OMIS 115 Final Project

Drug Consumption

By Lola Lewis

Table of Contents

01

Background

Drug Use in the United States

- Over 101,000 Americans died from drug-involved overdoses in 2024

- Approximately 95,000 alcohol-related deaths in the United states annually (drunk driving, health failure)

- In 2023 among 134.7M people aged 12 or older who used alcohol in 2023, 61.4M had engaged in binge drinking in the past month

How can predicting drug use be helpful?

Targeted Intervention

Creating targeted intervention plans for those suffering from drug abuse

Resource Allocation

Distributing healthcare and treatment resources where they are most needed

Improving Education

Promoting awareness and education on the risks of drug use

Further Research

Finding potential solutions to drug abuse and its risks

02

Data

Drug Consumption Dataset

https://www.kaggle.com/datasets/mexwell/drug-consumption-classification/data

- Data from Kaggle last updated April 2024
- Total number of respondents: 1885
- For each respondent, 12 attributes are known
    - Personality type and demographics
    - NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking)
- Measures usage of 18 different legal and illegal substances
    - i.e. Alcohol, cannabis, ketamine, nicotine, * chocolate, * caffeine
    - Recency of usage levels (CL0-CL6)

* The U.S. Food and Drug Administration (FDA) classifies caffeine as both a food additive and a drug due to its psychoactive properties. Chocolate, while not classified as a drug, contains theobromine and caffeine, which are stimulants that produce mild psychoactive effects, as noted by the National Institutes of Health (NIH).

Data Cleaning

- Converted column values for easier legibility and clearer relevance using attribute information given in data card (i.e. age, gender, education level, country of origin, ethnicity, usage)

```
age_col = {
          -0.95197: '18-24',
          -0.07854: '25 - 34',
          0.49788: '35 - 44',
          1.09449: '45 - 54',
          1.82213: '55 - 64',
          2.59171: '65+'
          }
data['Age'] = data['Age'].replace(age_col)
```

```
education_col = {
          -2.43591: 'Left School Before 16 years',
          -1.73790: 'Left School at 16 years',
          -1.43719: 'Left School at 17 years',
          -1.22751: 'Left School at 18 years',
          -0.61113: 'Some College,No Certificate Or Degree',
          -0.05921: 'Professional Certificate/ Diploma',
          0.45468: 'University Degree',
          1.16365: 'Masters Degree',
          1.98437: 'Doctorate Degree',
          }
data['Education'] = data['Education'].replace(education_col)
```

```
usage_col = {
    'CL0': 'Never Used',
    'CL1': 'Used over a Decade Ago',
    'CL2': 'Used in Last Decade',
    'CL3': 'Used in Last Year',
    'CL4': 'Used in Last Month',
    'CL5': 'Used in Last Week',
    'CL6': 'Used in Last Day',
    }
data['Alcohol'] = data['Alcohol'].replace(usage_col)
data['Amphet'] = data['Amphet'].replace(usage_col)
data['Amyl'] = data['Amyl'].replace(usage_col)
data['Benzos'] = data['Benzos'].replace(usage_col)
data['Caff'] = data['Caff'].replace(usage_col)
data['Cannabis'] = data['Cannabis'].replace(usage_col)
data['Choc'] = data['Choc'].replace(usage_col)
data['Coke'] = data['Coke'].replace(usage_col)
data['Crack'] = data['Crack'].replace(usage_col)
data['Ecstasy'] = data['Ecstasy'].replace(usage_col)
data['Heroin'] = data['Heroin'].replace(usage_col)
data['Ketamine'] = data['Ketamine'].replace(usage_col)
data['Legalh'] = data['Legalh'].replace(usage_col)
data['LSD'] = data['LSD'].replace(usage_col)
data['Meth'] = data['Meth'].replace(usage_col)
data['Mushrooms'] = data['Mushrooms'].replace(usage_col)
data['Nicotine'] = data['Nicotine'].replace(usage_col)
data['Semer'] = data['Semer'].replace(usage_col)
data['VSA'] = data['VSA'].replace(usage_col)
```

Feature Description

Value	Age ranges
-0.9517	18 - 24
-0.07854	25 - 34
0.49788	35 - 44
1.09449	45 - 54
1.82213	55 - 64
2.59171	65+

Value	Country
-0.09765	UK
-0.57009	USA
-0.28519	Other
0.24923	Canada
-0.09765	Australia
0.21128	Republic of Ireland
-0.46841	New Zealand

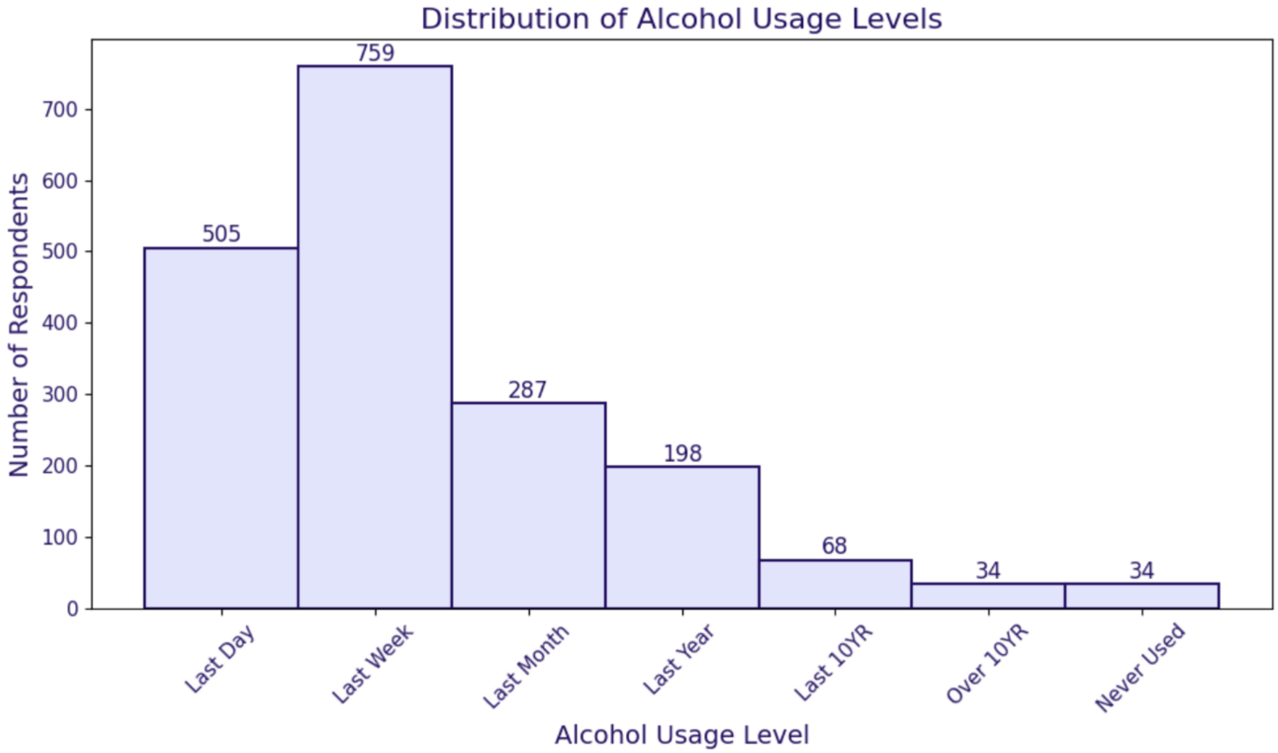
Feature Description

Value	Ethnicity
-0.31685	White
0.11440	Other
-1.10702	Black
-0.50212	Asian
0.12600	Mixed-White/Asian
-0.22166	Mixed-White/Black
1.90725	Mixed-Black/Asian

Drug Usage	
CL0	'Never Used'
CL1	'Used over a Decade Ago'
CL2	'Used in Last Decade'
CL3	'Used in Last Year'
CL4	'Used in Last Month'
CL5	'Used in Last Week'
CL6	'Used in Last Day'

