Classifying Vegan Foods

Soumya Mahavadi

IB Computer Science HL

Period 4B

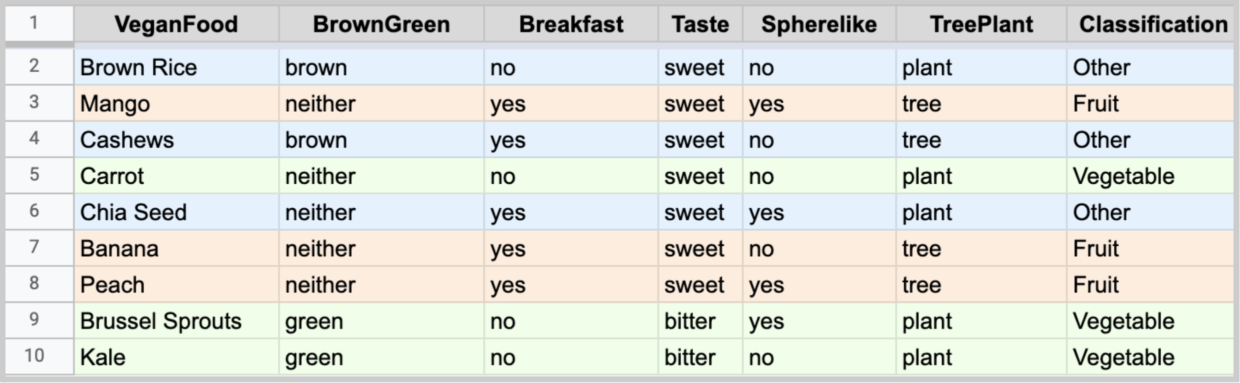
Approaching the Topic

As an individual who has been a vegetarian their entire life, I was instinctively drawn to choosing the topic of vegan foods for this decision tree project. I knew that the two diets were quite similar, with the exception that dairy and eggs are included in a vegetarian diet and not in a vegan one. Prior to my background research of the topic, I knew that fruits and vegetables could be classified both botanically and culinarily, and these classifications did not always match for each respective vegan food. For example, a tomato is commonly considered to be a vegetable—especially while cooking—but in reality, it is classified as a fruit. I was extremely intrigued by this difference in classification and wanted to learn how fruits and vegetables are botanically defined and classified. Hence, through research, I discovered that fruits are the products of trees or other plants and that they have seeds while vegetables are edible parts of plants and are usually eaten as a part of a larger meal. Within my research, I was most surprised to learn that an eggplant is actually a fruit and not a vegetable like I had believed for my entire life, up until this point. It makes sense since eggplants contain seeds, but I still find this classification fascinating.

Columns/Questions

The columns in my dataset were mainly centered around features to identify fruits and vegetables specifically, where any foods that did not match such features would be classified as “other.” I made this decision since “other” is more of a miscellaneous category, meaning that it does not have many defining traits. In total, I came up with five possible columns, or questions, for my decision tree. Two of the questions answered either “yes” or “no,” but in reality, four out of five of my questions were binary. The only question that was nonbinary in my dataset dealt with the color of the vegan foods since they could be labeled as brown, green, or neither.

Furthermore, for this column specifically, I was able to give the "other" category the defining trait of being brown in color since I noticed that many foods in this category—including brown rice, walnuts, and almonds—were in fact brown.

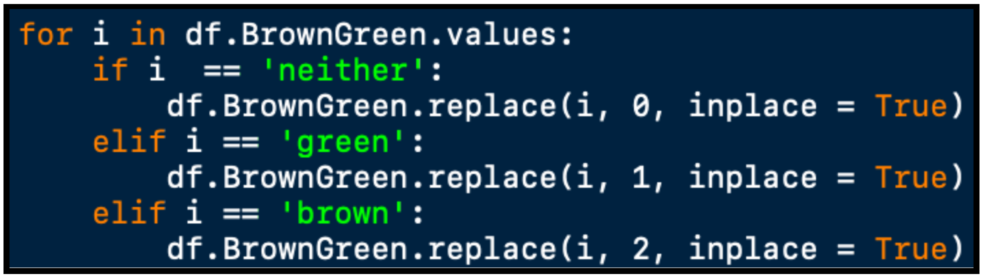


1	VeganFood	BrownGreen	Breakfast	Taste	Spherelike	TreePlant	Classification
2	Brown Rice	brown	no	sweet	no	plant	Other
3	Mango	neither	yes	sweet	yes	tree	Fruit
4	Cashews	brown	yes	sweet	no	tree	Other
5	Carrot	neither	no	sweet	no	plant	Vegetable
6	Chia Seed	neither	yes	sweet	yes	plant	Other
7	Banana	neither	yes	sweet	no	tree	Fruit
8	Peach	neither	yes	sweet	yes	tree	Fruit
9	Brussel Sprouts	green	no	bitter	yes	plant	Vegetable
10	Kale	green	no	bitter	no	plant	Vegetable

