Predicting Drug Abuse Susceptibility

Springboard Capstone - 2

Rishab Ghose

Table of Contents

Introduction

Drug use and abuse is often attributed to poor life decisions rather than the mental health issues that precipitate addiction. Despite the deserved increase in attention that mental health has gotten in the past few decades, addiction continues to be a taboo subject. In this project, I would like to explore some of the key indicators of susceptibility to drug use and build a model to predict an individual's abuse of certain drugs/substances - specifically: alcohol, cocaine, and benzodiazapine use. Such a model could help individuals take protective measures early on to reduce their risk of drug use and addiction if they were to know they were particularly susceptible. Furthermore, understanding the personality mechanisms underpinning drug use may be helpful to further therapeutic approaches to preventing and treating addiction.

Datasource

I will be using the Drug Consumption dataset from the UCI Machine learning repository to build and test my model. This dataset includes 32 features, 12 of which are personality/descriptive attributes about the individual, and 1885 instances. The data itself was collected using an online survey tool from Survey Gizmo in which the respondents, all of whom were over 18 years of age, were anonymously recruited over a 12 month period.

The survey employed three standardized, widely-used, and reliable psychology questionnaires to assess the different personality traits. The Revised NEO Five-Factor Inventory (NEO-FFI-R) was used to measure the "Big Five" basic personality domains (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), the Barratt Impulsiveness Scale (BIS-11) was used to measure impulsiveness in the respondents, and the Impulsiveness Sensation Seeking questionnaire (ImpSS) was used to measure a general sensation-seeking trait. The participants were then questioned on their history of use for 18 legal and illegal drugs and substances as well as one fictious drug which was included to identify those who lied or over-claimed their drug use. Rather than simply discerning whether an individual was a user or non-user of the drug, the questions asked the respondents to specify the recency of last use of the drug as this seemed to give more meaning and freedom to the data and classification of drug user or not.

The dataset can be found here:
https://archive.ics.uci.edu/ml/machine-learning-databases/00373/drug_consumption.data

Exploratory Data Analysis

The dataset includes 7 designations for our target variable:

CL0 = Never Used

CL1 = Used over a Decade Ago

CL2 = Used in Last Decade

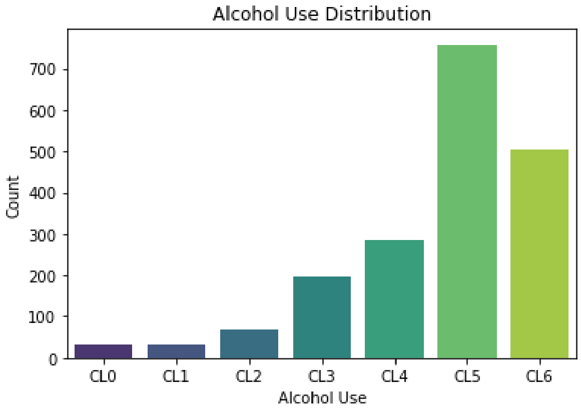
CL3 = Used in Last Year

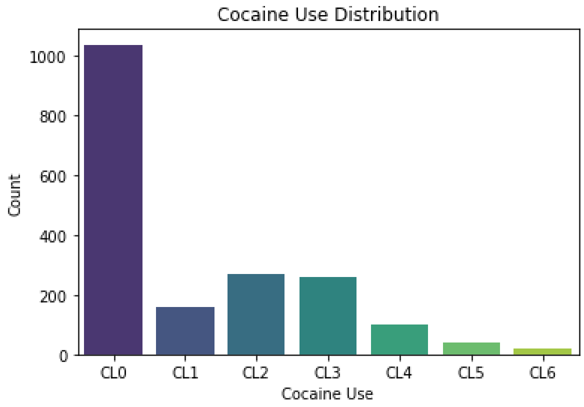
CL4 = Used in Last Month

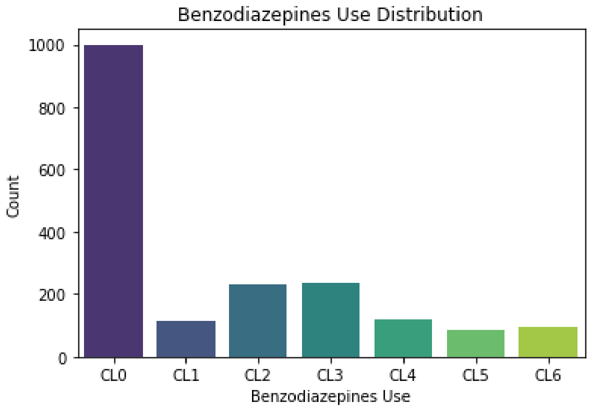
CL5 = Used in Last Week

CL6 = Used in Last Day

Firstly, from the below distributions of alcohol, cocaine, and benzodiazepines, I learned that the majority of our respondents consume alcohol quite frequently (within the last day, week, or month). Very few respondents have never tried alcohol, while the majority of the respondents have never tried cocaine or benzos. There are slightly more regular users of benzos than cocaine in our sample, likely due to the fact that benzos are legal prescription drugs while cocaine is illegal in most countries.

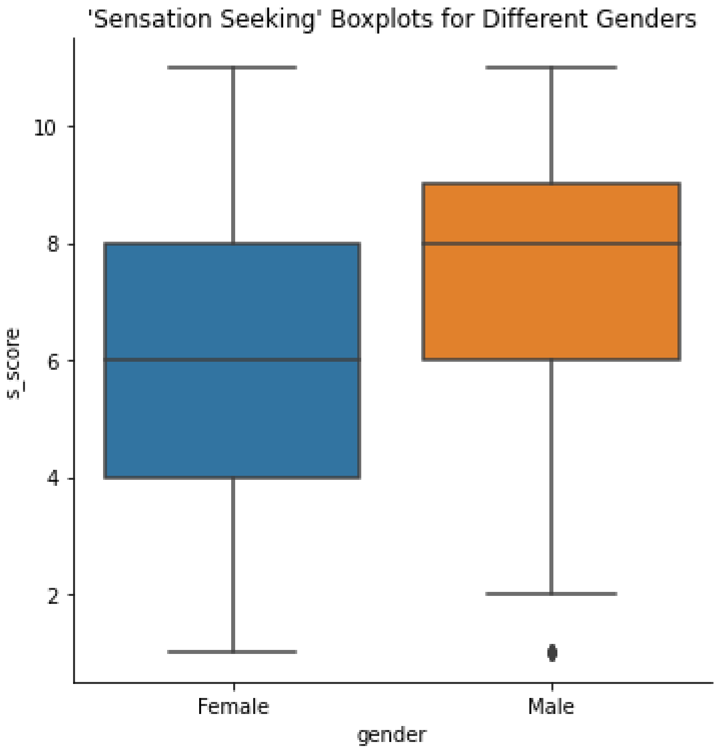






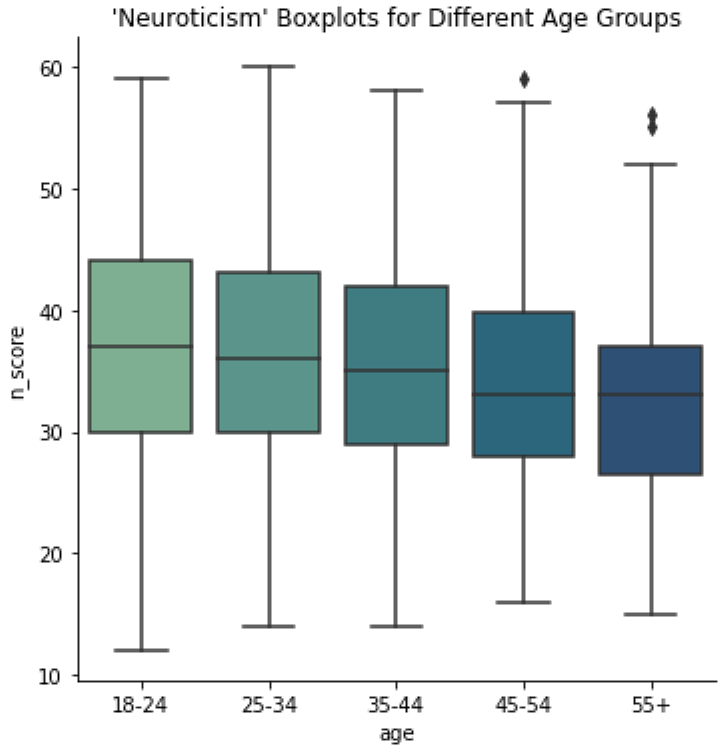
Personality Variables - Analysis

After finding no strong linear relations between any of our variables, I first decided to look closer into the personality variables, i.e. the numerical features. I first grouped them by gender and looked at their distributions. Throughout this report, I conducted 10 statistical tests in total, so using the Bonferroni Correction and a significant value of 0.05, the corrected p-value going forward will be 0.005. A frequentist t-test showed that there is a statistically significant difference in s-score (or sensation-seeking) between men and women ($p < 0.005$), in that men seem to have a higher score for this personality trait due to some underlying reason. A box and whisker plot for S-score for each gender is shown below.