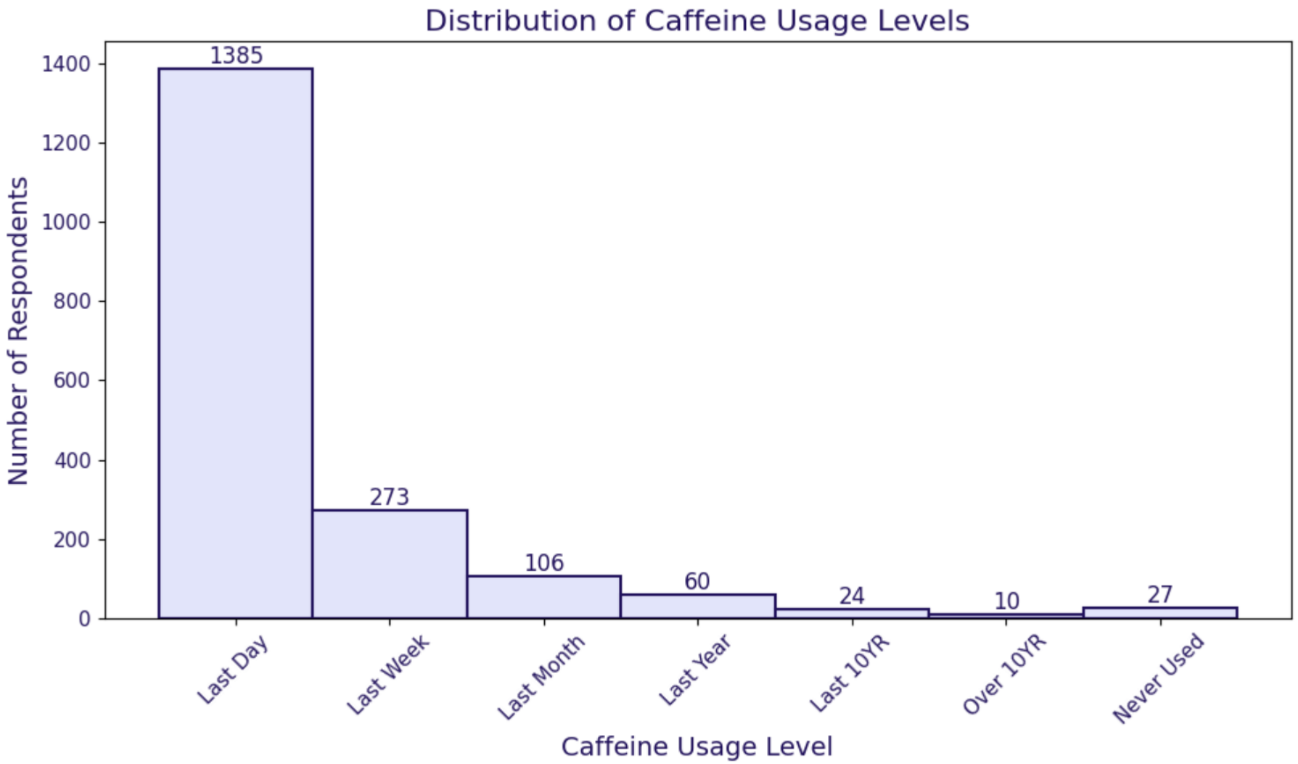
Drug Usage Distribution

Alcohol



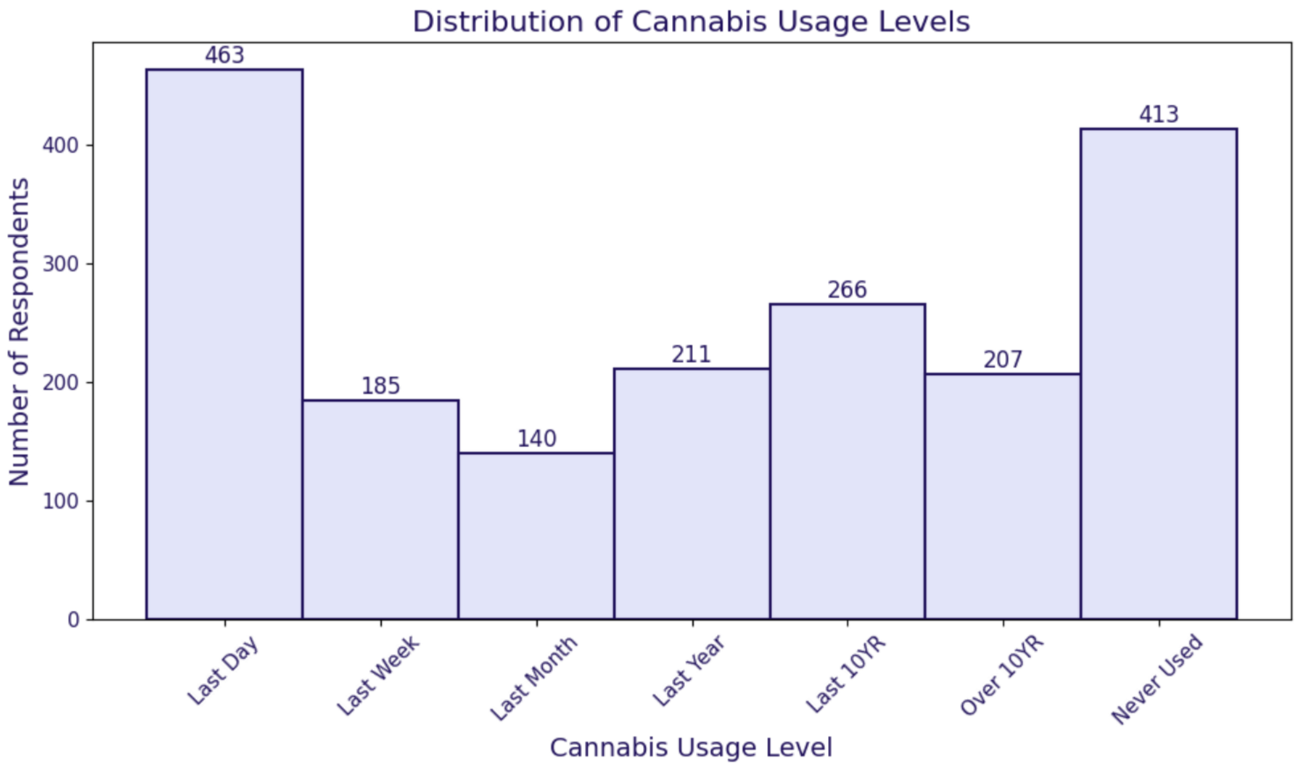
Distribution of Alcohol Usage Levels
Number of Respondents
0
100
200
300
400
500
600
700
505
759
287
198
68
34
34
Last Day
Last Week
Last Month
Last Year
Last 10YR
Over 10YR
Never Used
Alcohol Usage Level

