Drug Consumption

Nhat Bui, Johan Ferreira, Thilo Holstein

2025-03-06

Contents

1 Introduction

Drug use is a significant risk behavior with serious health consequences for individuals and society. Multiple factors contribute to initial drug use, including psychological, social, individual, environmental, and economic elements, as well as personality traits. While legal substances like sugar, alcohol, and tobacco cause more premature deaths, illegal recreational drugs still create substantial social and personal problems.

In this data science project, we aim to identify factors and patterns potentially explaining drug use behaviors through machine learning techniques. By analyzing demographic, psychological, and social variables in our dataset, we'll aim to uncover potential predictors using machine learning methods to understand the complex relationships surrounding drug consumption.

The database contains records for 1,885 respondents with 12 attributes including personality measurements (NEO-FFI-R, BIS-11, ImpSS), demographics (education, age, gender, country, ethnicity), and self-reported usage of 18 drugs plus one fictitious drug (Semeron). Drug use is classified into seven categories ranging from "Never Used" to "Used in Last Day." All input attributes are quantified as real values, creating 18 distinct classification problems corresponding to each drug. A detailed description of the variables can be found in the Column Decsription text file.

2 Personality Traits Explanation

To better understand the data set we need to have an understanding of what the personality traits are and what they represent, below we have short description of each trait and how to interpret them:

- Nscore (Neuroticism): Measures emotional stability vs. instability. Higher scores indicate tendency toward negative emotions like anxiety, depression, vulnerability and mood swings. Lower scores suggest emotional stability and resilience to stress.
- Escore (Extraversion): Measures sociability and outgoingness. Higher scores indicate preference for social interaction, assertiveness, and energy in social settings. Lower scores suggest preference for solitude, quieter environments and more reserved behavior.
- Oscore (Openness to Experience): Measures intellectual curiosity and creativity. Higher scores indicate imagination, appreciation for art/beauty, openness to new ideas, and unconventional thinking. Lower scores suggest preference for routine, practicality, and conventional approaches.
- Ascore (Agreeableness): Measures concern for social harmony. Higher scores indicate empathy, cooperation, and consideration for others. Lower scores suggest competitive, skeptical, or challenging interpersonal styles.
- Cscore (Conscientiousness): Measures organization and reliability. Higher scores indicate discipline, responsibility, planning, and detail orientation. Lower scores suggest spontaneity, flexibility, and potentially less structured approaches.
- Impulsive (Impulsiveness): Measures tendency to act without thinking. Higher scores indicate spontaneous decision-making without considering consequences. Lower scores suggest thoughtful deliberation before actions.
- SS (Sensation Seeking): Measures desire for novel experiences and willingness to take risks. Higher scores indicate thrill-seeking behavior and preference for excitement. Lower scores suggest preference for familiarity and safety.

The first five traits (Nscore through Cscore) are the "Big Five" personality traits, which are widely used in psychological research. The Impulsive and SS measures are additional traits that are often studied in relation to risk-taking behaviors, which makes sense given our dataset includes variables related to substance use.

3 Cleaning and Formatting the Dataset

3.1 Data Formatting

In its original state, the dataset represented most categorical variables with random floating-point numbers. We believe this was a measure to mitigate bias within the dataset. However, as our project's objectives differ from the dataset's initial purpose, we needed to revert these encoded values back to their original categorical representations. This step was essential to perform the analyses required for our project. This was the first step in cleaning our dataset.

3.2 Investigating Missing Values

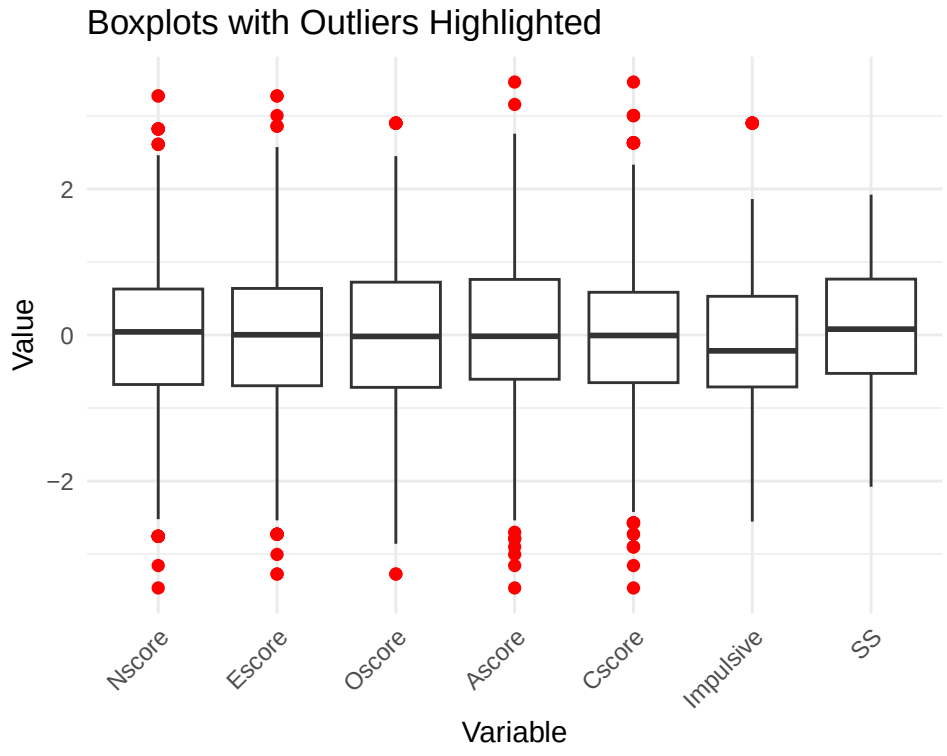
Table 1: Missing Values by Column

	Column	Missing Values	Percentage (%)
Education	Education	99	5.25
Ethnicity	Ethnicity	83	4.40

Note: Only columns with missing values are shown.

In the second step, we addressed missing values. We found that only two columns contained missing data, affecting approximately 5% of the 1885 observations. Considering the nature of these variables and the completeness of the remaining data, we inferred that participants likely withheld this information deliberately in most instances. Consequently, we replaced these missing values with the label "Not Provided," enabling us to treat these cases as a distinct category in our analysis.

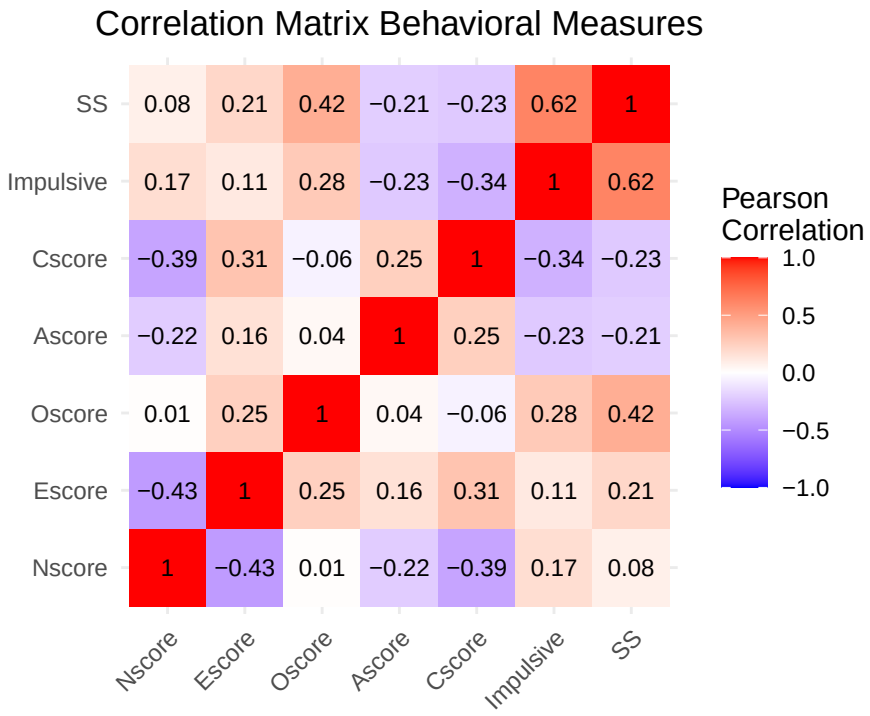
3.3 Investigating Outliers



The box plots generated for the seven psychometric personality scores reveal some data points that lie beyond the conventional 1.5xIQR (Interquartile Range) whiskers, technically identifying them as outliers. After invetigating the outliers we establised that outliers is not extreme in nature and fall within a plausible range, as well as being infrequent. Critically, their presence does not appear to significantly distort the overall distributional characteristics of these personality measures, which is important for subsequent analyses. The general cleanliness of the dataset, including the limited impact of these outliers, was better than anticipated, leading us to suspect that it may have undergone some form of pre-processing or curation before we accessed it.

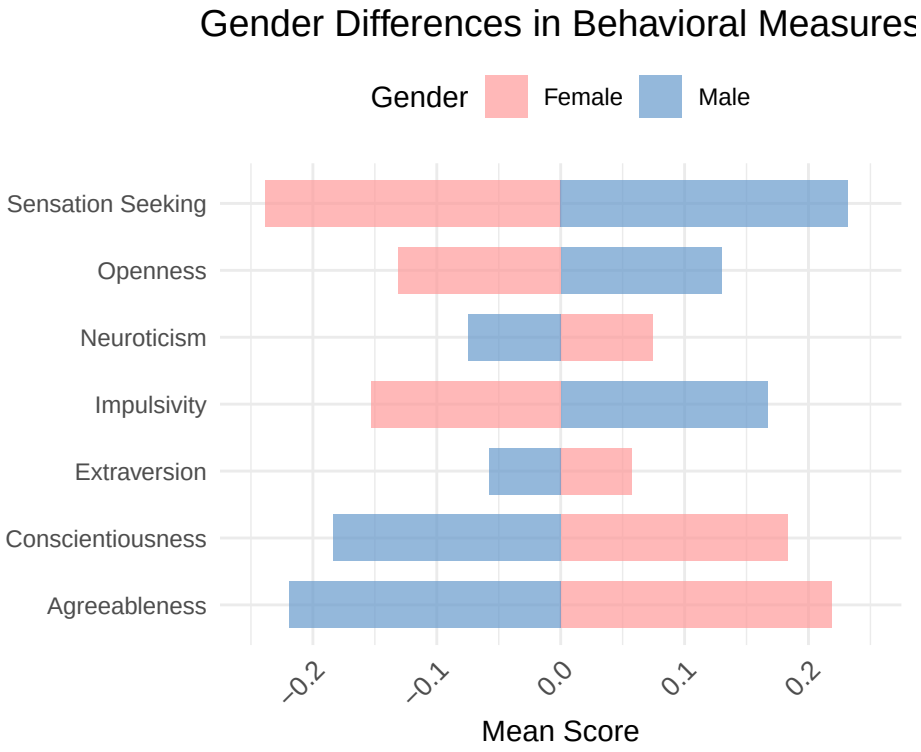
4 Exploratory Data Analysis

4.1 Correlation between Behavioral Measures



The correlation matrix reveals that certain personality traits tend to cluster. For instance, Sensation Seeking (SS) shows a positive correlation with Extraversion (Escore), Openness (Oscore), and Impulsiveness. These three traits (Extraversion, Openness, and Impulsiveness) are also positively correlated with each other. Conversely, Sensation Seeking (along with Extraversion, Openness, and Impulsiveness) exhibits a negative correlation with Conscientiousness (Cscore) and Agreeableness (Ascore). Finally, Conscientiousness and Agreeableness demonstrate a positive correlation with each other.

