CS 418 Final Project

Manasa Kandimalla, Shyam Patel, Carlos Antonio McNulty

December 3, 2019

1. Problem Selection

For our final project, we would like to solve the problem: which personality traits (e.g., neuroticism, extraversion, openness to experience, agreeableness, conscientiousness, impulsiveness and sensation) and other factors (e.g., age, gender and education) make one susceptible to the usage of various legal and illegal drugs. These drugs include alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse (VSA).

2. Data Collection

The data for our project consisted of a single dataset collected from the UCI Machine Learning Repository.

3. Data Preparation

The data provided came without a header, so for ease of use with the pandas library we added a header. To allow us to perform some analysis of the dataset we needed to convert the values in the dataset to something more interpretable. For instance the values for female and male were provided as the values -0.48246 and 0.48246. We converted these into categorial variables, at least for the duration of our analysis. We also had to label encode the drug usage responses and one-hot encode the gender category for later use with classifiers.

4. Data Exploration

The database contains records for 1,885 respondents. For each respondent, 12 attributes are known: measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity. Each respondent also provided their usage for 18 legal and illegal drugs, including the fictitious drug Semeron. The categories for drug usage consisted of: never used, used over a decade ago, used in the last decade, used in the last year, used in the last month, used in the last week, and used in the last day.

