(Machine) Learning About Cannabis

Matthew Kowal

Abstract

Cannabis is a popular plant used around the world for many purposes. While many know it for its recreational uses, it is known to produce many beneficial physical and mental effects [1] [2]. Recreational cannabis was legalized in Canada on October 17th. Due to its illegality prior to this date, there has been very little research done on the different effects and qualities of various cannabis strains. In this project, I analyze the relationship between multiple features of cannabis strains such as type, effect and flavour. I show that numerous features in the dataset are highly correlated and can be used to predict the features of new strains as well as recommend existing strains for recreational and medical purposes.

I. Introduction

Cannabis, part of the Cannabaceae plant family, is accepted as originating from Central Asia and has three species: Cannabis sativa (sativa), Cannabis indica (indica), and Cannabis ruderalis (used mainly for industrial hemp purposes) [3]. With the recent legalization of recreational marijuana in Canada, research involving cannabis can be done more freely and in detail in order to better understand the plant. Due to its long-standing prohibition, there have been limited studies on marijuana in the past involving the differences between species. Although these studies explain some of the mechanisms in how the different species differ, there is a lack of datasets that list numerous features of various strains, until recently.

Fig. 1. A female cannabis plant. One can see the leaves (red), used for industrial purposes, and the flowers (pink), used for medical and recreational purposes.

II. Dataset

The dataset I used is a publicly released dataset found on Kaggle which was acquired from the website www.Leafly.com. The dataset consists of 2350 unique strains, with features consisting of type, rating, effects, flavour and descriptions. The rating (0-5), effects and flavour are taken from the average reviews of each strain from the public which are collected via the website listed above. Type consists of three species of plant: sativa, indica and hybrid (hybrid refers to a strain cross breaded from both an indica and sativa). There are 14 effects listed such as sleepy, creative, talkative, hungry and

	Strain	Type	Rating	Effects	Flavor	Description
0	100-Og	[hybrid]	4.0	[Creative, Energetic, Tingly, Euphoric, Relaxed]	[Earthy, Sweet, Citrus]	$100 OG is a 50/50 hybrid strain that packs a ...
1	98-White-Widow	[hybrid]	4.7	[Relaxed, Aroused, Creative, Happy, Energetic]	[Flowery, Violet, Diesel]	The 98 Aloha White Widow is an especially pote...
2	1024	[sativa]	4.4	[Uplifted, Happy, Relaxed, Energetic, Creative]	[Spicy/Herbal, Sage, Woody]	1024 is a sativa-dominant hybrid bred in Spain...
3	13-Dawgs	[hybrid]	4.2	[Tingly, Creative, Hungry, Relaxed, Uplifted]	[Apricot, Citrus, Grapefruit]	13 Dawgs is a hybrid of G13 and Chemdawg genet...
4	24K-Gold	[hybrid]	4.6	[Happy, Relaxed, Euphoric, Uplifted, Talkative]	[Citrus, Earthy, Orange]	Also known as Kosher Tangie, 24k Gold is a 60%...
5	3-Kings	[hybrid]	4.4	[Relaxed, Euphoric, Happy, Uplifted, Hungry]	[Earthy, Sweet, Pungent]	The 3 Kings marijuana strain, a holy trinity o...

Fig. 2. The first five strains listed in the database. Features include type, rating, effects, flavour and description.

giggly. There are 47 different flavours, such as apple, berry, sweet, diesel, and spicy. Most strains have four or five effects and three flavours.

A. Data Cleanup and Preprocessing

Initially, reading the dataset failed due to non-UTF-8 characters existing in the dataset. This occurred during the download process, where many irregular symbols appeared in the dataset. The beginning of the supplementary code deals with this issue and removes all non-UTF-8 characters and unwanted symbols.

B. Caution

While this is a useful dataset that can be leveraged for many tasks, I would not claim that this dataset is guaranteed to be accurate and thus I would not suggest using this project/ dataset for someone who is seriously considering using cannabis recreationally or medically.