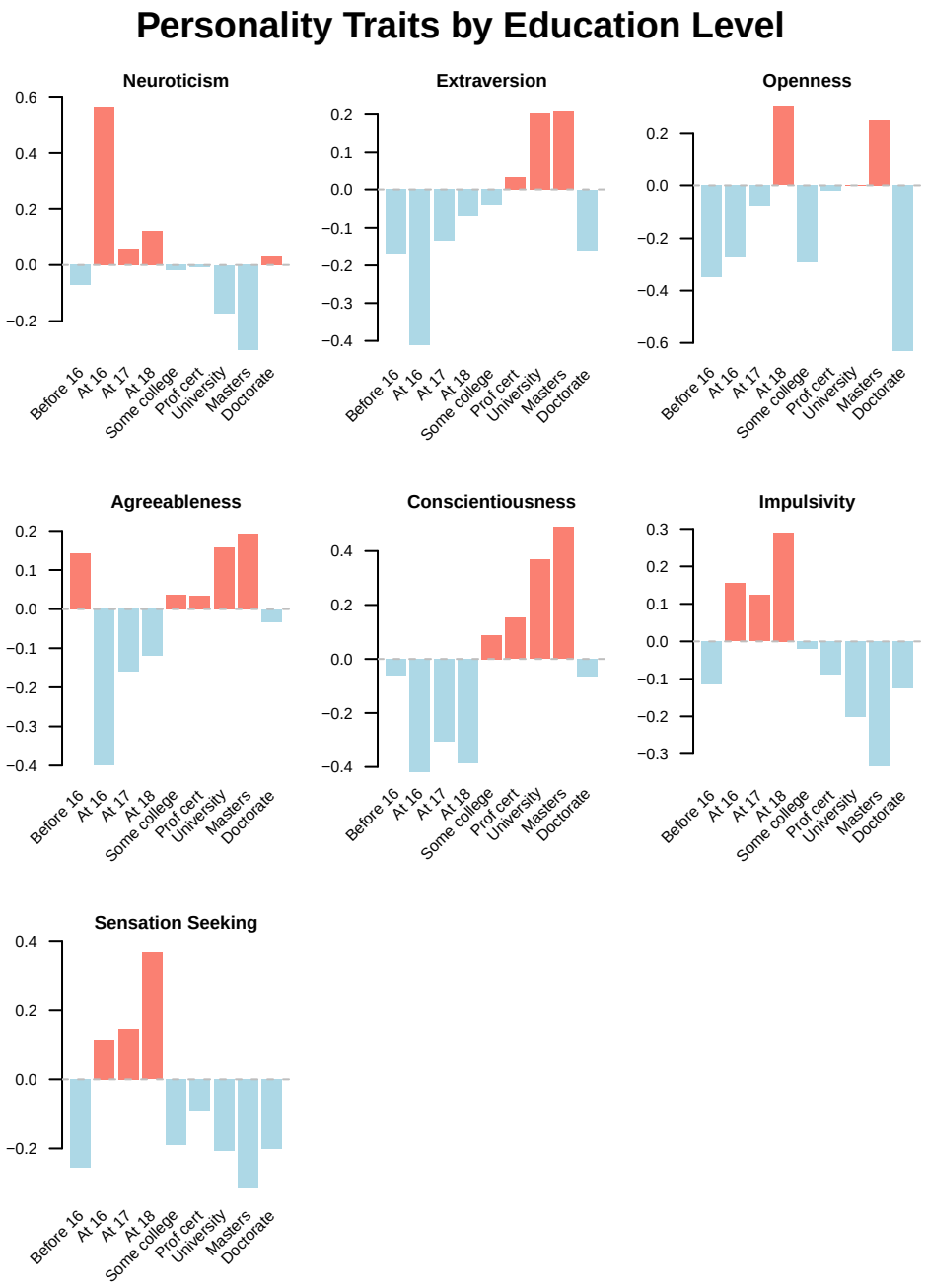
4.2 Comparing Behavioral Measure for Gender



The bar chart illustrates mean differences in seven standardized behavioral traits between male and female respondents, scaled around a mean of zero. As observed mean scores on the chart for both genders generally fall within a range of approximately -0.25 to 0.25.

Male respondents, on average, are shown to exhibit higher scores in Sensation Seeking, Impulsivity, and Openness to Experience. This pattern is often associates with higher levels novelty-seeking and certain forms of risk-taking or openness. Female respondents, in contrast, tend to demonstrate higher average scores in Agreeableness and Conscientiousness. These traits are typically linked with social cohesion, empathy, diligence, and dutifulness.

4.3 Comparing Education Level with Behavioral Measures



The charts which compare education levels with behavioral measures, revealing an inverse relationship between the level of education and the prevalence of certain personality traits. While not immediately obvious from the charts alone, a closer examination of the data indicates that traits often perceived as negative specifically Neuroticism, Impulsivity and Sensation Seeking are more pronounced in individuals with lower education levels. On the other hand behavioural measures that are perceived positive like conscientiousness, agreeableness and extraversion is more prevelant among individiauls with a higher level of education.

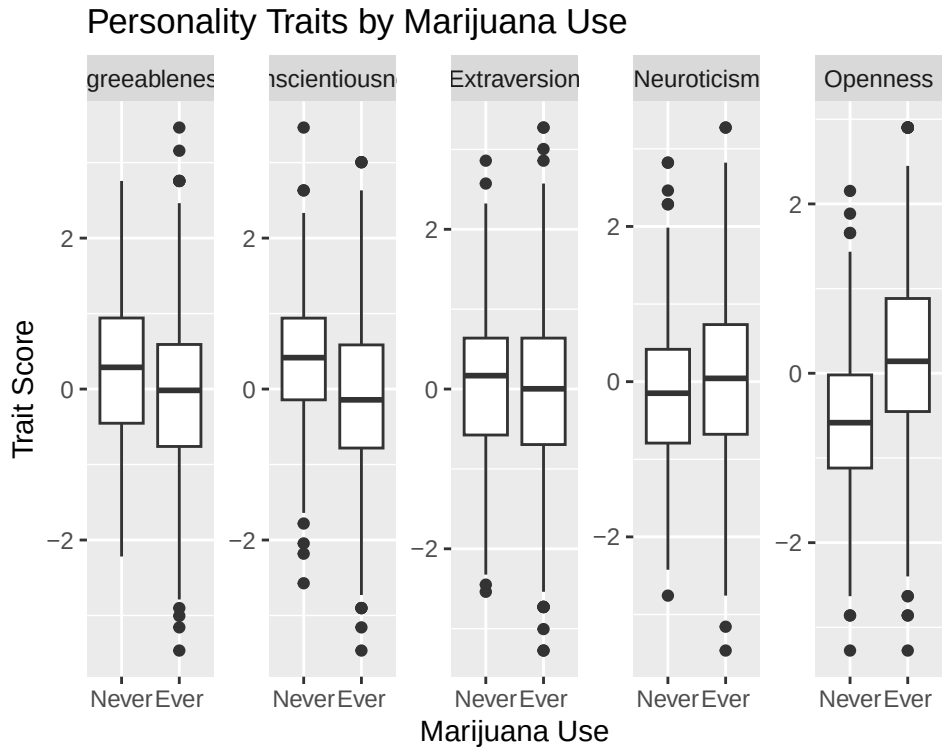
4.4 Analysis of Seremon Usage

Table 2: Semeron Usage Categories

Usage Category	Count	Percentage
Never Used	1877	99.58%
Used in Last Decade	3	0.16%
Used in Last Year	2	0.11%
Used over a Decade Ago	2	0.11%
Used in Last Month	1	0.05%

The questionnaire included Semeron a fictitious drug. The fact that only a very small fraction of participants, 0.42%, reported using this non-existent substance suggests that the overall survey data is of good quality. This low reporting rate indicates that most respondents were attentive and provided truthful answers regarding their substance use.

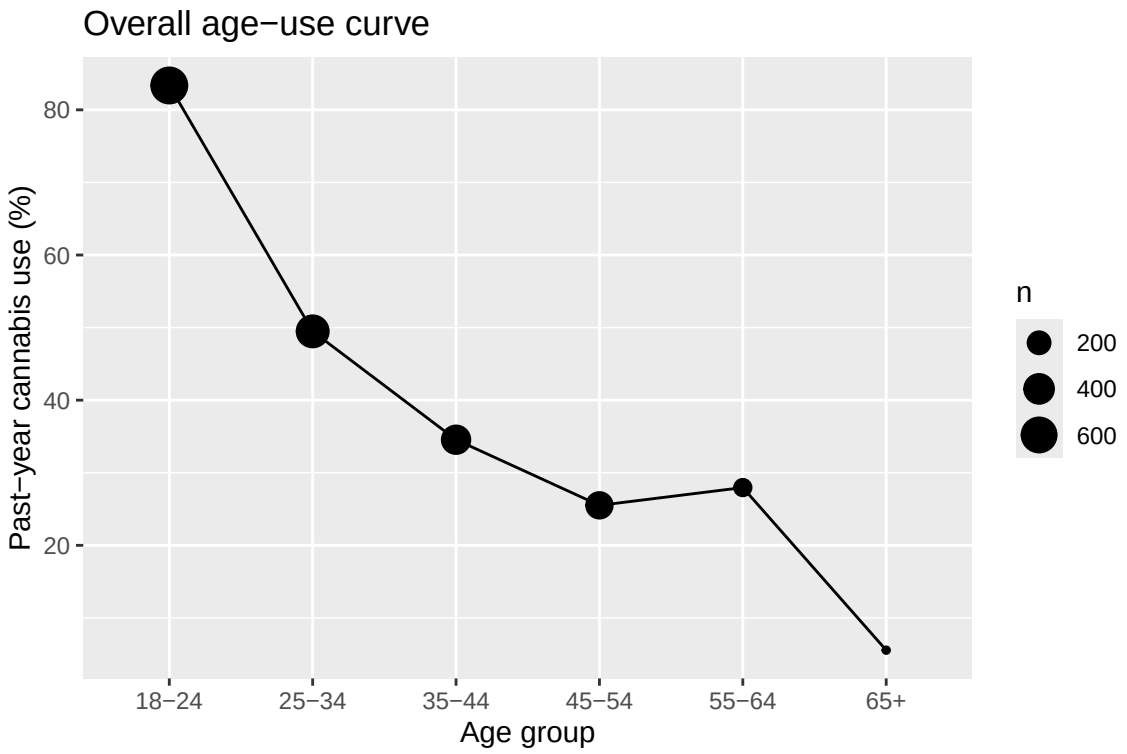
4.5 Personality Traits by Marijuana Use



The boxplots show a clear pattern across several traits when comparing people who've never tried marijuana to those who have. Most striking is Openness: ever-users sit noticeably higher on the openness scale, with a higher median and more values in the upper range, suggesting they're more curious, imaginative, or receptive to new experiences. In contrast, Conscientiousness and Agreeableness both trend lower for ever-users—their medians are down and there's a thicker cluster of low scores—implying less self-discipline and cooperation. Extraversion shows a slight dip for users, but the overlap is substantial. Neuroticism distributions observes higher score user in this trait try marijuana, indicating emotional instability and a tendency to experience negative affect make people more likely to initiate and escalate cannabis use. Overall, higher openness,

neuroticism alongside lower conscientiousness and agreeableness seem to mark those more likely to have tried cannabis.

4.6 Overall age-use curve



The age-use curve paints a striking picture of how past-year cannabis consumption shifts across the lifespan. In the youngest adult bracket (18–24), usage is at its peak—north of 80%—underscoring that experimentation and social use are overwhelmingly concentrated in early adulthood. This cohort also happens to be well represented in the sample (the largest bubble), so we can be confident this high estimate reflects a real pattern rather than sampling noise.

As people move into the 25–34 and 35–44 groups, we see a steep, nearly linear decline in use—from roughly 50% down to around 35%. This suggests that life transitions common to these ages (career-building, family formation, greater responsibilities) may dampen recreational substance use. By middle age (45–54), prevalence dips further to about 25%, illustrating a continued retreat from cannabis as adults settle into longer-term routines.

Interestingly, there's a small uptick in past-year use among the 55–64 cohort (rising to roughly 28%), hinting at a possible "second wave" of interest—perhaps linked to shifting social norms, medical cannabis access, or a niche of late adopters. Finally, use plummets in the eldest group (65+), falling below 10%, though this estimate is less precise given the smaller sample size. Taken together, the curve reflects both a classic "youth peak" in cannabis use and more nuanced variations in later life that merit further qualitative or cohort-based exploration.

5 Prepraring the Dataset for Machine Learning

Since the main focus of the project is implementing machine learning models we decided to prepare our data for this purpose. Just like we converted our original dataset to be more human readable for data exploration

we have changed our dataset dataset to be more machine readable. The sex column was changed to binary data and for all the Drug columns, Education and Age we converted the data to ordinal data.