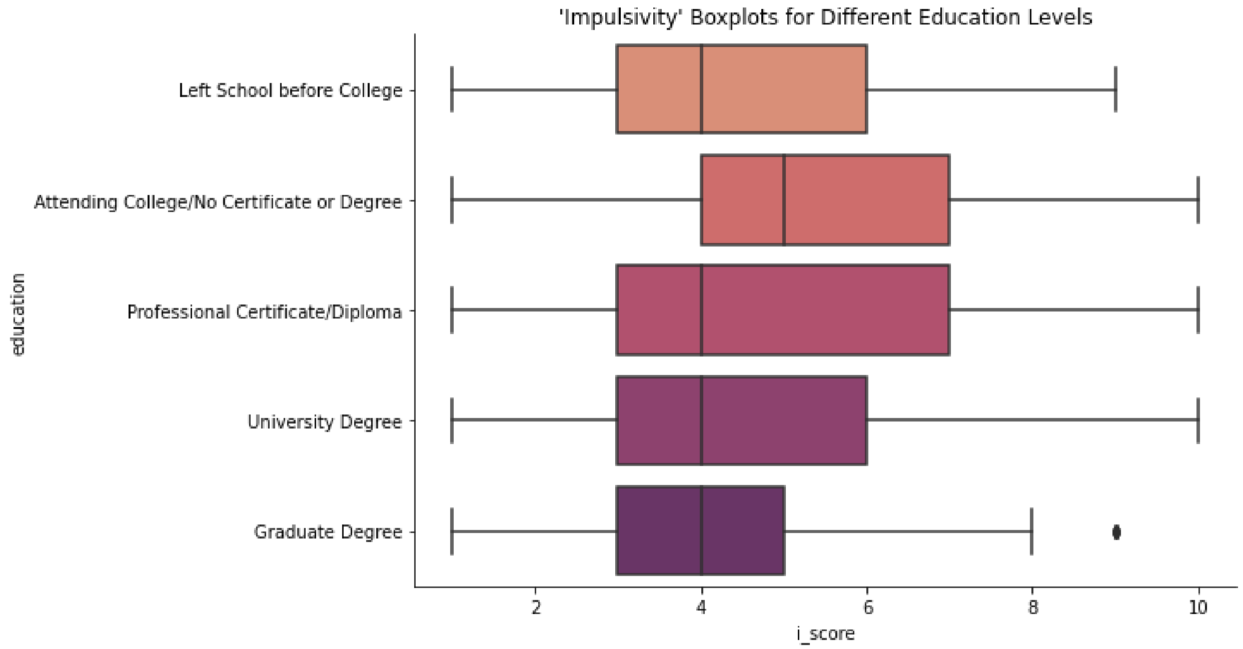


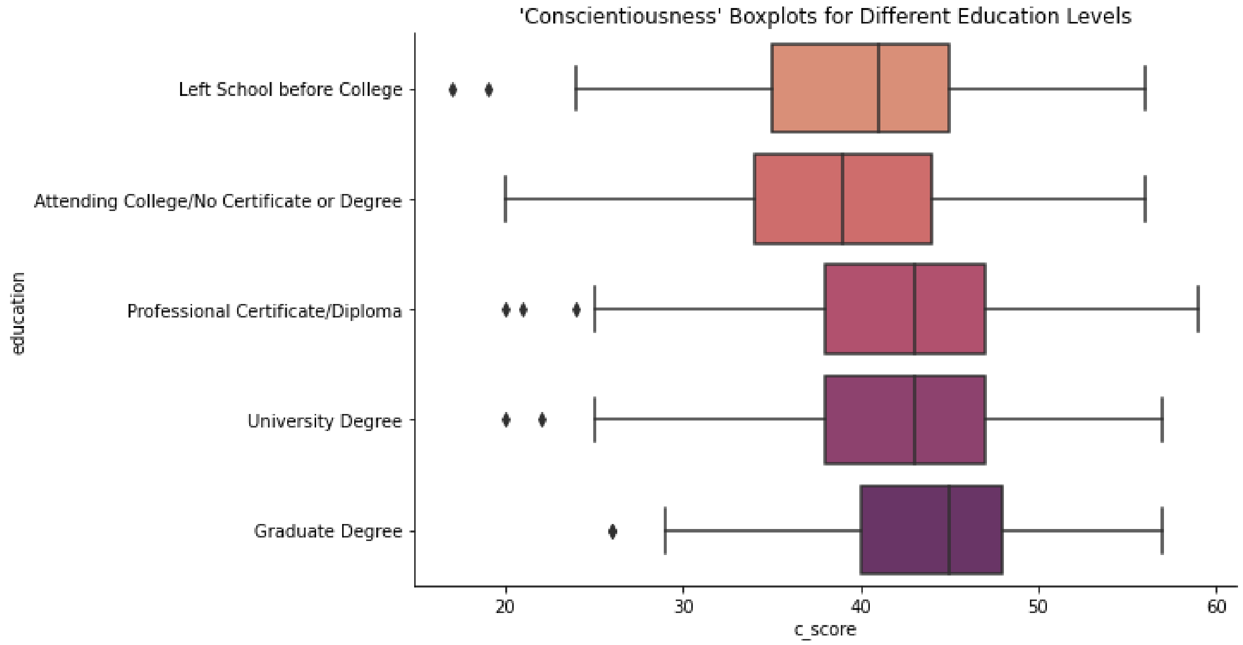
After conducting a few more t-tests, I found there to be a statistically significant difference in the means of impulsivity between genders (males were higher) as well as conscientiousness (females were higher).

I then grouped the data by age and did similar analyses on the personality variables. First, I saw that the oldest age group (65+) only had 18 respondents (less than 1% of the dataset). After finding similar summary statistics between this group and the 55-64 group, I decided to combine these two age groups to create one 55+ age group. I ran a few more t-tests to find that there were statistically significant differences in the means of neuroticism, sensation seeking, and impulsivity between the youngest age group (18–24) and the oldest age group (55+).