III. Background

A. Cannabinoids and Terpenes

Cannabinoids are a class of various chemicals contained in cannabis that are known for producing desired effects such as pain relief, euphoria, or appetite [4]. They act with the cannabanoid receptors in human cells which alter the neurotransmitters released in the brain [5]. The most well known cannabinoids are Tetrahydrocannabinol (THC) and Cannabidiol (CBD). These cannabinoids make up between 0.05%-30% (m/m) of the marijuana leaf. In most plants it is normal to see higher contents of THC than CBD however multiples studies have recently shown the powerful pain-relief and inflammation reduction effects of pure CBD [6] [7] and so the demand for higher CBD strains is increasing. Although these are the most well known cannabinoids in the marijuana plant, there are many others such as cannabidiolic acid, tetrahydrocannabinolic acid,

cannabigerolic acid, and other chemicals called terpenes, hydrocarbon based resins that give off strong aromas and flavours, that could also produce psychoactive or other types of effects [8] [9]. Hopefully due to the recent legalization, the public can obtain a deeper understanding of these chemicals, cannabinoids and terpenes, and use them accordingly.

B. Species and Effects

It is generally thought that indica's leave you feeling 'in-da-couch' which is ideal for night-time use, while sativa's are better for daytime consumption as they produce a more wakeful and alert high. Hybrid's range from sativa dominated, indica dominated, or 50-50 sativa-indica. The effects of hybrids are thought to range widely depending on the parent strains as well as the THC and CBD content. Although the differences between indica, sativa's and hybrids are frequently talked about socially and all over the internet, researchers and scientists have questioned the origin of these theories and no scientific evidence has ever shown that indica's and sativa's have unique effects. I show in this project, despite the non-existing evidence of type-effect dependancies, it is possible to predict above 62% accuracy, the type of weed based off of the effects of a strain for *this particular dataset*.

IV. Analysis of Data

A. Methods

The methods I used involved Naive Bayes classification, logistic regression, linear regression, and support vector machine's (SVM). The linear regression model was used only for the cases involving the prediction of rating, as this was the only continuous feature in the dataset. Logistic regression and SVM's are two methods that use a 'one-vs-one' approach, in that they create $\frac{C(C-1)}{2}$ classifiers and then base the final prediction off of the accumulation of the binary classifiers. Naive Bayes uses a 'one-vs-all' approach in which it designates the probability of a certain feature occurring given a class. For

example, if the strain we are examining has the effect 'sleepy', than the classifier has examined the dataset and knows the percentage (probability) of all types that have 'sleepy' as an effect. From this probability as well as the probabilities of all other features it can decide on the most likely type.

Due to the small number of data points, the dataset was divided into only training and testing sets, with validation set not being used due to the simplistic nature of the classifiers as well as the short training time. After pre-processing and cleaning up the data, I was left with 2163 unique strains. This was split into training and testing sets of 85% and 15% respectively.

B. Type and Rating Prediction

Two main objectives I had for the predictive portion of the project was to be able to predict the type and rating of a strain given the effects and flavours. One would expect, given the realization that no conclusive evidence has ever been provided distinguishing the effects of indica's from sativa's, that it would not be possible to accurately predict the type based on the effects. On the other hand, there should be a correlation between effects, flavours and rating, as the average person might enjoy a particular effect more such as 'euphoric'. The results are quite confusing however but I believe they can be explained due to biases in the data. I found that it was possible to somewhat accurately predict the type based on the effects and flavours, as well as there being no correlation between the effects, flavours and rating.

1) Type Prediction Results: Here are the accuracy results for the functions predicting type from the effects, flavours, and both effects and flavours:

TABLE I

ACCURACY OF CLASSIFIERS PREDICTING TYPE FROM EFFECTS AND FLAVOURS

Features	Naive Bayes	SVM	Logistic Regression
Effects	60.3%	61.9%	**62.2** %
Flavours	26.2%	51.4%	**52.3** %
Effects, and Flavours	28.9%	60.0%	**63.4**%

Logistic regression is the best classifier in all three cases with SVM a close second. Naive Bayes only performs decently in the case of using only effects for the prediction. This is because there are only three effects per strain and 14 effects, which is much simpler in a 'one-vs-all' approach. When flavours are added, it immediately jumps up to 61 one-hot encoded features which is much more difficult for the classifier to fit.

