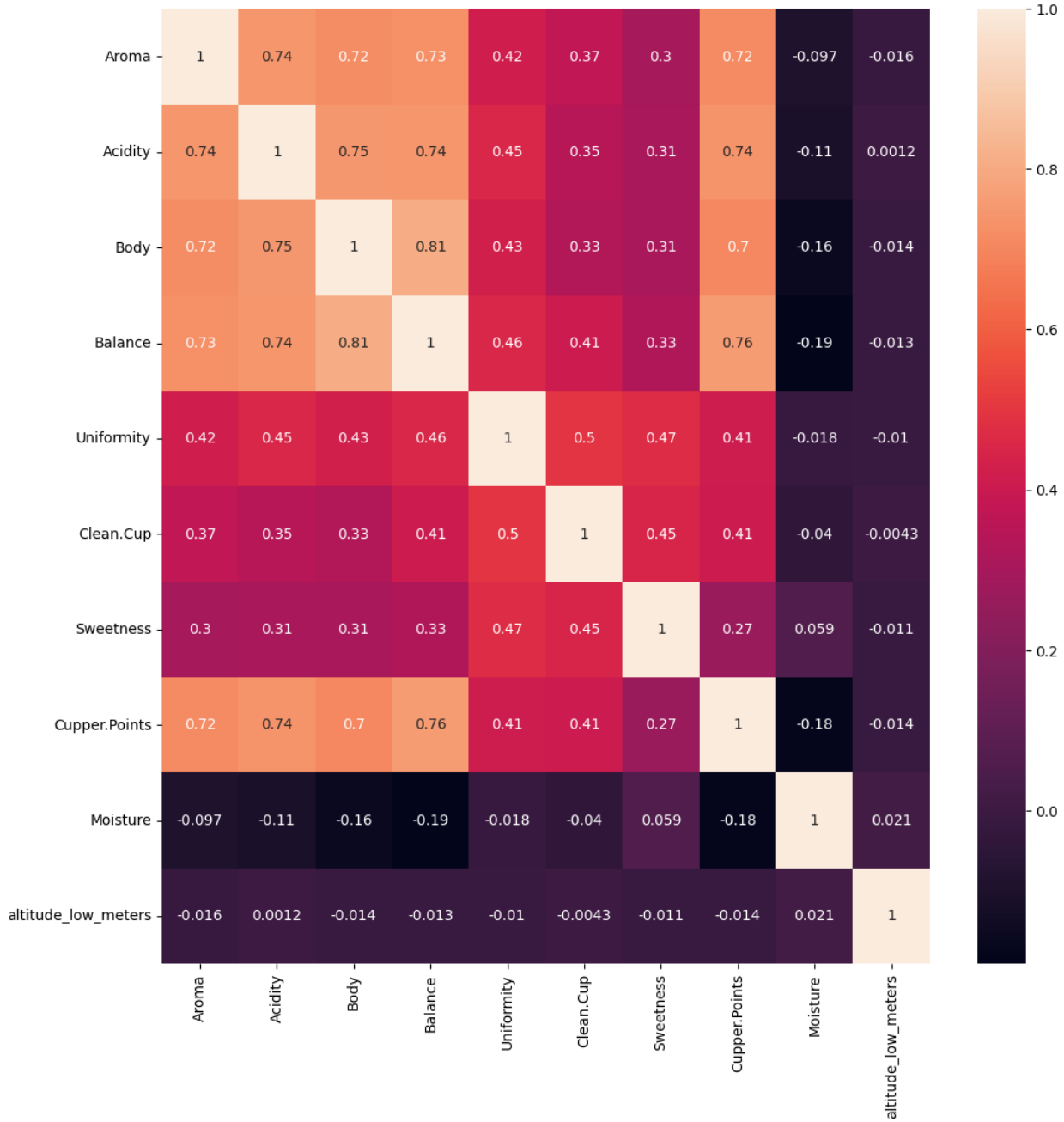

Figure 2: Correlation heat map of how linearly related each feature is to each other.

