Thanh Nguyen

Capstone Two Final Report:
"Personality Traits and Drug Consumption"

1) Introduction

Understanding an individual's risk of drug consumption is critical. Most of us would like to avoid drugs, either legal or illegal. Even more important, we would like to raise drug-free children. In this project, we look at the association between personality traits and drug consumption using a sample of 1885 individuals from diverse demographic backgrounds. Knowing the personality traits that are highly associated with drug consumption will not only help ourselves but also help us to help people around us to live a drug-free life. Our predictive models can be used by school staff when dealing with students as well as by healthcare professionals when dealing with patients.

2) Dataset

Our data is taken from the University of California, Irvine's Machine Learning Repository. The dataset owners collected the data using an online survey methodology and employed it in a study of personality and drug consumption risk[1]. The dataset contains data for 1885 respondents. The survey asked participants questions concerning their use of 18 legal and illegal drugs. We use this information to construct the outcome variables (the targets), whether the participant was user or non-user of a certain drug, for our project. We focus only on two drugs: amyl nitrite (amyl hereafter) and cannabis.

For each respondent, demographic factors and personality measurements were collected. These factors and measurements constitute the independent variables (the features) for our project. Demographic factors include age, gender, level of education, country of residence, and ethnicity. Personality measurements include Big Five personality traits, impulsivity and sensation seeking.

[1] UCI Machine Learning Repository: Drug consumption (quantified) Data Set.

The Big Five features are Nscore, Escore, Oscore, Ascore and Nscore. Fehrman et. al. (2015) summarizes these traits as:

1. *Neuroticism* (N) is a long-term tendency to experience negative emotions such as nervousness, tension, anxiety, and depression.
2. *Extraversion (E)* is manifested in outgoing, warm, active, assertive, talkative, cheerful, and in search of stimulation characteristics.
3. *Openness to experience (O)* is a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests.
4. *Agreeableness (A)* is a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness.
5. *Conscientiousness (C)* is a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient.

For each of the five traits, participants were asked to read 12 statements and indicate on a five-point Likert scale how much a given item applied to them (i.e., 0 = 'Strongly Disagree', 1 = 'Disagree', 2 = 'Neutral', 3 = 'Agree', to 4 = 'Strongly Agree'). The other two personality features are Impulsive and SS. Impulsive is impulsiveness measured using 30-item self-report questionnaire. SS is sensation seeking measured using 11 statements in true-false format. For each personality trait, the answers are aggregated and quantified into the scores recorded in the dataset.

3) Exploratory data analysis

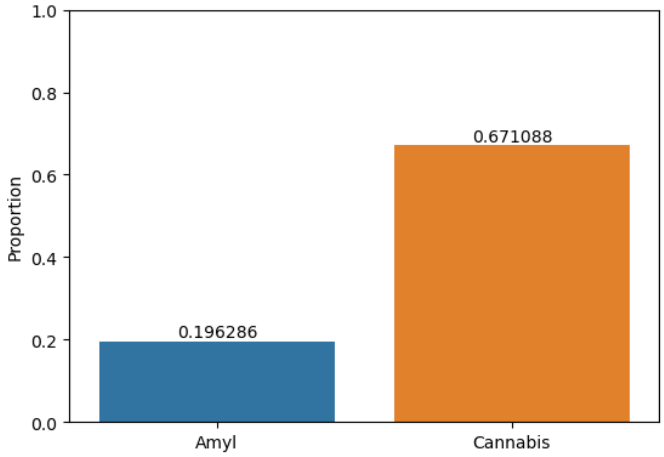
In this section, we explore the distributions of the features and the targets as well as the relationships among them. We also carry out feature importance analysis using Random Forest algorithm.

3.1) Targets

For each drug, the participants selected one of the answers: "Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day". We group these answers into two categories: non-user and user. Non-user category includes "Never Used" and "Used over a Decade Ago"; user category includes the

remaining. Out of 1885 respondents, 370 or 19.63% were Amyl_users. The numbers for Cannabis were 1265 or 67.11%.

Figure 1: Proportion of Amyl and Cannabis Users



3.2) Features

Age:

Participant ages were classified into six levels and quantified as shown in Table 1. In training models, we can either use this feature as a categorical variable (Age group) or as numerical variable (the quantified values - Age_value). We will do both approaches to see what one gives better performance.

Table 1: Number of Participants by Age Groups

Age group	Age_level	Age_value	Count
age18-24	0	-0.95197	643
age25-34	1	-0.07854	481
age35-44	2	0.49788	356
age45-54	3	1.09449	294
age55-64	4	1.82213	93
age65+	5	2.59171	18

Gender:

We have a balanced dataset between genders. The number of females was 942 and the number of males was 943.

Education:

Participant education was classified into nine levels as shown in Table 2. We believe it makes more sense to merge these levels into five groups: 'No high school degree' group includes education levels 1 to 3; 'High school degree' includes level 4; 'Some college experience' includes levels 5 and 6; 'College degree' includes level 7; and 'Graduate degree' includes levels 8 and 9. The numbers of participants in each education group are shown in Table 3.

Table 2: Number of Participants by Education Groups

Education level	Description	Count
1	Left school before 16	28
2	Left school at 16	99
3	Left school at 17	30
4	Left school at 18	100
5	Some college but no degree	506
6	Professional certificate	270
7	University degree	480
8	Masters degree	283
9	Doctorate degree	89

Table 3: Number of Participants by Education Groups

Education_group	Count
No high school degree	157
High school degree	100

Education_group	Count
Some college experience	776
College degree	480
Graduate degree	372

Country:

The numbers of participants in each country are shown in Table 4. UK has 1044 (55.38% of total 1885 participants). The number for USA is 557 (29.55%). All remaining countries account for only 15% of the samples.

