ITEC 3040 M - Winter 2023

Drug Consumption Dataset Final Report

By The Blue Group

Amanuel Demeke [redacted]
Gabrielle Campbell [redacted]
Jisuk Kang [redacted]
Sophia Maxine Villarosa [redacted]

Introduction

Motivation for our Problem

Magic mushroom use has been a growing topic in Canada for the last few years. More Canadians than ever before are interested in its use recreationally and medically. Currently, mushroom use is illegal in Canada, and there are only a few rare situations where it can be prescribed legally for medical use. As demand rises for this drug, people are taking to illegal methods to obtain mushrooms, and are even opening up un-regulated dispensaries - despite risking serious legal repercussions. This is reminiscent to marijuana's legalization in 2018, where excitement and demand had grown so high that people were operating unlicensed dispensaries, despite pushback from law enforcement.

We, the Blue Group, predict that in the coming years, demand for magic mushrooms will skyrocket, similarly to marijuana. In anticipation of that boom, we are going to conduct analysis on the Drug Consumption Dataset, provided by the UCI Machine Learning Repository, to better understand who these potential customers are.

Police raid of Cafe Dispensary on Harbord street, July 2019

Problem Definition

Our aim is to be able to predict future magic mushroom users in order to understand the demographic that have a higher chance of being consumers of this drug. We will use classification methods to figure out how to distinguish these people, with data on nicotine use, cannabis use, age, openness and gender.

Data Preprocess

The original dataset contains 32 attributes and 1885 instances. This is a lot of data to work with, so for the purposes of our analysis, we decided to do some thorough cleaning of the data. Our first step was to determine the attributes that we wanted to work with, and remove the ones we did not want to work with. This included removing attributes such as chocolate consumption, meth use, neuroticism etc. In the end we decided to work with 7 attributes: ID, Age, Gender, Oscore, Cannabis, Nicotine and Mushrooms, ID being used as a simple tuple identifier, and Mushrooms being our class attribute.

The original dataset was heavily simplified (presumably to conserve memory), with data being presented in decimal form, and unrelated strings (i.e., CL0, CL1, CL2 etc., to represent drug use). Our attributes did not even have proper labels, so we started there, and then moved on to converting every decimal value to a corresponding legible value.

In our initial plans, we wanted to use the Semer attribute (a fictitious drug) as a way to remove individual tuples from our dataset. Our thought process was that people who claim to use a fake drug cannot be trusted to give accurate answers on other attributes. Our personal machines did not have enough memory to remove these tuples as they were indexed at above 730. We were left with no choice but to leave these tuples in the data (total of 8 positive semer users).

All drug attributes contained 7 possible states correlating to the last time a person consumed a drug. We decided to turn our class attribute into a binary attribute (Yes or No). We made this decision to simplify the classification problem, and because it is more relevant to our goal. If someone used mushrooms 10 years ago, they would still be considered a potential user if mushrooms were to be legalized as the legality to the drug may have prohibited them from using it frequently. We kept the 7 possible states for the other drugs in our data. With that, our training set was ready.

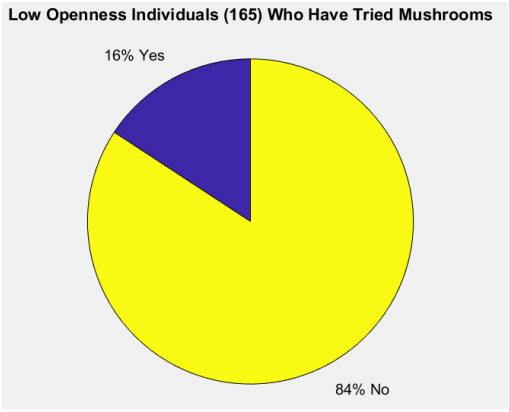
ID	Age	Gender	Oscore	Cannabis	Mushrooms	Nicotine
1	{' 35-44' }	{'Female'}	{'high' }	{'Never' }	{'No' }	{'Last dcd' }
2	{' 25-34' }	{'Male' }	{' high' }	{'Last mnth'}	{'No' }	{'Last mnth'}
3	{' 35-44' }	{'Male' }	{'moderate' }	{'Last yr' }	{'Yes'}	{'Never' }
4	{' 18-24' }	{'Female'}	{'moderate' }	{'Last dcd' }	{'No' }	{'Last dcd' }

What our training set looks like.

