Jota Yamaguchi
26th January, 2023

Project: weed_stuff

The Intro

I know I'm fascinated by data and analytics. The way I live is very analytical, very calculated. I think through every possible outcome to plan for the best and prepare for the worst. I get it from my Mom. This way of thinking helps me avoid things that could harm me. It also makes me a very anxious kid. I find it very difficult to try new, difficult things. For the past couple years, I was stuck in a comfort zone. But this changes. I'm going to make myself as uncomfortable as possible, pushing the boundaries of my comfort zone to test my limits. See what I'm really made of. These projects are to challenge myself. I want to solve real problems instead of homework questions. These aren't tasks or assignments. These projects are an expression of my genuine curiosity for the world.

Given this is my first project of the year, I wanted to start off with a topic that I was very familiar with. Allow me to introduce one of my best friends from the past three years. Her name is Marijuana, but she goes by Weed. Weed and I have been having an on-and-off relationship for a while now. It used to be super toxic. I would call her up when I needed a sweet escape from reality and she would provide it. The sweet escape was so relieving that she started to have control over me and I had no idea. Maybe I did and I just didn't care.

The Why

I don't want you to think I was the type of guy that was high 24/7. I had the bare-minimum self control to only use her about four times a week. I still prioritized my work to make sure I was being productive. I was just never present. It took me a while to finally break free from her grasp. I met a girl that helped me evaluate myself and why Weed had such a grasp on my life. With her help, I was finally able to take control of Weed and now I can use Weed to help myself..

The reason I have an interest in Marijuana is because I believe that: IF (emphasis on <u>**if**</u>) you have control over her, you can reap the benefits provided by her. My main curiosity is in the area of mental health benefits, but I have read that there are some physiological benefits as well.

With my analytical skills, curiosity, and personal experience, I want to help those that may benefit from controlled use of Marijuana. Furthermore, I know there are plenty of people that are under her control. I'm hoping that progressing and sharing the understanding of Marijuana will prevent more victims from falling into her grasp while also helping those break free.

The Question

From my understanding, there are two types of marijuana: Indica and Sativa. To simplify their differences, Indica has a more relaxing effect while Sativa has an uplifting, euphoric effect. What contributes to these differences are the amounts and ratios of certain chemical compounds in Marijuana. The most notable compounds are THC and CBD. Indicas tend to have a higher amount of CBD while Sativa tends to have a higher amount of THC. Based on this, I think I could create a machine learning algorithm that can tell whether a certain Marijuana product is Sativa or Indica based on its chemical compounds. Therefore my question for this project will be:

Could a Marijuana product be classified as Indica or Sativa based on its chemical compounds using ML?

The Beginning

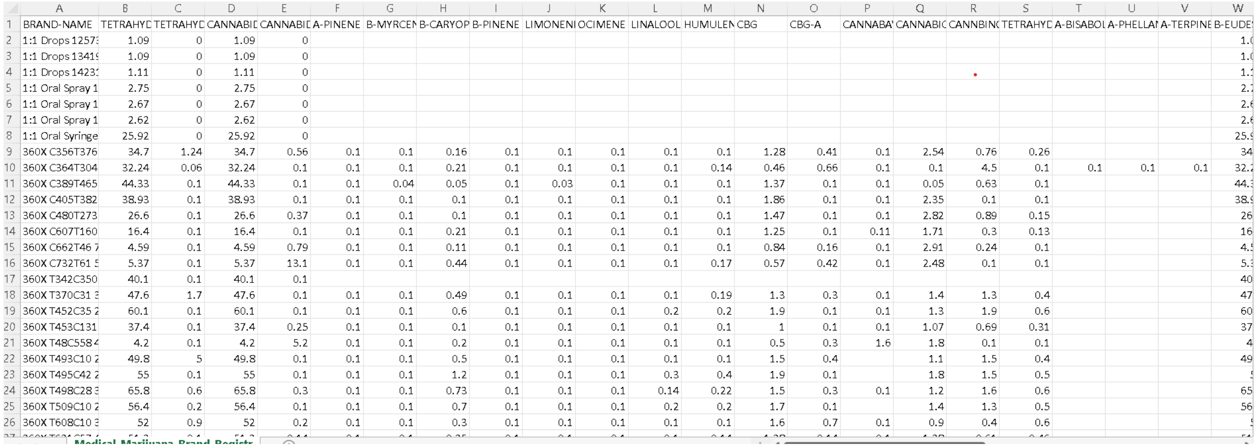
First things first, research. Luckily I already have a decent understanding of how Marijuana, THC and CBD . Without diving too deep into the neurochemistry, THC is responsible for the "high" feeling in your head, while CBD is responsible for more relaxing feelings across your whole body. As stated earlier, Indica plants will tend to have higher concentrations of CBD (Indica relaxes) while Sativa plants tend to have higher concentrations of THC.