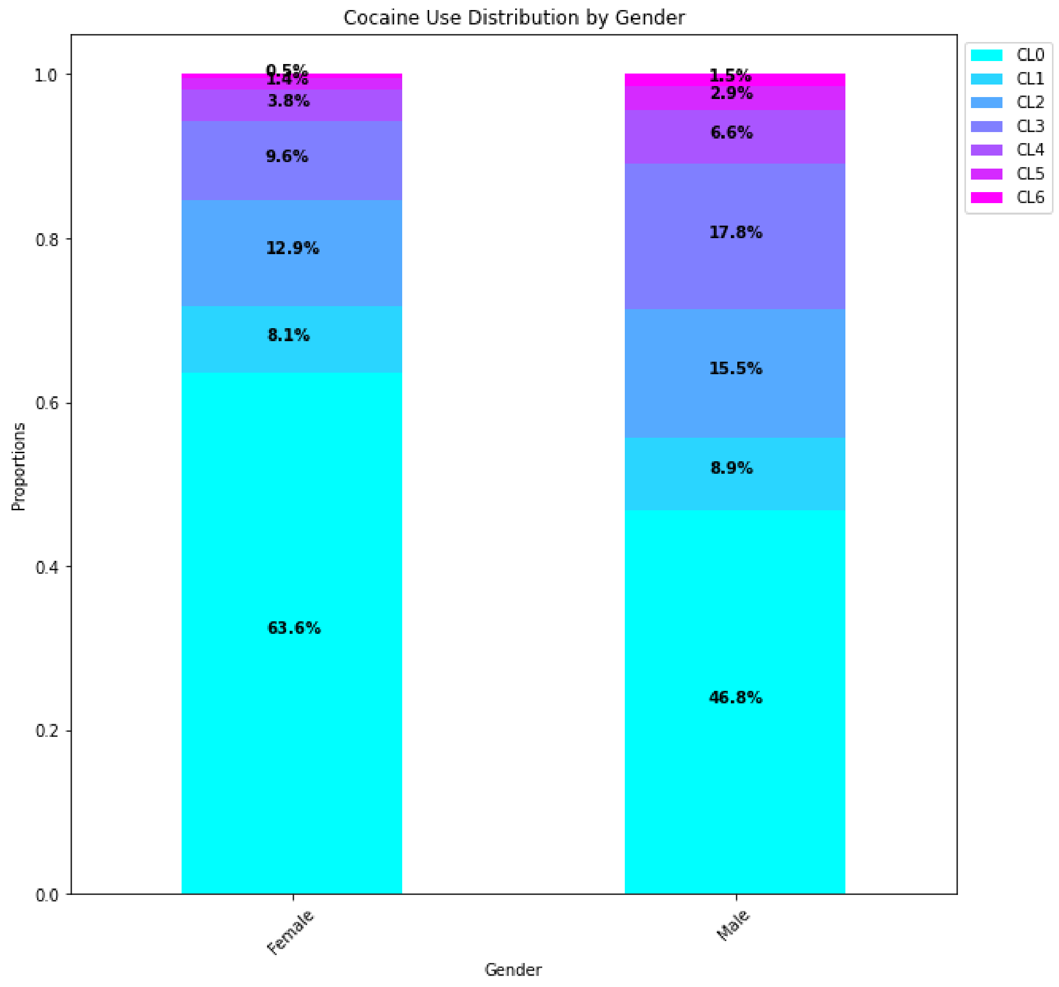
Lastly, I grouped the data by education groups and conducted similar analyses. Before this, however, I again saw very small numbers of respondents for several of the groups, and combined a few to create broader categories. Further t-tests showed that there is a significant difference in the means of openness to experience and impulsivity between those currently attending college and those who left school before college. I also found significant differences in agreeableness and conscientiousness between those who left school before college and those with a graduate degree.


'Impulsivity' Boxplots for Different Education Levels
Left School before College
Attending College/No Certificate or Degree
Professional Certificate/Diploma
University Degree
Graduate Degree
education
2
4
6
8
10
i_score

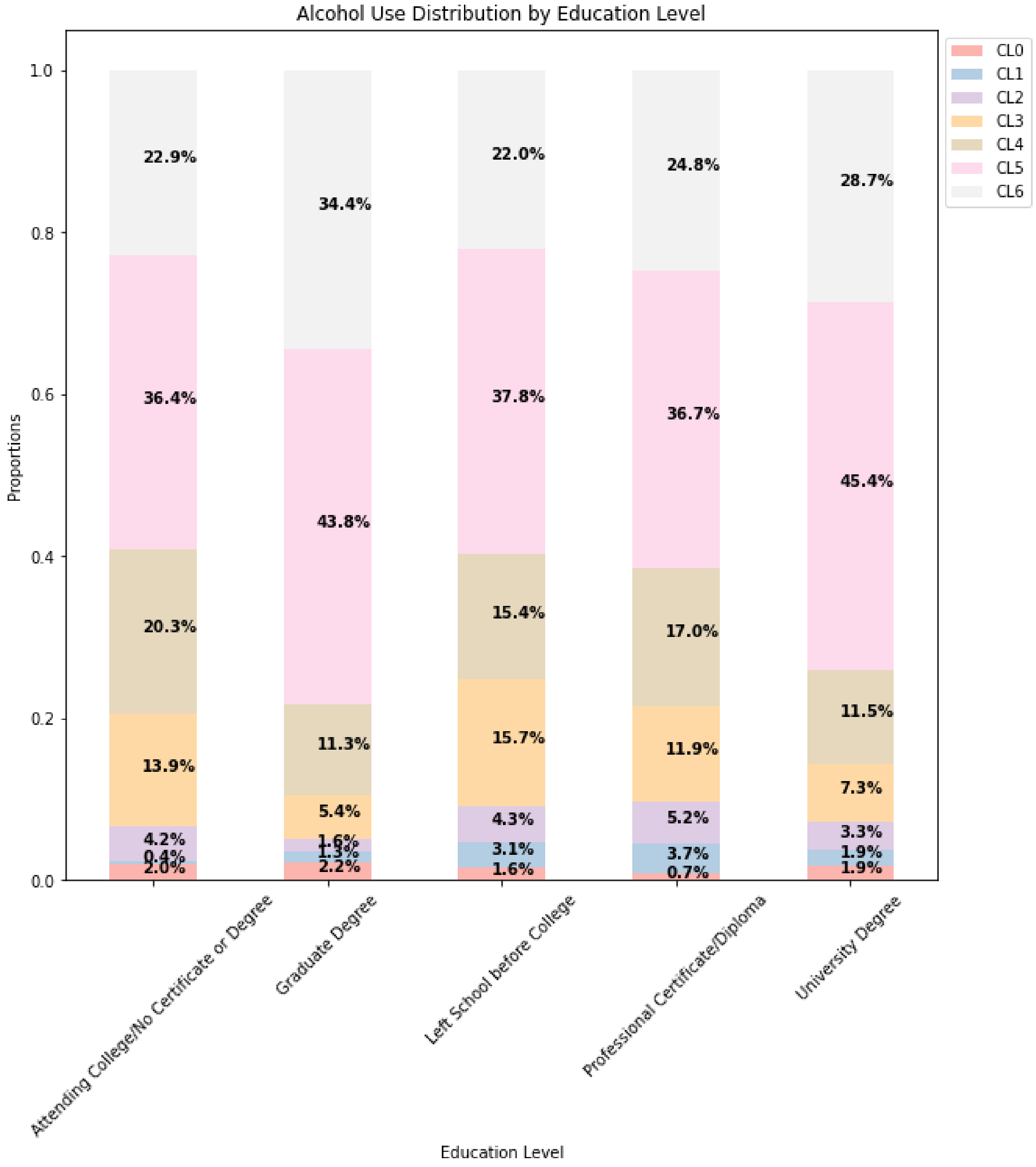

'Conscientiousness' Boxplots for Different Education Levels
Left School before College
Attending College/No Certificate or Degree
Professional Certificate/Diploma
University Degree
Graduate Degree
education
20
30
40
50
60
c_score

Drug Consumption Variables - Analysis

Next step was to explore the three drug use categories that we are interested in (alcohol, cocaine, and benzodiazepines). I found in our dataset that slightly more men consumed alcohol in the last day than women, while slightly more women consumed alcohol in the last week. While the total for these two categories were about the same for men and women, this may suggest that men are more likely to consume alcohol slightly (and possibly even abuse) alcohol more frequently than women. I found that there were far more women who never tried cocaine or benzos than men. With benzos, however, those who used in the last day were about the same between both genders, possibly due to the fact that it is a legal, prescription drug.