Data Exploration

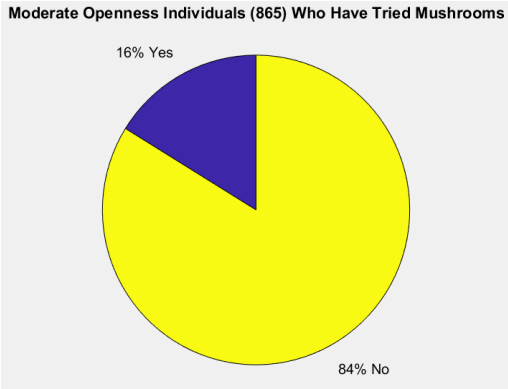
Before creating our classification models, we wanted to conduct some data exploration. In our progress report we noted our interest in looking at the *extraversion* attribute as a way of adding personality traits in our analysis. We decided that *openness* (Oscore) would be a more appropriate personality metric to include in our report as it is more fitting for drug data. Originally, Oscore was represented in the data as a number between 24-60. We took this range, divided it into three equal parts to make three ordinal values associated with a specific Oscore. 24-36 is considered low, 37-48 is moderate, and 49-60 is considered high (this can be seen in the figure above). Transforming the attribute values made it easier to distinguish high and low scoring people.

Low-Openness



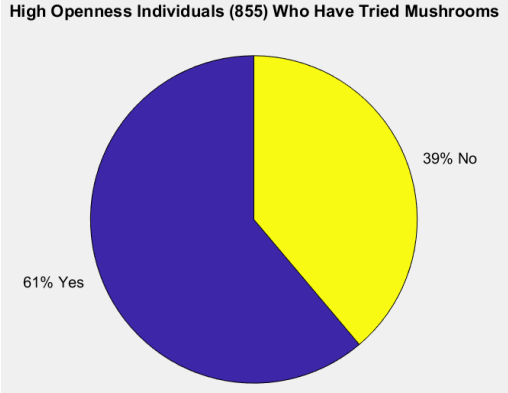
This pie chart is a representation of how many individuals who fall within the low openness personality trait group and have consumed magic mushrooms. The chart shows that there is a low correlation between persons who have this personality and how likely they are to have used mushrooms.

Moderate-Openness



This pie chart is a representation of how many individuals who fall within the moderate openness personality trait group have consumed magic mushrooms. Similar to the low openness personality pie chart, we see that the majority of individuals who are classified as moderate openness have not used mushrooms.

High-Openness



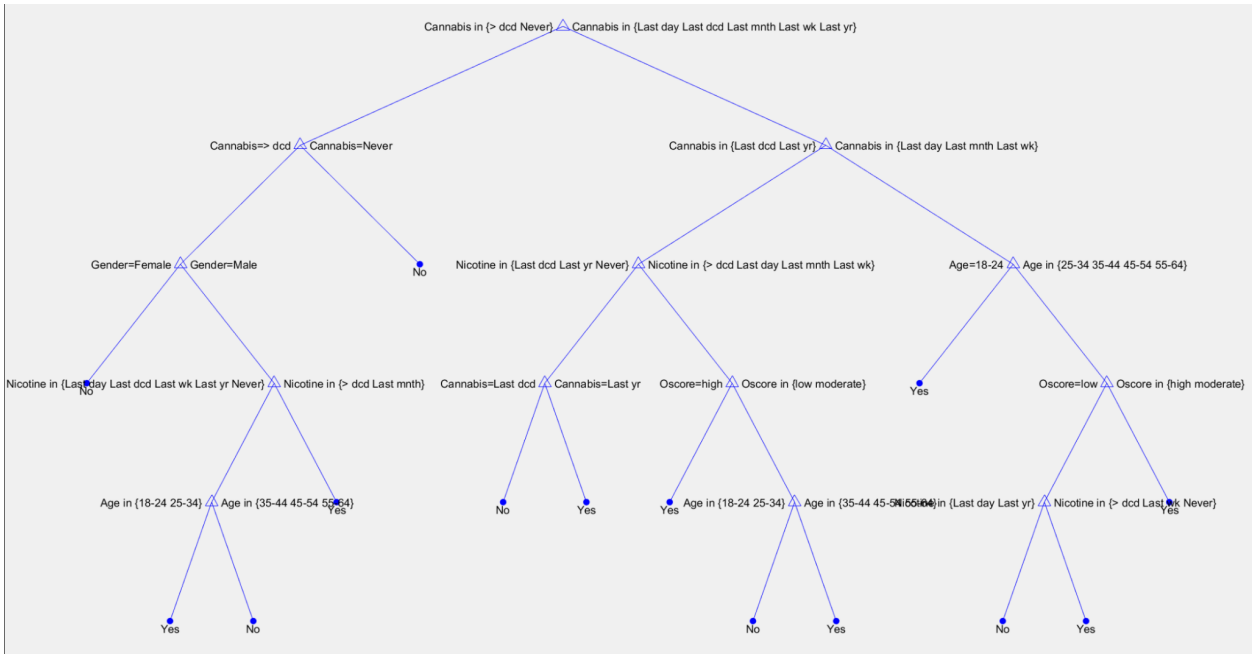
This pie chart is a representation of how many individuals who fall within the high openness personality trait group have consumed magic mushrooms. Majority who fall within this personality group have tried magic mushrooms. This clearly shows that there is a strong correlation between individuals who are more open to new experiences and magic mushroom consumption. We can infer that this also tells us that the majority of current and future users of magic mushrooms will most likely come from persons who score highly in this personality type.