The next step I need to take is getting the data. Now normally this step is already taken care of since homework assignments and projects are assigned with a dataset to work with. However in this case, I would have to find my own. I knew I was looking for a dataset that had the amount of CBD, THC, and any other chemical compound in a certain product. I figured I would have three options. 1) I could deep dive into the web and pray that I find a dataset containing exactly what I am looking for. 2) I could find a website containing the desired data and try to web scrape it. 3) Go on Kaggle to find a dataset.

I chose to Kaggle it. I understand how important the quality of the data is but I couldn't find a website that had the data I wanted and I also don't have the web scraping skills to collect the necessary data without wasting a significant amount of time (maybe I'll do a project on web scraping soon). Luckily, I found a dataset titled " Marijuana Brand Registry Data" that contains data on Marijuana products and its chemical compounds. The data is collected by the department of Consumer Protection which makes me believe that it should be a reliable source. The only issue is that the datapoints are not labeled with either Sativa or Indica. Therefore, I would have to use an unsupervised ML model in order to find patterns in the data to help me cluster them.

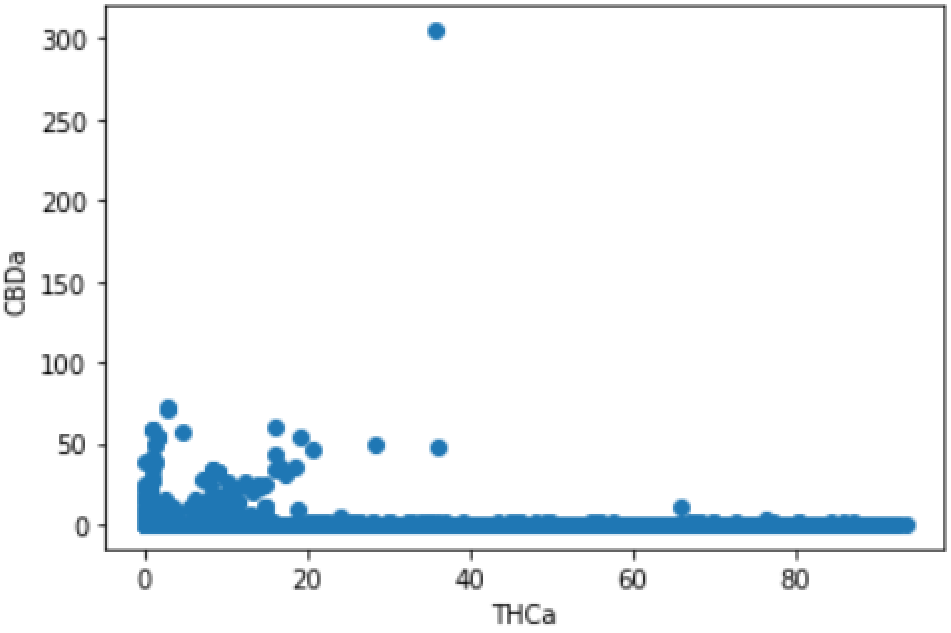
The Preparation

The data looks messy. Some values are represented as decimals while others as percentage and there are plenty of typos ("0..32" or "0.2."). The first thing to do is get rid of any columns I know I won't need such as columns containing urls, images, dates, etc. This helps me simplify each datapoint into a product and its chemical compounds. Next, I want to clean each datapoint so that it's usable. I turned all values into a float and dropped any row with a typo. I can now load it into a Pandas Dataframe and take a look at its first 5 values. There are a lot of NaN values. What to do? To solve this I used the following intuition: If a certain product has an unlabeled amount of a certain chemical, that would mean that the product does not contain that chemical. Therefore, NaN = 0. With that, the data is ready for use.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	BRAND-NAME	TETRAHYD	TETRAHYD	CANNABID	CANNABID	A-PINENE	B-MYRCEN	B-CARYOP	B-PINENE	LIMONENI	OCIMENE	LINALOOL	HUMULEN	CBG	CBG-A	CANNABA'	CANNABIC	CANNBIN(	TETRAHYD	A-BISABOL	A-PHELLAI	A-TERPINE	B-EUDE