Table 4: Number of Participants by Country

Country	Count
Australia	54
Canada	87
New Zealand	5
Other	118
Republic of Ireland	20
UK	1044
USA	557

Ethnicity:

The participants come from diverse ethnicities. However, majority of them (91.25%) are 'White'. Thus, we will not include this feature in our analysis.

Big Five traits (Nscore, Escore, Oscore, Ascore, Cscore):

These features have been normalized (i.e., means equal zero and standard deviations equal one). Their distributions are very close to normal and there is no concern about outliers.

Impulsive and SS:

Impulsive (Sensation Seeking) is measured at 10 (11) levels. As shown in figure 2 and 3, Impulsive has positive skewness and SS has negative skewness. We will treat these features as numerical variables.

Figure 2: Distribution of Impulsive Values

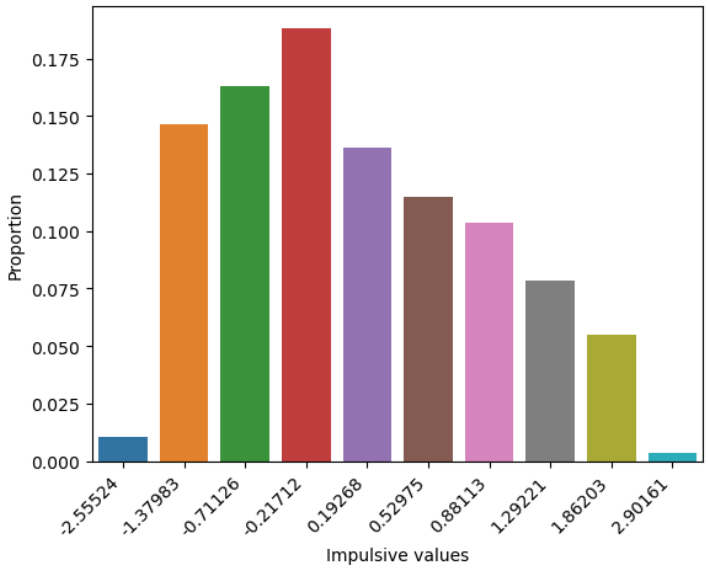
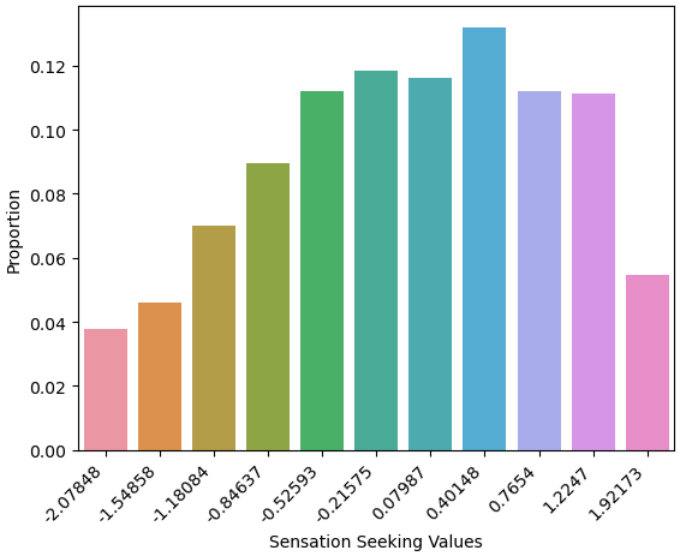


Figure 3: Distribution of Sensation Seeking (SS) Values



3.3) Relationship between targets and features

3.3.1) Drug consumption across age groups

As shown in figure 4 and 5, both amyl and cannabis consumption seem to decrease with ages.

Figure 4: Amyl Consumption Across Age Groups

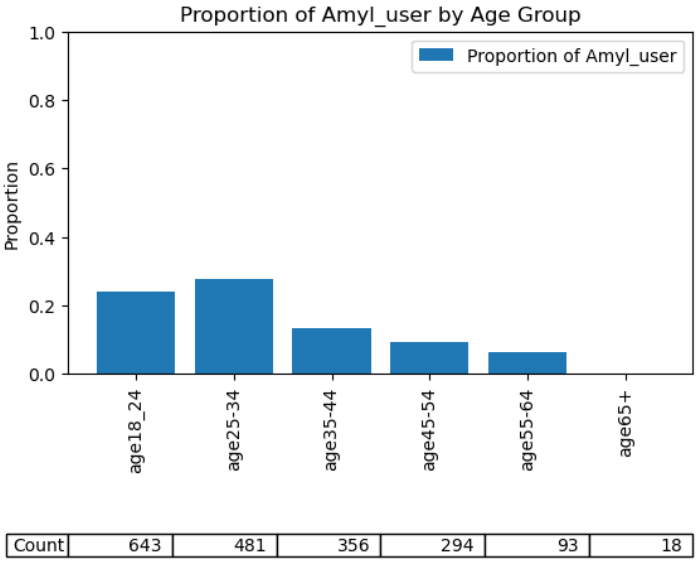
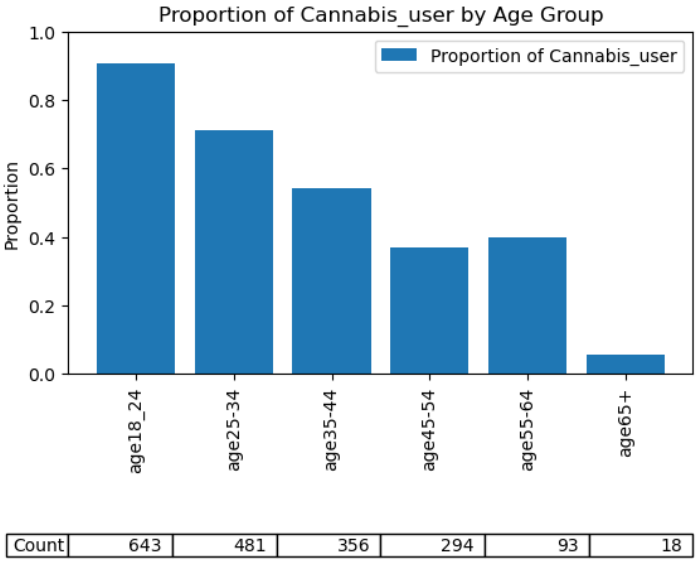