How does this information further our knowledge on the goal? Although it may be difficult to obtain data on a person's openness score from a marketing perspective, this data can be used to help target the *location* of advertisements. For example, poster ads for mushroom dispensaries could be put up near establishments that attract high openness individuals such as exotic restaurants, artistic shops, and even skydiving venues.

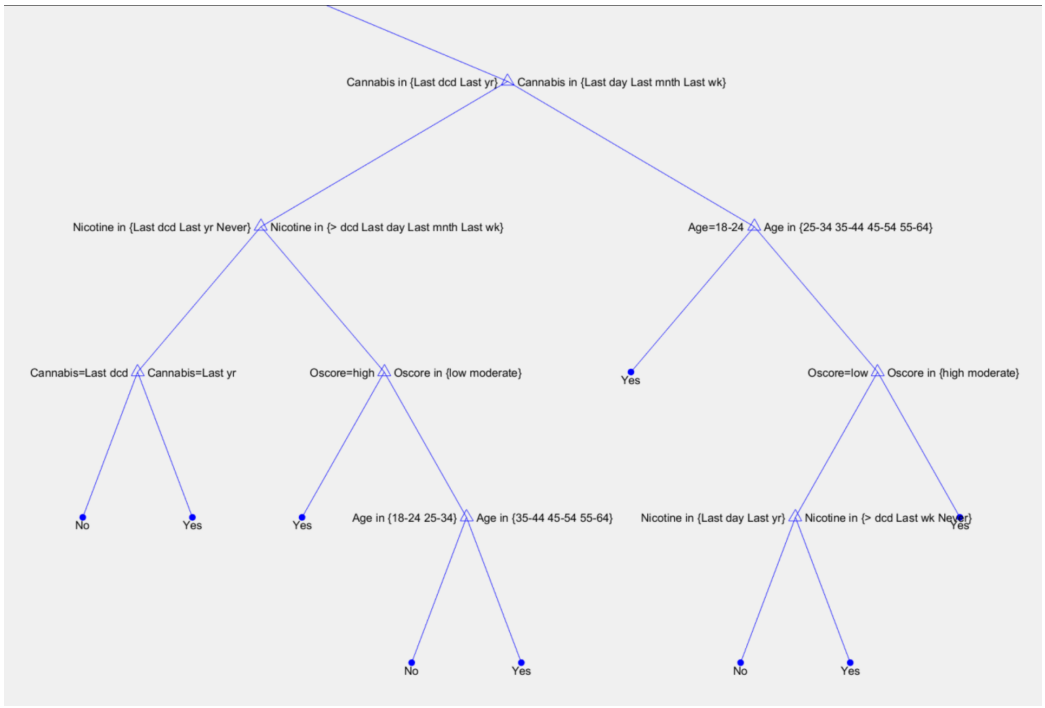
Classification Models

Decision Tree

Note: *This tree was created using MATLAB, nodes are split in a binary fashion. Some internal nodes continue off of the previous node to continue this split pattern (i.e., the root node and the second left node). Language is simplified and commas are not present in the tree, labels such as* “Cannabis in {>dcd Never}” *refer to* “Cannabis in {Last used over a decade ago, Never Used}”



Bottom right corner of tree:



Thoughts on the Decision Tree

Looking at the tree, there is some interesting information to extrapolate regarding the branches to leaf nodes. The fastest way to determine if someone has never used magic mushrooms is to see if they have used cannabis, if not, they have not used mushrooms. From a marketing perspective this is valuable information, this means that people reluctant to try cannabis will not see magic mushrooms as a more approachable alternative. Money should not be spent on advertisements trying to convince non cannabis users to try mushrooms.

Naive Bayes Classifier

There are two classes for this bayes classifier.

- mushroom = yes
- mushroom = never

P(mushroom = yes) = 903/1885 = 0.479

P(age = 18 – 24 | mushroom = yes) = 368/903 = 0.408

P(age = 25 – 34 | mushroom = yes) = 229/903 = 0.254

P(age = 35 – 44 | mushroom = yes) = 158/903 = 0.175

P(age = 45 – 54 | mushroom = yes) = 114/903 = 0.126

P(age = 55 – 64 | mushroom = yes) = 32/903 = 0.035

P(age = 65+ | mushroom = yes) = 2/903 = 0.002

P(gender = Male | mushroom = yes) = 569/903 = 0.630

P(gender = Female | mushroom = yes) = 334/903 = 0.370

P(cannabis = never | mushroom = yes) = 12/903 = 0.013

P(cannabis = yes | mushroom = yes) = 891/903 = 0.987

P(nicotine = never | mushroom = yes) = 82/903 = 0.091

P(nicotine = yes | mushroom = yes) = 821/903 = 0.909

This classifier is based on people who have ever used mushroom before.
Age, gender, cannabis use or not, nicotine use or not, attributes were computed with (mushroom = yes) class.