Caffeine



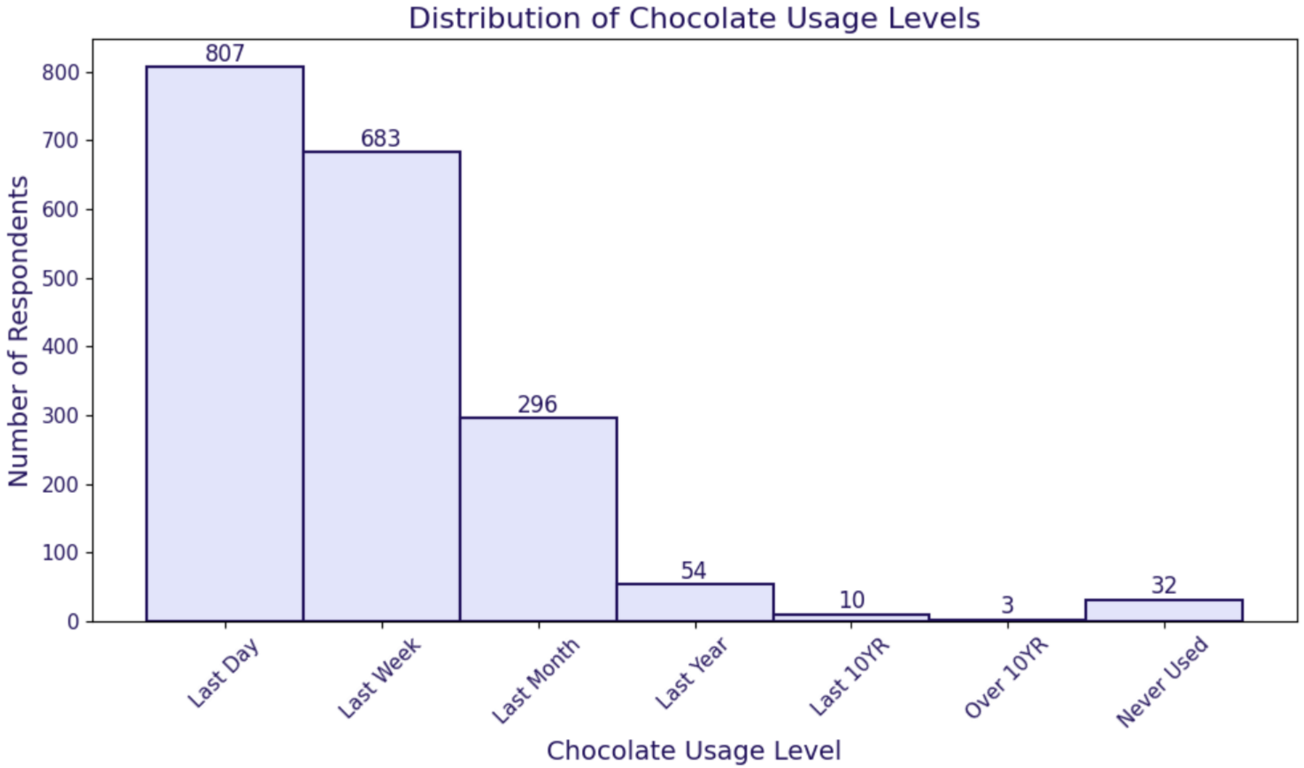
Distribution of Caffeine Usage Levels
Number of Respondents
1400
1200
1000
800
600
400
200
0
1385
273
106
60
24
10
27
Last Day
Last Week
Last Month
Last Year
Last 10YR
Over 10YR
Never Used
Caffeine Usage Level

Cannabis



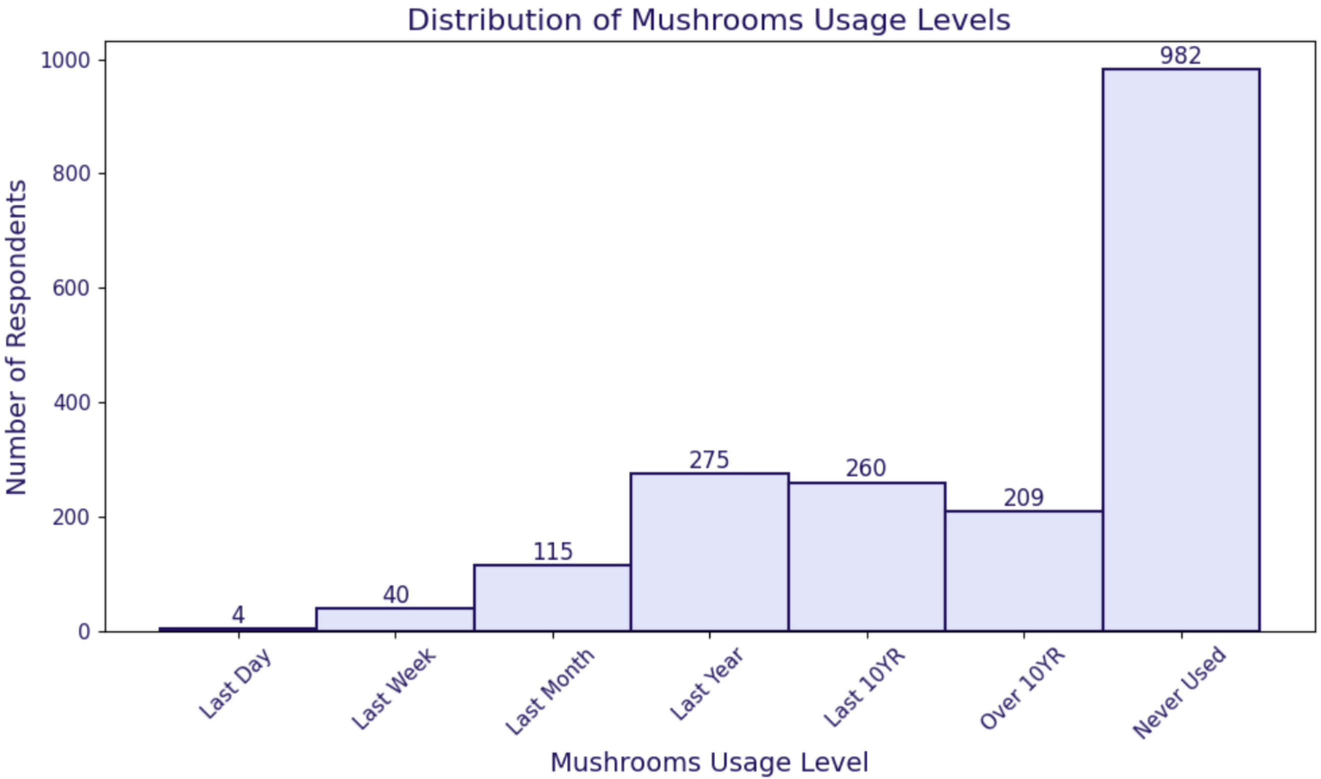
Distribution of Cannabis Usage Levels
463
185
140
211
266
207
413
Number of Respondents
0
100
200
300
400
Last Day
Last Week
Last Month
Last Year
Last 10YR
Over 10YR
Never Used
Cannabis Usage Level

Chocolate



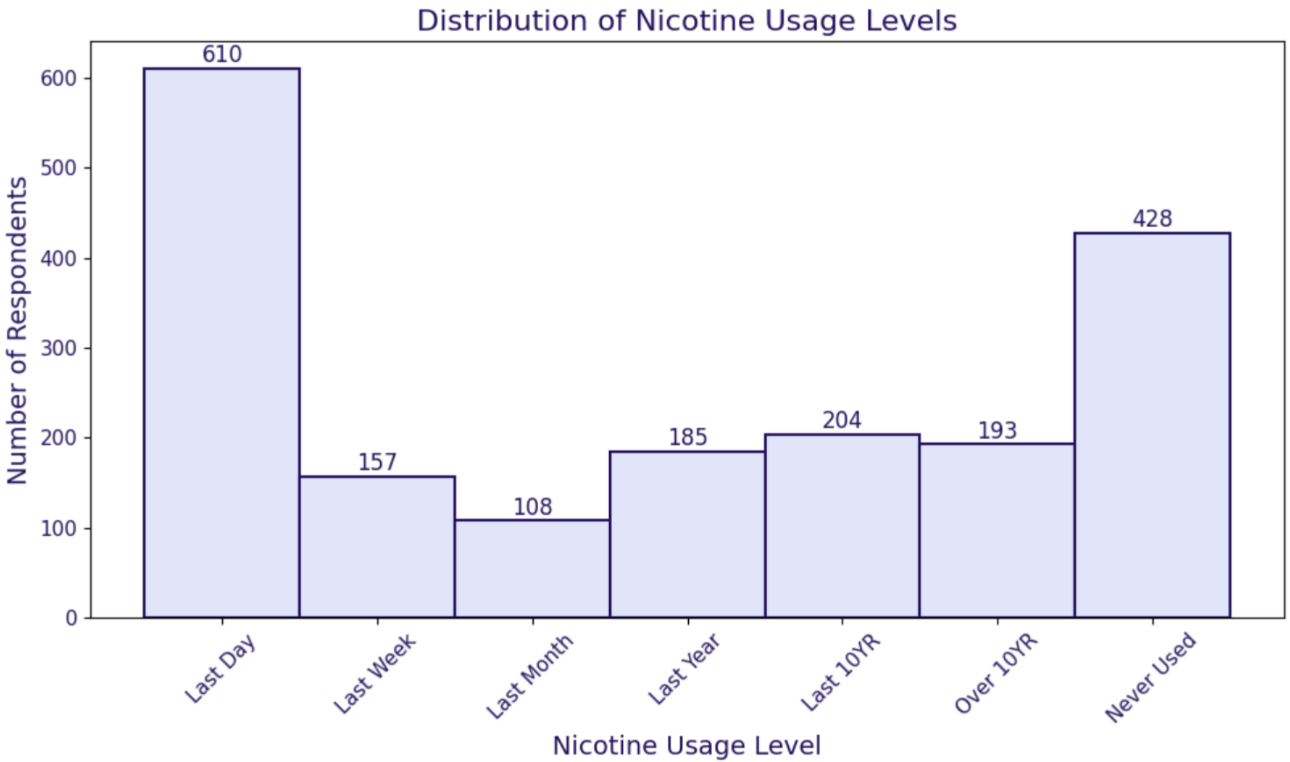
Distribution of Chocolate Usage Levels
Number of Respondents
800
700
600
500
400
300
200
100
0
807
683
296
54
10
3
32
Last Day
Last Week
Last Month
Last Year
Last 10YR
Over 10YR
Never Used
Chocolate Usage Level