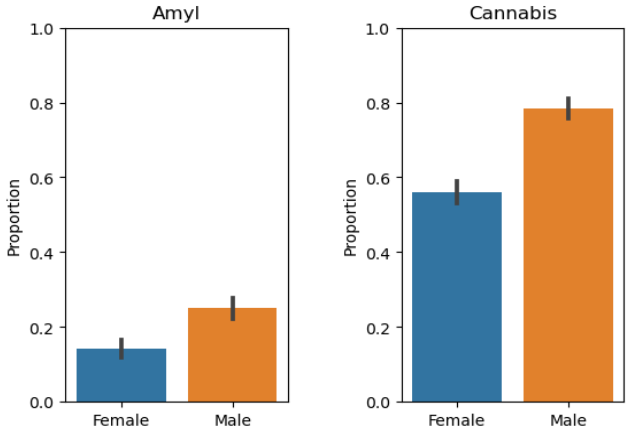
Figure 5: Cannabis Consumption Across Age Groups



3.3.2) Drug consumption across gender

Figure 6 shows proportions of males and females who are drug users. Males are more likely to be users for both amyl and cannabis. This pattern makes sense since genders correlate with some personality traits that are expected to predict drug usage such as impulsiveness and sensation seeking.

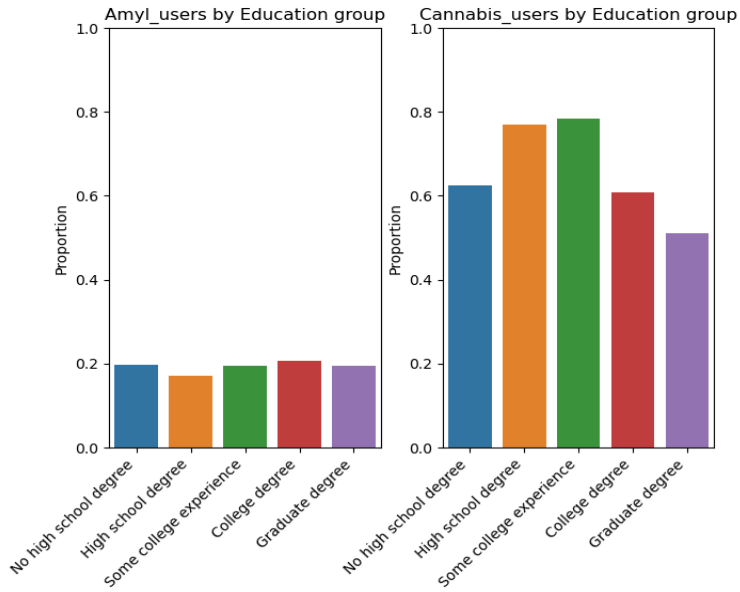
Figure 6: Drug Consumption Across Gender



3.3.3) Drug consumption across education groups

Figure 7 shows while there is no clear relationship between amyl consumption and education. Cannabis consumption and education seem to have a hump-shaped relationship.

Figure 7: Drug Consumption Across Education Groups



3.3.4) Drug consumption across countries

UK and USA constitute majority of the samples. It is interesting to compare the drug consumptions between these two countries. It is clear from figure 8 and 9 that UK has higher proportion of amyl users (21.7%) than USA does (12.4%), while USA has higher proportion of cannabis users (94.6%) than UK does (48%). T-tests show the difference in the proportions are statistically significant at p-value less than 1%.

Other interesting evidence is UK seems to have much lower proportion of cannabis users than all other countries. One should note that other than the USA, other countries have very few samples in the dataset. Thus, the estimated proportions for these countries are less reliable. It is interesting to investigate what factors that make people in the USA more likely to consume cannabis than people living in the UK.

Figure 8: Amyl Consumption Across Countries

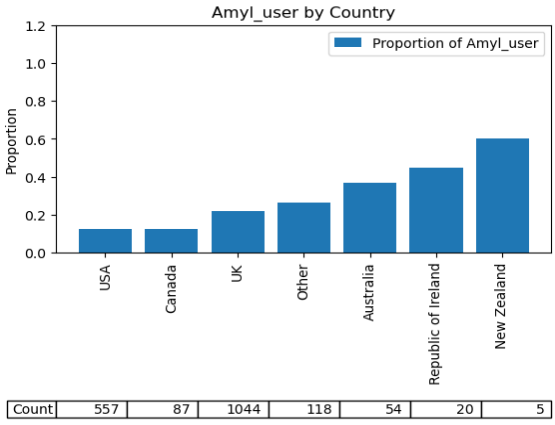
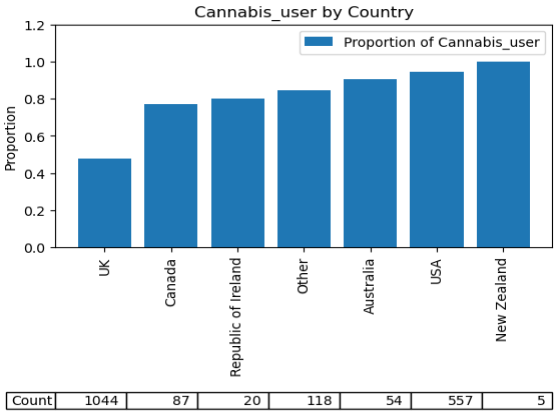


Figure 9: Cannabis Consumption Across Countries



3.3.5) Drug consumption and personality measurements

Figures 10 through 14 show relationship between drug consumption and each of the Big Five personality scores. For each score, we divide the sample into quintiles and plot the proportion of drug users in the quintiles. Overall, there are no clear relationship between amyl consumption and personality traits. On the other hand, there are some clear relationships for cannabis consumption.