The highest accuracy is, as expected, in the three feature case (effects, flavours). Logistic Regression achieves 63.4% accuracy in predicting the type from the effects and flavours. This is a fairly good score but not remarkable. If you were to use this model to try and determine the type of an unknown strain, the classifier would be correct about 2/3's of the time.

While it is possible that there is an unfound correlation between types and effects, I believe the reason for this high accuracy is the placebo effect. In my opinion, the vast majority of users believe that certain types have certain effects (e.g. indica's make you sleepier) and so they are biased towards believing that they are feeling those effects indeed. Although I cannot prove this, the overwhelming lack of evidence pointing towards these correlations is impossible to ignore. A study that could help understand this would be a placebo controlled study with people receiving 'random-typed' strains and then asking them to list the effects they feel the most fits this strain.

2) Rating Prediction Results: The first thing examined was the average rating of each type:

TABLE II
AVERAGE RATING OF EACH TYPE.

	Sativa	Indica	Hybrid
Average Rating (/5)	4.41	4.44	4.44

The types have basically the same rating, with sativa being rated only 0.6% less than the other types. I next used Linear Regression to try and predict the rating of a strain given its effects and flavours. The r^2 coefficient is listed with the features used for its prediction in table 3:

TABLE III
r^2 COEFFICIENT VALUE FOR PREDICTING RATING FROM EFFECTS, FLAVOURS, AND ALL FEATURES.

	Effects	Flavours	Type, Effects and Flavours
r^2	-0.02	-0.18	-0.36

It is clear that there is no linear relation between the effects, flavours and rating. This is in contrast to most of the discussion surrounding cannabis strains and types seen online. One would assume that some effects or flavours are generally more enjoyed (and therefore have a higher rating) than other effects and flavours but this does not appear to be the case. There are a few possibilities for why this is so but I believe the main reason is that all of the effects and flavours are positive. The 13 non-outlier

effects are: 'Creative', 'Relaxed', 'Uplifted', 'Tingly', 'Happy', 'Energetic', 'Euphoric', 'Focused', 'Sleepy', 'Hungry', 'Talkative', 'Aroused', and 'Giggly'. These are basically all beneficial effects if used in the right way (e.g. if you don't use sleepy cannabis when you want to be awake). Because of this, one would assume the highest rated strain would have all or most of the effects! Since each strain only has four to five effects, the result is essentially the model making random guesses because all effects can increase the rating and thus there is no correlation found between the effects and rating. The same can be said for flavour although this is slightly more subjective. Finally, all types have the same average rating so unless particular sativa's which have certain effects/ flavours are rated higher than other sativa's for example, there is no correlation between any feature and rating. This further points to the realization that effects, flavour and rating are a function of the cannabinoids and terpenes instead.

V. Recommender Systems

For both recreational users and people medicating with cannabis, a recommender system that helps provide similar strains to ones that the users have tried would be very useful. It would not only supply numerous strains that could achieve similar results as various strains they have had success with but also help users understand what effects, flavours and attributes of strains they desire. I develop a system that takes as input a description of a strain, and outputs the closest description and corresponding strain in the database. There are existing systems which also recommend similar strains, such as PotBotics, however this is the first one for this dataset.

A. Methods

The main method I used for the analysis of the 'description' feature is term frequency–inverse document frequency (TF-IDF) [10] . TF-IDF is a popular way of reflecting how important individual words are in a document. It is essentially a weighting value

that increases when a word is used frequently in a single document but not frequently in the overall collection of documents. For example, if the word 'blast-off' is used only twice in all of the descriptions of the strains, than it would have a very high weighting compared to the word 'the' because 'the' appears in every single description. The model would then associate both strains with the word 'blast-off' as similar.

Scikit-learn has a built in TF-IDF function that I used for this analysis. This function fits the data and produces a matrix of all of the documents (descriptions) and all of the unique words in all of the documents combined. I then use the cosine similarity metric between the TF-IDF matrix and itself in order to assign a similarity measure between all of the descriptions. The cosine similarity is defined as

$$cosine(x, y) = \frac{x \cdot y^T}{||x|| \cdot ||y||} \tag{1}$$

where x and y in my case are both the TF-IDF matrix. This results in the cosine similarity between every document and every other document. Expectedly, the most similar document is always itself which is simply skipped over when returning lists of similar documents.