Mushrooms



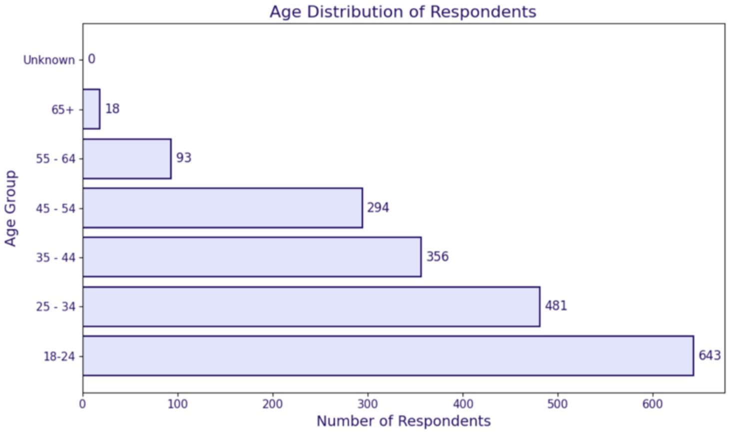
Distribution of Mushrooms Usage Levels
Number of Respondents
1000
800
600
400
200
0
4
40
115
275
260
209
982
Last Day
Last Week
Last Month
Last Year
Last 10YR
Over 10YR
Never Used
Mushrooms Usage Level

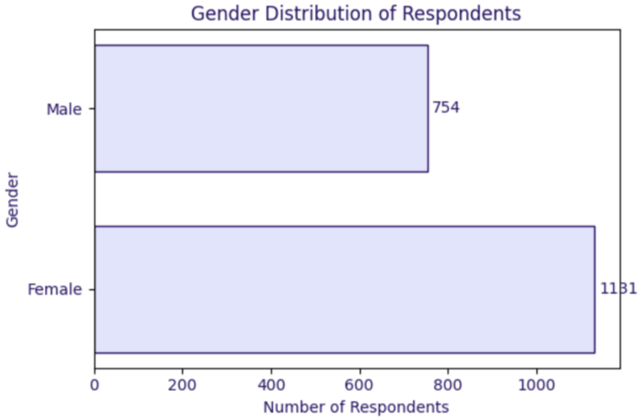
Nicotine

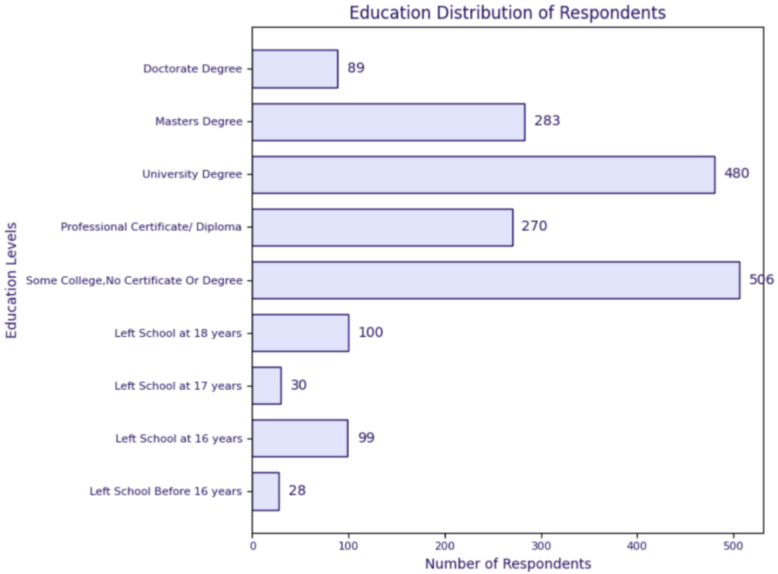


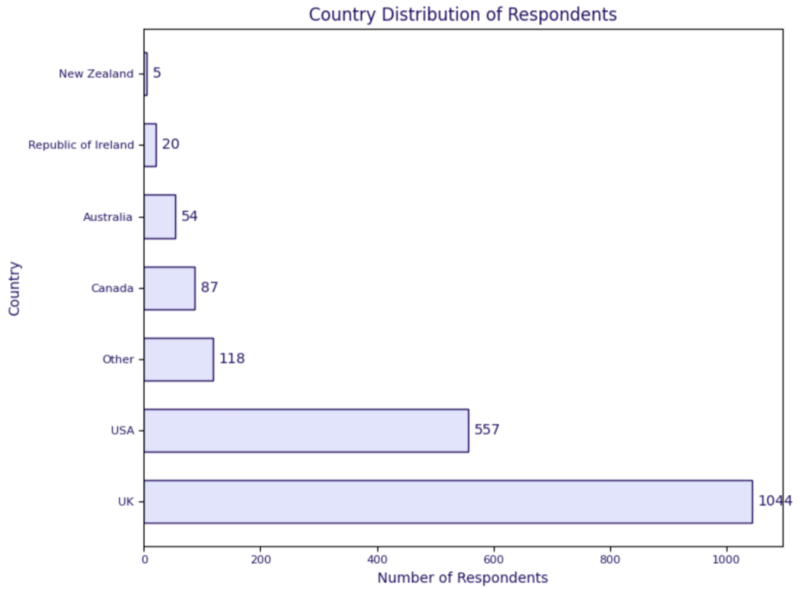
Distribution of Nicotine Usage Levels
Number of Respondents
0
100
200
300
400
500
600
610
157
108
185
204
193
428
Last Day
Last Week
Last Month
Last Year
Last 10YR
Over 10YR
Never Used
Nicotine Usage Level