Figure 1: Drug Usage

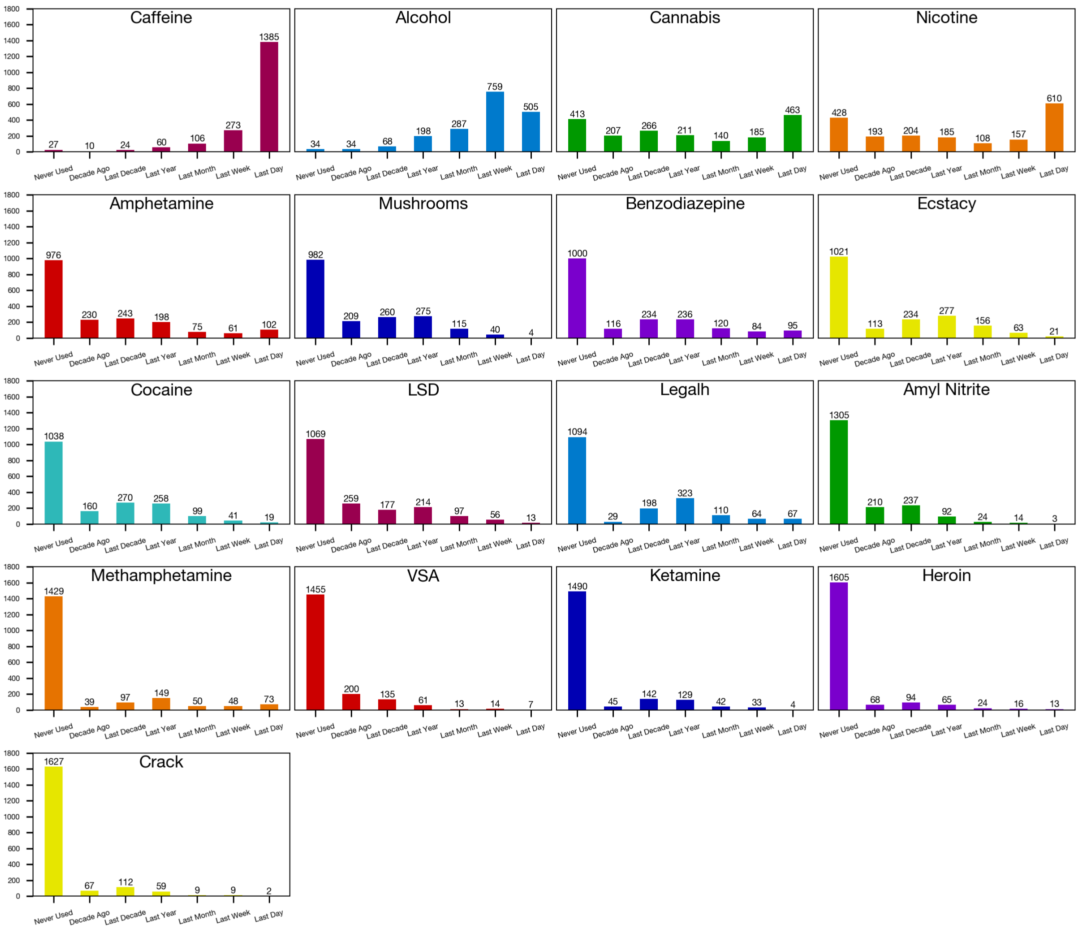


Figure 2: Illegal Drug Usage and Frequency by Gender

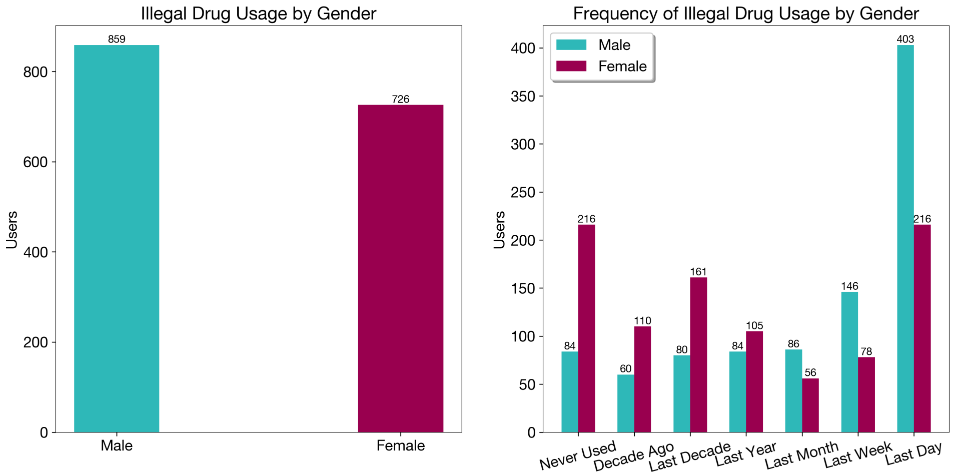


Figure 3: Illegal Drug Usage by Personality Trait

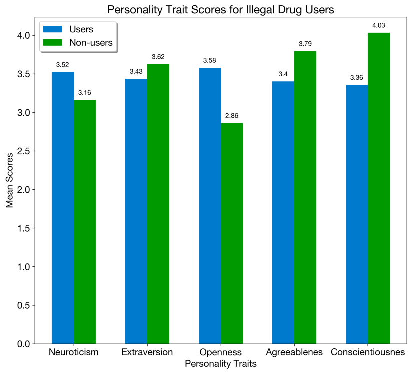
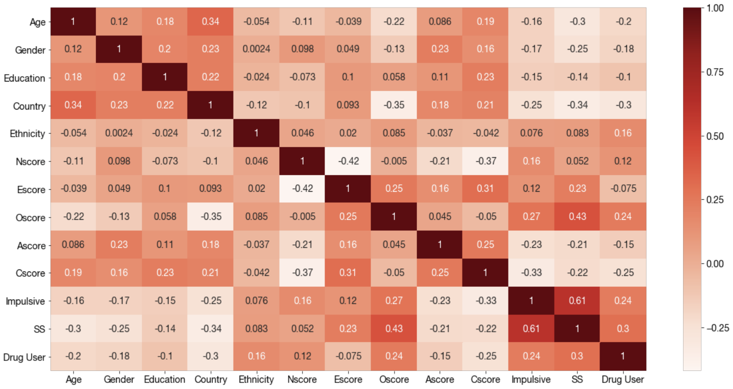


Figure 4: Heatmap of Correlation Matrix



5. Data Modeling

Decision Tree Classifier

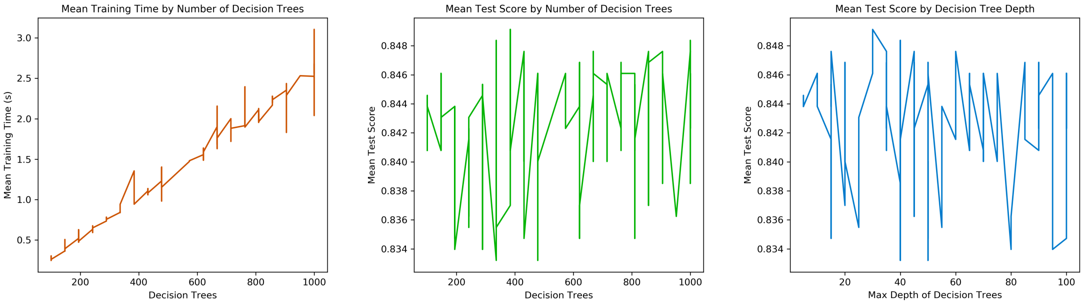
Random Forest Classifier

Feature Selection

To select the features for our binary classifier we consulted the pearson correlation matrix in Figure 4. In addition a feature selection tool from sklearn was used that determines a models features importances and removes those that fall below a certain threshold. Our analysis indicated that the features that might be best for our model were Age, Education, Country, Nscore, Escore, Oscore, Ascore, Cscore, Impulsive, and SS.

Grid Search

Figure 5: Hyperparameter Tuning with Grid Search



We performed grid search on 30,240 random forests by varying the number of estimators and the configuration for the base estimators provided to each random forest. The exhaustive search then chose the best performing model when tested with 10-fold cross validation. This model had an F1 score of 0.844 and an accuracy of 0.913. When the model was applied to the test set it resulted in an accuracy of 0.827.

Clustering

6. Results