Sahana Dhar & Madelyn Redick

# The Five Factor Model of Personality and its Relationship to Drug Habits

# Objectives

**Big Five**

| | Low scorers | High scorers |
|---|---|---|
| 1 Extraversion | Loner<br>Quiet<br>Passive<br>Reserved | Joiner<br>Talkative<br>Active<br>Affectionate |
| 2 Agreeableness | Suspicious<br>Critical<br>Ruthless<br>Irritable | Trusting<br>Lenient<br>Soft-hearted<br>Good-natured |
| 3 Conscientiousness | Negligent<br>Lazy<br>Disorganized<br>Late | Conscientious<br>Hard-working<br>Well-organized<br>Punctual |
| 4 Neuroticism | Calm<br>Even-tempered<br>Comfortable<br>Unemotional | Worried<br>Temperamental<br>Self-conscious<br>Emotional |
| 5 Openness to experience | Down-to-earth<br>Uncreative<br>Conventional<br>Uncurious | Imaginative<br>Creative<br>Original<br>Curious |

Using a dataset that consists of individuals' self-appointed scores on the 5 factors of personality, we analyzed their drug habits to distinguish significant relationships and create a fitting predictive model from these correlations. We employed the use of several classification models to look at individuals with higher frequencies of different drug habits, then examined their personality traits and the importance of these features when predicting drug use. Amongst the models tested were KNN and Logistic Regression.

## Main Questions

- Which drugs, and their frequency of use, have strong relationships with an individual's personality type?
- Which personality traits most influence one's drug habits?
- If fitting, could the model could potentially help identify individuals that may be at the greatest risk of unhealthy drug habits?

# Dataset

## FFM Traits

The 5-factor model of personality focuses on 5 traits:

- Nscore: neuroticism
- Escore: extraversion
- Oscore: openness to experience
- Ascore: agreeableness
- Cscore: conscientiousness

These traits were ranked on a continuous scale containing positive integers. The extremes of the score ranges for each trait have opposite identifiable personality characteristics.

## Drugs

There were 18 drugs recorded in this dataset: alcohol, amphetamine, amyl, benzos, caffeine, cannabis, chocolate, cocaine, crack, ecstasy, heroin, ketamine, legalh, LSD, methamphetamine, mushrooms, nicotine, and VSA. Participants rated their frequency of use on a scale of C0-C6. (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). For our model purposes, we primarily focused on individuals with scores of CL4-CL6, labelling these the unhealthy habits.

## Source

The data collectors and authors of the original study are E. Fehrman, A.K. Muhammad, E.M. Mirkes, V. Egan, and A. N. Gorban.

The dataset was obtained from UC Irvine's Machine Learning Repository.

| | ID | age | sex | education | country | ethinicity | Nscore | Escore | Oscore | Ascore | ... | ecstasy | heroin | ketamine | legalh | LSD | meth | mushrooms | nicotine | semer | VSA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0.49788 | 0.48246 | -0.05921 | 0.96082 | 0.12600 | 0.31287 | -0.57545 | -0.58331 | -0.91699 | ... | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL2 | CL0 | CL0 |
| 1 | 2 | -0.07854 | -0.48246 | 1.98437 | 0.96082 | -0.31685 | -0.67825 | 1.93886 | 1.43533 | 0.76096 | ... | CL4 | CL0 | CL2 | CL0 | CL2 | CL3 | CL0 | CL4 | CL0 | CL0 |
| 2 | 3 | 0.49788 | -0.48246 | -0.05921 | 0.96082 | -0.31685 | -0.46725 | 0.80523 | -0.84732 | -1.62090 | ... | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL1 | CL0 | CL0 | CL0 |
| 3 | 4 | -0.95197 | 0.48246 | 1.16365 | 0.96082 | -0.31685 | -0.14882 | -0.80615 | -0.01928 | 0.59042 | ... | CL0 | CL0 | CL2 | CL0 | CL0 | CL0 | CL0 | CL2 | CL0 | CL0 |
| 4 | 5 | 0.49788 | 0.48246 | 1.98437 | 0.96082 | -0.31685 | 0.73545 | -1.63340 | -0.45174 | -0.30172 | ... | CL1 | CL0 | CL0 | CL1 | CL0 | CL0 | CL2 | CL2 | CL0 | CL0 |
| 5 | 6 | 2.59171 | 0.48246 | -1.22751 | 0.24923 | -0.31685 | -0.67825 | -0.30033 | -1.55521 | 2.03972 | ... | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL6 | CL0 | CL0 |
| 6 | 7 | 1.09449 | -0.48246 | 1.16365 | -0.57009 | -0.31685 | -0.46725 | -1.09207 | -0.45174 | -0.30172 | ... | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL6 | CL0 | CL0 |
| 7 | 8 | 0.49788 | -0.48246 | -1.73790 | 0.96082 | -0.31685 | -1.32828 | 1.93886 | -0.84732 | -0.30172 | ... | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 |
| 8 | 9 | 0.49788 | 0.48246 | -0.05921 | 0.24923 | -0.31685 | 0.62967 | 2.57309 | -0.97631 | 0.76096 | ... | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL6 | CL0 | CL0 |
| 9 | 10 | 1.82213 | -0.48246 | 1.16365 | 0.96082 | -0.31685 | -0.24649 | 0.00332 | -1.42424 | 0.59042 | ... | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL0 | CL6 | CL0 | CL0 |