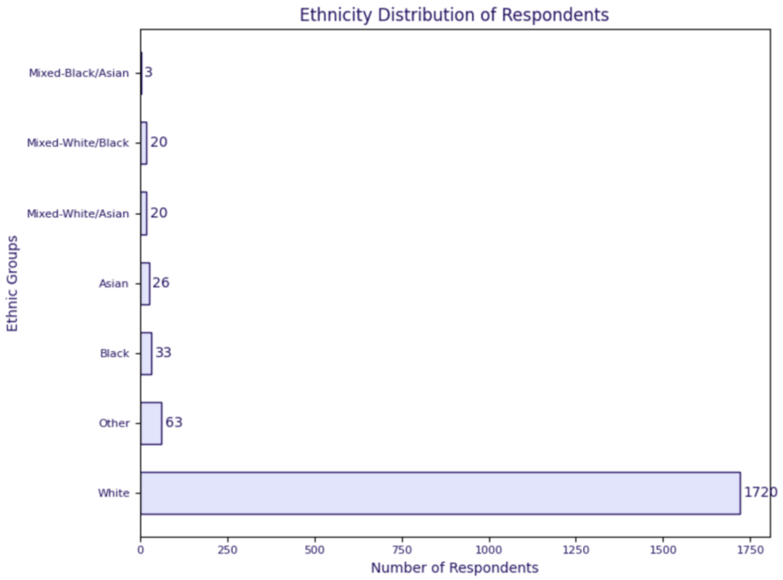
Demographic Distribution


Age Distribution of Respondents
Unknown
65+
55 - 64
45 - 54
35 - 44
25 - 34
18-24
0
18
93
294
356
481
643
Age Group
0
100
200
300
400
500
600
Number of Respondents


Gender Distribution of Respondents
Male
Female
754
1131
Gender
0
200
400
600
800
1000
Number of Respondents


Education Distribution of Respondents
Doctorate Degree
Masters Degree
University Degree
Professional Certificate/ Diploma
Some College,No Certificate Or Degree
Left School at 18 years
Left School at 17 years
Left School at 16 years
Left School Before 16 years
89
283
480
270
506
100
30
99
28
Education Levels
0
100
200
300
400
500
Number of Respondents


Country Distribution of Respondents
New Zealand
Republic of Ireland
Australia
Canada
Other
USA
UK
5
20
54
87
118
557
1044
Country
0
200
400
600
800
1000
Number of Respondents


Ethnicity Distribution of Respondents
Mixed-Black/Asian
Mixed-White/Black
Mixed-White/Asian
Asian
Black
Other
White
3
20
20
26
33
63
1720
Ethnic Groups
0
250
500
750
1000
1250
1500
1750
Number of Respondents

03

Models

Problem 1

Predicting Drug Usage Levels: Random Forest & Logistic Regression

Preprocessing – Multiclass Random Forest

```
categorical = ['Age', 'Gender', 'Education', 'Country', 'Ethnicity']
numeric = ['Neuroticism', 'Extraversion', 'Openness', 'Agreeableness',
           'Conscientiousness', 'Impulsiveness', 'SensationSeeking']


X = data[categorical + numeric]
y = data['Alcohol']


preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numeric),
        ('cat', OneHotEncoder(handle_unknown='ignore'), categorical)
    ]
)
```

Preprocessing – Multiclass Random Forest

```
X_preprocessed = preprocessor.fit_transform(X)


X_train, X_test, y_train, y_test = train_test_split(X_preprocessed, y, test_size=0.3, random_state=42)


rf_classifier = RandomForestClassifier(
    bootstrap=True,
    max_depth=10,
    min_samples_leaf=1,
    min_samples_split=2,
    n_estimators=100,
    random_state=42
)
```

Best hyperparameters from Grid Search for Alcohol where cv = 10

Problem 1 – Multiclass Results

Alcohol Usage Random Forest Accuracy: 39.22%
Alcohol Usage Logistic Regression Accuracy: 38.87%

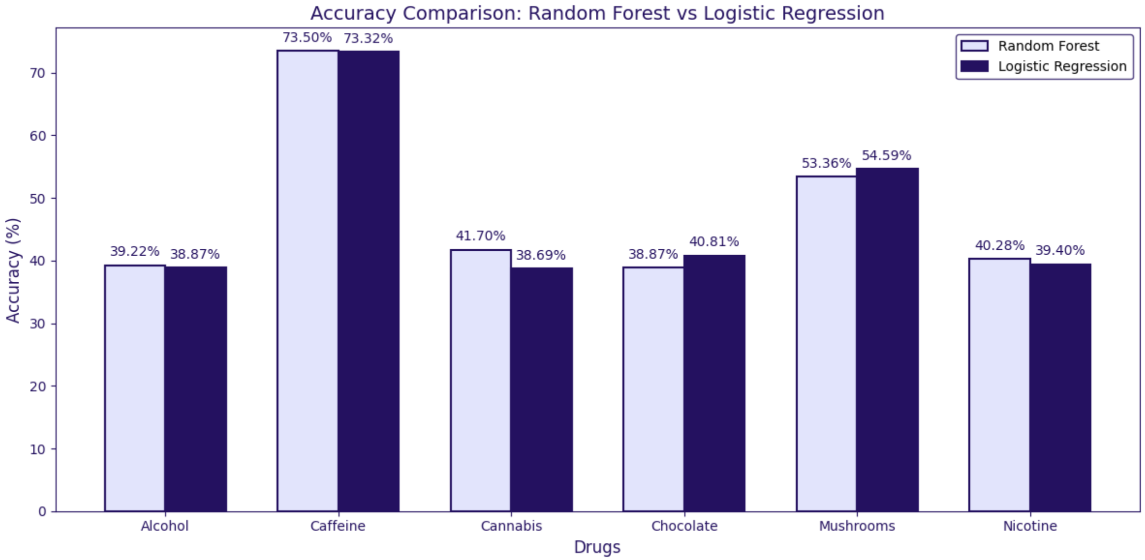
Caffeine Usage Random Forest Accuracy: 73.50%
Caffeine Usage Logistic Regression Accuracy: 73.32%

Cannabis Usage Random Forest Accuracy: 41.70%
Cannabis Usage Logistic Regression Accuracy: 38.69%

Chocolate Usage Random Forest Accuracy: 38.87%
Chocolate Usage Logistic Regression Accuracy: 40.81%

Mushroom Usage Random Forest Accuracy: 53.36%
Mushroom Usage Logistic Regression Accuracy: 54.59%

Nicotine Usage Random Forest Accuracy: 40.28%
Nicotine Usage Logistic Regression Accuracy: 39.40%



Problem 1 – Binary Results

Alcohol:
Random Forest Accuracy (Train): 84.84%
Random Forest Accuracy (Test): 81.63%
Random Forest Precision: 82.67%
Random Forest Recall: 98.28%

Logistic Regression Accuracy (Train): 82.71%
Logistic Regression Accuracy (Test): 81.45%
Logistic Regression Precision: 82.64%
Logistic Regression Recall: 98.07%

Caffeine:
Random Forest Accuracy (Train): 93.78%
Random Forest Accuracy (Test): 93.64%
Random Forest Precision: 93.64%
Random Forest Recall: 100.00%

Logistic Regression Accuracy (Train): 93.56%
Logistic Regression Accuracy (Test): 93.64%
Logistic Regression Precision: 93.64%
Logistic Regression Recall: 100.00%

Cannabis:
Random Forest Accuracy (Train): 93.10%
Random Forest Accuracy (Test): 77.56%
Random Forest Precision: 74.55%
Random Forest Recall: 70.46%

Logistic Regression Accuracy (Train): 81.88%
Logistic Regression Accuracy (Test): 77.21%
Logistic Regression Precision: 74.55%
Logistic Regression Recall: 69.20%

Chocolate:
Random Forest Accuracy (Train): 95.07%
Random Forest Accuracy (Test): 94.70%
Random Forest Precision: 94.70%
Random Forest Recall: 100.00%

Logistic Regression Accuracy (Train): 94.77%
Logistic Regression Accuracy (Test): 94.70%
Logistic Regression Precision: 94.70%
Logistic Regression Recall: 100.00%

Mushrooms:
Random Forest Accuracy (Train): 93.63%
Random Forest Accuracy (Test): 91.52%
Random Forest Precision: 0.00%
Random Forest Recall: 0.00%

Logistic Regression Accuracy (Train): 91.58%
Logistic Regression Accuracy (Test): 91.52%
Logistic Regression Precision: 0.00%
Logistic Regression Recall: 0.00%