B. Recommending Similar Strains

My goal for this system was to have a function that takes as input a strain name, and outputs another strain name with the most similar description (based on the TF-IDF matrix and cosine similarity).

1) Results: The strains that are returned are to be expected in most cases. It is very common for a strain to be recommended when a descendant of it is inputted. For example take a look at the recommendation for these common ancestors:

TABLE IV

RECOMMEND_A_STRAIN FUNCTION RESULTS. THE INPUT STRAIN IS IN BOLD. THE RESULTS ARE LISTED AS THE TOP FIVE MOST SIMILAR STRAINS, WITH THE MOST SIMILAR RESULT AT THE TOP OF THE LIST.

Recommend_a_Strain			
Blueberry	**Girl Scout Cookies**	**Haze**	**Og Kush**
Arctic Blue	Platinum Gsc	Green Haze	Berry Og
Crystalberry	La Kookies	Dutch Haze	Sfv Og Kush
Blue Hash	Moon Cookies	Royal Haze	Ig Kush
DJ Short Berry	Pineapple Cookies	Haze Wreck	Hellfire Og
Blackberry X Blueberry	Cherry Cookies	Thai Haze	Rudeboi Og

One can clearly see that the descriptions of descendant strains are scored as most similar to the input strain. We can look at the actual descriptions of a few examples in order to qualitatively understand the systems reasoning.

In the first example in table V, between Blueberry and Arctic Blue, one should note the similarities between the descriptions. Importantly, for medical users, Blueberry helps to 'suppress pain and relieves stress' while Arctic Blue 'helps suppress anxiety and relieve pain". It also will have a high similarity score due to the words 'DJ Short', 'indica' and 'sweet' which may not be as important for returning similar medical effects. In the second example, between Haze and Green Haze, there is less effect-based words in common. It is here where one can see the potential flaws in the recommendation system as the only words shared between the strains are 'Haze', 'South India' and 'sativa' which, while would share a high score due to the limited uses of the word 'Haze' in the database', do not necessarily mean these strains are alike. However it is not a coincidence that descendants of strains have very similar effects as their ancestors which means that, while for the wrong reason, the system will still return strains that offer very similar effects to the input strain.

TABLE V

THE FIRST RESULT FOR EXAMPLES 'BLUEBERRY' AND 'HAZE' FOR THE RECOMMEND_A_STRAIN FUNCTION

Strain	Description of Strain
Blueberry	A true A List cannabis strain. Blueberry's legendary status soared to new heights after claiming the High Times Cannabis Cup 2000 for Best Indica. The long history of the strain is traced back to the late 1970s when American breeder DJ Short was working with a variety of exotic landrace strains. However, throughout the decades of Blueberry's cultivation the genetics have been passed around, due in large part to DJ Short working with multiple seed banks and breeders. The sweet flavours of fresh blueberries combine with relaxing effects to produce a long-lasting sense of euphoria. Many medical patients appreciate Blueberry for its ability to suppress pain and relieve stress, while connoisseurs and growers admire the strain for its colourful hues and high THC content.
Arctic Blue	ArcticBlue is a 60/40 indica-dominant hybrid cannabis strain cultivated by ArcticBlue Farms. Bred using DJ Short's Blueberry and another Blueberry indica, you are immediately struck with the sweet and fruity scent of ripe blueberries. Patients may look to this strain to help suppress anxiety and relieve pain.
Haze	The illustrious Haze sativa first took root in Santa Cruz, California during the 1960s where long growing seasons accommodated her lengthy flowering cycle. Since then, Haze has become the proud parent of countless hybrids around the globe, passing on its genetics from Colombia, Mexico, Thailand, and South India. Although Haze cultivators must wait patiently for Haze flowers to reach full maturation, few strains can match the high-energy, creative buzz provided by this sativa staple. Haze's aroma is typically characterized by a spicy scent accented by hints of citrus and earthy sweetness.
Green Haze	Green Haze by A.C.E. Seeds is another version of their sativa Old Timers Haze and takes after her Thai and South Indian ancestry. Like its Haze parentage, Green Haze is also upbeat and verging on psychedelic for some. This sativa plant is truly for someone who loves growing as it can take a full 16-20 weeks to flower. Green Haze grows tall and is a lookalike of Purple Haze except for, of course, the colour. The flowers will have a woody smell mixed with fruit. Green Haze has a high potency and is good for tackling pain, though it may also inspire some serious snacking.