Figure 10 show both Amyl and Cannabis consumption seems to increase in Nscore (Neuroticism). The relationship is much stronger for Cannabis.

Figure 10: Drug Consumption Across Nscore Quintiles

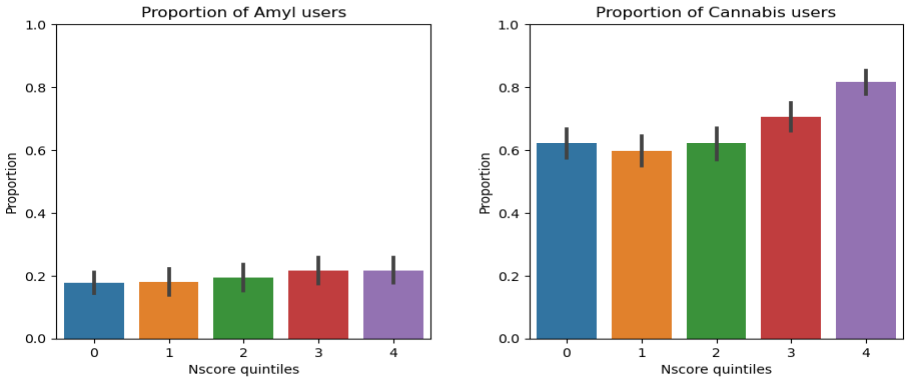


Figure 11 shows while Escore (Extraversion) has no clear relation with Amyl consumption, it has a U-shaped relationship with Cannabis usage.

Figure 11: Drug Consumption Across Escore Quintiles

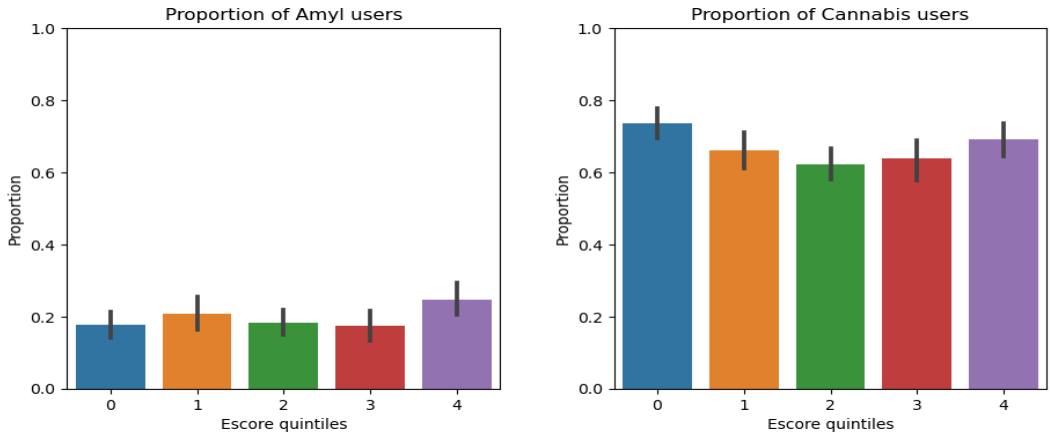


Figure 12 shows while Oscore (Openness to experience) has no clear relation with Amyl consumption, it seems to have a strong positive correlation with Cannabis usage.

Figure 12: Drug Consumption Across Oscore Quintiles

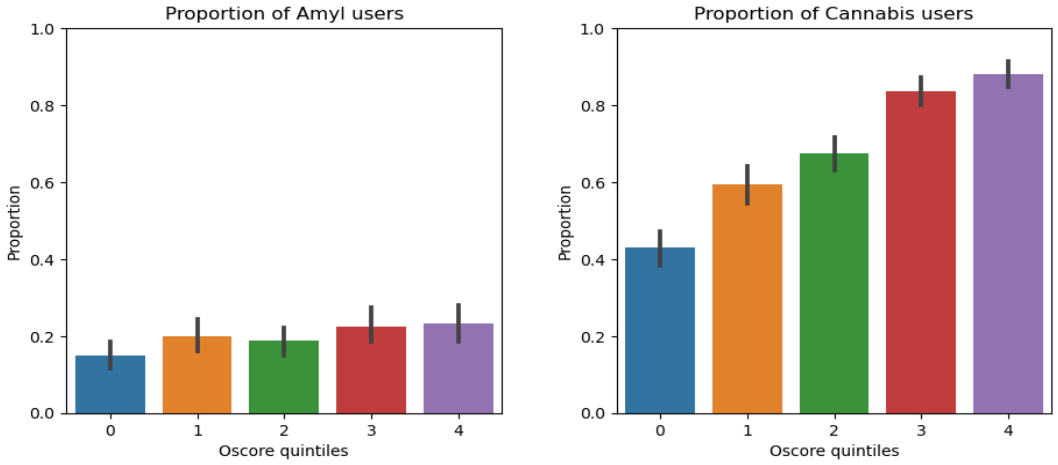


Figure 13 show Ascore (Agreeableness) has a slight negative correlation with Amyl consumption and a strong negative correlation with Cannabis consumption. The negative correlations seem counter-intuitive. One would expect people with high Ascore are more influenced by peer pressure and thus are more likely to try new things including drugs.