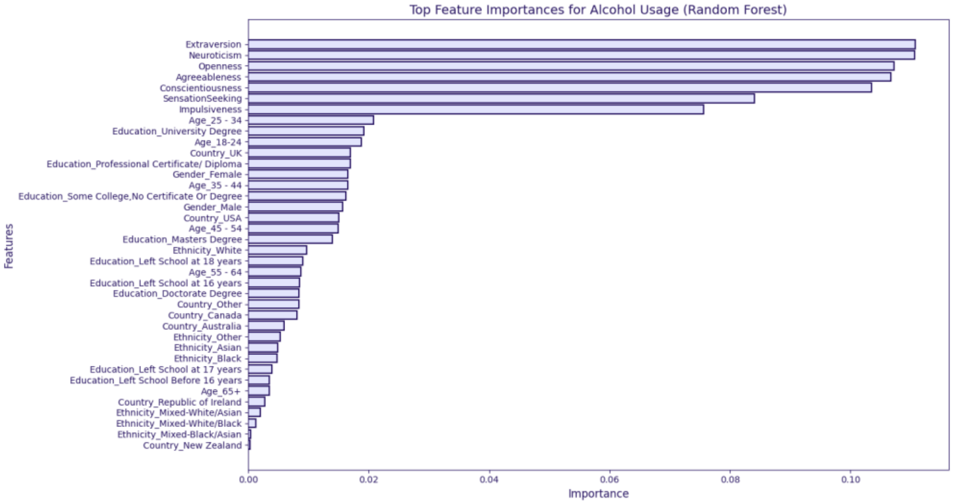
Nicotine:
Random Forest Accuracy (Train): 91.58%
Random Forest Accuracy (Test): 66.25%
Random Forest Precision: 64.17%
Random Forest Recall: 61.98%

Logistic Regression Accuracy (Train): 68.31%
Logistic Regression Accuracy (Test): 65.19%
Logistic Regression Precision: 63.10%
Logistic Regression Recall: 60.46%

04

Empirical Results

Problem 1 – Alcohol



Top 5 Features

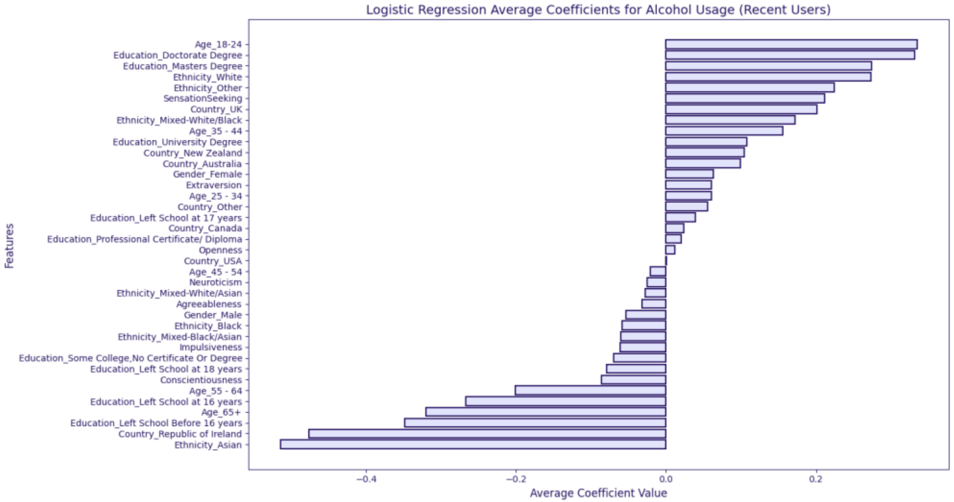
- Extraversion
- Neuroticism
- Openness
- Agreeableness
- Conscientiousness

Most Positive Coefficients

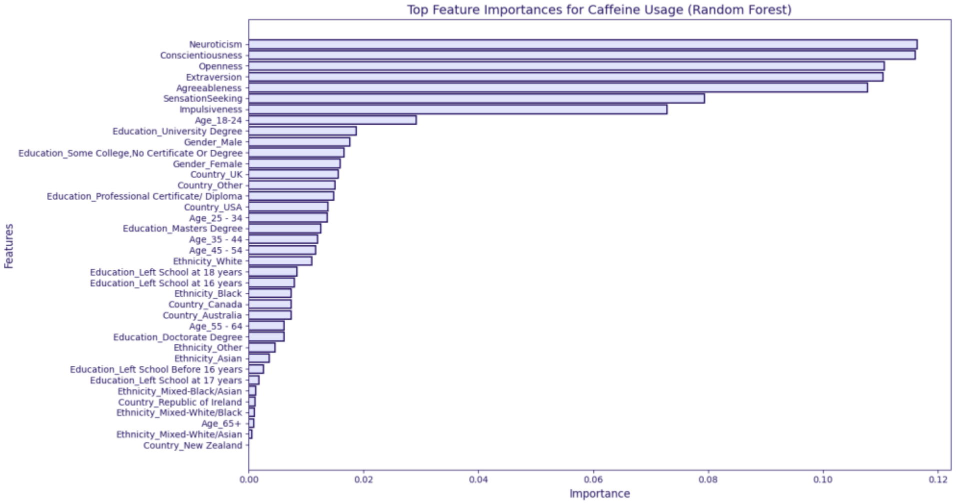
- Age_18-24
- Education_Doctorate Degree
- Education_Masters Degree

Most Negative Coefficients

- Ethnicity_Asian
- Country_Republic of Ireland
- Education_Left School Before 16 years



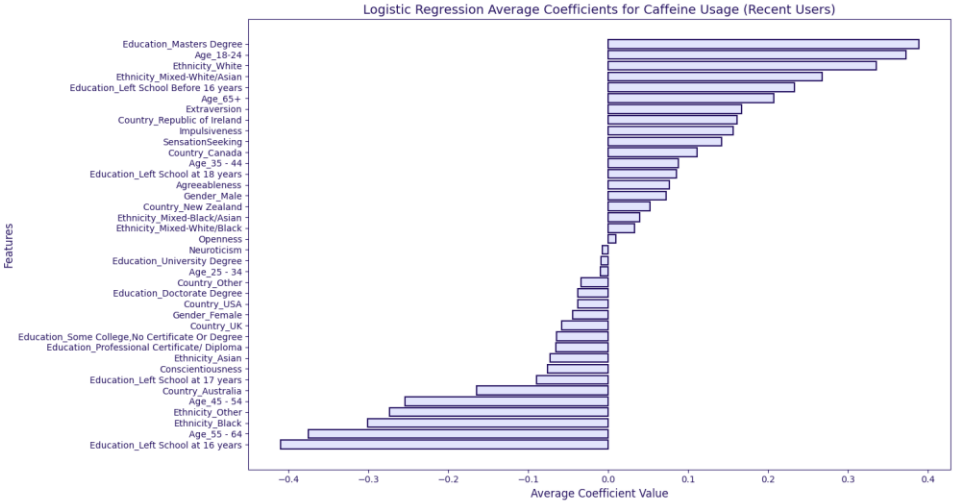
Problem 1 – Caffeine



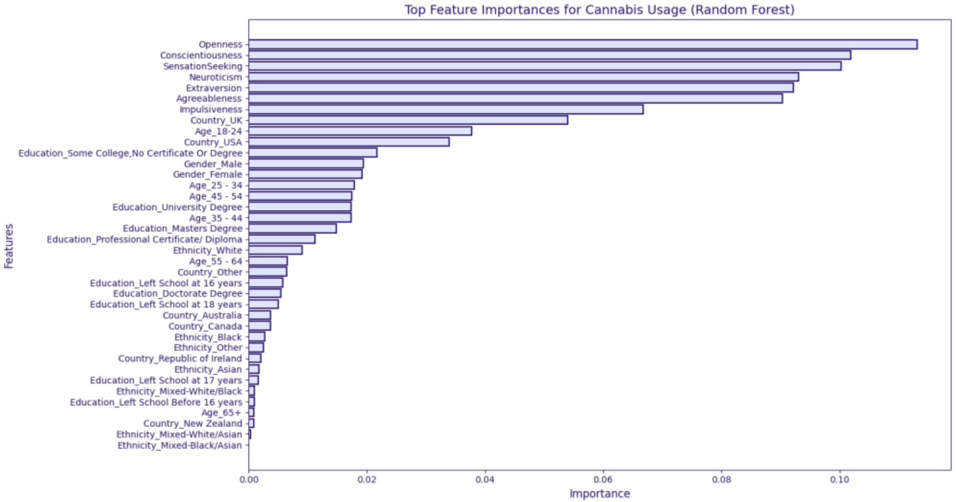
Top 5 Features
- Neuroticism
- Conscientiousness
- Openness
- Extraversion
- Agreeableness