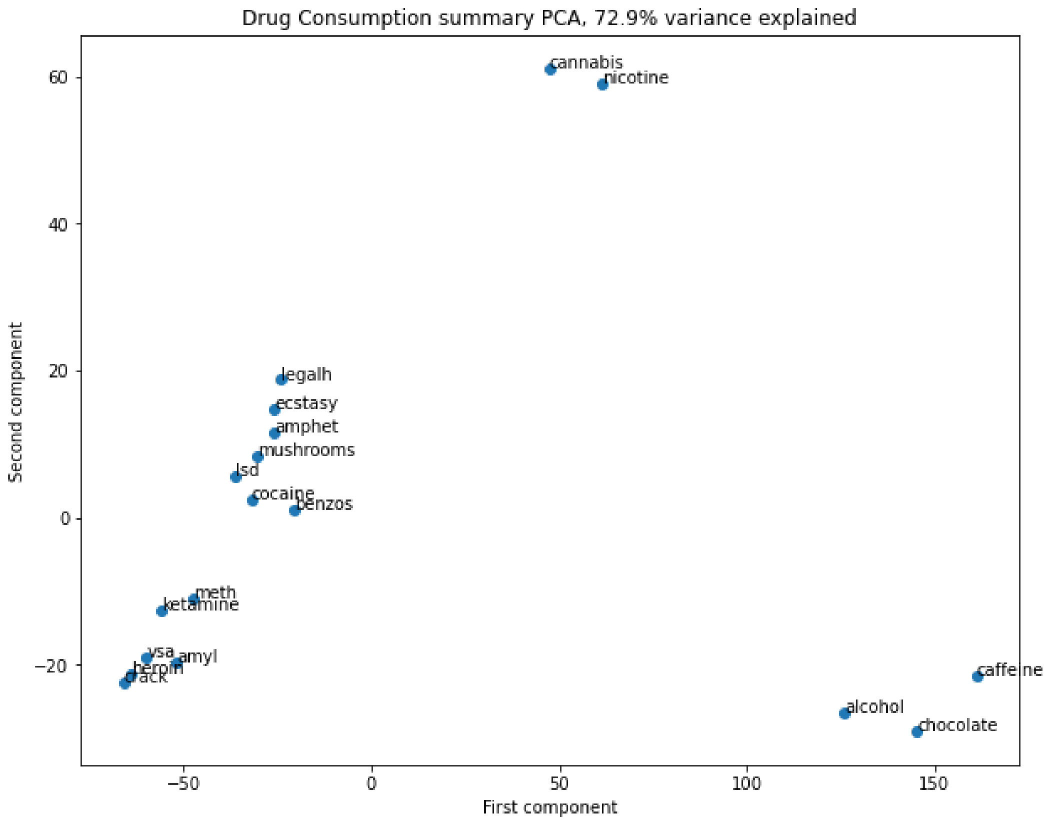


Grouping by education level, I found that those who hold a graduate degree tend to drink alcohol the most, followed closely by those who hold a university degree. Those attending college hold the largest percentage of past or present cocaine and benzo users.



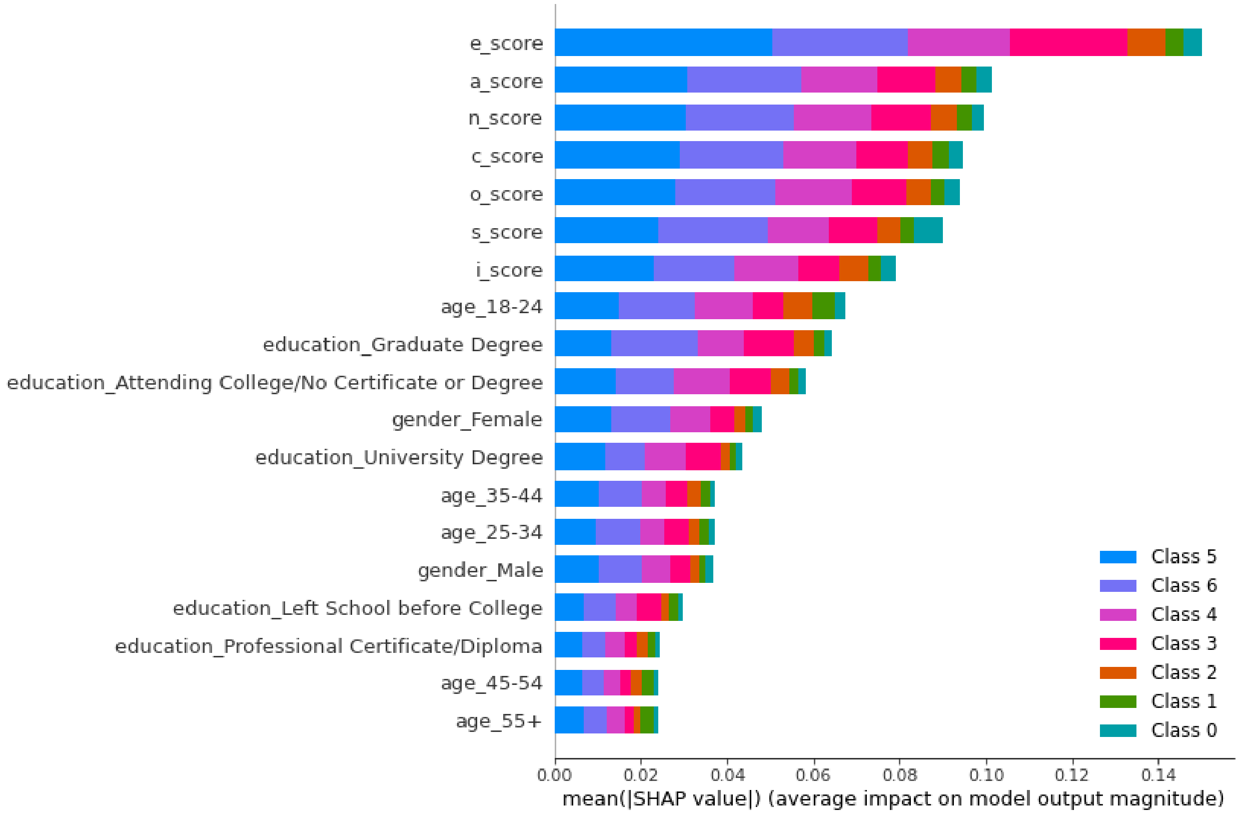
Principal Components Analysis

After conducting Principal Components Analysis, I found 4 major categories of drug users from 2 components explaining about 73 percent of the variance between the drug consumption variables. One for cannabis and nicotine, one for the perceived least dangerous substances and most widely accepted (alcohol, chocolate, caffeine), one for the party/college/prescription/hallucination drugs (LSD, ecstasy, cocaine, benzos, etc.), and one for the most dangerous and possibly most addicting drugs (heroin, crack, methadone, etc.). It is interesting to see alcohol and chocolate grouped so closely to each other, showing just how widely used alcohol is.