The correlation heat map above shows the correlations between features. The correlations are related from -1 to 1, with values close to the extremes (i.e., 0.8) signifying “high or strong” correlations, and the values close to 0 (i.e., 0.2) signifying “low or weak” correlations. Those with high correlations to each other shouldn’t generally be used as features, due to the likelihood of dependence to be mostly on the highly correlated variable. In this heatmap, the higher correlations between *Body* and *Balance* points suggest that *Body* and *Balance* can typically be used to predict each other, likely because if the body scores high enough, it could be improving

the balance of the coffee considerably. Conversely, *Balance* and *Moisture* are not strongly correlated, so the *Moisture* in the beans may have little to no obvious effect on the Balance, or vice versa.

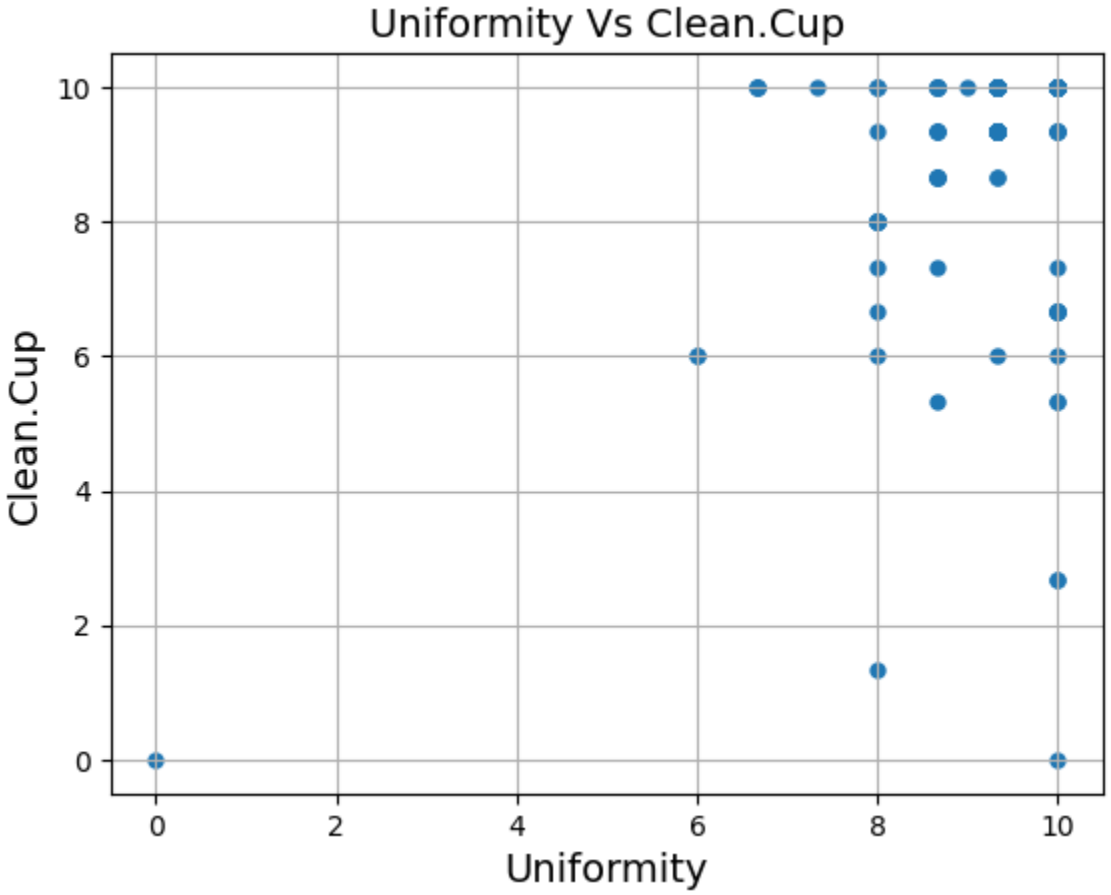

Figure 3: Scatter plot displaying the relationship between Clean.Cup and Uniformity