# Data Cleaning

- **Removed rows in which individuals claimed they used the drug 'semer'.**
  Semer is a fictitious drug made up by the data collectors in order to determine the honesty of each individual they interviewed.
  Therefore, if someone claimed to have used semer, we removed the rest of the information they contributed to make the dataset as authentic as possible.

- **Initially, we wanted to include demographics as a part of our ML analysis.**
  However, initial investigations into the data showed that over 91% of the participants in the dataset were white.
  Since this is not representative of all real population samples and cannot be applied to any individual, demographics were removed from our consideration.

- **FFM scores are all positive integers, but those present in the dataset were decimal and nonpositive.**
  Web scraping was done from the data source in order to collect the actual.
  The actual values are the most useful when analyzing the meaning of the results, but the decimal values are useful for ML and visualizations because they clearly represent the extremes of each FFM attribute.

- **In order to make graphing and calculations easier, the 'CL' from each drug use frequency value was dropped.**
  These values were still treated as the categorical values they represent

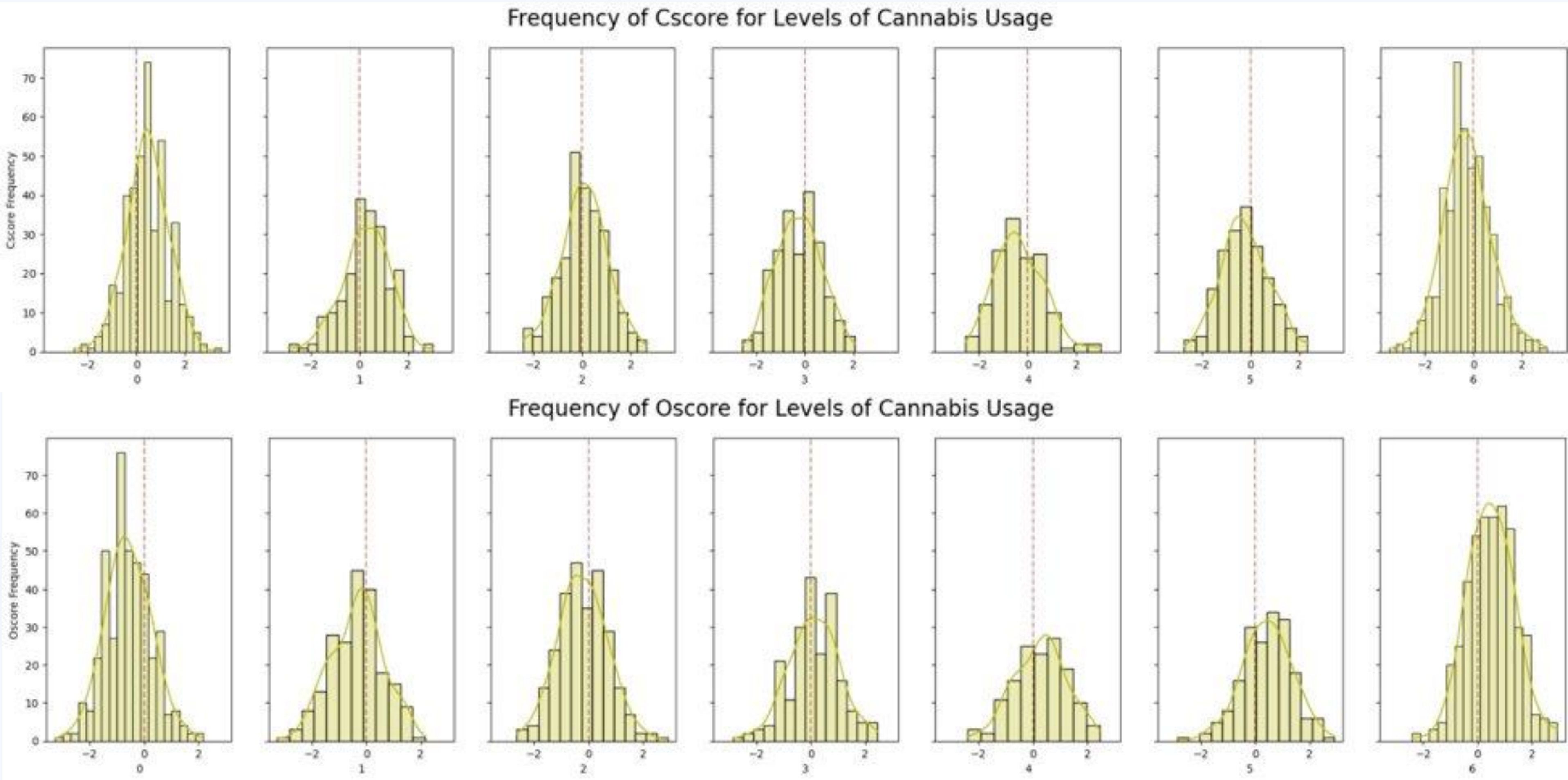
| | ID | Nscore | Escore | Oscore | Ascore | Cscore | alcohol | amphet | amyl | benzos | ... | crack | ecstasy | heroin | ketamine | legalh | LSD | meth | mushrooms | nicotine | VSA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0.31287 | -0.57545 | -0.58331 | -0.91699 | -0.00665 | 5 | 2 | 0 | 2 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 1 | 2 | -0.67825 | 1.93886 | 1.43533 | 0.76096 | -0.14277 | 5 | 2 | 2 | 0 | ... | 0 | 4 | 0 | 2 | 0 | 2 | 3 | 0 | 4 | 0 |