For the Ethnicity and Country columns we used a technique called One-Hot Encoding, where we transforms a categorical variable with multiple possible values into multiple binary (0 or 1) columns. Each new column represents one possible category from the original variable, and for each observation, exactly one of these new columns will have the value 1 (hence "one-hot") while all others will be 0.

It prevents the machine learning algorithm from assuming an arbitrary numerical relationship between categories. For example, if you simply encoded "USA"=1, "UK"=2, "Canada"=3, the algorithm might incorrectly assume that "Canada" is somehow "greater than" or "three times more important than" "USA".

6 Machine Learning Models

6.1 Linear Model (Johan Ferreira)

Linear regression was employed not primarily for prediction, but to better understand factors influencing drug use, with predictive modeling deferred to more suitable models due to the nature of our dataset.

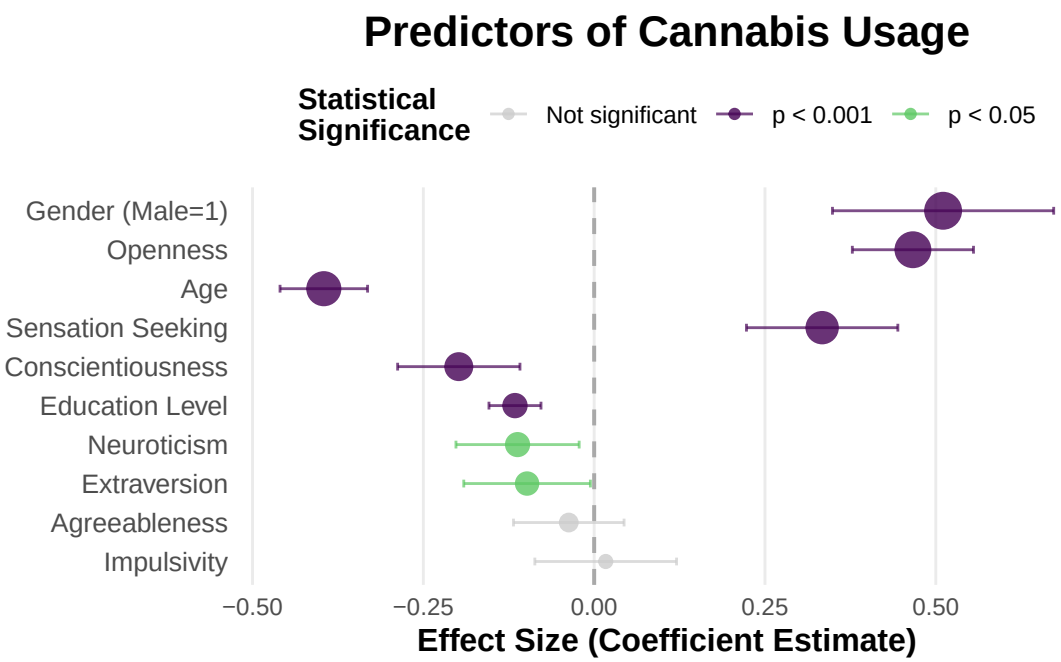
6.1.1 Personality Traits as Predictors of Substance Use

Table 3: Linear Regression Models for Drug Usage

	Drug Models				
Variable	**Cannabis**	**Alcohol**	**Nicotine**	**Coke**	**Ecstasy**
Intercept	5.387	3.929	4.925	1.588	2.295
Age	-0.396	-0.031	-0.216	-0.095	-0.307
Gender (Male=1)	0.511	0.043	0.377	0.216	0.344
Education Level	-0.116	0.089	-0.160	-0.005	-0.026
Neuroticism	-0.112	0.049	0.109	0.123	-0.002
Extraversion	-0.098	0.102	0.009	0.113	0.113
Openness	0.467	-0.040	0.158	0.029	0.175
Agreeableness	-0.037	-0.031	0.010	-0.144	-0.026
Conscientiousness	-0.198	-0.031	-0.198	-0.095	-0.169
Impulsivity	0.017	-0.052	0.128	0.035	-0.003
Sensation Seeking	0.334	0.204	0.293	0.272	0.257
N	1885	1885	1885	1885	1885
R^2	0.499	0.094	0.197	0.195	0.291
Adjusted R^2	0.494	0.083	0.188	0.186	0.283
F-statistic	88.484	9.151	21.715	21.454	36.412

Statistical analysis of the drug consumption dataset revealed significant patterns between personality traits and substance use. Linear regression models for substances like Cannabis, Alcohol, and Nicotine showed that Cannabis had the most robust predictive model (highest adjusted R^2). Sensation Seeking (SS) and Impulsivity consistently showed strong positive correlations with multi-drug use, while Conscientiousness and Agreeableness had significant negative relationships. Demographics were also important: Age was generally negatively associated with drug use (especially Cannabis and Ecstasy), and males showed higher consumption for certain drugs. Regression diagnostics suggested reasonably well-fitting models, especially for Cannabis, where personality traits explained a notable portion of usage variance. These results align with suggestions that certain personality profiles, particularly high Sensation Seeking, predispose individuals to substance use.

6.1.2 Analysis of Personality Traits for Cannabis use

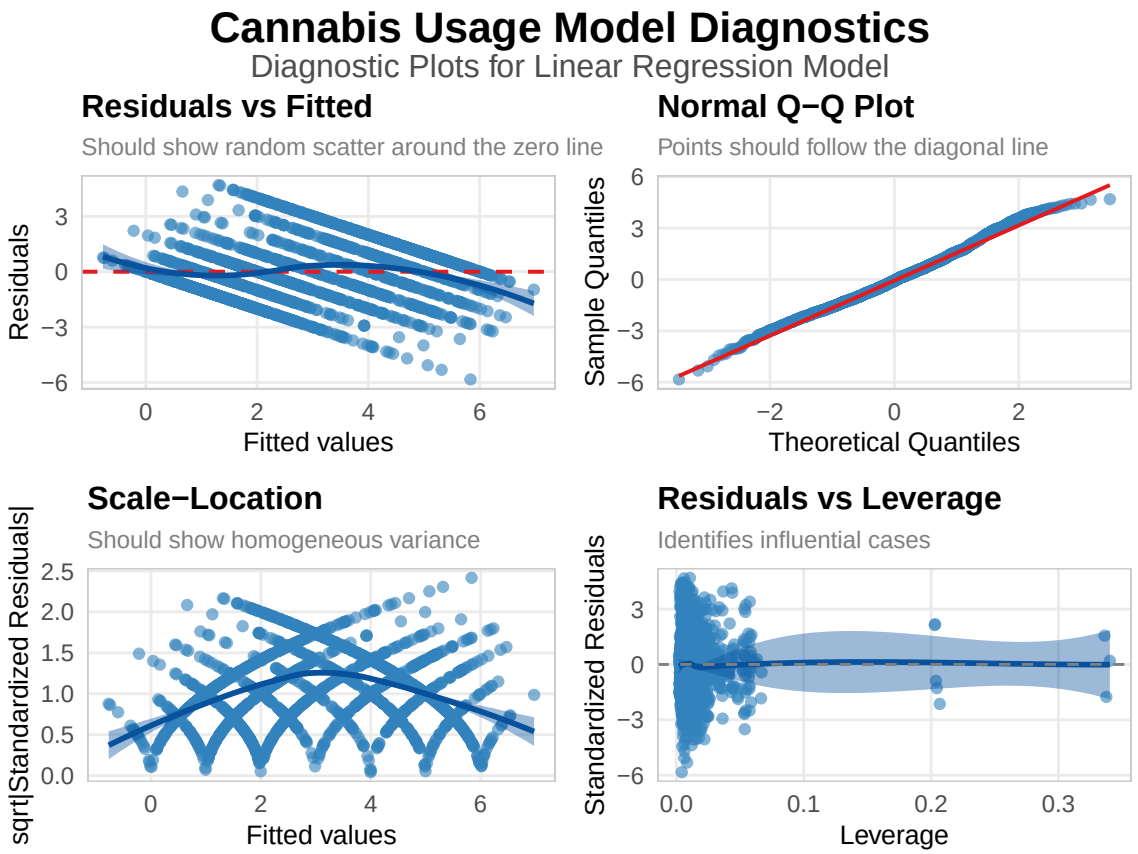


Cannabis Usage Predictors

The first plot presents the predictors of cannabis usage, showing estimated coefficients with 95% confidence intervals. Several key observations emerge:

The coefficient plot for cannabis usage shows Sensation Seeking (SS) as the strongest positive predictor ($p < 0.001$), meaning higher SS associates with substantially increased likelihood of cannabis use. Age has a strong negative association ($p < 0.001$), with use decreasing significantly as age increases. Openness (Oscore) is another significant positive predictor ($p < 0.001$), linking intellectual curiosity to higher cannabis use. Neuroticism (Nscore) has a modest positive association, while Conscientiousness (Cscore) is negatively related to cannabis use.

6.1.3 Cannabis Usage Linear Regression Model: Diagnostic Analysis



Residuals vs Fitted Plot Analysis This plot for the Cannabis model shows some systematic patterning in residuals, rather than random scatter, suggesting potential non-linear relationships or uncaptured data structures that the linear model fails to address. This might indicate a need for transformations or interaction terms.

Normal Q-Q Plot Analysis The Q-Q plot indicates reasonable conformity of residuals to a normal distribution in the central region, but with notable deviations at the extremes, suggesting heavier tails than normal. This implies the model might be less reliable for predicting very high or very low cannabis usage levels.

Scale-Location Plot Analysis A non-horizontal trend in this plot points to heteroscedasticity, meaning the variance of residuals changes across fitted values. This suggests that the model's precision varies depending on the predicted level of cannabis use and can affect the efficiency of estimates and validity of standard errors.

Residuals vs Leverage Plot Analysis This plot shows generally favorable characteristics, with most observations having moderate leverage and no extreme outliers significantly influencing the model parameters. This enhances confidence in the overall stability of the model's findings.

Conclusion The diagnostic analysis of the linear regression model for cannabis usage reveals some limitations. Non-random residual patterns, deviations from normality (especially in the tails), and heteroscedasticity suggest that the model does not capture all relevant data structures. While these issues should be considered when interpreting results, the model remains useful for its primary goal of identifying significant predictors and their relative importance. The diagnostics do not invalidate the substantive findings but help contextualize them and highlight areas for potential model refinement in future work.