P(mushroom = never) = 982/1885 = 0.521

P(age = 18 – 24 | mushroom = never) = 275/982 = 0.280

P(age = 25 – 34 | mushroom = never) = 252/982 = 0.257

P(age = 35 – 44 | mushroom = never) = 198/982 = 0.202

P(age = 45 – 54 | mushroom = never) = 180/982 = 0.183

P(age = 55 – 64 | mushroom = never) = 61/982 = 0.062

P(age = 65+ | mushroom = never) = 16/982 = 0.016

P(gender = Male | mushroom = never) = 374/982 = 0.381

P(gender = Female | mushroom = never) = 608/982 = 0.619

P(cannabis = never | mushroom = never) = 401/982 = 0.408

P(cannabis = yes | mushroom = never) = 581/982 = 0.592

P(nicotine = never | mushroom = never) = 346/982 = 0.352

P(nicotine = yes | mushroom = never) = 636/982 = 0.648

This classifier is based on people who have never used mushrooms before.
Age, gender, cannabis use or not, nicotine use or not, attributes were computed with (mushroom = never) class

Evaluation Methods

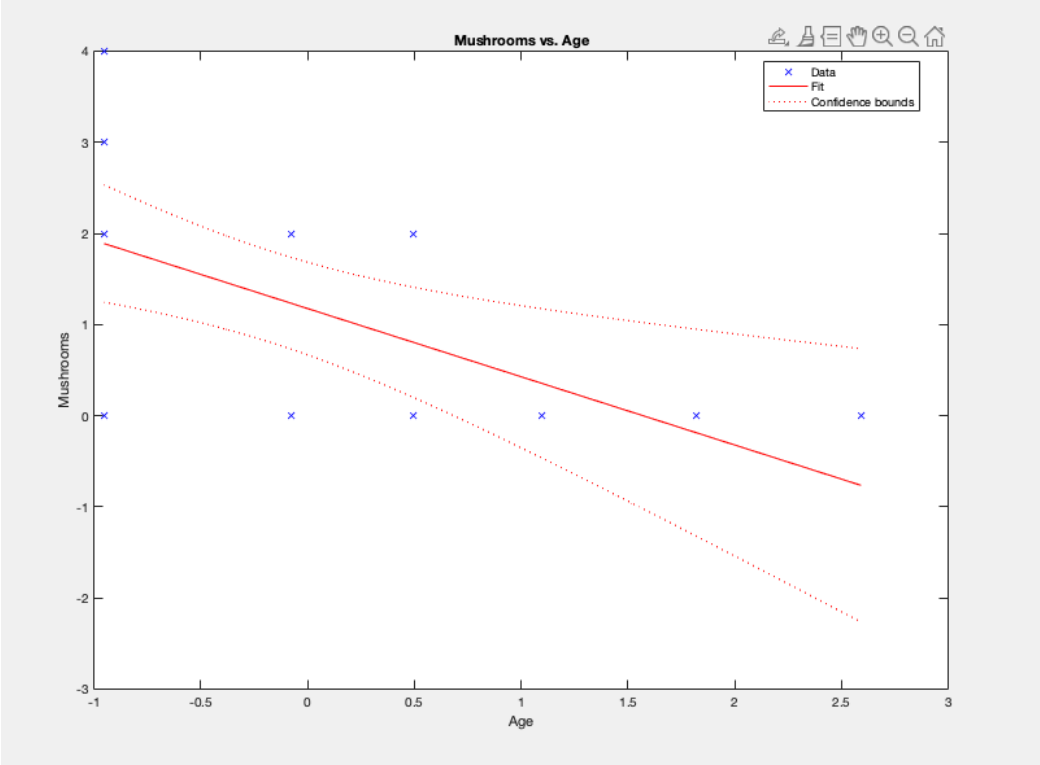
The sampleTable is used to perform the Linear Regression analysis and the Scatterplot Analysis. Note: The sample contains 30 entities, that were picked by random (and without repetition) from the original training set. The indexes of the selected entities were generated using an external vector generator.

```
sampleTable =

  30×7 table
```