2	1:1 Drops 1257	1.09	0	1.09	0																		1.0
3	1:1 Drops 1341	1.09	0	1.09	0																		1.0
4	1:1 Drops 1423	1.11	0	1.11	0																		1.1
5	1:1 Oral Spray 1	2.75	0	2.75	0																		2.7
6	1:1 Oral Spray 1	2.67	0	2.67	0																		2.6
7	1:1 Oral Spray 1	2.62	0	2.62	0																		2.6
8	1:1 Oral Syringe	25.92	0	25.92	0																		25.9
9	360X C356T376	34.7	1.24	34.7	0.56	0.1	0.1	0.16	0.1	0.1	0.1	0.1	0.1	1.28	0.41	0.1	2.54	0.76	0.26				34
10	360X C364T304	32.24	0.06	32.24	0.1	0.1	0.1	0.21	0.1	0.1	0.1	0.1	0.14	0.46	0.66	0.1	0.1	4.5	0.1	0.1	0.1	0.1	32.2
11	360X C389T465	44.33	0.1	44.33	0.1	0.1	0.04	0.05	0.1	0.03	0.1	0.1	0.1	1.37	0.1	0.1	0.05	0.63	0.1				44.3
12	360X C405T382	38.93	0.1	38.93	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1.86	0.1	0.1	2.35	0.1	0.1				38.9
13	360X C480T273	26.6	0.1	26.6	0.37	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1.47	0.1	0.1	2.82	0.89	0.15				26
14	360X C607T160	16.4	0.1	16.4	0.1	0.1	0.1	0.21	0.1	0.1	0.1	0.1	0.1	1.25	0.1	0.11	1.71	0.3	0.13				16
15	360X C662T46 7	4.59	0.1	4.59	0.79	0.1	0.1	0.11	0.1	0.1	0.1	0.1	0.1	0.84	0.16	0.1	2.91	0.24	0.1				4.5
16	360X C732T61 5	5.37	0.1	5.37	13.1	0.1	0.1	0.44	0.1	0.1	0.1	0.1	0.17	0.57	0.42	0.1	2.48	0.1	0.1				5.3
17	360X T342C350	40.1	0.1	40.1	0.1																		40
18	360X T370C31 3	47.6	1.7	47.6	0.1	0.1	0.1	0.49	0.1	0.1	0.1	0.1	0.19	1.3	0.3	0.1	1.4	1.3	0.4				47
19	360X T452C35 2	60.1	0.1	60.1	0.1	0.1	0.1	0.6	0.1	0.1	0.1	0.2	0.2	1.9	0.1	0.1	1.3	1.9	0.6				60
20	360X T453C131	37.4	0.1	37.4	0.25	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1	0.1	0.1	1.07	0.69	0.31				37
21	360X T48C558 4	4.2	0.1	4.2	5.2	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.5	0.3	1.6	1.8	0.1	0.1				4
22	360X T493C10 2	49.8	5	49.8	0.1	0.1	0.1	0.5	0.1	0.1	0.1	0.1	0.1	1.5	0.4		1.1	1.5	0.4				49
23	360X T495C42 2	55	0.1	55	0.1	0.1	0.1	1.2	0.1	0.1	0.1	0.3	0.4	1.9	0.1		1.8	1.5	0.5				5
24	360X T498C28 3	65.8	0.6	65.8	0.3	0.1	0.1	0.73	0.1	0.1	0.1	0.14	0.22	1.5	0.3	0.1	1.2	1.6	0.6				65
25	360X T509C10 2	56.4	0.2	56.4	0.1	0.1	0.1	0.7	0.1	0.1	0.1	0.2	0.2	1.7	0.1		1.4	1.3	0.5				56
26	360X T608C10 3	52	0.9	52	0.2	0.1	0.1	0.3	0.1	0.1	0.1	0.1	0.1	1.6	0.7	0.1	0.9	0.4	0.6				5