| 2 | 3 | -0.46725 | 0.80523 | -0.84732 | -1.62090 | -1.01450 | 6 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 3 | 4 | -0.14882 | -0.80615 | -0.01928 | 0.59042 | 0.58489 | 4 | 0 | 0 | 3 | ... | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 0 |
| 4 | 5 | 0.73545 | -1.63340 | -0.45174 | -0.30172 | 1.30612 | 4 | 1 | 1 | 0 | ... | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 2 | 2 | 0 |
| 5 | 6 | -0.67825 | -0.30033 | -1.55521 | 2.03972 | 1.63088 | 2 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 |
| 6 | 7 | -0.46725 | -1.09207 | -0.45174 | -0.30172 | 0.93949 | 6 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 |
| 7 | 8 | -1.32828 | 1.93886 | -0.84732 | -0.30172 | 1.63088 | 5 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 9 | 0.62967 | 2.57309 | -0.97631 | 0.76096 | 1.13407 | 4 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 |
| 9 | 10 | -0.24649 | 0.00332 | -1.42424 | 0.59042 | 0.12331 | 6 | 1 | 0 | 1 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 |

# Data Visualization

Even with the narrowed-down dataset, there was still a lot of data left to analyze. In order to compare FFM scores with drug use frequency, we used histograms for each FFM trait and each drug.

For each drug, 5 FFM trait histogram subplots were created where each subplot was a drug use frequency.

We were looking for curves that were not centered around 0 on the x-axis. This shows that a certain FFM score has a relationship with a drug use frequency for that drug.