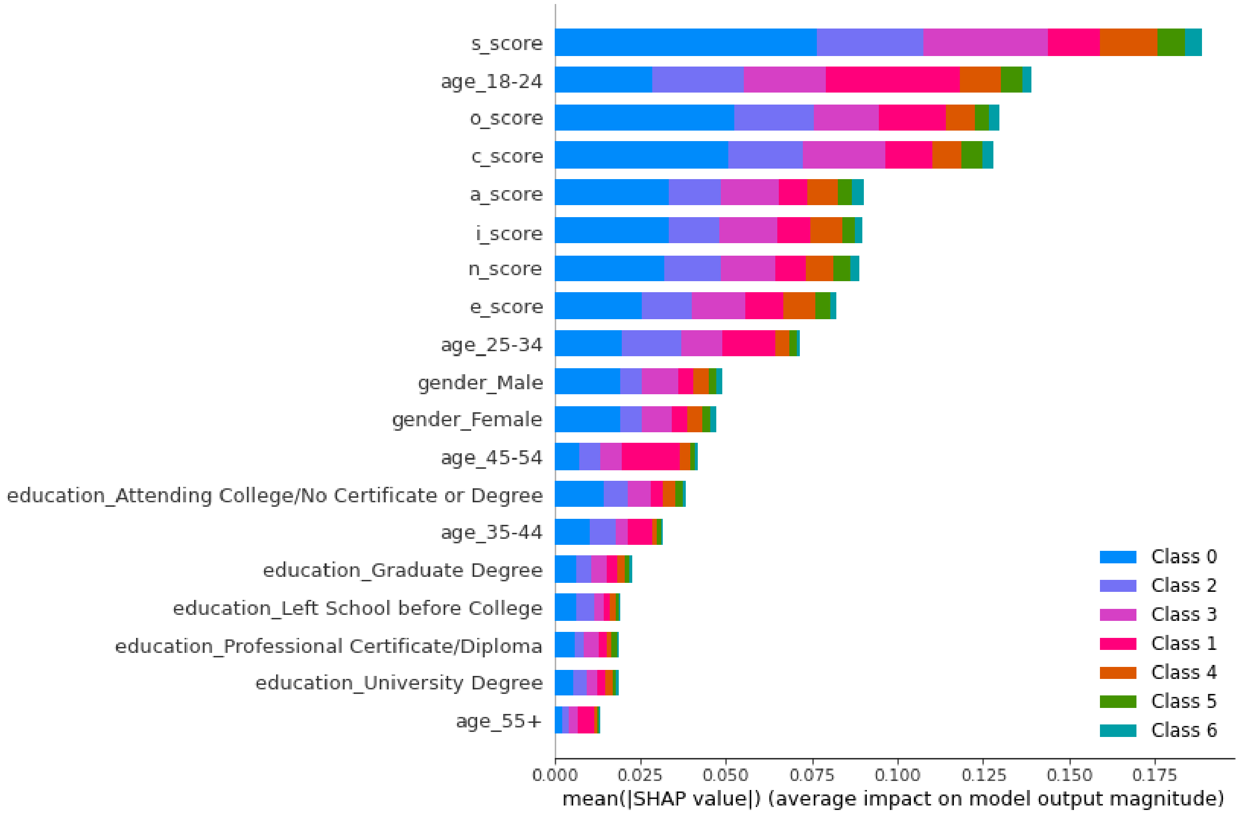


SHAP Feature Importances - Random Forest

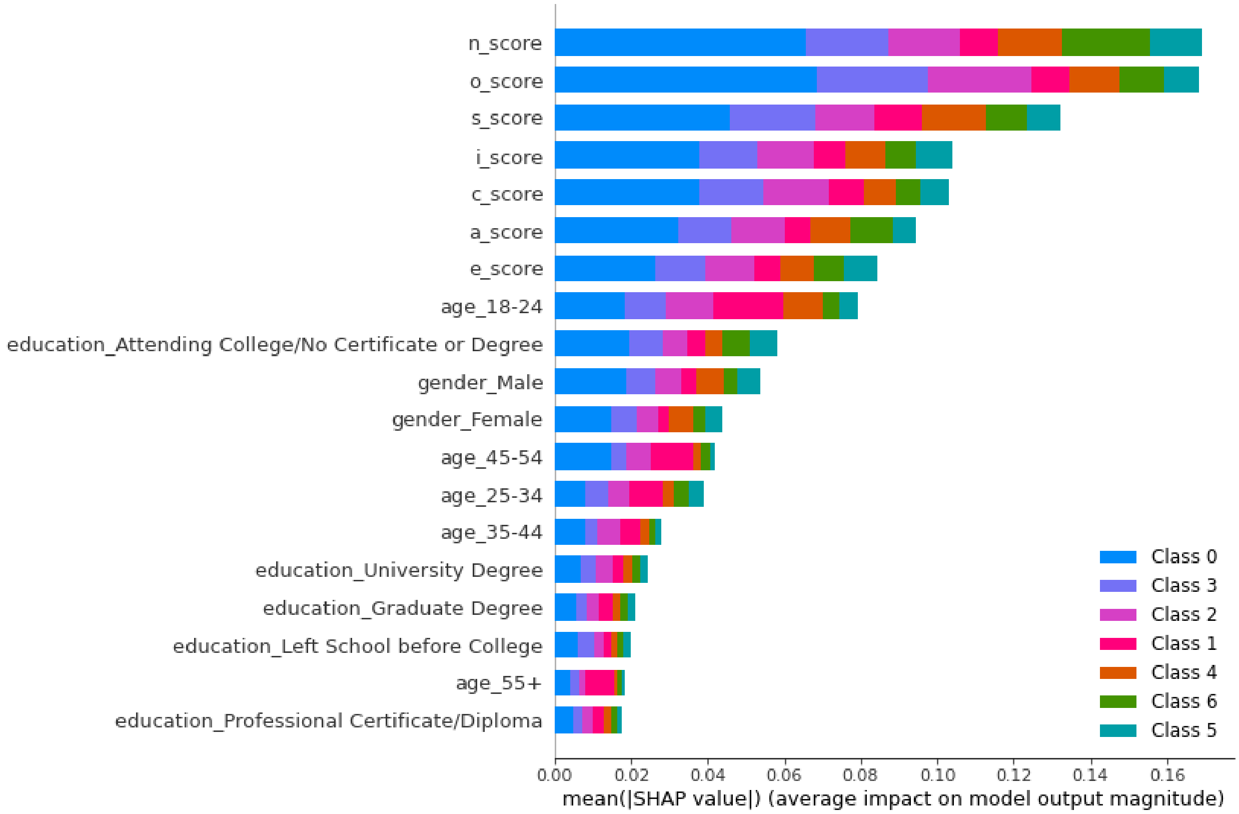
I next used a random forest classifier from scikit learn as well as the shap library to find feature importances for each of our drug categories of interest. For alcohol, openness to experience holds the most significance followed by agreeableness and neuroticism, while the younger age groups also hold some significance.



For cocaine use, we see sensation seeking holds the most significance, followed by the youngest age group, openness to experience, and conscientiousness.



For benzo use, the most significant personality variables are neuroticism, openness to experience, and sensation seeking.



For all three drugs/substances, the oldest age group holds the least significance. For cocaine and benzos, we see that either leaving school before college or holding some sort of degree holds little significance, while attending college holds much more significance. For alcohol, however, those attending college as well as those holding a university degree (undergraduate or graduate) holds more significance.

Machine Learning

Pre-Processing / Feature Selection

I began by removing the “Race” feature, as 90% of the respondents were white, so the data was not entirely representative. I also combined some categories for the other features as I did during the exploratory analysis. Again, the target variables values have the below format:

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

Since alcohol had such a high number of respondents who used in the last day, I labeled CL6 and CL5 as a 1 to represent “User”, and changed all other values to 0 to represent “Non-User”. For cocaine, however, I labeled CL6, CL5, CL4, and CL3 as “User” and the rest as “Non-User” as this seemed like a useful distinction between the two groups for this particular drug. For benzodiazepines, I just labeled CL6, CL5, and CL4 as “Users” and the rest as “Non-Users” since these are prescription drugs so people may be taking these drugs more frequently on doctors orders. Lastly, I converted all categorical features to dummy variables, dropping the largest category from each as this would contain repetitive information.

Train-test split

I split the source data set by setting aside 60% of data for training and the remaining 40% for testing (while stratifying the data for our target variables to ensure similar proportions of User vs non user in each set). I then standardized the personality variables in the training set by fitting a StandardScaler() object to this data and using this to transform the feature sets for both the training set and the test set.

Modeling

Modeling was done separately for each substance, in which I implemented logistic regression, random forest, and k-nearest neighbors models for each and determined the best model using ROC-AUC scores. The Receiver Operator Characteristic (ROC) curve is typically used to evaluate binary classification problems. It is defined as a probability curve that plots the True Positive Rate against the False Positive Rate at various threshold values to see how well the model can separate the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the classifier's ability to distinguish between classes and is used as a summary statistic of the ROC curve. We want a higher ROC-AUC score, as this would mean the model performed better at distinguishing between the two classes across all probability thresholds.