I wanted to see if I can visualize the clusters before using any math. The easiest way to do that would be to create a scatter plot with THC and CBD as the axis. This is where I came across a big issue. For some reason, the THC and CBD values for each data point are identical. This must be an error on the part of the creator of this dataset. I couldn't use CBD and THC to visualize clusters so I looked through the column names to see if there might be another pair of chemical compounds I could use. Luckily, I found THCa and CBDa. I had no idea what THCa and CBDa were, but from my research I learned that THCa is the inactive version of THC (same concept for CBD and CBDa). Below is a visualization of the two. I saw two things from this graph. 1) I cannot see any noticeable trends so I will need to scale the data and 2) I will need to get rid of outliers. So that's what I did. I used the Z-score and a threshold of 3 to get rid of outliers and used MinMaxScaler to scale the dataset.



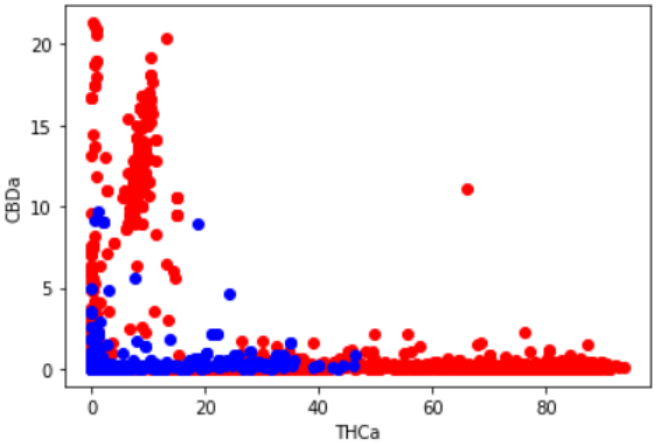
The K-Means Clustering

I decided to begin with K-Means Clustering. I figured if I should start with the simplest one and change the algorithm if I need to. However, I know that clustering is a very sensitive algorithm. There are a lot of rows with very little information and I didn't want it to mess-up my algorithm. To deal with this, I decided to run Principal Component Analysis (PCA) on the dataset to reduce the dimensions while retaining the important information. Using the elbow method, I was able to determine the number of components that worked best is 5. With the dataset with reduced dimensions, I was ready to go.

To find the best combination of hyperparameters of KMeans, I ran GrindSearchCV. The following is the best performing combination of hyperparameters.

KMeans(n_clusters= 2, init= 'k-means++', n_init= 10, max_iter= 300, tol= 0.0001)

Running this model in a 5-fold Cross Validation results in an average Silhouette Score of 0.73. Decent, but not good enough. Visualizing the clusters makes it obvious that the clusters are not separated in a way that would separate Indica from Sativa.



I tried tweaking certain parts of the data preparation process but was unable to get the score above 0.73. This leads me to believe that this is the limit to the K-Means Clustering on this dataset.