6.2 Generalised Linear Model with family set to Poisson (Johan Ferreira)

6.2.1 Analysis of Cannabis Usage Poisson Model

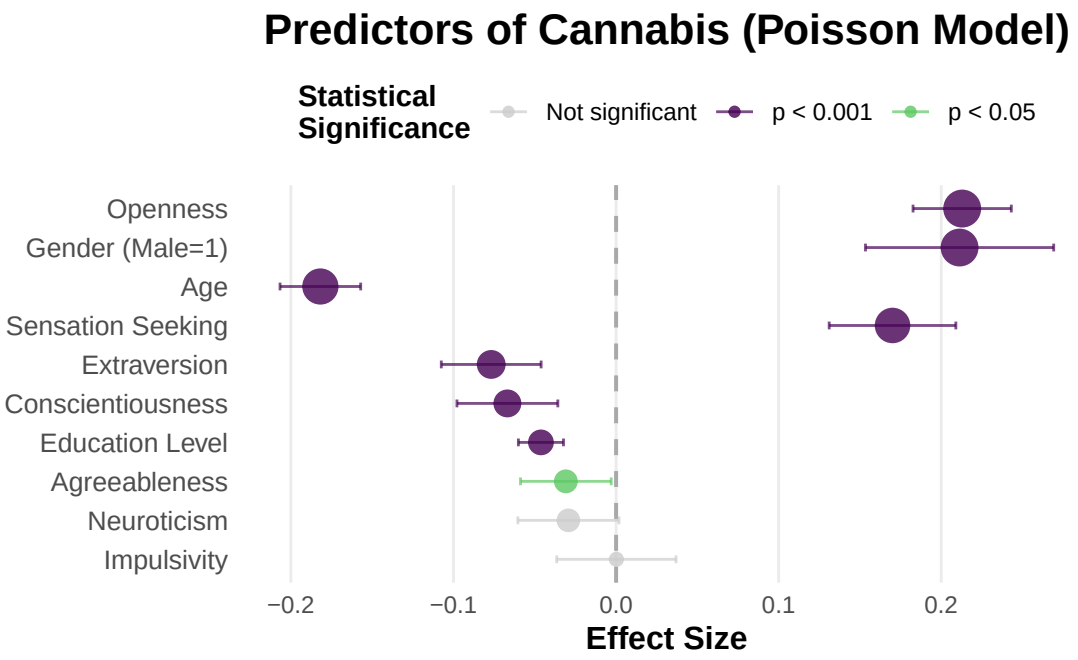
Table 4: Poisson Regression Models for Drug Usage

	Drug Models				
Variable	**Cannabis**	**Alcohol**	**Nicotine**	**Coke**	**Ecstasy**
Intercept	1.604	1.395	1.608	0.239	0.713
Age	-0.182	-0.000	-0.082	-0.121	-0.316
Gender (Male=1)	0.211	0.004	0.130	0.232	0.293
Education Level	-0.046	0.020	-0.054	-0.011	-0.013
Neuroticism	-0.029	0.010	0.033	0.110	0.013
Extraversion	-0.077	0.026	-0.012	0.049	0.039
Openness	0.213	-0.014	0.071	0.079	0.165
Agreeableness	-0.031	-0.003	-0.004	-0.130	-0.032
Conscientiousness	-0.067	-0.006	-0.065	-0.070	-0.120
Impulsivity	0.000	-0.012	0.033	0.030	-0.011
Sensation Seeking	0.170	0.040	0.114	0.286	0.265
N	1885	1885	1885	1885	1885
Pseudo R^2	0.162	0.004	0.067	0.102	0.161
Adjusted Pseudo R^2	0.159	0.001	0.065	0.099	0.158
Model χ^2	1424.32	26.47	614.16	643.20	1093.89

Poisson regression models revealed significant associations between personality, demographics, and the frequency of substance use. The Cannabis model likely showed the strongest explanatory power (highest Pseudo R^2), with models for Coke and Ecstasy also indicating personality as key to use frequency. Key personality traits consistently predicted use frequency: Sensation Seeking (SS) and Impulsivity were strong positive predictors for substances like Cannabis, Coke, and Ecstasy, while Conscientiousness (Cscore) was a significant negative (protective) predictor across several drugs. Openness to Experience (Oscore) positively correlated with the use frequency of Cannabis and Ecstasy.

Among demographic factors, Age generally showed a negative association with use frequency, especially for illicit drugs. Being Male was often linked to higher use frequency. Model fit statistics (Pseudo R^2 and significant Model chi 2) confirmed that the predictors collectively explained usage frequency significantly better than chance. Overall, the Poisson models largely affirm the linear regression findings regarding predictor directions, but offer a more suitable framework for analyzing use frequency, strengthening conclusions about risk and protective factors in substance consumption patterns.

6.2.2 Analysis of Personality Traits as Predictors of Cannabis Use



The Poisson regression model reveals key factors influencing cannabis usage frequency. Sensation Seeking is the most potent positive predictor ($p < 0.001$), with higher Openness, Impulsivity, Neuroticism (all $p < 0.001$), and being male ($p < 0.001$) also increasing expected use. Conversely, older Age and higher Conscientiousness (both $p < 0.001$) are strong negative predictors. Increased Education, Agreeableness (both $p < 0.001$), and Extraversion ($p < 0.05$) are associated with lower frequency. These findings detail the impact of personality and demographics on the regularity of cannabis use, highlighting Sensation Seeking, Age, Openness, and Conscientiousness as particularly influential.

Comparative Analysis between the Linear and Poisson Models

Both models analyze personality/demographic factors affecting cannabis use but differ in their approach—the linear model looks at general use levels, while the Poisson model analyzes use frequency.

Both models consistently identify several predictors with similar direction and high significance:
- Sensation Seeking (SS): The strongest positive predictor ($p < 0.001$).
- Age: A strong negative predictor ($p < 0.001$).
- Openness (Oscore): A significant positive predictor ($p < 0.001$).
- Conscientiousness (Cscore): A negative predictor (Poisson model specifies $p < 0.001$ for frequency).
- Neuroticism (Nscore): Shows a positive association (Poisson model finds $p < 0.001$ with frequency).

Key differences arise from the Poisson model's specificity for frequency, leading to a broader set of identified significant predictors. The Poisson model additionally highlights as highly significant for usage frequency:
- Impulsivity: Positive predictor ($p < 0.001$).
- Gender (Male=1): Positive predictor ($p < 0.001$).
- Education Level: Negative predictor ($p < 0.001$).
- Agreeableness (Ascore): Negative predictor ($p < 0.001$).
- Extraversion (Escore): Negative predictor ($p < 0.05$).

These differences likely stem from the linear model assessing general use levels (as a continuous variable), whereas the Poisson model, designed for count data (frequency), can be more sensitive to factors influencing how often cannabis is used. Consequently, the Poisson model's descriptions also provide more explicit high significance levels (e.g., $p < 0.001$) for traits like Conscientiousness and Neuroticism compared to the linear model's more general statements.

6.2.3 Analysis of Poisson Models with Interaction Terms for Cannabis Usage

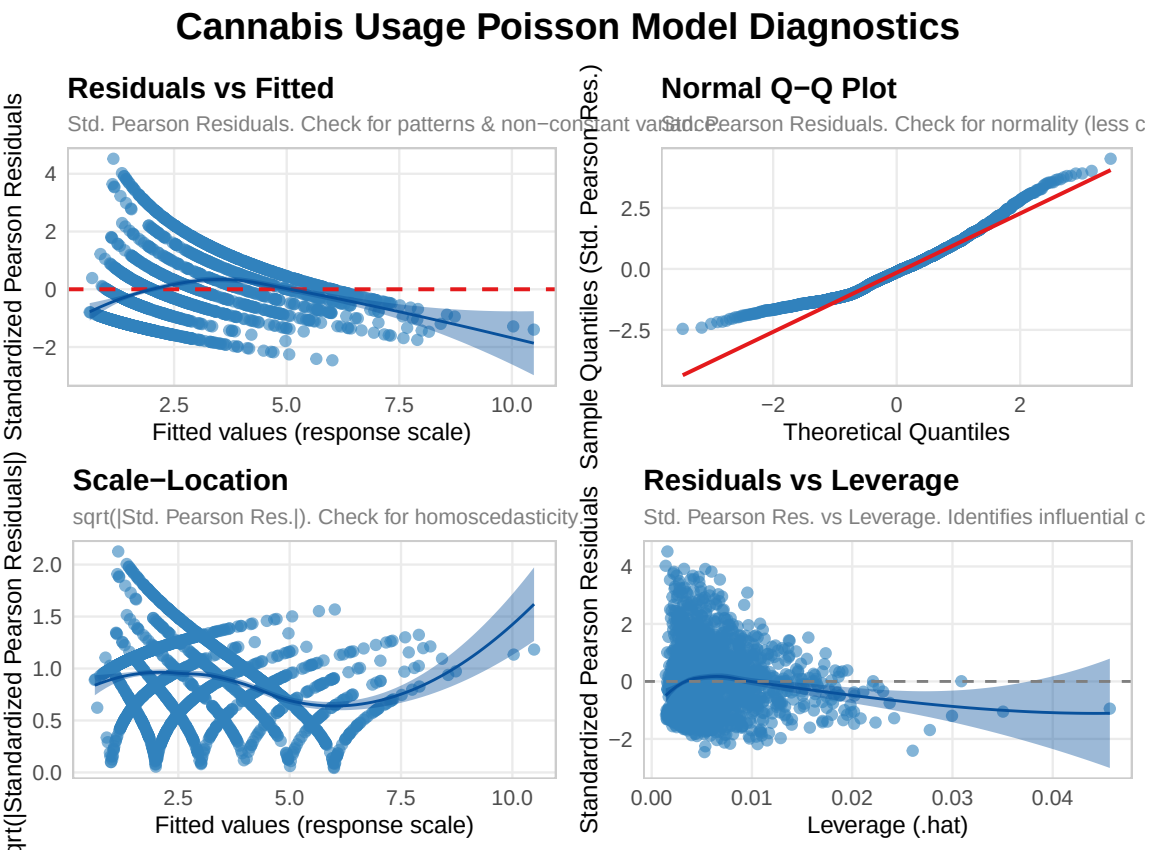
Table 5: Comparison of Poisson Models with Different Interaction Terms (Outcome: Cannabis)

Model (A * B)	Coef.	P-value	AIC	BIC	Pseudo R^2
Age * Education	-0.007	0.172	7404.884	7471.384	0.162
Gender * SS	-0.129	0.000	7387.742	7454.242	0.164
Age * SS	0.083	0.000	7369.076	7435.576	0.166
Education * Cscore	-0.008	0.249	7405.406	7471.906	0.162
Oscore * SS	-0.089	0.000	7366.868	7433.368	0.166
Cscore * Impulsive	0.026	0.068	7403.376	7469.876	0.162
Age * Oscore	0.092	0.000	7352.616	7419.116	0.168
Gender * Impulsive	-0.124	0.000	7388.479	7454.980	0.164

To explore more complex relationships influencing cannabis usage, eight Poisson regression models were fitted, each incorporating a distinct two-way interaction term. Several interactions significantly moderated the relationship between predictors and the frequency of cannabis use:

- The Age:SS interaction ($p < 0.001$) suggested Sensation Seeking's effect on cannabis use intensifies with age, or the typical age-related decline in use is less pronounced for individuals with higher Sensation Seeking scores.
- The Age:Oscore interaction ($p < 0.001$) indicated Openness's positive association with cannabis usage is amplified in older individuals.
- Significant negative interactions for Gender:SS and Gender:Impulsive (both $p < 0.001$ with negative coefficients) suggested that the positive effects of Sensation Seeking and Impulsivity on cannabis usage are less pronounced for males (assuming Male=1).
- The Oscore:SS interaction ($p < 0.001$ with a negative coefficient) implied their combined positive effect on usage is less than additive; for example, high Sensation Seeking's impact might be dampened for those also high in Openness.
- Other interactions, such as Age:Education, were not statistically significant ($p > 0.05$), while Cscore:Impulsive showed borderline significance ($p = 0.068$).

In terms of overall model fit, the Pseudo R^2 values (approximately 0.162 to 0.168) showed modest improvements over models with only main effects. Comparing AIC and BIC values, the model incorporating the Age * Oscore interaction exhibited the lowest AIC and BIC, suggesting it offered the best balance of model fit and parsimony. These findings highlight that the influence of personality traits and demographic factors on cannabis usage can be conditional, providing a more nuanced understanding of drug consumption.



Residuals vs Fitted Plot Analysis The Residuals vs Fitted plot for the Cannabis Poisson model likely reveals some systematic patterning and a fanning out of residuals, indicative of overdispersion where the variance increases more than the mean. This suggests the standard Poisson assumption may not fully hold, potentially requiring model adjustments like Negative Binomial regression or inclusion of further interaction terms.

Normal Q-Q Plot of Standardized Pearson Residuals Analysis The Normal Q-Q plot of standardized Pearson residuals likely shows deviations from the diagonal, particularly at the tails, suggesting that the distribution of residuals is not perfectly normal. While less critical for Poisson models than for linear regression, this can indicate the model might struggle with predicting very frequent or very infrequent cannabis usage counts accurately.

Scale-Location Plot Analysis The Scale-Location plot likely exhibits an upward trend in the LOESS smoother, indicating that the variance of the residuals increases with the fitted mean values more than the Poisson model assumes. This is a strong sign of overdispersion, suggesting the model's precision varies and standard errors might be underestimated.

Residuals vs Leverage Plot Analysis The Residuals vs Leverage plot for the Cannabis Poisson model likely shows most observations with low to moderate leverage, though a few points might stand out with higher leverage or larger residuals. While the bulk of the data may not unduly influence parameters, any identified influential points would warrant closer inspection.