ID	Age	Gender	Oscore	Cannabis	Mushrooms	Nicotine
1	-0.07854	-0.48246	-1.68062	1	0	1
2	1.82213	0.48246	0.44585	1	0	3
3	-0.07854	0.48246	0.44585	2	0	2
4	0.49788	0.48246	-0.71727	0	0	0
5	0.49788	0.48246	-0.84732	6	0	6
6	0.49788	0.48246	-0.01928	5	2	6
7	-0.07854	0.48246	-0.31776	3	0	6
8	-0.95197	0.48246	0.58331	3	3	5
9	-0.95197	0.48246	0.14143	2	0	6
10	-0.95197	-0.48246	-0.01928	4	2	4
11	-0.07854	-0.48246	-2.39883	3	2	6
12	-0.95197	0.48246	0.44585	4	4	3
13	-0.95197	-0.48246	-0.31776	3	0	2
14	-0.95197	-0.48246	-1.27553	6	3	0
15	-0.95197	-0.48246	1.24033	6	0	5
16	-0.07854	0.48246	0.44585	6	2	6
17	-0.95197	-0.48246	0.7233	6	3	5
18	1.09449	-0.48246	-0.17779	5	0	1
19	0.49788	0.48246	0.44585	0	0	0
20	-0.95197	-0.48246	-0.17779	6	0	4
21	1.82213	-0.48246	0.88309	1	0	0
22	2.59171	-0.48246	-0.58331	0	0	6
23	-0.95197	0.48246	0.58331	5	3	3
24	-0.95197	-0.48246	1.24033	5	4	3
25	-0.95197	-0.48246	0.29338	4	3	2
26	-0.95197	-0.48246	1.65653	5	0	3
27	-0.95197	-0.48246	2.15324	4	2	4
28	0.49788	-0.48246	0.14143	5	2	1
29	-0.07854	-0.48246	2.90161	6	0	6
30	-0.95197	-0.48246	0.58331	5	4	6

Based on the results of the linear regression analysis, we can assume that the findings obtained from the sample table are representative of the complete drug consumption data table with 1885 instances. Therefore, the conclusions drawn from the sample table are likely to be applicable to the larger dataset.

This graph shows the relationship between Age and Mushroom use



We kept the Age values as it is in the data because the matlab function would only work with doubles or floats

-0.95197 represents the ages 18-24
-0.07854 represents the ages 25-34
0.49788 represents the ages 35-44
1.0945 represents the ages 45-54
1.8221 represents the ages 55-64
2.5917 represents the ages 65+

And the mushroom use is set to

CL0 = 0 which represents those who NEVER used Mushrooms
CL1 = 1 which represents those who used Mushrooms over a decade ago
CL2 = 2 which represents those who used Mushrooms in the last decade
CL3 = 3 which represents those who used Mushrooms in the last year
CL4 = 4 which represents those who used Mushrooms in the last month
CL5 = 5 which represents those who used Mushrooms in the last week
CL6 = 6 which represents those who used Mushrooms in the last day