Figure 3 deptics *Uniformity* and *Clean.Cup* with a correlation of 0.5 (shown in Figure 2), and that is evident in the moderate to weak, positive correlation present. The data is not strongly correlated due to how spread out the data points are, so any pattern between *Clean.Cup* and *Uniformity* is harder to discern. The positive direction indicates that as *Uniformity* points increase, *Clean.Cup* points do as well, The *Uniformity* mostly ranged from 6 to 10, with one unusual value at 0, while the *Clean.Cup* had most of the data between 5 and 10, with 4 unusual values, all of which were not close to the main cluster of data points. There is one main cluster to the scatter plot.

To preprocess the data, any rows with empty species or feature columns were removed, to remove any bias coming from a lack of information. Over 97% of the rows described Arabica beans, so to prevent models from favoring or only returning “Arabica”, the labels were balanced. To balance the labels “Arabica” and “Robusta”, an oversampling technique was used, bringing the balance to an even split between the two. Oversampling was used rather than undersampling because of the limited size of the dataset.

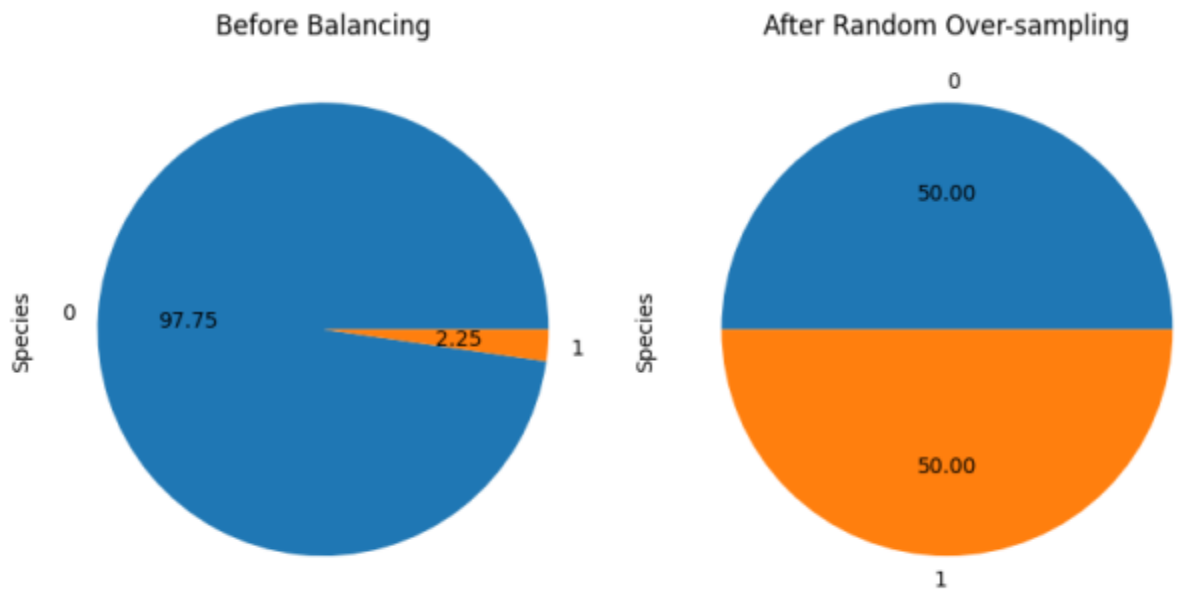

Figure 4: Charts showing the percent of the data with each of the labels, before and after label-balancing.

Note: Arabica is represented as 0, and Robusta as 1 in the pie charts above.

After the preprocessing, the models were trained. The following section includes descriptions for the 8 models trained.