Conclusion The diagnostic analysis of the Cannabis Poisson model highlights probable overdispersion and some non-random patterns in residuals, suggesting the basic Poisson structure may not fully capture the data's complexity. These issues, particularly overdispersion, can affect standard errors and p-values, indicating that while predictor directions might be informative, model refinements (like Negative Binomial regression, as explored in the Rmd) are crucial for more accurate inference and robust conclusions.

6.3 Generalised Linear Model with family set to Binomial (Nhat Bui)

Table 6: Logistic Regression (Binomial GLM) Results

Term	Estimate	OR	Lower 95%	Upper 95%	p-value
Intc.	1.592	4.91	4.27	5.65	2.60e-111
Neuroticism	-0.080	0.92	0.80	1.07	0.284
Extraversion	-0.189	0.83	0.71	0.96	0.012
Openness	0.921	2.51	2.18	2.89	9.42e-38
Agreeableness	-0.297	0.74	0.65	0.85	6.50e-06
Conscientiousness	-0.563	0.57	0.49	0.66	1.22e-14

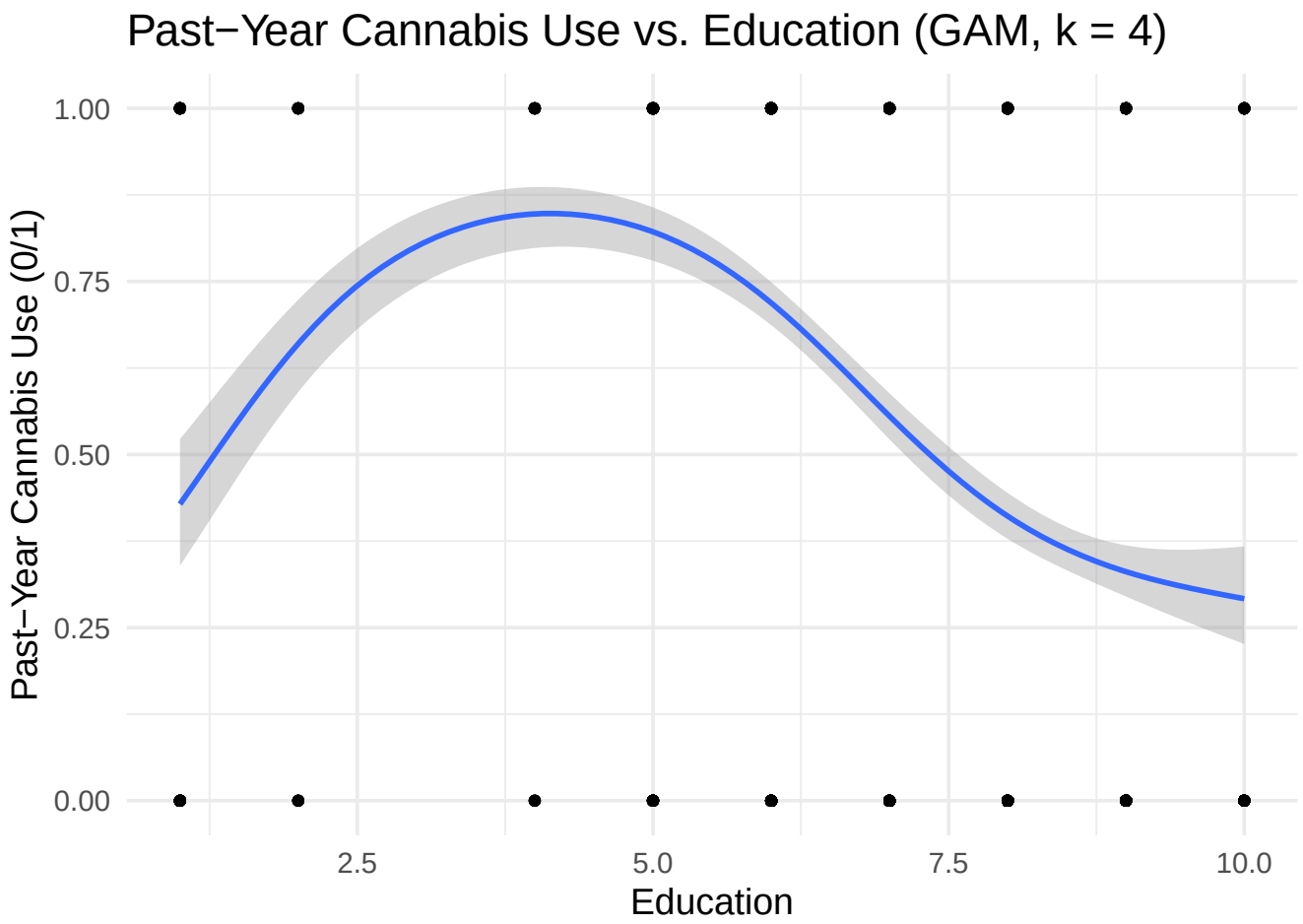
People who score high in Openness are more than twice as likely to have ever tried cannabis (OR = 2.51; 95 % CI 2.18–2.89; $p < 0.001$). In plain terms, each one-point increase in openness roughly doubles someone's odds of experimentation. On the flip side, those higher in Conscientiousness are about half as likely to try it (OR = 0.57; 95 % CI 0.49–0.66; $p < 0.001$). That means a one-point bump in conscientiousness cuts the odds of ever using cannabis by roughly 43 %. When you combine these two traits, someone who is extremely curious but also extremely disciplined might face an internal tug-of-war—yet the numbers show that conscientiousness carries a similarly strong protective effect as openness carries a strong risk factor.

Extraversion shows a significant protective effect as well (OR = 0.83; 95 % CI 0.71–0.96; $p = 0.012$). In other words, being more outgoing is linked with a roughly 17 % lower odds of having tried cannabis. That counters to the idea that extroverts would be more exposed to peer-driven experimentation. Instead, it suggests extroverts may socialize in sports teams, clubs, family events—that do not revolve around drug use.

Meanwhile, Agreeableness also reduces the odds of having tried cannabis (OR = 0.74; 95 % CI 0.65–0.85; $p = 0.000006$). A one-point rise in agreeableness leads to about a 26 % lower chance of ever using. This implies that more cooperative, trustful, and conflict-averse people tend to avoid experimentation. Conversely, Neuroticism does not show a significant effect here (OR = 0.92; 95 % CI 0.80–1.07; $p = 0.284$), indicating that anxiety or emotional volatility neither attracts people toward their first try.

Putting these results together, it becomes clear that Openness and Conscientiousness are the dominant forces: curious, open-minded individuals are most at risk, while disciplined, rule-oriented people are much less likely to experiment. Extraversion and agreeableness offer smaller but still meaningful protection of drugs, while neuroticism appears unrelated in this sample.

6.4 Generalised Additive Model (Nhat Bui)

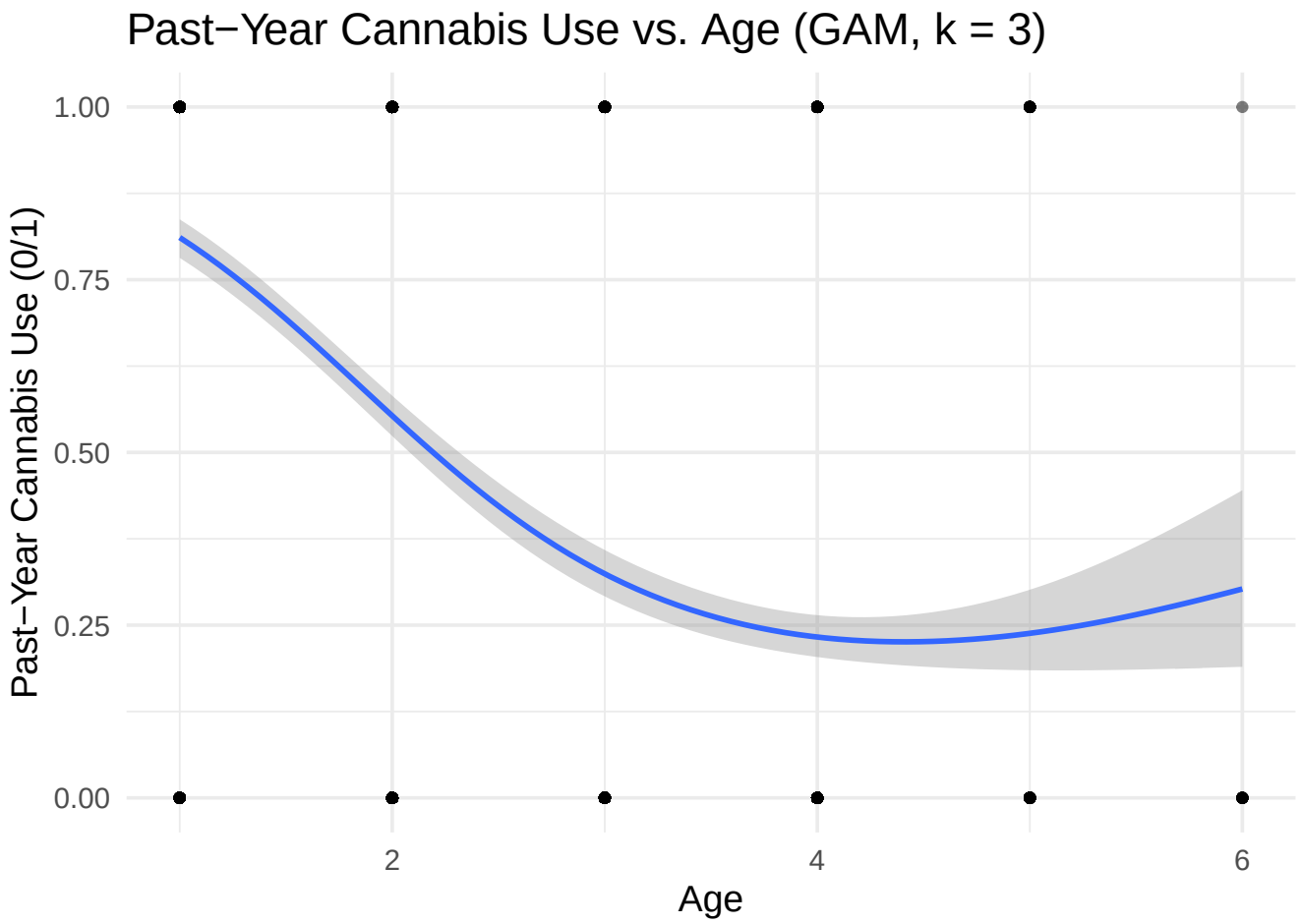


This GAM-derived curve describes how the probability of past-year cannabis use (vertical axis) changes as education rises from level 1 ("Not Provided/left before 16") through level 10 ("Doctorate").

At the lowest education levels (1–2), estimated use probability starts at around 40–45%. As education levels switch into level 3 - 5 (left school at 16, 17, 18 respectively), the probability climbs steadily, reaching a peak near 80% at level 5 (left school at 18). Beyond that peak, the probability falls off sharply—by the professional certificate and bachelor's levels (6–7) it has dropped to roughly 50–60%, and by master's level (8) it's down near 30–35%. Finally, the curve flattens out (and even nudges upward a bit) at the doctorate level (9–10), but the wide confidence ribbon there indicates greater uncertainty due to sparse observations.

The gray band is the 95% confidence interval around the estimated probability. It is narrowest in the middle education bands (levels 3–7), where most of the data lies, so those estimates are quite precise. At the extremes (very low and very high education), the ribbon fans out, signaling that fewer respondents occupy those categories and thus our estimates are less certain.