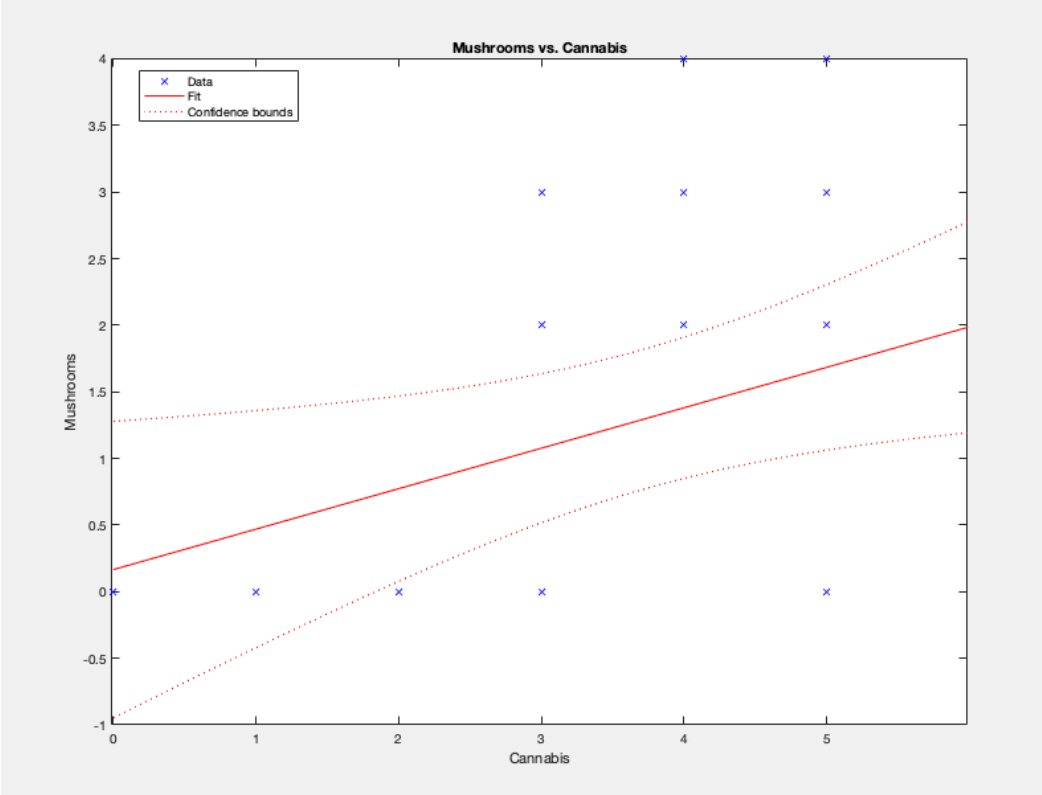
Conclusion from Analysis

The usage of magic mushrooms is highest among individuals aged 18-24, and gradually decreases with age. Those aged 25-34 and 35-44 also report higher usage rates than older age groups. Based on the graph, it can be inferred that younger individuals are more likely to consume mushrooms than their older counterparts. This analysis suggests that age is a significant factor in predicting mushroom consumption.

Matlab code for finding the linear regression between Mushrooms (dependent variable) and Age (independent variable)

```
>> model2 = fitlm(sampleTable, 'Mushrooms ~ Age');
>> figure;
>> plot(model2)
>> model2

model2 =


Linear regression model:
    Mushrooms ~ 1 + Age

Estimated Coefficients:
                       Estimate              SE              tStat              pValue
                   ________________    ________________    ________________    ____________________

    (Intercept)       1.17694294525925    0.248049778206521     4.7447853159512    5.57200132666037e-05
    Age              -0.748684658314432    0.250460280476635    -2.98923508705515     0.00576891032735078


Number of observations: 30, Error degrees of freedom: 28
Root Mean Squared Error: 1.34
R-squared: 0.242,  Adjusted R-Squared: 0.215
F-statistic vs. constant model: 8.94, p-value = 0.00577
```

This graph shows the relationship between Cannabis use and Mushrooms use



For both Mushrooms and Cannabis use

CL0 = 0 which represents those who NEVER used the drug
CL1 = 1 which represents those who used the drug over a decade ago
CL2 = 2 which represents those who used the drug in the last decade
CL3 = 3 which represents those who used the drug in the last year
CL4 = 4 which represents those who used the drug in the last month
CL5 = 5 which represents those who used the drug in the last week
CL6 = 6 which represents those who used the drug in the last day