C. Predicting Strain from Description Input

The purpose of this function was to give users the ability to input the description of a strain, rather than the strain name, and have the output be strains which have descriptions similar to the users input. This would let people search for appropriate strains if they are a new user, try to find strains they have used but do not remember the name of, or users who are searching for certain effects of a strain outside of the listed 'Effects' column. This function is beneficial as many people are not aware of every strain they consume nor do they often know of a strain that produces the desired effects. (Note: I left out the function which simply searches for strains with a certain effect or flavour as it has already been done with a nice interface on Leafly.com directly)

1) Results:

TABLE VI

STRAIN_FROM_DESCRIPTION FUNCTION RESULTS. EXAMPLE USER DESCRIPTIONS ARE GIVEN WITH THE FIRST OUTPUT STRAIN AND ITS DESCRIPTION.

Input Description	Strain and Description of Strain
'giggly and hungry, perfect for video games'	**Sour Bubba**: 'Sour Bubba blends two famous strains that are robust all on their own but, when unified,show unique flavours and effects that speak to their long, potent lineages. With forward-leaning mental stimulation brought on by Diesel and a high-end Kush oriented body aura brought on by Bubba, Sour Bubba relaxes while providing ample mental energy to enjoy stationary activities like board games, video games, and movies.'
'before bed body high'	**Blueberry Kush**: 'Blueberry Kush is a strong indica strain that slowly brings on a heavy body sensation, helping patients forget their pain and relax. Originally from Oregon, this strain is a hybrid of Blueberry and OG Kush, which is evident in its fresh berry aroma with notes of earthy herbalness. This indica is best for evenings or before bed.'
'relaxing and uplifting with pain relief'	**Sour Flower**: 'Sour Flower provides an uplifting sativa high and is good for relieving pain and anxiety.'
'high CBD with clear head'	**MediHaze**: 'MediHaze (or CBD Medi Haze) is an 80/20 sativa-dominant strain bred by CBD Crew that crosses genetics from Super Silver Haze, Neville's Haze, and an undisclosed CBD-rich parent. Its THC to CBD ratio typically comes out 1:1, but some seeds will offer a doubled CBD content. THC and CBD's synergy provide relief to a variety of symptoms, some of which include pain, inflammation, and anxiety. Pine, mint, and spice aromas burst from MediHaze buds in a fragrant introduction to its clear-headed, uplifting effects.'

It is clear from table VI that the results from this function are substantially better than the results from the Recommend_a_Strain function in terms of how close the description is based on relevant factors such as desired effect. This is due to the size of the input in comparison to the other function. The Strain_from_Description function can take a small sentence or phrase as an input and as a result, the outputs are only those with the input in them. For the Recommend_a_Strain function, where the input is the entire description, there is more noise and unrelated words included in the search.

One small flaw of this function is that the input should be what the description of the cannabis would be, not a first person description of what the user desires. For example, ”pungent and thick smoke, makes you sleepy” is preferable over ”I want the weed to make me feel sleepy and have thick smoke”. This is because the words ’I want the weed to’ are irrelevant words that can add noise to the decision process.

VI. Conclusion

I have shown that despite there being no scientific evidence that the effects and flavours of a cannabis strain are a function of its type (indica, sativa, hybrid), we are able to predict a new strains type with 63.4% accuracy using Logistic Regression. This brings into question whether or not we will see more studies done to determine if this relationship is meaningful, or if the dataset used is biased in its creation. It may be that users are succumbing to the placebo effect, and are feeling the effects that they expect to experience, based off of the strain type.

On the other hand, no relationship could be found between the effects, flavours, type and rating. Each type had the same average rating and the r^2 value for the linear regression model was approximately 0 for all cases. This implies that a strain will have a rating independent of its type, effects or flavours. One possible explanation for this seems to be a bias in the data. There are only positive effects given and mostly positive flavours. This would imply that the strain with the most effects will have the highest rating, but since each strain only has three effects, the relationship between the effects and ratings is uncorrelated.