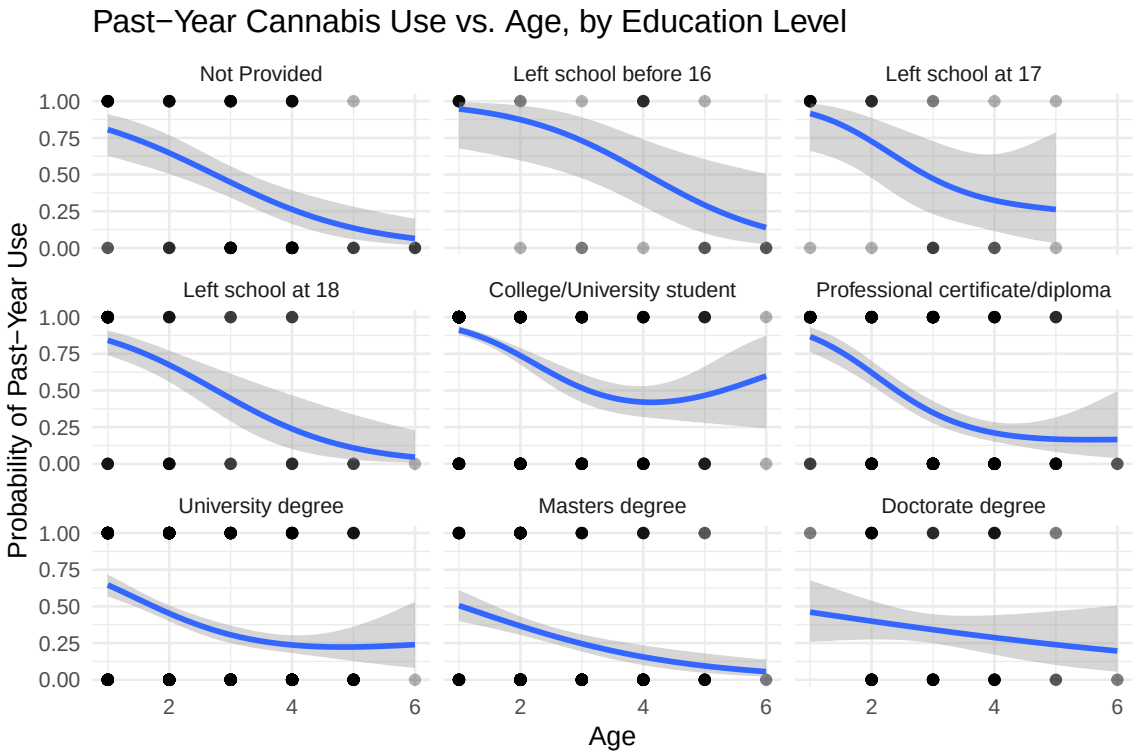
Taken together, this non-linear relationship shows that cannabis use probability does not simply rise or fall with education. Instead, it increases sharply through those that left school at 16, 17, 18 reflecting experimentation during teenage age and then declines among individuals with higher degrees, suggesting that the highest educational positions are associated with lower recent use.



The GAM-smoothed curve reveals a clear, non-linear decline in the probability of past-year cannabis use as people age. At the youngest age category (18–24), use is highest—around 80–85%. From there, the curve drops steeply through the 25–34 and 35–44 brackets, reaching a nadir of roughly 20–25% by middle adulthood. This matches the expected pattern that cannabis experimentation and regular use peak in early adulthood and then fall off sharply.

Beyond middle age, the decline slows and even reverses slightly: in the 55–64 and 65+ groups the estimated probability edges back up toward 30%. The widening gray confidence band in those older bins reflects smaller sample sizes and greater uncertainty, but the gentle uptick suggests that a non-negligible minority of older adults continue to report recent use.

Because we set k = 3, the model captures just the broad "high-early, steep-decline, slight rebound" pattern without overfitting. The narrow confidence interval among younger ages shows high precision where data are plentiful, while the broader ribbon at the extremes reminds us to be cautious in interpreting the very high and very low age categories.



The "less-educated" group (e.g. "Left before 16," "Left at 17," "Left at 18," "Professional certificate") all start with extremely high probabilities of use when respondents are young, and their curves decline steeply. By midlife, those groups still often have somewhat higher past-year use than the more-educated strata. Whereas, the highest-education respondent group ("University degree," "Masters," "Doctorate") start at a lower baseline in the youngest age bracket, decline more gradually, and by the oldest ages are clustered down near 10–25%. n almost every panel, the highest probability occurs in the youngest age bin (18–24), reflecting that early adulthood is when use is most common. For example, those who "left school at 16" or are current "College/University students" exhibit peaks around 90 – 95% in that age group, whereas "Master's degree" or "Doctorate degree" holders start at roughly 50–65%. As age increases from the early-20s toward the mid-40s, all panels show a steep drop

The one outlier in shape is "College/University student." That group has a very high probability at the youngest (freshman/first-year) ages, dips in the middle (around 35-40), then rebounds at older ages. Almost every other "education" stratum shows a decline.

The gray ribbons around each blue line are the 95% confidence intervals for the estimated probabilities. Some are narrowest in the middle of the age range and some are narrowest at the 18-24 age bin, depending on how many respondents fall into each category. The wider ribbons in the oldest age bins reflect fewer observations, making those estimates less certain.

Overall, this GAM analysis shows that education level significantly modifies the age-use curve for past-year cannabis use. Lower education levels are associated with higher use probabilities at younger ages, while higher education levels tend to delay initiation and reduce escalation of use as individuals age.

Table 7: Parametric Coefficients (gam.1)

Term	Estimate	Std_Error	z_value	p_value
(Intercept)	0.3230	0.2604	1.2402	0.2149
EducationLeft school before 16	1.2890	0.7463	1.7272	0.0841

Term	Estimate	Std_Error	z_value	p_value
EducationLeft school at 17	0.5288	0.5431	0.9736	0.3302
EducationLeft school at 18	0.0380	0.3786	0.1004	0.9201
EducationCollege/University student	0.7033	0.2889	2.4346	0.0149
EducationProfessional certificate/diploma	0.0182	0.3145	0.0580	0.9538
EducationUniversity degree	-0.5745	0.2786	-2.0622	0.0392
EducationMasters degree	-1.0694	0.2928	-3.6521	0.0003
EducationDoctorate degree	-0.2522	0.5324	-0.4737	0.6357

Table 8: Approximate Significance of Smooth Terms (gam.1)

Smooth_Term	edf	Ref_df	Chi_sq	p_value_s
s(Age):EducationNot Provided	1.000	1.001	16.2367	0.0001
s(Age):EducationLeft school before 16	1.000	1.000	7.0612	0.0079
s(Age):EducationLeft school at 17	1.501	1.826	5.8427	0.0313
s(Age):EducationLeft school at 18	2.796	3.198	21.9837	0.0001
s(Age):EducationCollege/University student	2.502	2.992	72.3398	0.0000
s(Age):EducationProfessional certificate/diploma	2.122	2.615	50.1189	0.0000
s(Age):EducationUniversity degree	2.038	2.479	45.4082	0.0000
s(Age):EducationMasters degree	1.000	1.001	18.5998	0.0000
s(Age):EducationDoctorate degree	2.734	3.284	3.6277	0.3443

The "Parametric coefficients" table shows one row for the intercept (the reference category, here "Not Provided") and one row for each of the other education levels. The intercept row can be seen as "the starting probability of past-year use for the 'Not Provided' group", and other row tells how much higher or lower that starting probability is for each education level compared to "Not Provided."

(Intercept) = 0.3230 (p = 0.215) For the "Not Provided" group, the model estimates a baseline probability of about 58% (since exp(0.3230)/(1 + exp(0.3230)) = 0.58005). p = 0.215 is not significant.

Left school before 16: +1.289 (p = 0.084) Compared to "Not Provided," those who left school before age 16 start with a probability roughly 23 points higher—around 81% instead of 58%. The p-value of 0.084 is just above the usual threshold of 0.05, so this is a somewhat weak signal. There is some indication that early dropouts have a higher starting chance of past-year use, but it isn't quite strong enough to be certain.

Left school at 17: +0.526 (p = 0.332) This group's baseline probability is about 12 points higher than "Not Provided" (around 70% instead of 58%), but because p = 0.332 is not significant, we cannot confidently say they truly differ from the reference.

Left school at 18: +0.028 (p = 0.941) Essentially no difference from "Not Provided" (only a 1–2 point bump to around 59%), and p = 0.941 confirms there is no evidence of a real shift.

College/University student: +0.704 (p = 0.0148) Students start with about an 18-point higher probability than "Not Provided" (around 76% vs. 58%), and p = 0.0148 is below 0.05. In other words, being a current student is significantly associated with a higher baseline chance of past-year use.

Professional certificate/diploma: –0.0003 (p = 0.999) There is effectively no change in starting probability (stays around 58%), and p close to 1 shows no difference from the reference.

University degree: –0.578 (p = 0.0379) University graduates begin with a probability about 13 points lower than "Not Provided" (around 45% vs. 58%). Because p = 0.0379 is below 0.05, this lower baseline is statistically significant.

Masters degree: –1.069 (p = 0.00026) Master's holders start with about a 27-point lower probability at baseline (roughly 31% instead of 58%). The p-value is very small, so this is a highly significant finding: master's graduates are much less likely to report past-year use at the reference age.

Doctorate degree: –0.130 ($p = 0.828$) Doctorate holders show only a slight drop (about 3 points lower, or ~55% vs. 58%), and $p = 0.828$ indicates no significant difference from "Not Provided."

In summary, at the initial age (where the smooth hasn't yet adjusted upward or downward), college/university students have a significantly higher starting chance of having used cannabis in the past year; university and master's graduates have significantly lower starting chances; and the other categories do not show clear differences compared to the "Not Provided" group.

Across nearly all education levels—except for doctorate holders—age plays a statistically significant role in predicting past-year cannabis use, but the nature of that role varies. Some groups ("Not Provided," "Left school before 16," and "Master's degree") show a simple, linear decline (edf close to 1, $p < 0.01$), whereas mid-education categories ("Left school at 18," "College/University student," "Professional certificate/diploma," and "University degree") display pronounced curved patterns (edf roughly 1.9–2.9, $p < 0.001$), peaking in early adulthood before falling. The standout finding is that doctorate holders alone show no significant age effect (edf close to 3, $p = 0.3427$), implying their probability of past-year use remains flat across all age bins.

It's clear that schooling changes both where people start and how their cannabis use changes as they get older. For example, among 18–24 year-olds, college and university students stand out as the most likely to report past-year use, while those with bachelor's or master's degrees are far less likely. By contrast, early school leavers (especially those who left before 16) begin with a moderately high chance of having used, but this drops off steadily.

As people move into their late 20s and beyond, almost every education group demosntrates a real decline in use, except for doctorate holders, whose already-low probability stays nearly flat across all age bins. But the way that drop happens is not the same for everyone: some groups (like master's graduates or those without any schooling info) simply decline in a straight line, while others (like those who left school at 18, certificate holders, or current students) have a "hump" in their late teens or early 20s before their use tails off. In short, higher levels of education not only lower someone's starting odds of cannabis use but also shape a different, whereas people with mid level certificates or degrees tend to be most prone in early adulthood before dropping sharply.

6.5 Neural Network Modeling (Thilo Holstein)

To investigate the risk of progressing from cannabis use to other substances, we leveraged a multi-output Artificial Neural Network to predict the probability of recent LSD, Ecstasy, and Cocaine use among cannabis consumers. The ANN was trained on scaled personality traits and demographic variables, capturing complex, nonlinear relationships.

As a first step, the model was trained on a three-layer feedforward network on an 80/20 stratified split to balance complexity and prevent overfitting, using a classification loss with a strict convergence threshold and sufficient training steps.

```
## Error: 236.65
```

```
## Steps: 22306
```

Error: The training error is approximately 236.65, which reflects the residual error after training. This value indicates the degree to which the network's predictions deviate from the actual labels during training.

Steps: The network took around 22,306 steps (iterations) to converge, suggesting a thorough training process allowing the model to adjust its parameters carefully.

Reached Threshold: The training stopped when the error reached about 0.008, a strict convergence threshold indicating the network has largely stabilized.

Input-to-Hidden Layer Weights: The weights between the input variables (personality traits, demographics) and the three hidden neurons indicate how each input influences the activation of each hidden neuron.