Figure 13: Drug Consumption Across Ascore quintiles

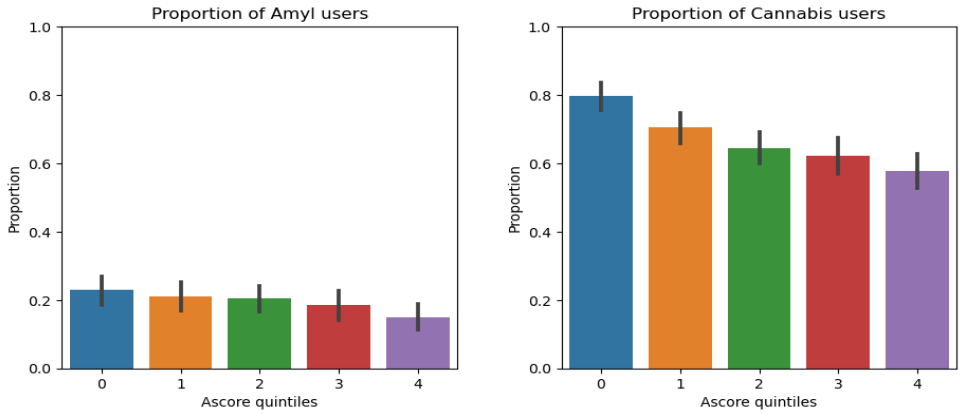
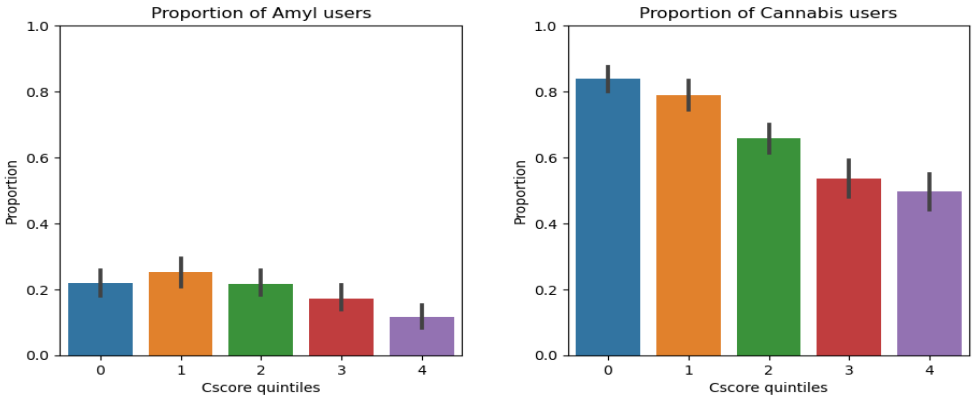


Figure 14 shows negative correlation between Cscore (Conscientiousness) and both Amyl and Cannabis consumption. This relationship is expected.

Figure 14: Drug Consumption Across Cscore quintiles



3.4) Correlation matrix and categorical association

3.4.1) Correlation matrix

Figure 15 shows correlation matrix among the numerical features and targets. Some of the takeaways here are:

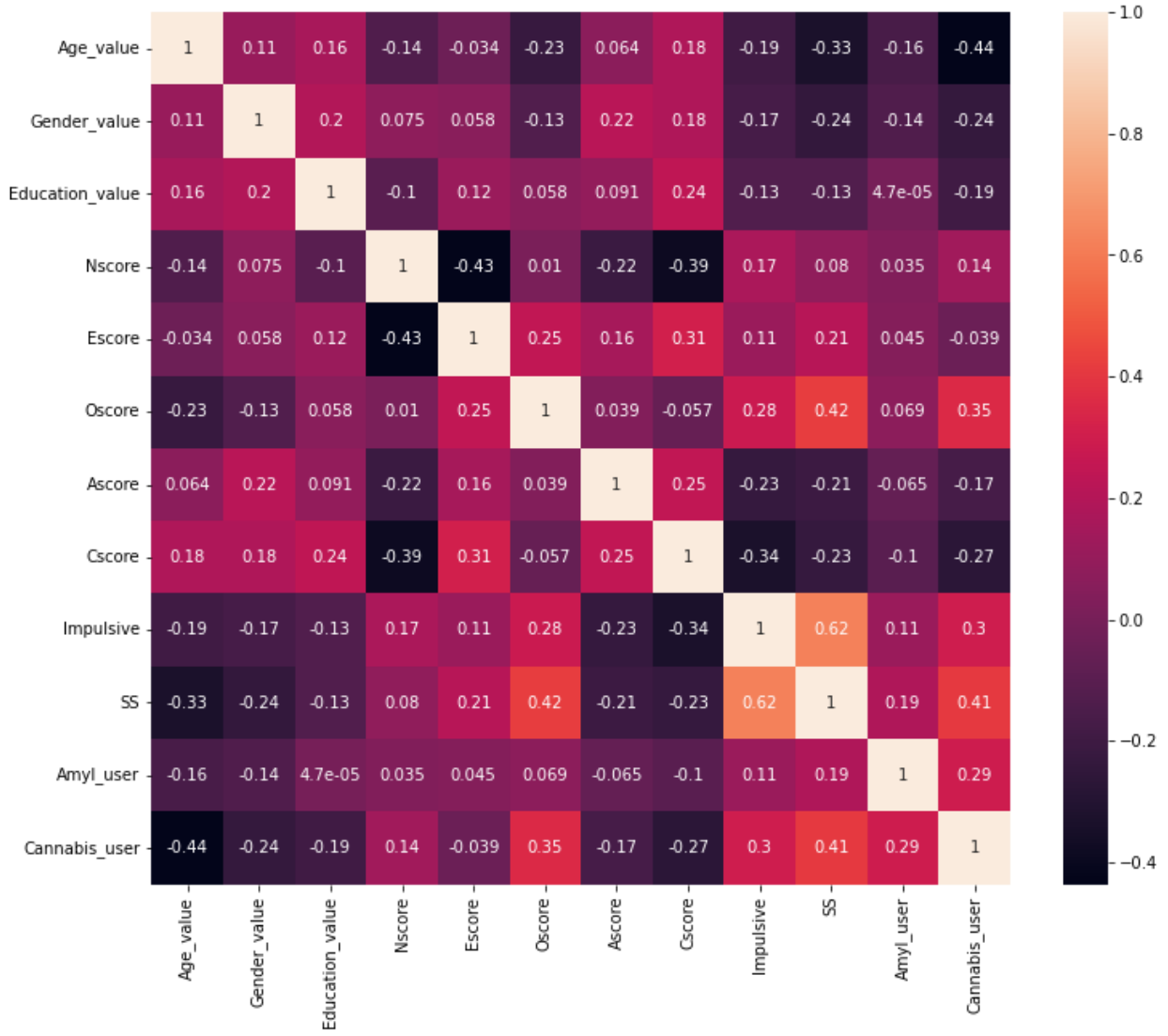
- There is some concern about multicollinearity. Some correlations among Big Five traits, Impulsive and SS are high. For example, correlation between impulsive and SS is 0.62; between Oscore and SS is 0.42.