The Gaussian Mixture Model

I wanted to try at least one more model to see if I can get better clustering. I chose to use the Gaussian Mixture Model because the trends in the data may be too complicated for a K-Means algorithm to find.

I used very similar steps to the K-Means model to create the algorithm. I performed PCA and reduced the original dataset into one with 5 dimensions. From there, I used a GridSearchCV and found the best performing combination of hyperparameters as the following:

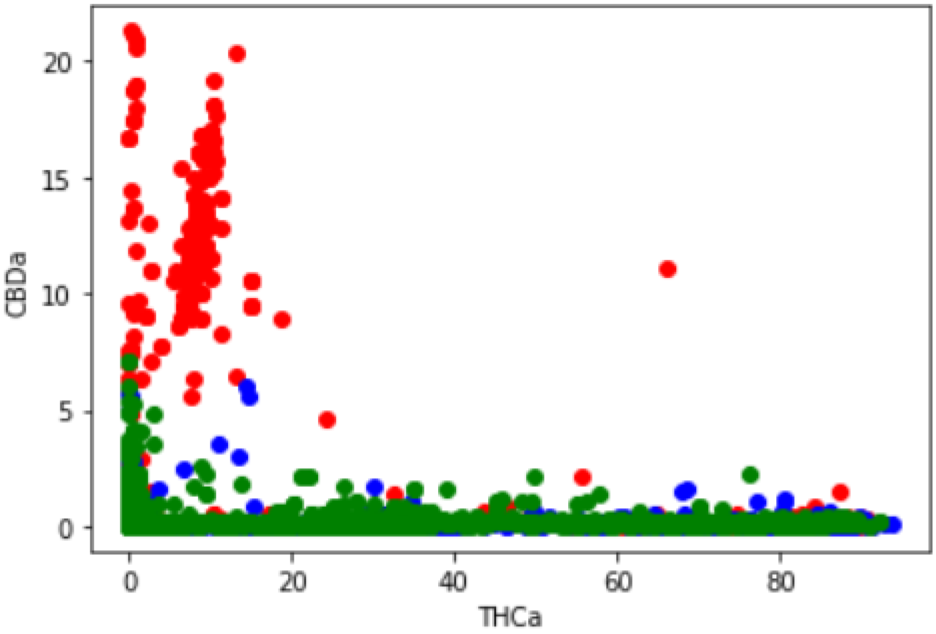
GaussianMixture(n_components=2 covariance_type = 'full',reg_covar = 0.001, init_params = 'random')

But again, the results were not satisfying. The clusters seemed random and inconsistent. I began to think I won't be able to solve this problem with this dataset.

I then remembered. It isn't just Indica and Sativa. There's a third player in this game. The middle-man of the two, Hybrid. I hypothesized that if I set the number of clusters/components to 3 (Indica, Sativa, Hybrid), my results will improve.