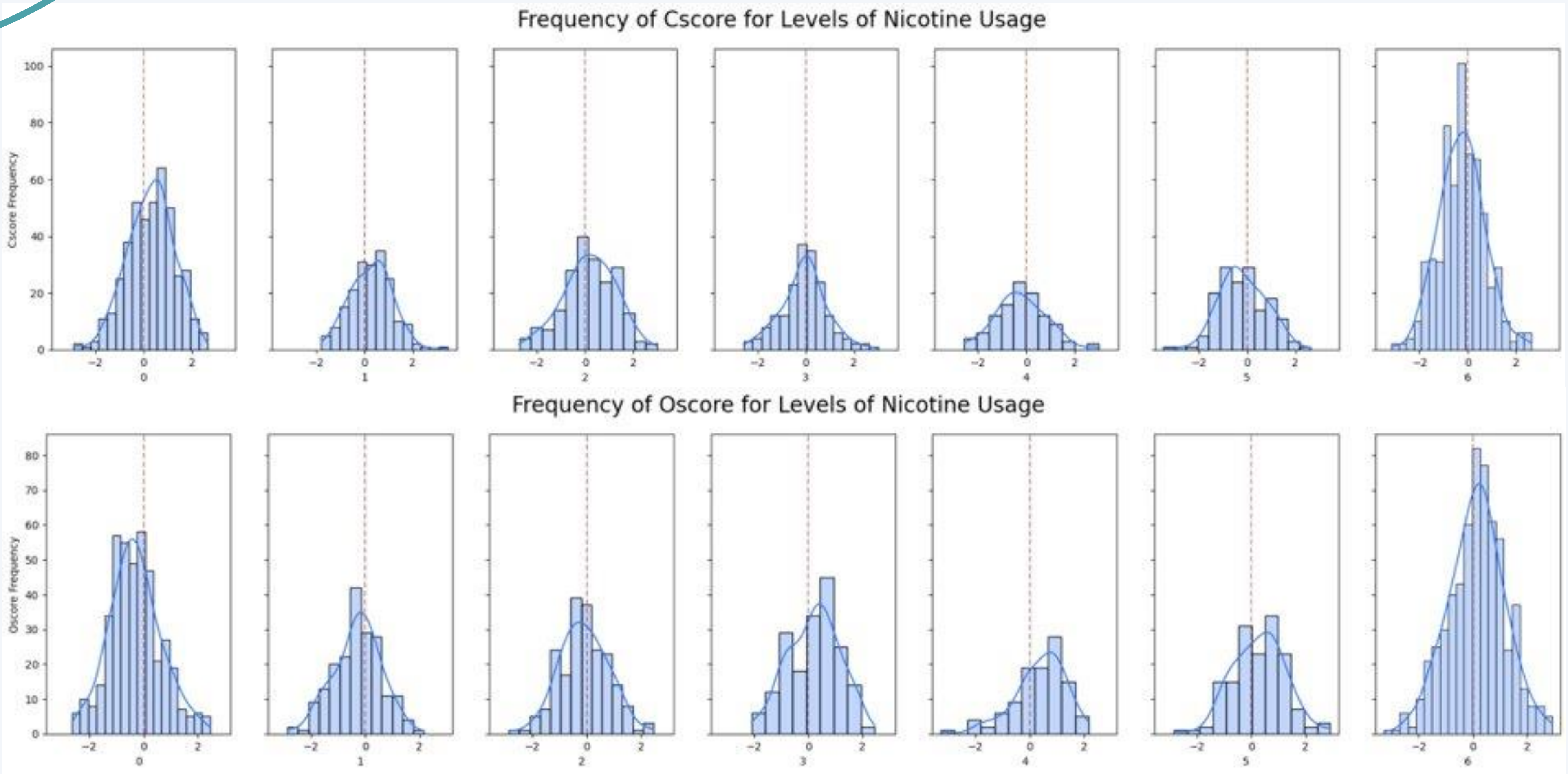

Both of these examples show that more people who do not use nicotine and/or cannabis have higher Cscores and lower Oscores.

High Cscores: hardworking, well-organized, punctual
Low Oscores: Down to earth, uncurious, conventional

More people who frequently use nicotine and/or cannabis have lower Cscores and higher Oscores.