Most Positive Coefficients
- Education_Masters Degree
- Age_18-24
- Ethnicity_White

Most Negative Coefficients
- Education_Left School at 16 years
- Age_55-64
- Ethnicity_Black



Problem 1 – Cannabis



Top 5 Features

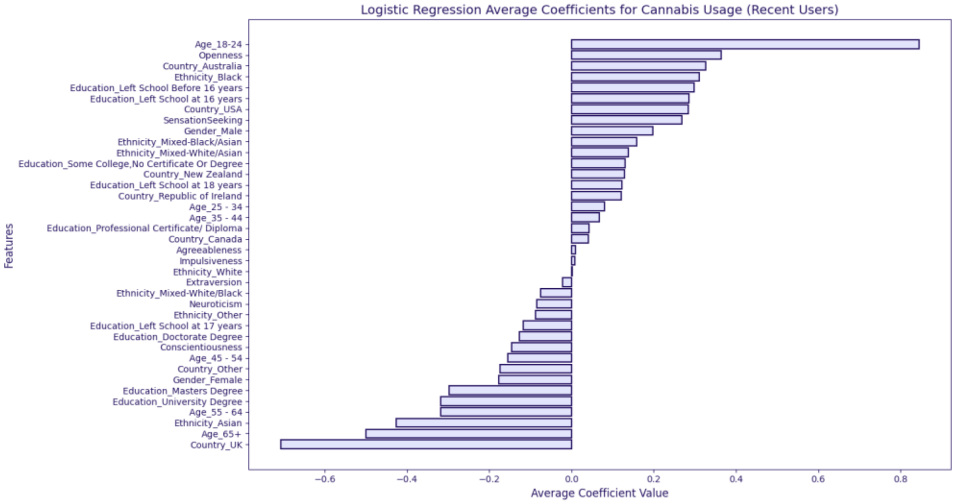
- Openness
- Conscientiousness
- Sensation Seeking
- Neuroticism
- Extraversion

Most Positive Coefficients

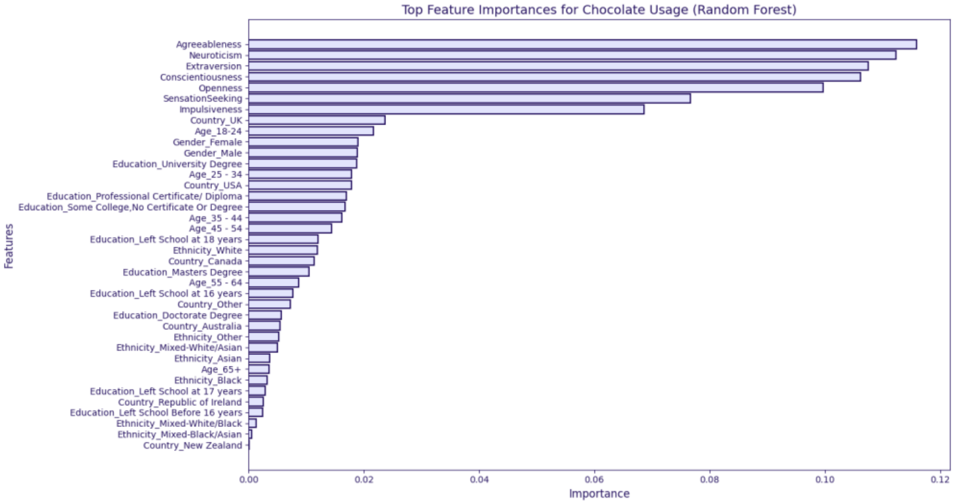
- Age_18-24
- Openness
- Country_Australia

Most Negative Coefficients

- Country_UK
- Age 65+
- Ethnicity_Asian



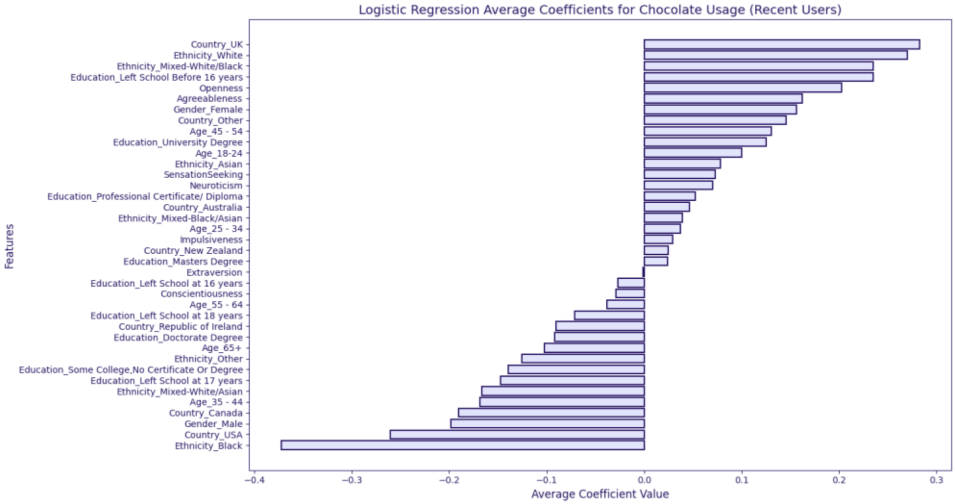
Problem 1 – Chocolate



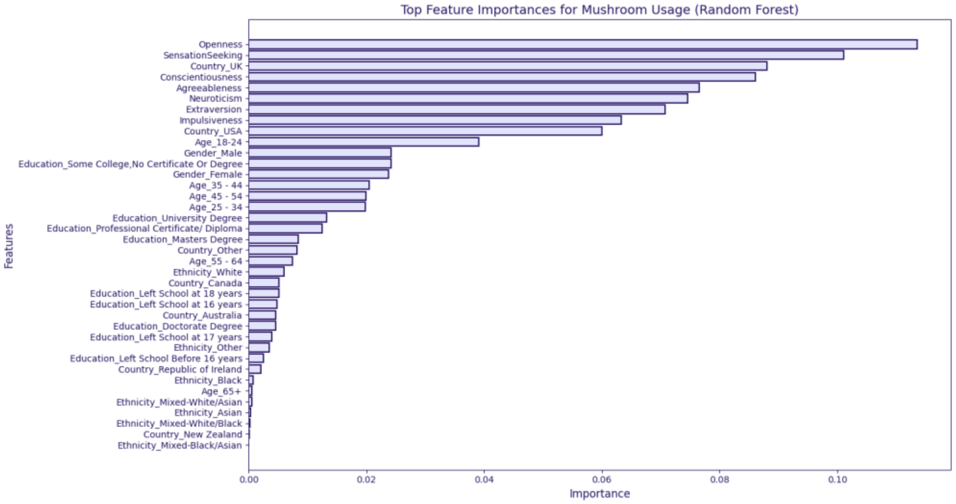
Most Positive Coefficients
- Country_UK
- Ethnicity_White
- Ethnicity_Mixed-White/Black

Most Negative Coefficients
- Ethnicity_Black
- Country_USA
- Gender_Male

Top 5 Features
- Agreeableness
- Neuroticism
- Extraversion
- Conscientiousness
- Openness