Figure 1: Screenshot of Dataset

Creating the Decision Tree

I used an automatic approach to create my decision trees as I believed that the computer would more efficiently classify the given vegan foods than I would. Rather than using the code provided in class—which utilized the decision tree classifier—to achieve this automatic approach, I instead used the code from another website that I found online. This decision was made because I ran into a few errors when running the Graphviz module on my macOS software. After browsing through the internet, I was able to find a program that did not utilize Graphviz, allowing it to run on my computer without any errors.



```
for i in df.BrownGreen.values:
    if i  == 'neither':
        df.BrownGreen.replace(i, 0, inplace = True)
    elif i == 'green':
        df.BrownGreen.replace(i, 1, inplace = True)
    elif i == 'brown':
        df.BrownGreen.replace(i, 2, inplace = True)
```

Figure 2: Snippet of Code to Convert from Qualitative to Quantitative Values

However, before running the program, I had to make a few edits to the code. The program uses the scikit-learn, which means that it processes quantitative values. My dataset consisted of qualitative descriptions, such as “yes,” “no,” “sweet,” “green,” etc. Therefore, I had to convert these qualitative descriptions into numeric values to create my decision trees (seen in Figure 2).

Entropy Value of Root Node	1.582
Gini Value of Root Node	0.665
Number of Columns/Questions Used (Entropy)	5
Number of Columns/Questions Used (Gini)	3
Accuracy Rate (Entropy)	83%
Accuracy Rate (Gini)	83%

Table 1: Values for Original Decision Trees with Training Data

When considering which tree to use to manually classify Potter’s additions, accuracy was my top priority. However, this was not a deciding factor between the two trees, since both Gini and entropy resulted in an 83% accuracy rate, or a 17% error rate. Thus, I turned to my next priority, which concerned the complexity and structure of the decision trees themselves. Ultimately, since the entropy tree employed more questions than the Gini tree, I decided to proceed with the former for the rest of my project.