Low Cscores: Negligent, lazy, disorganized
High Oscores: Imaginative, curious, creative

These descriptions fit the stereotypes often given to those who are frequent drug users and those who are not.



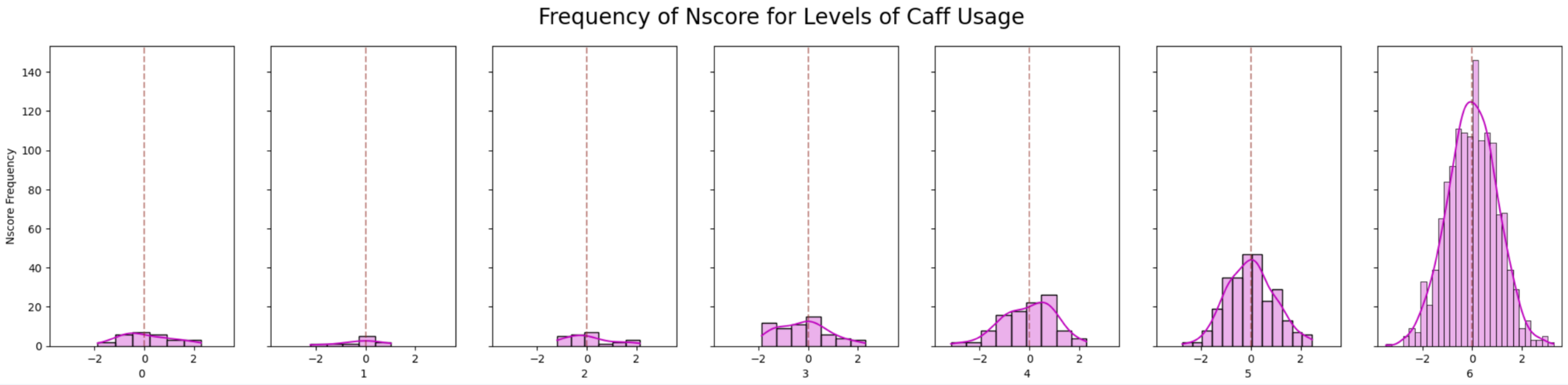
# Testing One Feature - Caffeine

In our initial exploration of the relationship between personality levels and drug frequency, we focused our attention on caffeine as a variable at first. Previous research (Bartol, 1975) has suggested a negative association between neuroticism and high caffeine use, so we employed a regression tactic to see if this connection stood within our dataset and if our model could capture it.

We utilized logistic regression and found a robust negative correlation coefficient of -0.42. This correlation, reflected in the regression coefficient, supports our hypothesis and strengthens our confidence in the logistic regression classifier method as a reliable tool for further exploring our dataset.

Notably, the histogram revealed a pattern:
as caffeine frequency increased, neuroticism levels exhibited a noticeable rightward skew. This observation also displayed the inverse relationship between neuroticism and high caffeine consumption.

| Feature Importance (Nscore) | Feature Importance (Ascore) | Feature Importance (Cscore) | Feature Importance (Escore) | Feature Importance (Oscore) |
|---|---|---|---|---|
| -0.426429 | -0.029863 | -0.132543 | -0.340332 | -0.034396 |



# Logistic Regression

## Terminology

*Accuracy:* The proportion of correctly classified instances out of the total instances.

*Precision:* The ratio of true positives to the sum of true positives and false positives.

*Recall:* The ratio of true positives to the sum of true positives and false negatives.

*F1 Score:* The weighted average of precision and recall.