- Some correlations between demographic and personality features are significant. These correlations should be expected since environmental factors affect people's personality and vice versa.

- It might be harder to predict Amyl users than to predict Cannabis users. The magnitude of correlations of features with Amyl usage is lower than with Cannabis usage. For example, Age_value and Cannabis_user have correlation of -0.44; while Age_value and Amyl_user correlation is only -0.16.

- There is a positive correlation between amyl and cannabis consumption. This evidence is quite expected. If one consumes one drug, one is more likely to consume the other.

Figure 15: Correlation Matrix



3.4.2) Categorical association

The only categorical feature we have is Country. To estimate the predicting power of Country, we calculate Cramer's V (measure of association) and Theil's U (uncertainty coefficient) between Country and the targets. Cramer's V between country and amyl usage (cannabis usage) is 0.15 (0.46). While Cramer's V is symmetric measure of association, Theil's U allows for asymmetry. Theil's U measures, given Country, how much we know about the drug consumption. The Theil's U for amyl is 0.10 and for cannabis is 0.19.

3.5) Feature importance analysis

We use Random Forest algorithm to find out what features are most relevant in predicting drug consumption. Figure 16 shows the relative feature importance in predicting Amyl consumption. The Big Five personality traits, impulsivity and sensation seeking are the top seven features. This result shows the importance of personality in determining whether a person is a drug user.

Figure 16: Relative Feature Importance in Predicting Amyl Consumption

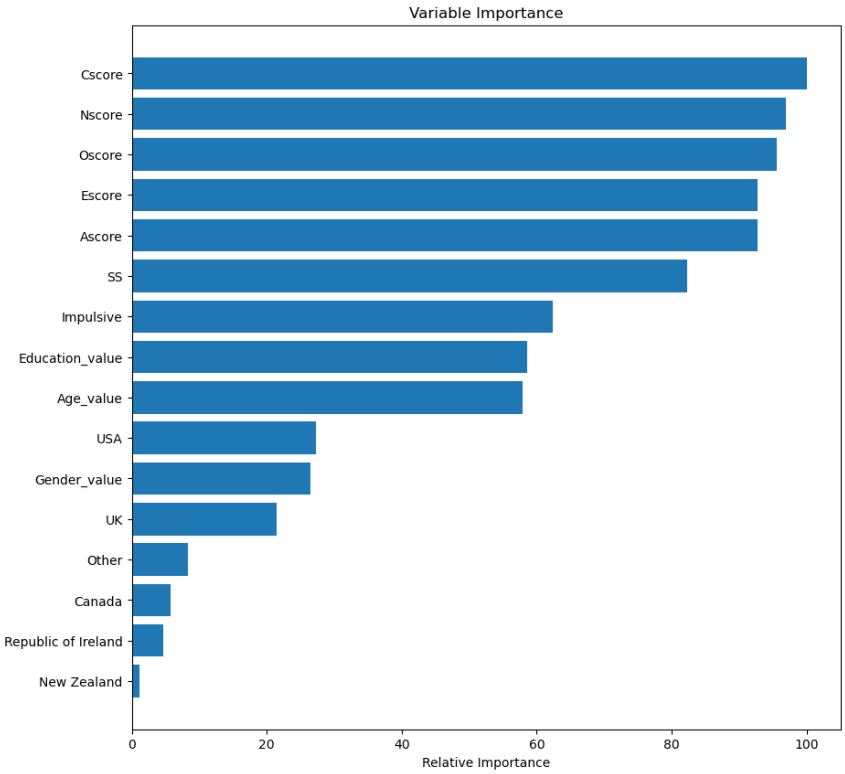


Figure 17 shows the feature importance in predicting Cannabis consumption. Age is the most important feature. It is quite different from the result for Amyl where age has only ninth place. In addition, consistent with the fact that UK has much lower cannabis consumption than all other countries, UK_dummy feature has the third place.