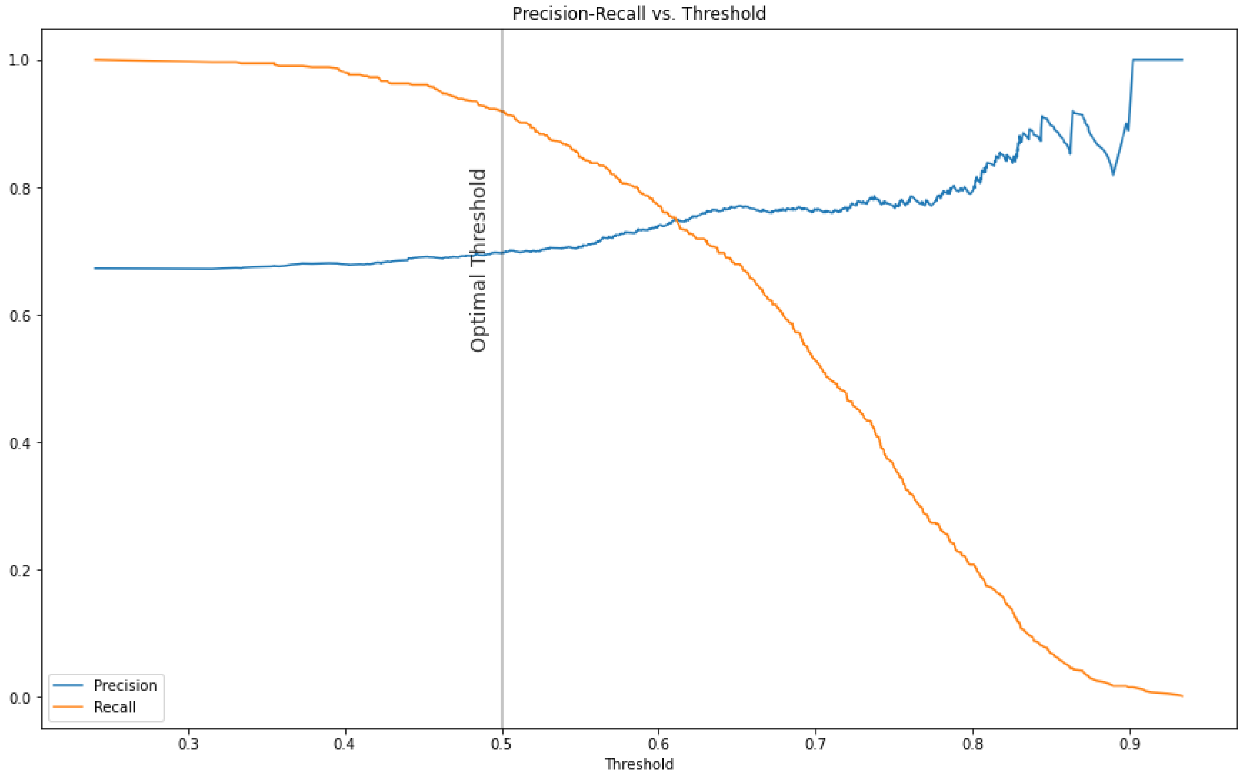
After determining the best model to proceed with for each substance, I then further looked at the performance of this model by finding the optimal threshold. In our business case, we would want to put more of an emphasis on maximizing recall while still maintaining a decent precision. This is because labeling a User as a non-user is more detrimental to the use of these models - we would rather be over-cautious! Once I determined the optimal threshold, I used that to produce confusion matrices and classification reports.

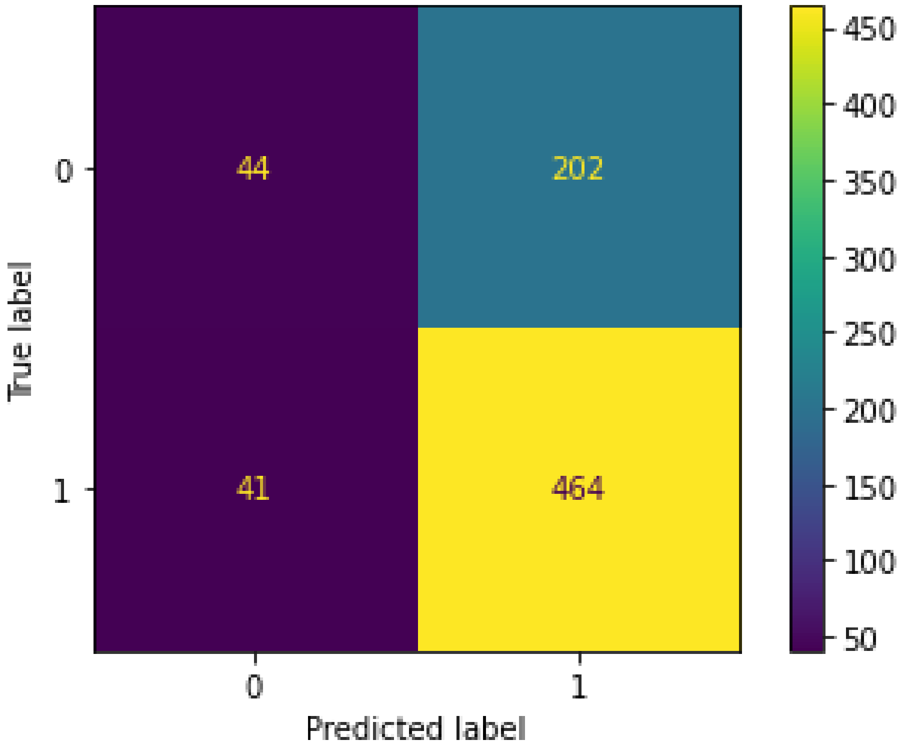
Alcohol:

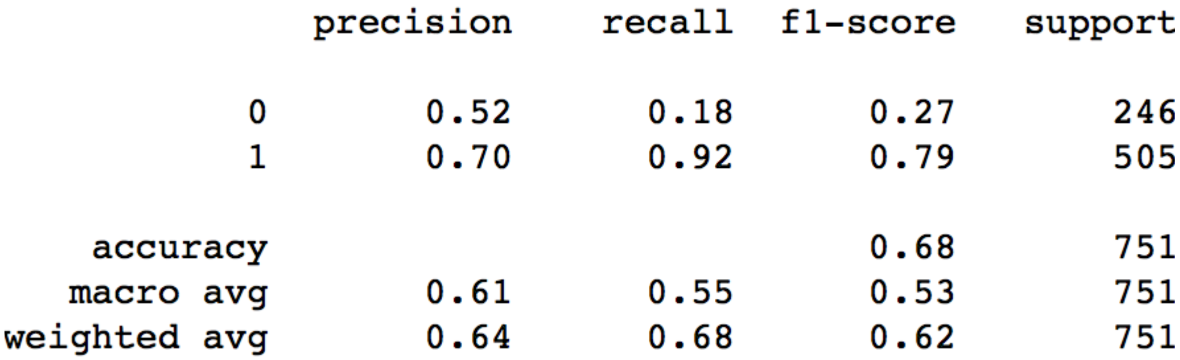
Our models for alcohol did not seem to perform too well in terms of ROCAUC scores. This is likely due to the fact that alcohol is so widely used and the users are so different from each other in terms of personality traits, ages, education levels, etc. Because of this, a model would have some trouble in trying to classify who is a user or not based on the limited set of features given in the data set. After tuning hyperparameters using cross validation and grid search, the below table shows the results for each model - logistic regression performed best.

Alcohol Classifier Performance		
Classifier	**ROCAUC Score**	**Best Hyperparameters**
Logistic Regression	0.643	‘C’: 100
Random Forest	0.598	'bootstrap': False, 'max_depth': 110, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 20
KNN	0.574	'n_neighbors': 47

To maintain at least a precision of 0.7, I chose an optimal threshold of 0.5, at which the below confusion matrix and classification reports were yielded.







```
              precision    recall  f1-score   support

           0       0.52      0.18      0.27       246
           1       0.70      0.92      0.79       505

    accuracy                           0.68       751
   macro avg       0.61      0.55      0.53       751
weighted avg       0.64      0.68      0.62       751
```