Model Descriptions

Decision Tree

Decision trees are essentially a map of checks, or "decisions" arranged such that the decision of one check influences the next check. They are often represented with nodes and paths branching from each node to new nodes, based on the decision at the original node. They can however, be very unstable, with a few data points changing the tree drastically. Furthermore, they are usually less accurate than other models on most datasets.

Random Forest

Random forest-based models use decision trees. These models make many trees, each trained on a certain group of data and features combination, with the intention that many trees together will likely make a good decision, even with a few trees being inaccurate. Forests are more robust than standalone trees, but the computation time increases with each additional tree linearly. They can be trained in parallel however, so the time can yet be decreased.

Gradient Boosting

Gradient Boosting works on a similar principle to random forests, except instead of having multiple independent trees, the trees are trained based on the errors of each subsequent tree. As a result, Gradient Boosting can be more accurate, but it has a high risk of overfitting with noisy data.

XGBoost

XGBoost, or eXtreme Gradient Boosting, is just like gradient boosting, except that it has a regularization technique applied to it for better accuracy and faster performance.

Naïve Bayes (Gaussian)

Gaussian Naive Bayes functions as a probabilistic classifier, leveraging Bayes' theorem. The crux lies in assuming feature independence and a Gaussian distribution for these features. It simplifies the classification process, treating each feature as uncorrelated, which aids computational efficiency.

The classifier determines the probability of a data point belonging to a particular class based on the likelihood of its features following a Gaussian distribution. Despite its simplicity and efficiency, Gaussian Naive Bayes may falter if the independence assumption is violated or if the data doesn't conform to Gaussian distributions.

In essence, it's a streamlined approach, assuming simplicity in feature relationships to make quick and informed classifications.

K-Nearest Neighbors

K-Nearest Neighbors (KNN) is an intuitive classification algorithm relying on the majority class of the k nearest neighbors in feature space to classify a data point. Its simplicity lies in its direct interpretation — a data point is akin to its closest neighbors. However, KNN's computational cost can be high for extensive datasets, and it's sensitive to irrelevant features.

Support Vector Machine

Support Vector Machines are classifiers seeking an optimal hyperplane to separate distinct classes in the feature space. They excel in high-dimensional spaces and capture complex relationships. However, they are sensitive to kernel and parameter choices, and training time can be substantial, particularly for large datasets.

Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is an iterative optimization algorithm for model training. It processes random subsets or individual data points to minimize the cost function. SGD's strength lies in its efficiency with large datasets, but its stochastic nature introduces randomness,

potentially leading to noisy updates. Proper tuning of learning rates is crucial for optimal performance.

Metric Descriptions

For each of the models, specific metrics were used to score how well the model did compared to the others.

Accuracy Score

The accuracy score is a measure of how many predictions were correct throughout the entire testing dataset.

F1 Score

The F1 Score works like the Accuracy score, but it works well for class-imbalanced data, because it calculates accuracy for each class, rather than for the entire dataset, so models which only predict the “positive” or only predict the “negative” classes can be easily discerned due to having a lower score.

AUC Score

The AUC score is a measure of how well a binary classification model separates the “positive” and “negative” classes. It stands for Area Under the ROC Curve. The ROC (receiver operating characteristic curve) plots the True Positive Rate against the False Positive Rate, and the AUC Score takes the area under the curve from 0 to 1.

Results

Model	Accuracy Score	F1 Score	AUC Score
Gradient Boosting	1.000000	1.000000	1.000000
Decision Tree	1.000000	1.000000	1.000000
Random Forest	1.000000	1.000000	1.000000
XGBoost	0.997696	0.997743	0.997653
Naive Bayes (Gaussian)	0.995392	0.995495	0.995305

Model	Accuracy Score	F1 Score	AUC Score
K-Nearest Neighbors	0.988479	0.988814	0.988263
Support Vector Machine	0.569124	0.294340	0.576498
Stochastic Gradient Descent	0.559908	0.666667	0.554192