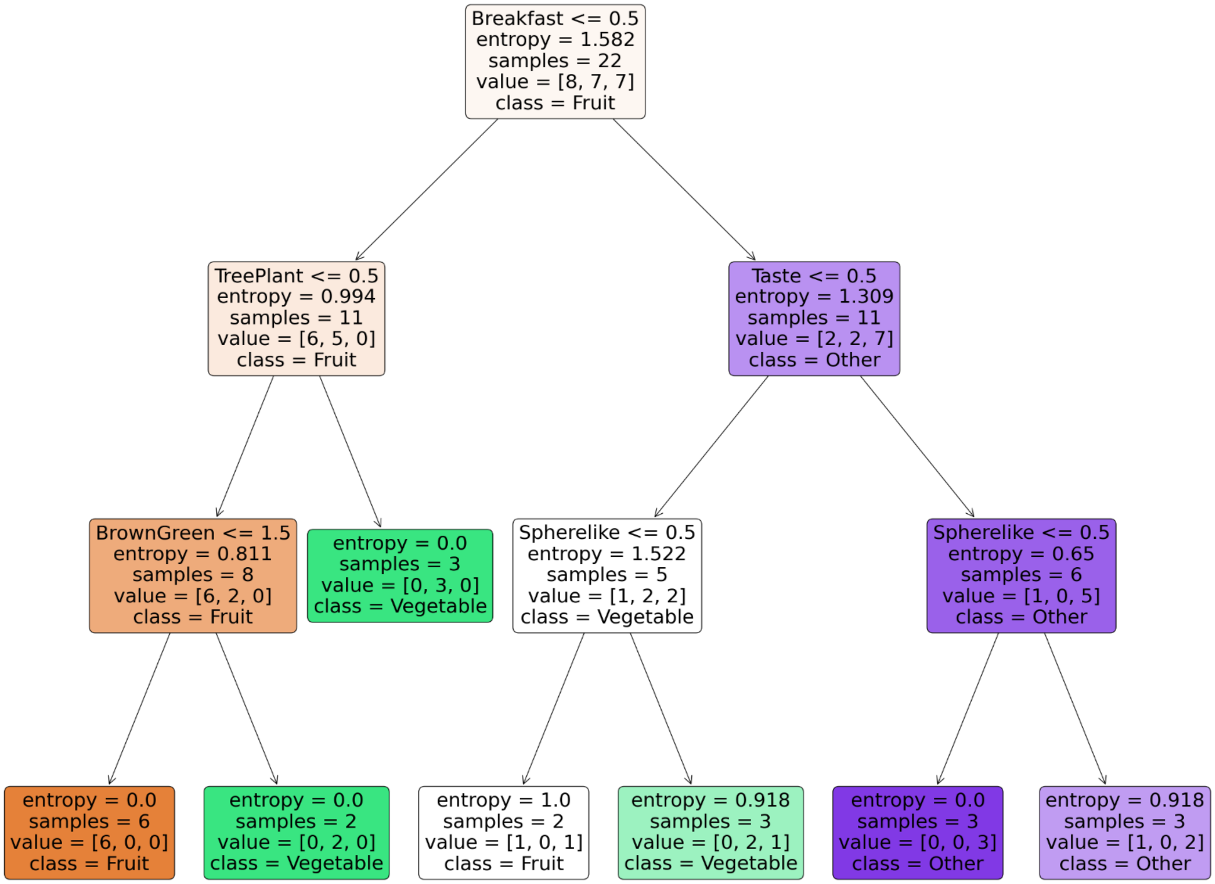


Figure 3: Original Entropy Decision Tree

Potter's Additions

Vegan Food	BrownGreen	Breakfast	Taste	Spherelike	TreePlant
Rhubarb	neither	no	bitter	no	plant
Spelt	brown	yes	sweet	no	plant
Olives	neither	no	sweet	yes	tree

Table 2: Collected Data for Potter's Additions

I used the search engine Google to research Potter's additions and accordingly collect data for the columns in my dataset. Most of the research I gathered seemed accurate, but I was a bit concerned with the information Google provided about the taste of olives. While I thought olives are more bitter, I read online that they are mildly sweet. As an individual who does not particularly eat olives, I decided to proceed with Google's evaluation of taste rather than my own opinion. Nonetheless, besides this small concern, I did not run into other challenges with Potter's additions.

Vegan Food	Entropy Tree Classification	Real Classification
Rhubarb	Other	Vegetable
Spelt	Vegetable	Other
Olives	Fruit	Fruit

Table 3: Entropy Tree vs Real Classifications of Potter's Additions

As shown in Table 3, the entropy decision tree was able to accurately classify only one out of Potter's three additions: olives. Looking at the two errors with more detail, it seems as if the tree had trouble discerning whether a food should be classified as a vegetable or as "other."

Entropy Value of Root Node	1.555
Gini Value of Root Node	0.653
Number of Columns/Questions Used (Entropy)	3
Number of Columns/Questions Used (Gini)	4
Accuracy Rate (Entropy)	57%
Accuracy Rate (Gini)	43%

Table 4: Values for New Decision Trees with Training and Test Data

When comparing the new entropy and Gini trees with their respective original trees, there was not a significant difference between their respective values of the root nodes. That being said, however, there was a considerable change in accuracy. While the Gini validation's accuracy dropped 40% with Potter's additions, the entropy validation only decreased by 26%, resulting in a higher accuracy rate and a lower error rate. Since accuracy was my top priority, I, once again, chose the entropy tree over the Gini tree.

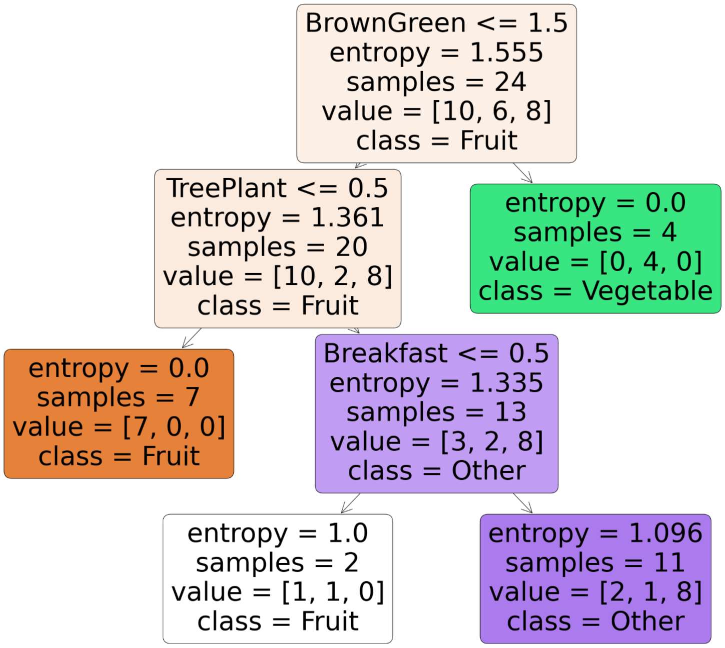


Figure 4: New Entropy Decision Tree with Potter's Additions

Conclusion

Through the completion of this project, I was able to delve into the botanical differences between fruits and vegetables—something I wasn't particularly aware of before. I also learned a lot about decision trees as a classification tool. In particular, I initially thought that automatic decision trees would be a very accurate tool, similar to the k-nearest-neighbors (KNN) algorithm; however, through this project, I realized that my presumption was wrong. There are a few errors in the new entropy tree, such as the fact that it classifies brown foods as vegetables—which can be seen in the root node—instead of "other." While I believe a decision tree that was manually created would be more effective, in terms of code, I have come to the conclusion that automatic decision trees are not the most accurate. Instead, I would rather use KNN as a classification tool.