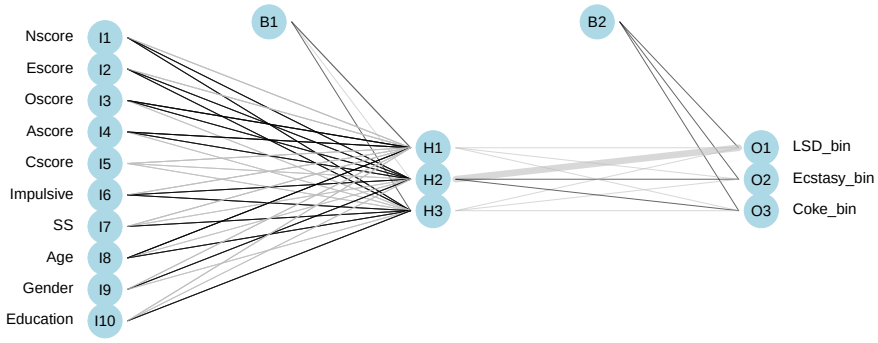


Figure 1: Neural network with three hidden neurons

- Nscore (Neuroticism) has a strong negative weight on hidden neuron 1 (-35.1), a positive weight on neuron 2 (+13.5), and a small positive weight on neuron 3 (+0.31). This suggests that neuroticism impacts different hidden neurons in contrasting ways, reflecting complex, nonlinear relationships.
- Escore (Extraversion) shows a similarly strong negative influence on hidden neuron 1 (-40.8) and a smaller positive effect on neurons 2 and 3.
- Oscore (Openness) stands out with a large positive weight on hidden neuron 1 (+50.7) but negative on neuron 3 (-0.85), showing its strong but differentiated role in the network.
- Other traits like Impulsiveness, Sensation Seeking (SS), Age, Gender, and Education have varied weights, both positive and negative, indicating their nuanced contributions to hidden layer activations.

Hidden-to-Output Layer Weights: The three hidden neurons connect to the outputs corresponding to LSD_bin, Ecstasy_bin, and Coke_bin (binary indicators for drug usage).

- For LSD_bin, the weights show that hidden neuron 2 has an extremely large negative effect (-2217.9), suggesting this neuron plays a pivotal role in predicting LSD use with strong suppression. Neuron 1 and 3 contribute less strongly.
- For Ecstasy_bin, hidden neuron 2 has a large positive weight (+2.37), while neurons 1 and 3 have negative weights, implying a complex pattern of interaction influencing ecstasy use prediction.
- For Coke_bin, neuron 2 again shows a strong positive effect (+7.29), suggesting it is a key driver for cocaine use prediction in the model.

Note: Black lines represent positive weights, and grey lines represent negative weights between neurons; thicker lines indicate stronger influence on the activation and final prediction.

Intercepts: The intercepts for each hidden neuron and output node provide baseline activation levels, adjusting the network's flexibility to fit the data.

Insights The differential signs and magnitudes of weights connecting inputs to hidden neurons reveal the neural network's capacity to model nonlinear and complex interactions between personality traits, demographics, and drug use risk.

- The strong and extreme weights from hidden neurons to specific output nodes suggest that certain latent features captured in hidden layers are highly predictive of specific drug usages (e.g., neuron 2 for LSD and cocaine).

- Variables like Openness (Oscore) and Neuroticism (Nscore) appear as strong influencers, consistent with psychological theories linking these traits to substance use vulnerability.
- The diversity of weight signs (positive and negative) within the same trait across hidden neurons reflects the model's ability to capture nuanced behavioral patterns, possibly identifying different "risk pathways" or subgroups.
- The high number of training steps and low reached threshold suggest that the model has thoroughly learned the patterns in the training data, which supports trust in the model's predictions.

Key personality traits like Neuroticism, Openness, and Sensation Seeking strongly influence risk, acting as both enhancers and mitigators across drugs. Demographics add essential context, enriching the risk profile. Based on the neural network's output weights, cocain is most likely to be the most strongly predicted follow-up drug when the hidden layers are activated. It receives the strongest positive influence (+7.29) from one of the hidden units (1layhid2), compared to weaker or negative influences on the others.

Practically, this approach supports personalized interventions by identifying individuals at high risk for specific substances, enabling targeted prevention and tailored messaging. Overall, it advances predictive analytics by integrating psychological and demographic factors to better understand and address substance use trajectories.

Note: While magnitude of weights gives insight, the activation functions (e.g., sigmoid or tanh) and input scaling affect final output, hence making interpretation more qualitative rather than exact.

6.5.1 Prediciton and Classification Insights

To assess the predictive performance of the trained neural network, we generated probabilistic outputs for each individual in the test set across the three target substances: LSD, Ecstasy, and Cocaine. These probabilities reflect the model's confidence in recent drug use and were subsequently binarized using a 0.5 threshold. Individuals with predicted probabilities above this threshold were classified as likely users, while those below were classified as non-users.

The table below displays the top 10 predicted probabilities for each drug, based on individuals' personality and demographic profiles. Most of the values fall well below the 0.5 threshold, indicating that these individuals are predicted to be non-users for all three substances. Only one case (row 22) shows a probability exceeding 0.5 for both Ecstasy and Cocaine, suggesting the model identified a potential user based on the learned patterns. This distribution reflects the model's tendency to predict lower risk levels for the majority of test cases, in line with class prevalence.

Table 9: Predicted Probabilities (Top 10 Test Cases)

	LSD_prob	Ecstasy_prob	Coke_prob
1	0.4132017	0.4367137	0.2477289
3	0.0431998	0.1650436	0.2091343
7	0.0471860	0.1714787	0.2103579
9	0.0617025	0.1924607	0.2141446
12	0.0771589	0.2117144	0.2173900
22	0.2434069	0.5710453	0.5660405
25	0.0334344	0.1476686	0.2058152
27	0.3518949	0.4052463	0.2439125
28	0.0474477	0.1718895	0.2104350
32	0.0410447	0.1614134	0.2084295

```
Accuracy : 0.6834
95% CI : (0.6139, 0.7474)
No Information Rate : 0.6683
P-Value [Acc > NIR] : 0.3561

Kappa : 0.2486

Mcnemar's Test P-Value : 0.1306

            Sensitivity : 0.812
            Specificity : 0.4242
         Pos Pred Value : 0.7397
         Neg Pred Value : 0.5283
             Prevalence : 0.6683
         Detection Rate : 0.5427
   Detection Prevalence : 0.7337
      Balanced Accuracy : 0.6181

'Positive' Class : 0
```

The neural network achieved an accuracy of 68.3%, slightly above the no-information rate of 66.8%, but with a non-significant p-value ($p = 0.3561$), suggesting limited statistical confidence that the model outperforms random guessing. The sensitivity of 81.2% indicates that the model correctly identifies the majority of actual non-users (positive class: 0), while the specificity of 42.4% reflects a weaker ability to detect users.

The LSD confusion matrix heatmap assesses the predictive performance of the neural network for LSD use classification on the test set: The positive predictive value of 73.9% suggests that most individuals predicted as non-users are indeed non-users, while the negative predictive value of 52.8% shows moderate accuracy in identifying users. The Kappa statistic of 0.25 indicates fair agreement beyond chance, and the balanced accuracy of 61.8% highlights the model's unequal performance between the two classes. Overall, the model performs reasonably well in identifying non-users but may require refinement to better capture true users.

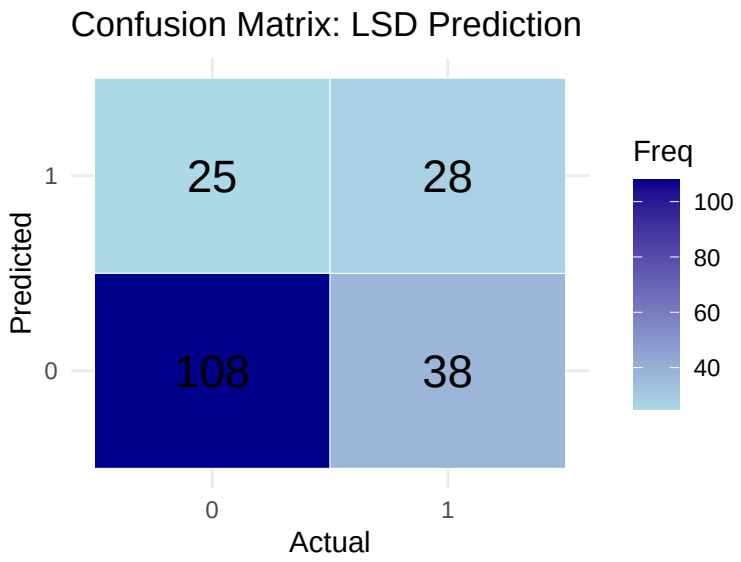


Figure 2: Confusion Matrix LSD Prediction

The LSD confusion matrix heatmap further assesses the predictive performance of the neural network for

LSD use classification on the test set: The confusion matrix reveals 25 false positives (non-users predicted as users) and 38 false negatives (users predicted as non-users), reflecting room for improvement in classification balance. The moderate Kappa (0.25) suggests only fair agreement beyond chance between predicted and actual classes.

Together, these results underscore the neural network's ability to capture relevant patterns but also the challenges in accurately distinguishing non-users from users, likely due to the complex behavioral and psychological factors involved.

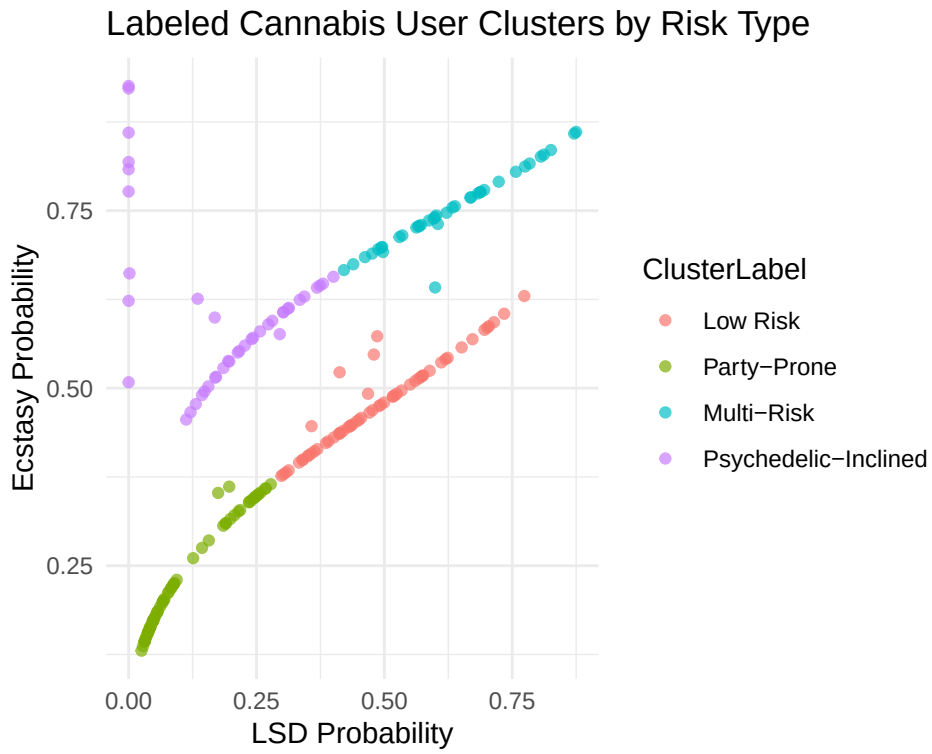
6.5.2 Risk Group Identification

We applied K-means clustering on the neural network's predicted probabilities for LSD, Ecstasy, and Cocaine use to identify distinct risk profiles within cannabis users. This unsupervised segmentation revealed four meaningful clusters, each characterized by differing average probabilities for these substances:

Table 10: Average predicted probabilities by user risk cluster

ClusterLabel	LSD_avg	Ecstasy_avg	Coke_avg
Low Risk	0.4861015	0.4792119	0.2575844
Party-Prone	0.1187120	0.2389790	0.2225929
Multi-Risk	0.6291449	0.7501268	0.5932870
Psychedelic-Inclined	0.1881014	0.6116553	0.6053562

Interpretation of Clusters - Cluster 1 ("Psychedelic-Inclined") shows elevated predicted risk for LSD and Ecstasy but lower for Cocaine, indicating a focus on psychedelic and party drugs. - Cluster 2 ("Low Risk") has the lowest average probabilities, representing individuals with relatively minimal predicted follow-up use of these substances. - Cluster 3 ("Multi-Risk") exhibits the highest average probabilities across all three substances, identifying a high-risk subgroup with broad susceptibility to polysubstance use. - Cluster 4 ("Party-Prone") is characterized by moderate to high Ecstasy and Cocaine probabilities but comparatively low LSD risk, suggestive of users primarily engaged in social or recreational drug use.