Figure 17: Relative Feature Importance in Predicting Cannabis Consumption

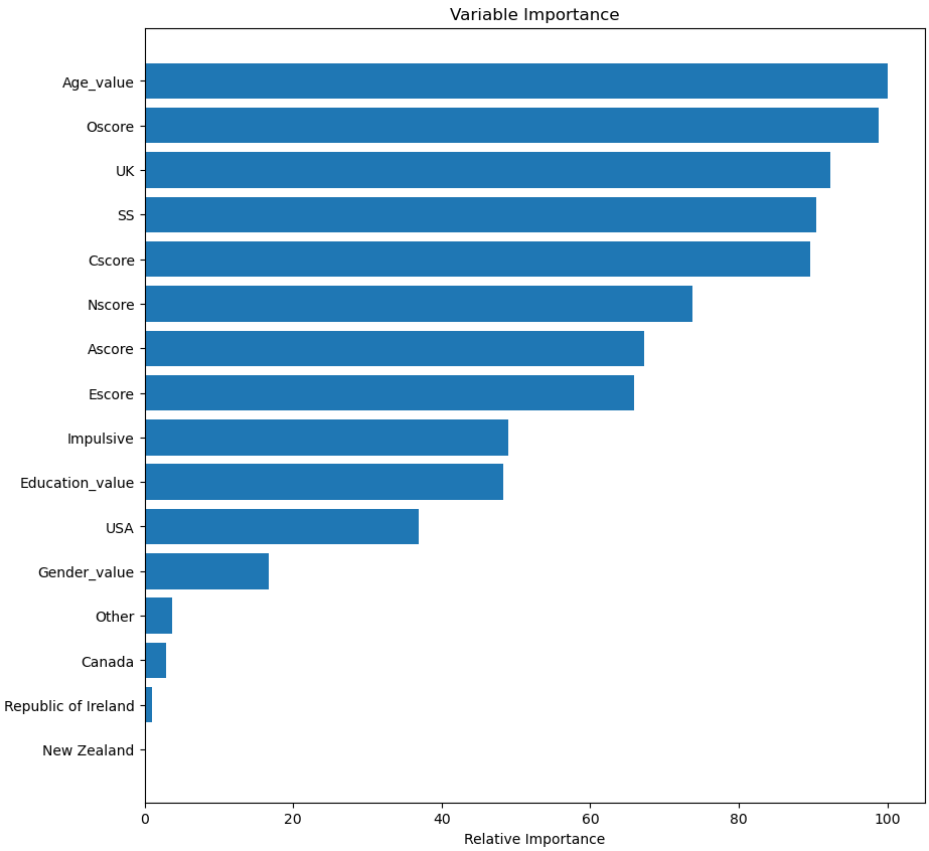


Table 5 shows the top five important features in predicting Amyl and Cannabis consumption. One might think that Cannabis is a "recreational" drug, while Amyl is an "illegal" drug. The consequence of consuming these two drugs is different. Thus, we would expect the important features to be different between the two drugs.

Table 5: Top Five Important Features

Amyl consumption	Cscore	Nscore	Oscore	Escore	Ascore
Cannabis consumption	Age	Oscore	UK	SS	Nscore

4) Modeling

We search the best model to predict Amyl and Cannabis users in three steps. The models we consider are random forest, gradient boosting, and logistic regression. First step, we split our data

into train set (80% of samples) and test set (20% of samples). For each model, we implement a 5-fold grid search on the train set to choose the best hyperparameters. We use area under the ROC curve (roc_auc) as our performance measure.

Once we have chosen the best model with the best hyperparameters we move to the second step - choosing the best threshold for classification to achieve the optimal precision-recall trade-off. We split the train set into a main set (70% of train set) and validation set (30% of train set). We refit the best model on the main set and use validate set to draw precision and recall curves that help us to choose the optimal threshold.

In the third step we evaluate the performance of our model on the unseen data, the test set. We refit the best model now on the whole train set. We use the optimal threshold to make prediction for the test set and report performance measures such as precision, recall and accuracy.

In the original dataset, Age variable is quantified and has 6 unique values. Each value corresponds to an age range, for example age from 25 to 34 (please see table 1). Thus, we have two approaches to treat Age variable: i) Consider Age as numerical type, and ii) Consider Age as categorical type. That is to create age group dummies, for example 'age25_34' or 'age35_44'. We found that treating Age as numerical type always has higher cross-validation performance.

Some of our features have relatively high correlation with each other. Thus, we also check if using only a subset of features can achieve better prediction. To search for the optimal subset of features, we use SelectKBest as well as manually choosing some sensible subsets. However, our result shows that using all features still deliver the best cross-validation performance. Probably, we have only a small number of features (only 19) and the correlations among them are not too high, thus the multicollinearity is not too severe.

4.1) Modeling Amyl users

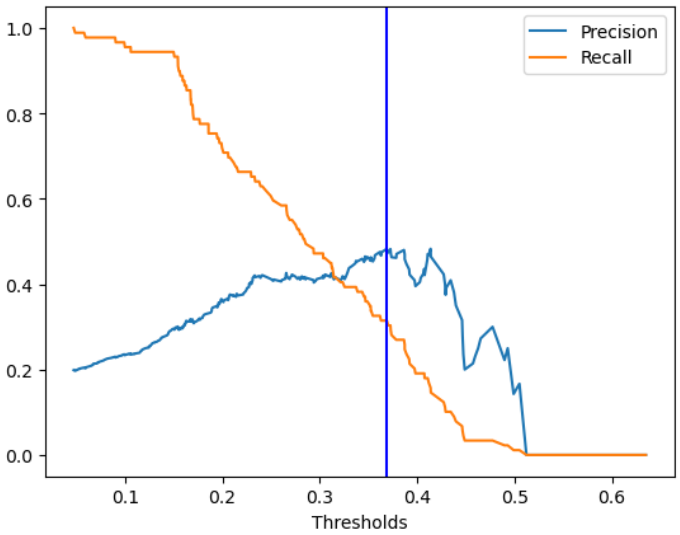
Table 6 shows the hyperparameter tuning result. The three models have very similar performances. Gradient boosting has the highest performance.