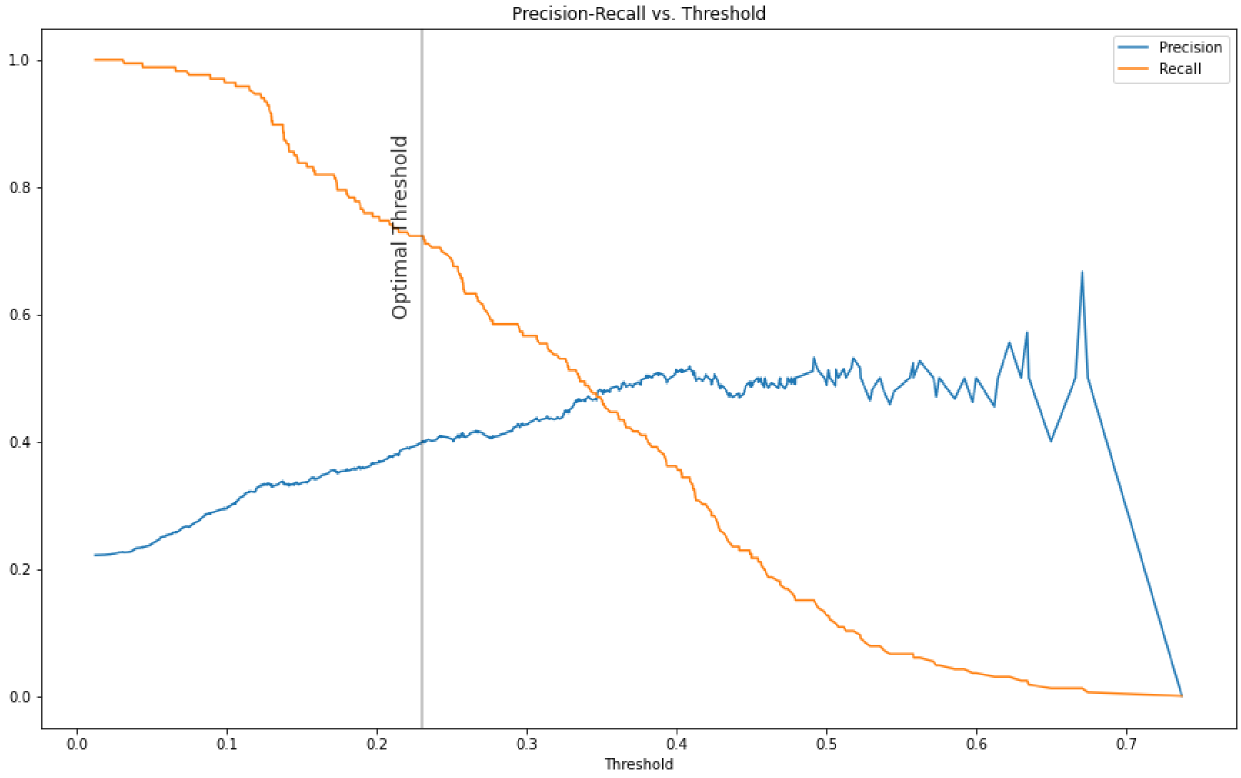
We can see that our model produced a very high number of false positives, and a much smaller number of false negatives. This is by design, as we would rather optimize the recall over the precision - there is more harm in falsely labeling someone as a potential non-user than a potential user.

Cocaine:

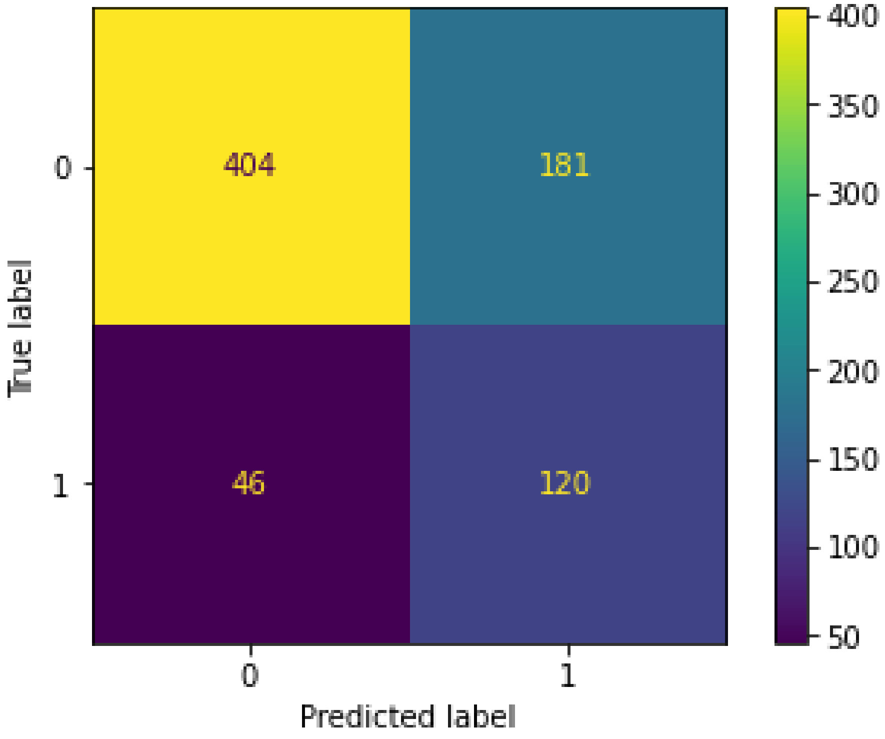
Models for cocaine use performed better than alcohol. After using grid search to tune parameters and find the best classifier, logistic reg. was again the top performer.

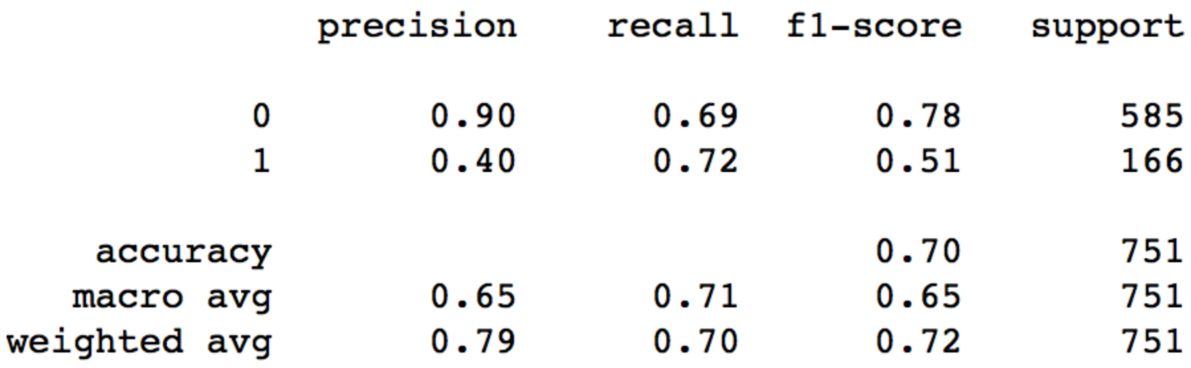
Cocaine Classifier Performance		
Classifier	**ROCAUC Score**	**Best Hyperparameters**
Logistic Regression	0.769	'C': 0.1
Random Forest	0.753	'bootstrap': True, 'max_depth': 110, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 6, 'n_estimators': 100
KNN	0.737	'n_neighbors': 10

Next step was to find the optimal threshold by maintaining a precision of at least 0.4, which was found to be 0.23 from the below graph.



At this optimal classification threshold, we can see that the recall was significantly higher than the precision for the model. This is again preferred as the purpose of the model is to inform one whether they are more susceptible to a certain substance abuse - cocaine, in this case. Since it is only giving a forewarning, or something to look out for, it is less harmful for the model to predict a false positive than a false negative. As we can see from the confusion matrix below, there were many non-user cases that were predicted as users.





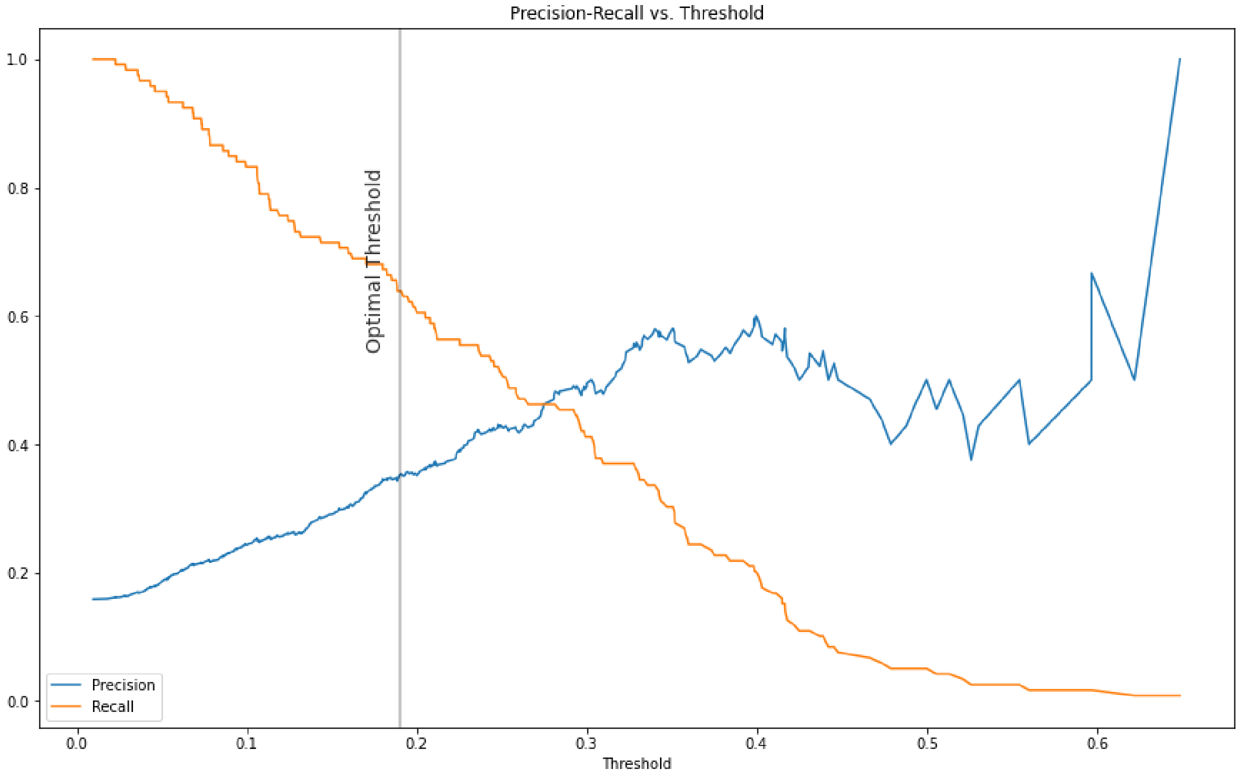
```
              precision    recall  f1-score   support

           0       0.90      0.69      0.78       585
           1       0.40      0.72      0.51       166

    accuracy                           0.70       751
   macro avg       0.65      0.71      0.65       751
weighted avg       0.79      0.70      0.72       751
```