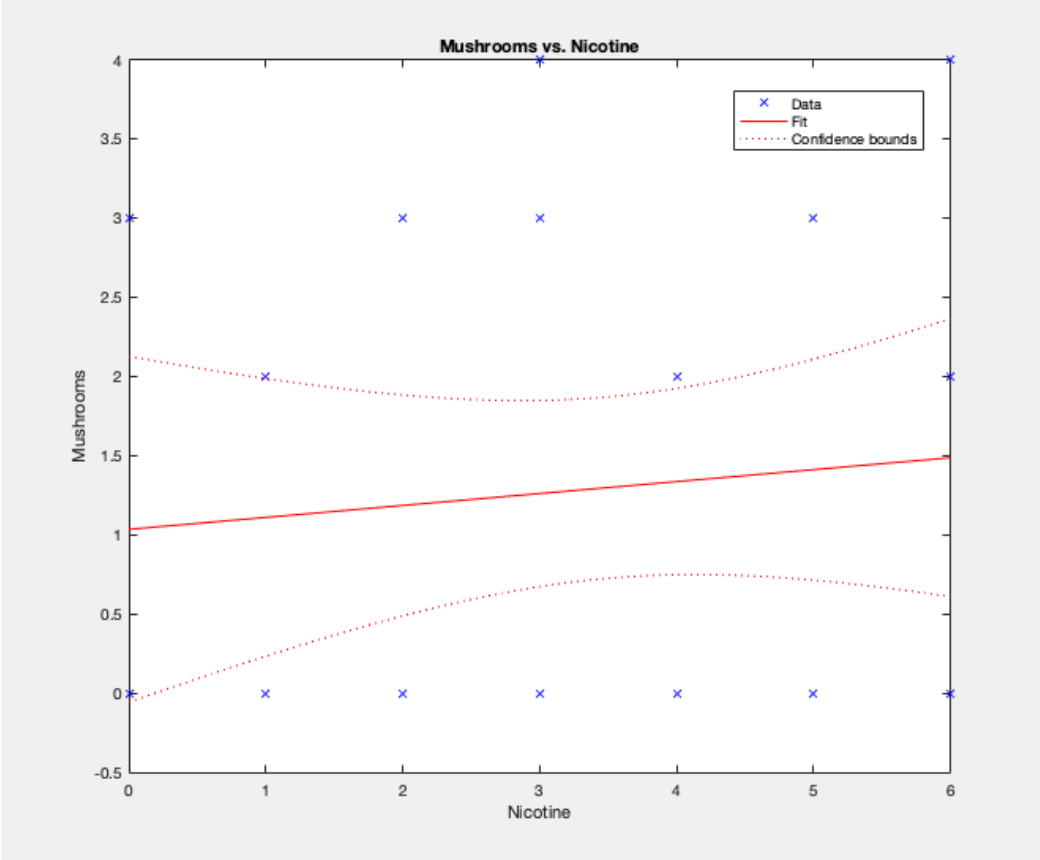
Conclusion from analysis

The graph indicates a positive correlation between recent and frequent cannabis usage and magic mushroom consumption. The data is more dispersed in the top right quadrant of the graph, suggesting that individuals who have used cannabis in the last year, month, week, or day are more likely to have also consumed mushrooms. This analysis implies that recent cannabis users are more prone to using mushrooms as well.

Matlab code for finding the linear regression between Mushrooms (dependent variable) and Cannabis (independent variable)

```
>> model3 = fitlm(sampleTable, 'Mushrooms ~ Cannabis');
>> model3

model3 =


Linear regression model:
    Mushrooms ~ 1 + Cannabis

Estimated Coefficients:
                    Estimate          SE          tStat          pValue
                    ________________  ________________  ________________  ________________

    (Intercept)     0.166295884315907  0.543318149598749  0.306074598168163  0.761812119118605
    Cannabis        0.303670745272525  0.128299127350892   2.36689642044095  0.0250872279178454


Number of observations: 30, Error degrees of freedom: 28
Root Mean Squared Error: 1.4
R-squared: 0.167,  Adjusted R-Squared: 0.137
F-statistic vs. constant model: 5.6, p-value = 0.0251
```

This graph shows the relationship between Mushroom use and Nicotine Use



For both Mushrooms and Nicotine use

CL0 = 0	which represents those who NEVER used the drug
CL1 = 1	which represents those who used the drug over a decade ago
CL2 = 2	which represents those who used the drug in the last decade
CL3 = 3	which represents those who used the drug in the last year
CL4 = 4	which represents those who used the drug in the last month
CL5 = 5	which represents those who used the drug in the last week
CL6 = 6	which represents those who used the drug in the last day

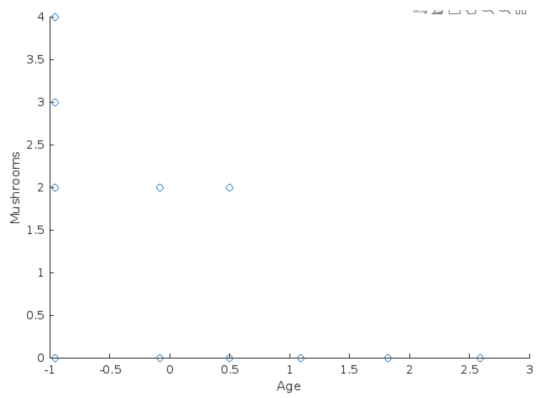
Conclusion from analysis

The usage of magic mushrooms and nicotine does not exhibit a clear correlation. The data points are scattered on the graph, and there is no linear relationship between them. Therefore, it is not possible to predict an individual's likelihood of consuming magic mushrooms based on their nicotine usage. In conclusion, there appears to be little to no relationship between the two substances.

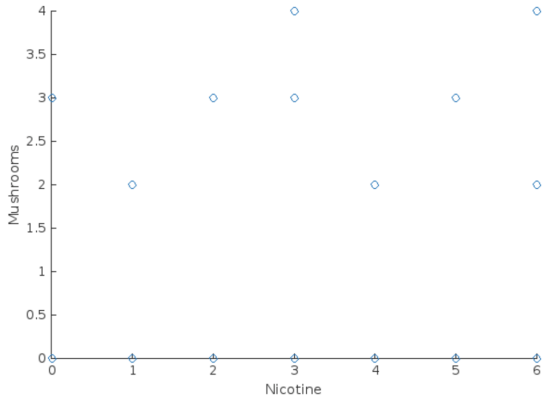
Matlab code for finding the linear regression between Mushrooms (dependent variable) and Nicotine (independent variable)

```
>> model5 = fitlm(sampleTable, 'Mushrooms ~ Nicotine');
>> model5

model5 =


Linear regression model:
    Mushrooms ~ 1 + Nicotine

Estimated Coefficients:
                    Estimate              SE              tStat              pValue
                  ____________________  ____________________  ____________________  ____________________

    (Intercept)     1.03655913978495    0.532390090531784     1.9469917983439    0.0616299321965507
    Nicotine       0.075268817204301    0.129505022291052    0.581203847331421       0.5657527464382


Number of observations: 30, Error degrees of freedom: 28
Root Mean Squared Error: 1.53
R-squared: 0.0119,  Adjusted R-Squared: -0.0234
F-statistic vs. constant model: 0.338, p-value = 0.566
>> figure;
>> plot(model5)
fx >>
```