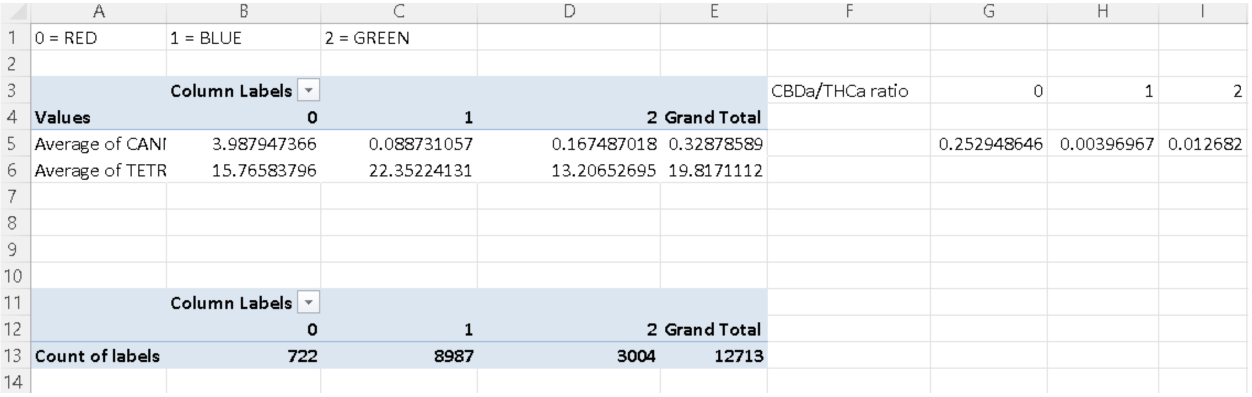
With the dimension-reduced dataset, I began the process again. After hours of tinkering, trial and error, I found that this still did not necessarily improve my algorithm performance. With that being said, there seemed to be more obvious patterns in the clustering that correlated to the CBDa/THCa ratio of a certain Marijuana product. Below is a visual along with the algorithm.

```
GaussianMixture(n_components=3, covariance_type = 'full', reg_covar = 0.001, init_params = 'random')
```



The Results

The plot above shows a rough trend in the data. The red cluster contains products that tend to have a higher CBDa:THCa ratio. This leads me to believe that this cluster would most likely be an Indica strain. The blue and green clusters are hard to distinguish so I exported the dataframe with the new labels back into an excel sheet to see if I can use a pivot table to distinguish the two. I found that the green cluster has a middle-of-the-pack CBDa:THCa ratio and the blue cluster has the lowest. Based on this information, I would suggest that the blue cluster would be a Sativa strain and the green cluster would be hybrid. But as we can see from the plot, the two are nearly indistinguishable.



	A	B	C	D	E	F	G	H	I
1	0 = RED	1 = BLUE	2 = GREEN						
2									
3		Column Labels				CBDa/THCa ratio	0	1	2
4	Values	0	1	2	Grand Total				
5	Average of CANI	3.987947366	0.088731057	0.167487018	0.32878589		0.252948646	0.00396967	0.012682
6	Average of TETR	15.76583796	22.35224131	13.20652695	19.8171112				
7									
8									
9									
10									
11		Column Labels							
12		0	1	2	Grand Total				
13	Count of labels	722	8987	3004	12713				
14									

The Conclusion

The clusterings are not perfect both mathematically and visually. However, there seems to be a pretty clear distinction between the red cluster and the green/blue cluster. If I could evaluate the green and blue cluster as one, then I could claim that the red cluster would be Indica products while the combined green/blue cluster would be Sativa products.

There are many areas to improve on. Primarily, I would have preferred if I could have used a supervised machine learning model to classify the products into its respective strain. This would require a dataset that includes strain labels of each product. This would also help me evaluate the accuracy of my model because I would have a target to compare my results against. By using an unsupervised clustering ML model, I could only find patterns in the data and use my prior knowledge or research to evaluate model accuracy.

Another area that could improve the results of this research is that much of the data is flawed. Looking at the original dataset provided from kaggle, we can see that the CBD and THC content for each product (the two most important chemical compounds when distinguishing Sativa vs Indica) are identical. This limited my ability to use CBD and THC to cluster the data points because each product would have a CBD:THC ratio of 1. I assume that it would have improved my results if the data in the THC and CBD columns were accurate. Furthermore, I couldn't use CBD and THC to visualize the clusters so I had to use CBDa and THCa. While the two are important when clustering the data, they do not have the significance that CBD and THC have.

Well, that was my first project. It was definitely a significant difference from any homework assignment I've done. I removed the training wheels for the first time, ate it,and was left with scratches and bruises. However, having to come up with your own problem, methodology, results and conclusions was challenging but exciting. I would love to be given the opportunity to tackle this problem with more accurate data. It would help me improve my model while also throwing new challenges for me to overcome.

The Kaggle Dataset:

https://www.kaggle.com/datasets/rahulgolder/marijuana-brand-registry-data