Benzodiazepine:

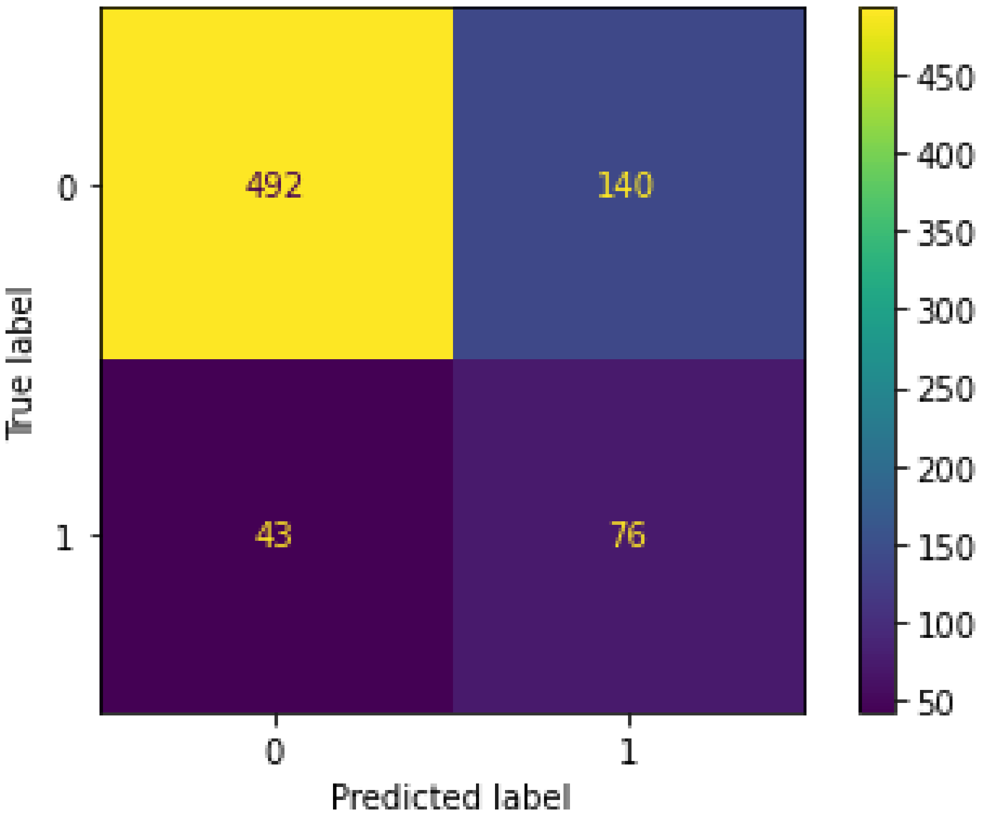
Lastly, out of our models for benzodiazepine logistic regression again performed the best.

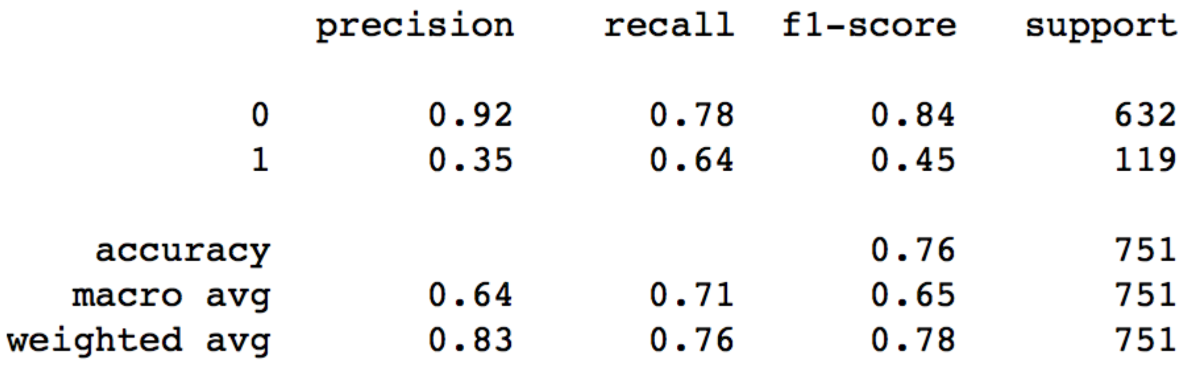
Benzodiazepine Classifier Performance		
Classifier	**ROCAUC Score**	**Best Hyperparameters**
Logistic Regression	0.766	‘C’: 0.1
Random Forest	0.733	'bootstrap': True, 'max_depth': 110, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100
KNN	0.746	'n_neighbors': 43

Maximizing recall while maintaining a precision of at least 0.35, we find the optimal threshold to be 0.19.



At the optimal threshold, recall is again given a greater emphasis for the same reasons as before.





```
              precision    recall  f1-score   support

           0       0.92      0.78      0.84       632
           1       0.35      0.64      0.45       119

    accuracy                           0.76       751
   macro avg       0.64      0.71      0.65       751
weighted avg       0.83      0.76      0.78       751
```

For all of these models, we can see that the precision was quite low. For this reason, it is important to understand that these models will be much more useful if they are used to predict the probabilities of being a User vs a Non-User, rather than simply classifying one as a User vs Non-User.

Conclusion

Some of the drawbacks from this data set includes lack of personality and lifestyle features describing each individual - including more information could be helpful to produce better performing models (especially for alcohol use). Another constraint may be the lack of descriptiveness of past drug use for each individual, such as how often/how many times one has taken a certain drug. The only descriptor for this is the recency that one took the drug. In the future, I would like to try out more classification models on the different substances such as Support Vector Machines or Naive Bayes classifiers. I would also like to create models for more of the substances listed in the original dataset. Another method I would also like to try is discerning between User and Non-User in a different way than I implemented in the preprocessing notebook.

Models like these can be very useful for mental health professionals and even education systems to determine if an individual is more at risk for a certain substance use/abuse. In doing so, they could recommend the individual (and their legal guardians if minors) to be extra cautious and take certain preventative measures such as after school programs, stronger emphasis on anti-drug awareness and education, etc. These models will be much more useful if they are used to create probabilities of susceptibility to the substance use rather than a strict binary classification of being a User or Non-user. Again, I would like to emphasize that these models should NOT be used as a direct indicator that one is certain to become a drug user without preventative measures, but rather be used as another tool in a mental health professional’s toolkit. These can certainly be useful in bringing awareness and heighten cautiousness early on - we saw from our dataset that so many of the users are of younger age groups. Efforts like these are crucial in strengthening our commitment to preventing the spells of addiction that plague so many communities and families.