| Column | Accuracy (%) | Precision (%) | Recall (%) | F1 Score (%) | Cross-Val Accuracy | Feature Importance (Nscore) | Feature Importance (Ascore) | Feature Importance (Cscore) | Feature Importance (Escore) | Feature Importance (Oscore) |
|---|---|---|---|---|---|---|---|---|---|---|
| alcohol | 44.290657 | 29.281356 | 44.290657 | 32.130783 | 0.430523 | -0.153767 | -0.050975 | -0.106647 | -0.404583 | -0.012827 |
| amphet | 47.222222 | 54.250356 | 47.222222 | 35.257822 | 0.451803 | -0.028086 | 0.217305 | 0.066953 | -0.089094 | 0.219954 |
| amyl | 70.454545 | 79.183884 | 70.454545 | 58.242424 | 0.684330 | 0.571651 | 0.027452 | 0.130931 | 0.417316 | 0.054807 |
| benzos | 41.242938 | 30.257779 | 41.242938 | 32.406453 | 0.441121 | -0.400080 | 0.139662 | 0.085490 | 0.049156 | 0.083627 |
| caff | 75.083056 | 81.291597 | 75.083056 | 64.397612 | 0.759321 | -0.361110 | 0.014686 | -0.130916 | -0.299941 | -0.024789 |
| cannabis | 44.545455 | 61.729798 | 44.545455 | 29.620793 | 0.472472 | 0.191660 | 0.120345 | 0.117593 | 0.049409 | -0.455823 |
| chocolate | 42.927632 | 51.916090 | 42.927632 | 36.980226 | 0.436957 | -0.027868 | 0.058177 | -0.103149 | -0.068489 | 0.084937 |
| cocaine | 59.420290 | 51.216545 | 59.420290 | 44.835310 | 0.618703 | -0.373940 | 0.005688 | -0.119886 | -0.240466 | 0.331891 |
| crack | 88.888889 | 89.173789 | 88.888889 | 87.145969 | 0.734167 | -0.436944 | 0.533338 | -0.205242 | 0.216591 | 0.248120 |
| ecstasy | 50.877193 | 75.149639 | 50.877193 | 34.941230 | 0.524122 | 0.159080 | 0.192427 | -0.016270 | 0.064720 | -0.089919 |
| heroin | 58.974359 | 67.428967 | 58.974359 | 45.213675 | 0.534058 | 0.093730 | 0.413470 | 0.267336 | 0.362100 | -0.096837 |
| ketamine | 59.420290 | 68.396739 | 59.420290 | 48.517520 | 0.610453 | 0.235514 | 0.182422 | 0.200832 | 0.273249 | -0.393276 |
| legalh | 59.358289 | 75.875776 | 59.358289 | 44.219933 | 0.569121 | 0.056286 | 0.089285 | 0.124177 | 0.126830 | -0.076825 |
| LSD | 50.000000 | 40.495392 | 50.000000 | 34.391534 | 0.557895 | -0.145134 | 0.123952 | 0.108512 | -0.183513 | -0.329110 |
| meth | 38.679245 | 55.668545 | 38.679245 | 24.628316 | 0.459375 | -0.079408 | 0.071716 | 0.048664 | -0.001679 | 0.258093 |
| mushrooms | 64.583333 | 62.796046 | 64.583333 | 56.849695 | 0.638279 | 0.339705 | -0.147813 | 0.252515 | 0.266559 | -0.292513 |
| nicotine | 55.714286 | 56.844044 | 55.714286 | 39.942227 | 0.575472 | -0.005718 | -0.036362 | 0.261517 | -0.222838 | 0.026616 |
| VSA | 59.375000 | 69.687500 | 59.375000 | 48.897059 | 0.610526 | 0.280426 | 0.526156 | 0.002683 | -0.235288 | -0.067272 |

## Interpretations & Analyses

The presented data frame provides a comprehensive assessment of predictive models for substance use across various drugs, encompassing metrics such as accuracy, precision, recall, and F1 score. Notably, certain drugs stand out with high accuracy scores, exemplifying the efficacy of the models in these cases. For instance, 'crack' exhibits an impressive accuracy of 88.89%, indicating a robust ability to correctly identify instances of crack cocaine use. Additionally, 'caff' (caffeine) demonstrates a notable accuracy of 75.08%, highlighting the model's proficiency in predicting caffeine consumption.

Examining feature importance scores sheds light on the personality traits that significantly influence substance use prediction for each drug. In 'amyl' (amyl nitrite), the feature importance values indicate that traits such as Nscore and Oscore play a substantial role in the model's ability to predict amyl nitrite usage. Conversely, for 'nicotine,' the influence of Cscore and Escore is apparent, underlining the relevance of conscientiousness and extraversion in predicting nicotine consumption.