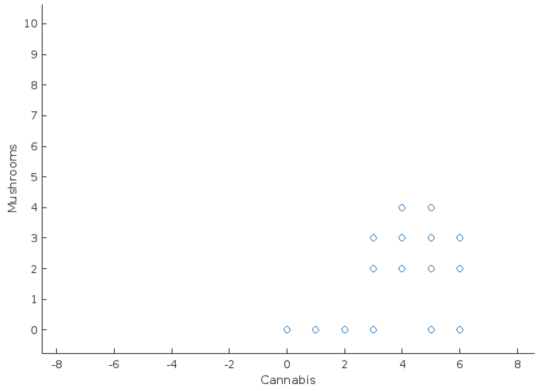
Scatterplot Analysis



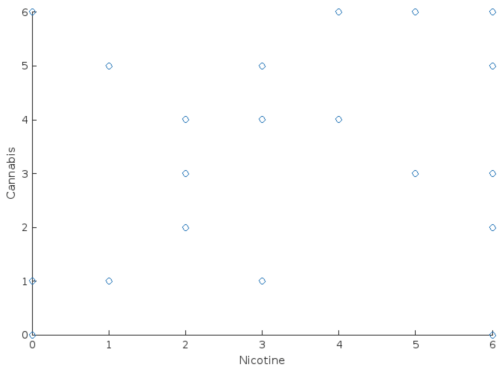
Mushrooms vs Age



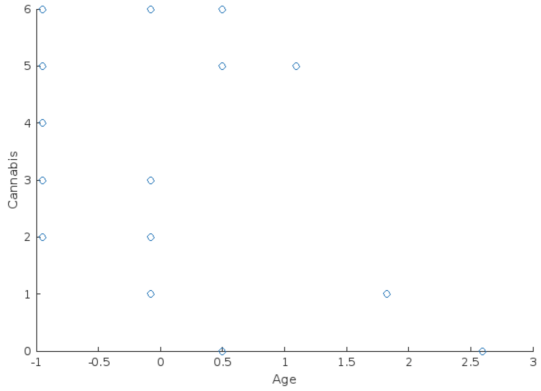
Mushrooms vs Nicotine



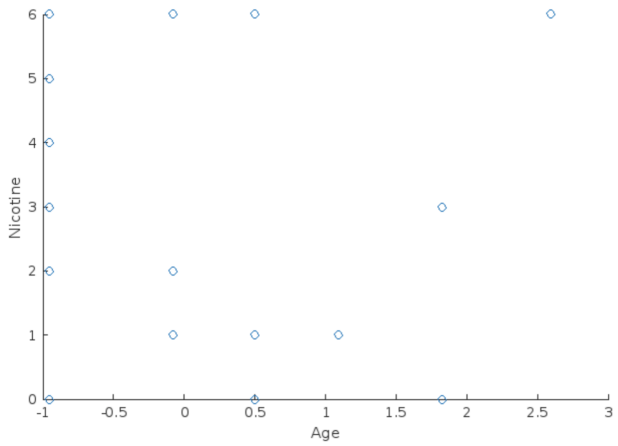
Mushrooms vs Cannabis



Cannabis vs Nicotine



Cannabis vs Age



Nicotine vs Age

The use of multiple models to aid in the prediction of the likelihood of someone being a magic mushroom user is helpful as we can take multiple possibilities and relationships into consideration. In our graphs we can see many different relationships being presented on the scatter plot diagram. An analysis of these relationships can allow us to see a possible deeper connection among our variables.

The Mushroom vs Age and Mushroom vs Nicotine graphs have a very weak relationship whilst the Mushroom vs Cannabis graph has a strong relationship. This allows us to see that Cannabis consumption plays a large role in Magic Mushroom usage, making Cannabis a suitable determining factor in deciding the likelihood of someone being a future consumer of magic mushrooms.

The Cannabis vs Age graph has a moderate relationship Cannabis vs Nicotine has a stronger relationship and has a positively linear correlation as well. This is due to most individuals who have used Nicotine have also used Cannabis with many being frequent users of both.

The Nicotine vs Age graph has a weak relationship. However, it also tells us that most Nicotine consumers fall within the age groups of 18 - 24 and 25 -34.

With all this information to consider, we can safely come to the conclusion that Nicotine, Cannabis and Age are suitable factors to take into consideration as they all strongly relate to each other and will aid in a more accurate determination of whether someone is likely to be a future magic mushroom consumer.

Division of work

Amanuel	Problem Motivation , Data Preprocessing, Data Exploration - Pie Charts, Decision Tree Model
Gabrielle	Data Exploration - Scatter Plots, Problem Definition, Evaluation Method
Sophia Maxine	Data Preprocessing, Evaluation Method (linear regression analysis)
Jisuk	Classification - Bayes