Problem 1 – Mushrooms



Top 5 Features

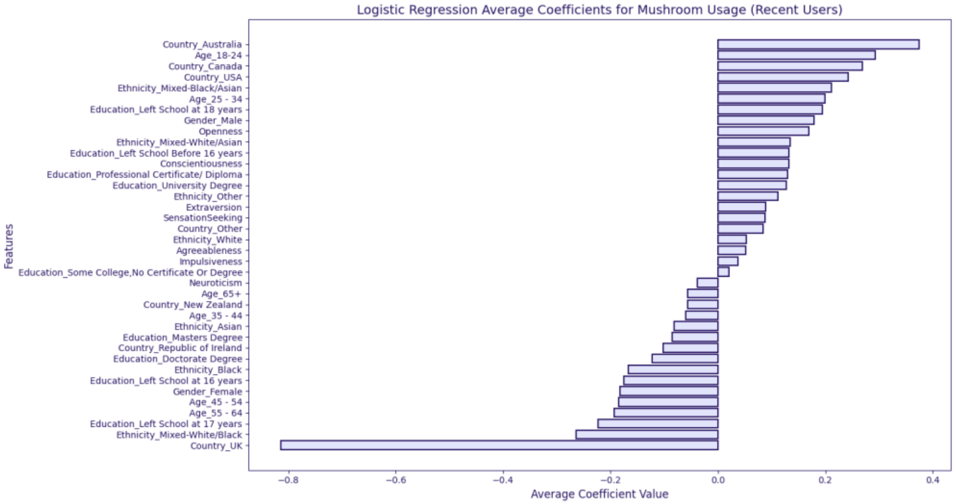
- Openness
- Sensation
- Country_UK
- Conscientiousness
- Agreeableness

Most Positive Coefficients

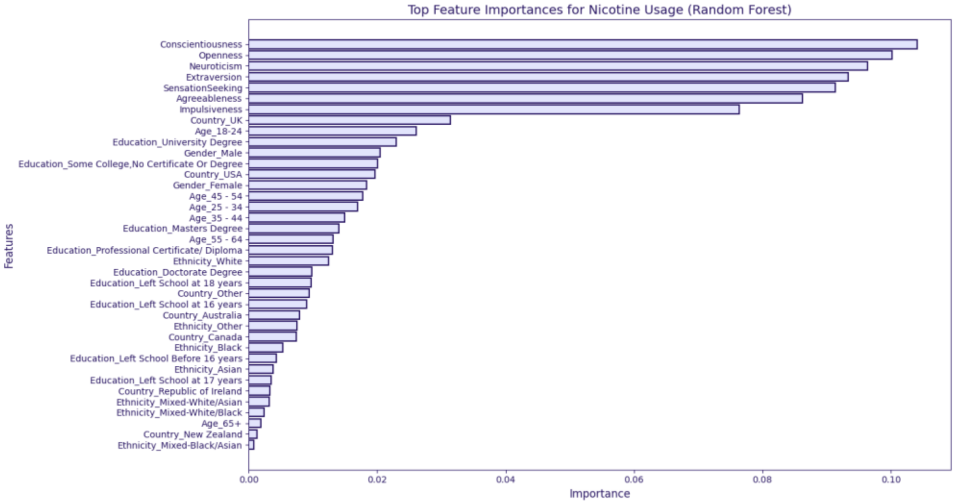
- Country_Australia
- Age_18-24
- Country_Canada

Most Negative Coefficients

- Country_UK
- Ethnicity_Mixed-White/Black
- Education_Left School at 17 years



Problem 1 – Nicotine



Top 5 Features

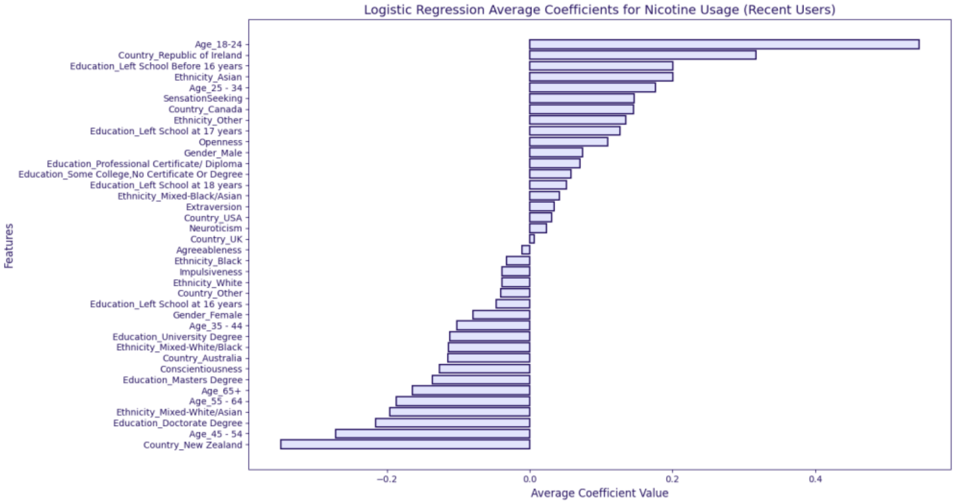
- Conscientiousness
- Openness
- Neuroticism
- Extraversion
- Sensation Seeking

Most Positive Coefficients

- Age_18-24
- Country_Republic of Ireland
- Education_Left School Before 16 years

Most Negative Coefficients

- Country_New Zealand
- Age_45-54
- Education_Doctorate Degree



05

Conclusion

Problem 1 – Key Findings

Random Classifier

- Higher feature importance corresponds to a feature being a stronger determinant of drug usage
- Personality types were the strongest determinants across drug types

Logistic Regression

- Positive coefficients indicate higher likelihood of being a frequent user of a drug
- Negative coefficients indicate lower likelihood of being a frequent user of a drug
- Categorical (binary) features: represent shifts in the log-odds of the outcome when moving from one category to another

    - Ex. Being female increases the log-odds of being a frequent drug user by 0.8 compared to being a male

- Numerical (continuous) features: represent the change in the log-odds of the outcome for every one-unit increase in the feature

    - Ex. For everyone one-unit increase in Neuroticism, the log-odds of being a frequent drugs user increases by 0.2

Problem 1 – Key Findings

Alcohol
- Multiclass Random Forest: Personality traits are strongest indicators of drug usage
- Logistic Regression: Age 18-24 most likely to be a frequent user, Ethnicity Asian least likely to be a frequent user

Caffeine
- Multiclass Random Forest: Personality traits are strongest indicators of drug usage
- Logistic Regression: Education Masters Degree most likely to be a frequent user, Education Left School at 16 years least likely to be a frequent user

Cannabis
- Multiclass Random Forest: Personality traits are strongest indicators of drug usage
- Logistic Regression: Age 18-24 most likely to be a frequent user, From UK least likely to be a frequent user

Chocolate
- Multiclass Random Forest: Personality traits are strongest indicators of drug usage
- Logistic Regression: From UK most likely to be a frequent user, Ethnicity Black least likely to be a frequent user

Mushrooms
- Multiclass Random Forest: Living in the UK is a strong indicator of drug usage
- Logistic Regression: From Australia most likely to be a frequent user, From UK least likely to be a frequent user

Nicotine
- Multiclass Random Forest: Personality traits are strongest indicators of drug usage
- Logistic Regression: Age 18-24 most likely to be a frequent user, From New Zealand least likely to be a frequent user

Future Use of Findings

- Targeting sensation seeking behaviors in solutions- providing constructive alternatives to risky behaviors
- Target ages 18-24 – improving education around drug usage & abuse
- Create age-specific risk assessment tools incorporating personality trait measurements for early intervention programs and preventive healthcare strategies
- Develop studies that track personality trait evolution and drug use patterns across age groups to identify critical transitionary moments
- Shifting focus of targeted intervention programs to women

Thank You!