# KNN, Decision Tree, Random Forest Classifier

The tables explore the performance of three other predictive models—Decision Tree, Random Forest, and K-Nearest Neighbors—across all drugs, providing a view of their efficacy through performance metrics and feature importance scores.
Observing the overall model suitability, it becomes evident that tree-based models, particularly Random Forest, frequently outperform K-Nearest Neighbors.

## Feature Importance

| | Drug | Model | Feature | Importance |
|---|---|---|---|---|
| 0 | alcohol | Decision Tree | Nscore | 0.276542 |
| 1 | alcohol | Decision Tree | Ascore | 0.178692 |
| 2 | alcohol | Decision Tree | Cscore | 0.189876 |
| 3 | alcohol | Decision Tree | Escore | 0.145966 |
| 4 | alcohol | Decision Tree | Oscore | 0.208924 |
| 5 | alcohol | Random Forest | Nscore | 0.210448 |
| 6 | alcohol | Random Forest | Ascore | 0.200145 |
| 7 | alcohol | Random Forest | Cscore | 0.205113 |
| 8 | alcohol | Random Forest | Escore | 0.188006 |
| 9 | alcohol | Random Forest | Oscore | 0.196288 |
| 10 | amphet | Decision Tree | Nscore | 0.225341 |

## Different Models and Accuracy for Each Drug

| | Drug | Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| 0 | alcohol | Decision Tree | 0.437768 | 0.390334 | 0.437768 | 0.402187 |
| 1 | alcohol | Random Forest | 0.444206 | 0.416341 | 0.444206 | 0.420072 |
| 2 | alcohol | K-Nearest Neighbors | 0.442060 | 0.420558 | 0.442060 | 0.425532 |
| 3 | amphet | Decision Tree | 0.208333 | 0.239672 | 0.208333 | 0.205688 |
| 4 | amphet | Random Forest | 0.291667 | 0.323133 | 0.291667 | 0.240393 |
| 5 | amphet | K-Nearest Neighbors | 0.375000 | 0.379808 | 0.375000 | 0.367835 |
| 6 | amyl | Decision Tree | 0.384615 | 0.769231 | 0.384615 | 0.448718 |
| 7 | amyl | Random Forest | 0.538462 | 0.807692 | 0.538462 | 0.579882 |
| 8 | amyl | K-Nearest Neighbors | 0.461538 | 0.415385 | 0.461538 | 0.437247 |
| 9 | benzos | Decision Tree | 0.355556 | 0.357392 | 0.355556 | 0.344101 |
| 10 | benzos | Random Forest | 0.422222 | 0.421178 | 0.422222 | 0.417457 |
| 11 | benzos | K-Nearest Neighbors | 0.377778 | 0.328591 | 0.377778 | 0.332689 |

# Future Improvements

## Demographics Consideration

Notably, our focus on personality scores rather than demographics in the initial analysis provided valuable insights into the psychological factors driving drug habits. Future investigations may benefit from incorporating demographic variables to enhance the comprehensiveness of the predictive models.

## Feature Improvement

The identification of Oscore as a consistently relevant and influential feature suggests a potential avenue for feature improvement. Fine-tuning models to better capture the differences of openness to experience could enhance the accuracy and specificity of predictions related to drug habits. Exploring additional personality traits and their interactions could also contribute to a more comprehensive understanding.

# Results

## Personality Trait Influence

Our analysis consistently revealed that personality traits, particularly openness to experience (measured by Oscore), played a pivotal role in predicting drug habits. Oscore demonstrated higher feature importance across various drugs, indicating a strong association between an individual's openness and their frequency of drug use.

## Predictive Accuracies

Caffeine, crack, and amyl emerged as standouts with higher predictive accuracies across all models. These substances displayed robust tendencies influenced by personality traits, showcasing the utilization of our classification models in uncovering nuanced patterns in drug habits.

## Model Comparison

The application of various classification models, including KNN and Logistic Regression, allowed for a comprehensive examination of the relationships between personality traits and drug habits. This diverse approach strengthened the robustness of our findings, providing an understanding of the complex interplay between individual characteristics and drug use.

## Objective Alignment

Our study successfully addressed the overarching objectives of identifying drugs with strong relationships to personality types and determining which personality traits exerted the most significant influence on drug habits. The results affirm the relevance of personality traits, particularly openness, in shaping an individual's proclivity for specific substances.