Table 6: Hyperparameter tuning for Amyl users prediction

Model Name	Best Parameters	Optimal Feature Set	Cross-Validation Score (roc_auc)
Random Forest	max_depth = 5 max_features = 'auto' n_estimators = 50	All features.	0.7533
Gradient Boosting	learning_rate = 0.05 max_depth = 3 max_features = 8 min_samples_split = 2 n_estimators = 50 subsample = 1	All features.	0.7571
Logistic regression	C = 1 penalty = ‘l1' solver = 'liblinear'	All features.	0.7486

Figure 18 shows the precision and recall curves. Consistent with the result from section 3, exploratory data analysis, that it would be challenging to predict Amyl users, here we do not achieve high precision on the validation set.

Figure 18: Precision and Recall curves for Amyl users prediction



Given the low precision, our model will not be used as a primary tool to predict Amyl users. Our model only can serve as a secondary tool in the toolboxes of school or healthcare professionals. We want people to use our model, thus we choose the threshold that gives us the highest possible precision. The blue vertical line in figure 18 shows the optimal threshold of 0.3687. At this threshold, the precision is 48.28% and the recall is 31.46%.

Now, we have chosen the best model with the best hyperparameters and the optimal threshold. We test our model on unseen data, the test set. The performance of our model is precision 50%, recall 23%, and accuracy 80.4%.

4.2) Modeling Cannabis users

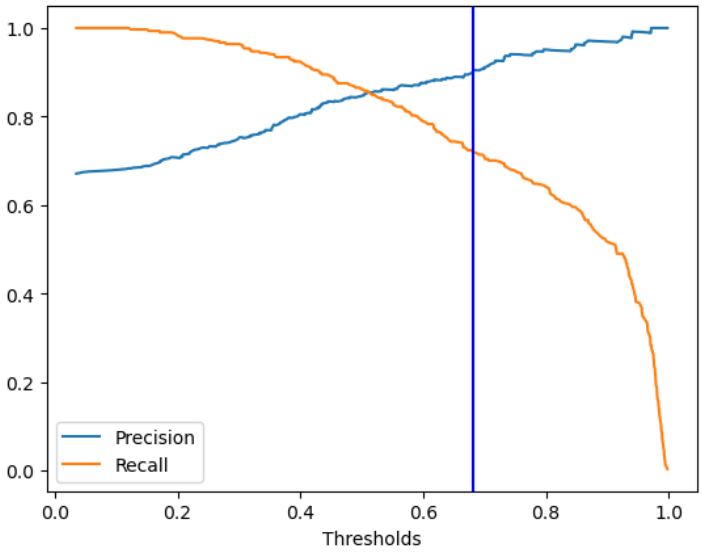
Table 7 shows the hyperparameter tuning result. Like the case of Amyl, the three models have very similar performances. This time, logistic regression has the highest performance.

Table 7: Hyperparameter tuning for Cannabis users prediction

Model Name	**Best Parameters**	**Optimal Feature Set**	**Cross-Validation Score (roc_auc)**
Random Forest	max_depth = 5 max_features = 0.3 n_estimators = 50	All features.	0.8778
Gradient Boosting	learning_rate = 0.02 max_depth = 3 max_features = 6 min_samples_split = 6 n_estimators = 100 subsample = 1.0	All features.	0.8780
Logistic regression	C = 0.1 penalty = 'l2' solver = ‘sag'	All features.	0.8781

Figure 19 shows the precision and recall curves. Consistent with the result from section 3, exploratory data analysis, that it would be easier to predict Cannabis users than to predict Amyl users, here we achieve relative high precision on the validation set.

Figure 19: Precision and Recall curves for Cannabis users prediction



Given the high precision, our model can be used as a primary tool for school or healthcare professionals to predict Cannabis users. For people to use the model, we want it to have high precision. However, there is not much need for precision to be higher than 90%. So, we choose the threshold that delivers precision of 90%. This way we still achieve a reasonably good recall. The blue vertical line in figure 19 shows the optimal threshold of 0.6812. At this threshold, the precision is 90.1% and recall is 72%.

Now, we have chosen the best model with the best hyperparameters and the optimal threshold. We test our model on unseen data, the test set. The performance of our model is precision 90.3%, recall 77.1%, and accuracy 79%.

5) Conclusion and future work

Table 8 summarizes our models' performance on unseen data, the test set. Our model predicting Amyl users does not have high precision. It only can serve as a secondary tool in the toolboxes of school or healthcare professionals. On the other hand, the model predicting Cannabis users has high precision and good recall. This model can be used as a primary tool for professionals.

Table 8: Models' performance on unseen data

	Precision	Recall	Accuracy
Predicting Amyl users	50%	23%	80.4%
Predicting Cannabis users	90.3%	77.1%	79%

There are three areas we can work on to improve our models. First, figure 18 "Precision and Recall curves for Amyl users prediction" shows that the precision curve goes down after the threshold reaches about 0.42. The reason is we have some samples in the validation set that have high predicted probability to be Amyl users, but are not. The validity of these samples needs to be investigated.

The issue described in the above paragraph also highlights a question of how to determine the optimal threshold. Here we determine the optimal threshold based on a single split of the train set (into main and validation sets) and ends up having some "suspicious" samples in the validation set. What would be the optimal threshold, if we split the train set using different random_state arguments. How many splits we should try to choose the optimal threshold.

The second area relates to variables Impulsivity and Sensation Seeking. Like the Age variable, these variables are quantified and have 10 and 11 unique values. In our analysis we consider these variables as numerical type. Alternatively, we can consider them as categorical type and check if we can improve the performance of our models.

Finally, the third area we can explore is how interaction terms among features, especially between individual and environmental factors, relate to drug consumption. Fehrman et al (2015) states "both individual and environmental factors predict substance use and different patterns of interaction among these factors may have different implications".

References:

E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," arXiv [Web Link], 2015