Figure 5: Performance of each of the tested models, with the aforementioned metrics

In Figure 5, a list of 8 different models are shown, along with their scores on the same data. Of them all, Decision Tree Model, Gradient Boosting and Random Forest models performed the best, with an accuracy score of 1. The F1 scores displayed a similar trend, and describe the lack of false positives and negatives in the aforementioned 3 models. The AUC scores, which shows how well the binary classification is, are also relatively high, with Random Forest, Gradient Boosting, and Decision Tree Models having a perfect score of 1, and the lowest being the Stochastic Gradient Descent and support vector machine models, with scores slightly greater than 0.55.

It must be noted that, even though the scores are good, they aren't very realistic, as accuracy scores being that high for such a small dataset may mean there is overfitting, as having perfect accuracy when predicting the taste of coffee is a peculiar result.

Of all 8 models, 3 showed similar diagnostics, and all showed promise for being the best model for the situation: the Decision Tree, Gradient Boosting, and Random Forest models. They scored 1 in the Accuracy, F1, and AUC scores. However, because overfitting is a potential problem, the Decision Tree may not be the best model because it's very responsive to changes, and thus is prone to overfitting when the tree is complex, though it is easiest to interpret. Gradient Boosting models also have a high overfitting risk (because noisy data is constantly accounted for with the subsequent "trees"), outlier sensitivity, and also have a lot of data preprocessing requirements, which makes Random Forest the best at classifying coffee beans based on their species, due to its balanced complexity, while still boasting powerful prediction capabilities.

Limitations

There are a number of limitations present with the methodology described. The original data was a very small sample size of all the coffee present in the world. Furthermore, it's not current data, so conclusions drawn cannot be simply related to coffee grown in 2023 or later. The collection methods of the data have not been specified, so the determination of each of the flavor-related

scores is unknown. This can pose a problem, especially if such scores were given by humans using their senses, as it introduces bias from human opinion. Additionally, the data for the Robusta species has been artificially balanced, with oversampling, so certain patterns may have been artificially magnified when they shouldn't have been.

Conclusion

All in all, most of the models performed well, with Random Forest being one of the highest performing, with the least disadvantages with its implementation. It had a 100 percent success rate in predicting coffee species based on flavor and regional data, though much of the data was artificially over-sampled, to balance the labels.

To expand on the project, tuning and adjusting each model so it performs its best would be the next logical step, as some of the algorithms perform better after heavy tuning, which was not completed in this study.

After that, finding more information on the data acquired, or even completely recollecting it, with documented processes for each step, would remove some of the uncertainties explained in the *Limitations* section.

Coffee is a complex beverage, and how it tastes cannot simply be drawn to one specific factor. The variance in region, species, and roasting all play a part in how each batch of beans tastes compared to others. Because of this however, finding coffee that appeals to a consumer's tastes is no simple task, and with further work done on this project, consumers can buy coffees in such a way that they'll know exactly what to expect in their next morning cup.

References

B, F. (n.d.). *Coffee Quality Data (CQI May-2023)*. Kaggle. Retrieved December 1, 2023, from https://www.kaggle.com/datasets/fatihb/coffee-quality-data-cqi

Coffee Regions and their Different Flavour Profiles. (n.d.). Perk Coffee. Retrieved December 1, 2023, from https://perkcoffee.co/my/coffee-regions-and-their-different-flavour-profiles/

Hanum, C., & Karim, A. (2023, March). *(PDF) The Taste of Arabica Coffee in Several Altitude and Shading Condition*. ResearchGate. Retrieved December 1, 2023, from https://www.researchgate.net/publication/369412029_The_Taste_of_Arabica_Coffee_in_Several_Altitude_and_Shading_Condition

Lee, T. (2023, October 27). *97+ Coffee Statistics For 2023 (US, UK, Worldwide & More)*. Tim's Coffee. Retrieved December 1, 2023, from https://timscoffee.com/blog/statistics/