The descriptions were analyzed using the TF-IDF method and resulted in two functions: Recommend_a_Strain and Strain_from_Description. Recommend_a_Strain takes another strain name as input and outputs the most similar strain based on the cosine similarity metric. This gave decent results but many of the recommended strains contained similar words that were undesired such as the country of origin. It also is

not very robust as any spelling mistake in the input will result in an error message. Strain_from_Description takes a user defined description as an input, which defines the flavours, effects, or attributes they want the strain to have, and outputs the corresponding strain which is most similar. This function achieves very nice results due to the small amount of noise in the input and multiple examples showed great qualitative results when searching for attributes like 'high CBD', 'good for video games' or 'pain relief'.

A. Future Work

Cannabinoids and terpenes have been clinically studied and shown to produce a multitude of effects ranging from anti-anxiety to anti-inflammation [11] [2]. It is of utmost importance to add the list cannabinoids and terpenes for each of the strains in this database. This would allow a proper analysis of the strains to see if there is a correlation between type and chemical content (e.g. do sativa's usually contain more THC than indica's?). We could then conclude as to whether or not the theory of type based effects is fact or fiction. It would also give a deeper understanding as to what chemicals, or combination of chemicals, produce desired effects. This could then be leveraged to supply people with a more accurate implementation of the functions I have produced in this work.

REFERENCES

[1] J. R. Johnson, M. Burnell-Nugent, D. Lossignol, E. D. Ganae-Motan, R. Potts, and M. T. Fallon, "Multicenter, double-blind, randomized, placebo-controlled, parallel-group study of the efficacy, safety, and tolerability of thc: Cbd extract and thc extract in patients with intractable cancer-related pain," *Journal of pain and symptom management*, vol. 39, no. 2, pp. 167–179, 2010.

[2] E. B. Russo, "Taming thc: potential cannabis synergy and phytocannabinoid-terpenoid entourage effects," *British journal of pharmacology*, vol. 163, no. 7, pp. 1344–1364, 2011.

[3] M. Colbert, "Indica, sativa, ruderalis – did we get it all wrong?," *The Leaf Online*, 2015.

[4] K. Nelson, D. Walsh, P. Deeter, and F. Sheehan, "A phase ii study of delta-9-tetrahydrocannabinol for appetite stimulation in cancer-associated anorexia.," *Journal of palliative care*, 1994.

[5] P. Pacher, S. Bátkai, and G. Kunos, "The endocannabinoid system as an emerging target of pharmacotherapy," *Pharmacological reviews*, vol. 58, no. 3, pp. 389–462, 2006.

[6] M. Serpell, S. Ratcliffe, J. Hovorka, M. Schofield, L. Taylor, H. Lauder, and E. Ehler, "A double-blind, randomized, placebo-controlled, parallel group study of thc/cbd spray in peripheral neuropathic pain treatment," *European journal of pain*, vol. 18, no. 7, pp. 999–1012, 2014.

[7] S. Burstein, "Cannabidiol (cbd) and its analogs: a review of their effects on inflammation," *Bioorganic & medicinal chemistry*, vol. 23, no. 7, pp. 1377–1385, 2015.

[8] A. T. Peana, P. S. D'Aquila, F. Panin, G. Serra, P. Pippia, and M. D. L. Moretti, "Anti-inflammatory activity of linalool and linalyl acetate constituents of essential oils," *Phytomedicine*, vol. 9, no. 8, pp. 721–726, 2002.

[9] S. Martin, E. Padilla, M. Ocete, J. Galvez, J. Jimenez, and A. Zarzuelo, "Anti-inflammatory activity of the essential oil of bupleurum fruticescens," *Planta medica*, vol. 59, no. 06, pp. 533–536, 1993.

[10] G. Salton and J. Michael, "Mcgill. 1983," *Introduction to modern information retrieval*, 1983.

[11] A. W. Zuardi, R. Cosme, F. Graeff, and F. Guimarães, "Effects of ipsapirone and cannabidiol on human experimental anxiety," *Journal of Psychopharmacology*, vol. 7, no. 1_suppl, pp. 82–88, 1993.