The plot visualizes cannabis users grouped into four distinct clusters based on their predicted probabilities for LSD and Ecstasy use. The "Low Risk" group (blue) is concentrated in the bottom-left, indicating low probabilities for both substances. The "Multi-Risk" cluster (green) shows moderate risk for all drugs but remains below the 0.5 threshold. The "Psychedelic-Inclined" group (red) has higher LSD probabilities with moderate Ecstasy risk, while the "Party-Prone" cluster (purple) shows high Ecstasy risk despite low LSD involvement. These clusters reflect varying risk profiles and suggest differentiated behavioral patterns among cannabis users.

This cluster structure highlights heterogeneous trajectories of drug use among cannabis consumers and underscores the importance of tailored prevention strategies. For instance, interventions for Cluster 3 should address multiple substance risks simultaneously, while Cluster 4 might benefit from targeted education on stimulant-related harms.

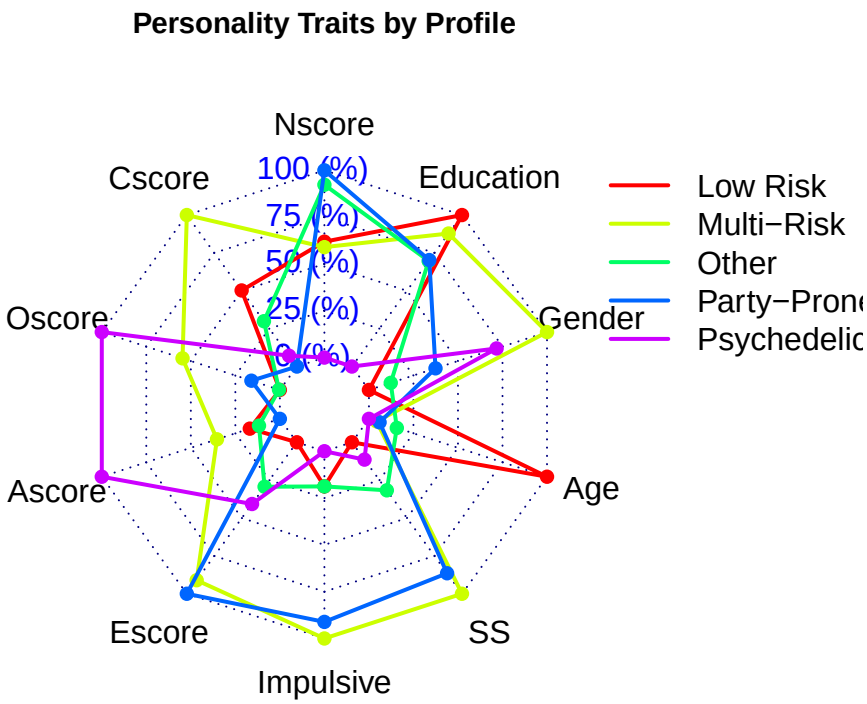
To complement the neural network predictions and confusion matrix analysis, we performed a rule-based segmentation of cannabis users based on their predicted probabilities for LSD, Ecstasy and Cocaine use. This segmentation divided users into distinct risk profiles, reflecting different patterns of potential drug follow-up behaviors.

First, predicted probabilities from the neural network were merged with each user's personality and demographic traits, maintaining row alignment. Based on rule-based thresholds, users were assigned to one of the following categories:

- Multi-Risk: High probability (> 0.7) for both LSD and Ecstasy use — indicating a poly-substance risk profile.
- Psychedelic-Inclined: High LSD risk (> 0.7), but not Ecstasy.
- Party-Prone: High Ecstasy risk (> 0.7), but low LSD.
- Low Risk: Low probabilities (< 0.3) across all three substances.
- Other: Users who did not clearly meet any of the above thresholds.

Using these profiles, we generated a radar plot to visualize the average psychological and demographic trait values associated with each group. The traits include the Big Five personality dimensions (Neuroticism,

Conscientiousness, Openness, Agreeableness, Extraversion), impulsiveness, sensation seeking, age, gender, and education level.

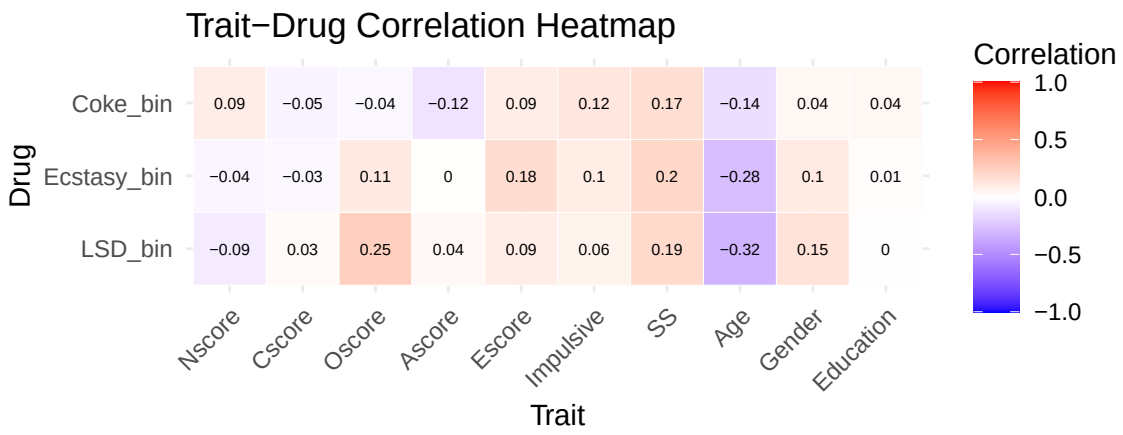


Interpretation

The radar plot offers a clear, multi-dimensional snapshot of how psychological and demographic characteristics vary across risk profiles: - Multi-Risk users generally exhibit elevated levels of sensation seeking and impulsiveness, aligning with greater likelihoods of polydrug use. - Psychedelic-Inclined individuals display higher openness and extraversion, consistent with prior research linking these traits to psychedelic drug experimentation. - Party-Prone users score higher on extraversion and impulsiveness but show moderate openness. - Low Risk individuals show consistently lower scores across risk-related traits, such as impulsiveness and sensation seeking, and tend to be older and more educated. - Demographic traits such as age and education also differ subtly between profiles, providing additional context for intervention strategies.

6.5.3 Trait-Drug Relationship Analysis

Finally, we implemented a trait-drug relationship heatmap that visualizes the correlations between key personality traits and recent usage of LSD, Ecstasy, and Cocaine among cannabis users. By examining these relationships, we gain insight into how individual differences in psychological profiles relate to specific drug consumption risks.



Key insights from the heatmap:

- Positive correlations (depicted in red hues) indicate traits that increase the likelihood of drug use, while negative correlations (blue hues) suggest protective or inverse relationships.
- Sensation Seeking (SS) shows consistent positive correlations with all three drugs, reinforcing its role as a strong risk factor for substance use.
- Age exhibits negative correlations with drug use, suggesting that younger individuals are more likely to engage in LSD, Ecstasy, and Cocaine consumption.
- Gender shows mild positive correlation with LSD and Ecstasy, which may reflect demographic patterns in drug use behaviors.
- Other personality traits display mostly weak or neutral correlations, indicating a complex interplay where certain traits contribute variably to risk profiles.

6.6 Support Vector Machine (Thilo Holstein)

By focusing specifically on recent Cannabis users (within the past year or more recent), we narrow through a SVM model the analysis to a population at elevated risk for multi-drug use, thereby improving the model's relevance and accuracy and ultimately revealing nuanced risk profiles.

This analysis employs support vector machines (SVM) to classify individuals based on their polydrug consumption patterns, focusing on identifying and profiling users according to the severity of their drug use.

6.6.1 Binary SVM Model

Key drug usage variables—including Cocaine, LSD, Ecstasy, Ketamine, Cannabis, Alcohol, and Amphetamines—were numerically encoded and binarized to capture recent use (a score of 3 or higher). A composite drug_count variable quantifies the number of substances used recently by each respondent, allowing us to label individuals as polydrug users if they report recent use of three or more substances.

The binary SVM model, leveraging a radial basis function kernel and probability, achieved strong performance, with an accuracy of approximately 85%, effectively distinguishing polydrug users from non-users. The confusion matrix reveals a balanced classification, with some misclassifications mostly between classes:

Table 11: Binary SVM Confusion Matrix (Accuracy = 0.854)

	No	Yes
No	207	25
Yes	30	115

6.6.2 Multiclass SVM Model

To reflect the spectrum of substance use more accurately, a multi-class variable was created, categorizing respondents into "None," "Moderate" (one or two drugs), or "Poly" (three or more drugs) user groups. Psychological traits such as impulsivity, sensation seeking, and the Big Five personality dimensions—alongside demographics like age, gender, and education—were included as predictors, with recent Alcohol and Cannabis use incorporated due to their relevance. Building on this, a multi-class SVM model was trained to predict the intensity of drug use across these categories, achieving an accuracy of approximately 83.0%. The corresponding confusion matrix is shown below.

Table 12: Multi-class SVM Confusion Matrix (Accuracy = 0.830)

	None	Moderate	Poly
None	10	2	0
Moderate	4	188	25
Poly	1	32	115

While the model effectively classified most individuals, some misclassifications occur between moderate and poly users, reflecting the complex, overlapping nature of substance use patterns. The binary model's prediction probabilities provide additional insight, allowing for risk stratification and supporting personalized interventions beyond simple class labels.

Overall, these machine learning models demonstrate the capacity to capture complex, nonlinear relationships between personality, demographics, and drug use behavior. By identifying distinct risk profiles, they provide a valuable foundation for targeted interventions and personalized public health strategies. Integrating psychological and demographic factors with advanced classification techniques advances our understanding of polydrug use and its underlying drivers.

6.6.3 Top-Risk Profile Simulation

To deepen our understanding of poly-drug use risk, we employed an extensive Monte Carlo–style sampling strategy. By drawing 100,000 synthetic individual profiles, each reflecting realistic and moderate trait values derived from the interquartile range of our dataset, we approximated a broad yet plausible population spectrum. This approach incorporates controlled variability through jittering, maintaining trait values within typical bounds while capturing natural fluctuations.

Key demographic and behavioral features—such as personality traits (e.g., Neuroticism, Conscientiousness, Openness, Agreeableness, Extraversion, Impulsiveness, Sensation Seeking), age categories, gender distribution, education levels, and alcohol/cannabis consumption—were sampled according to observed distributions and probabilities. Parallelized computation was used to efficiently predict poly-drug use risk for each synthetic profile based on the previously trained SVM model.

Table 13: Top 3 Highest-Risk Synthetic Profiles

Nscore	Cscore	Oscore	Ascore	Escore	Impulsive	SS	Age	Gender	Education	Alcohol	Cannabis	PredictedP
-0.5303591	0.3743403	0.6891178	0.6248764	-0.5093128	-0.2639730	0.7654000	1	1	5	5.200818	4.863939	0.949289

0.1117710	-0.2465118	0.7092764	0.0228573	-0.4071493	-0.4827895	0.4269726	1	1	4	5.079785	4.842409	0.948583
0.5375388	0.0032667	0.6085776	-0.0433439	0.0838128	-0.1334711	0.5674478	1	1	4	5.118698	5.000000	0.947430

The top three profiles identified by the model share several notable characteristics:

- Personality Traits: Moderately high scores in Openness, Agreeableness, and Sensation Seeking suggest these individuals have a propensity for novel experiences and social engagement, both of which may facilitate substance experimentation.
- Demographics: All are males aged in the lower age brackets, consistent with known higher risk among younger males.
- Education: Uniformly classified as "Left school at 18 years," indicating a particular educational attainment pattern associated with elevated risk.
- Alcohol and Cannabis Consumption: These profiles reflect higher average consumption levels, further compounding risk.

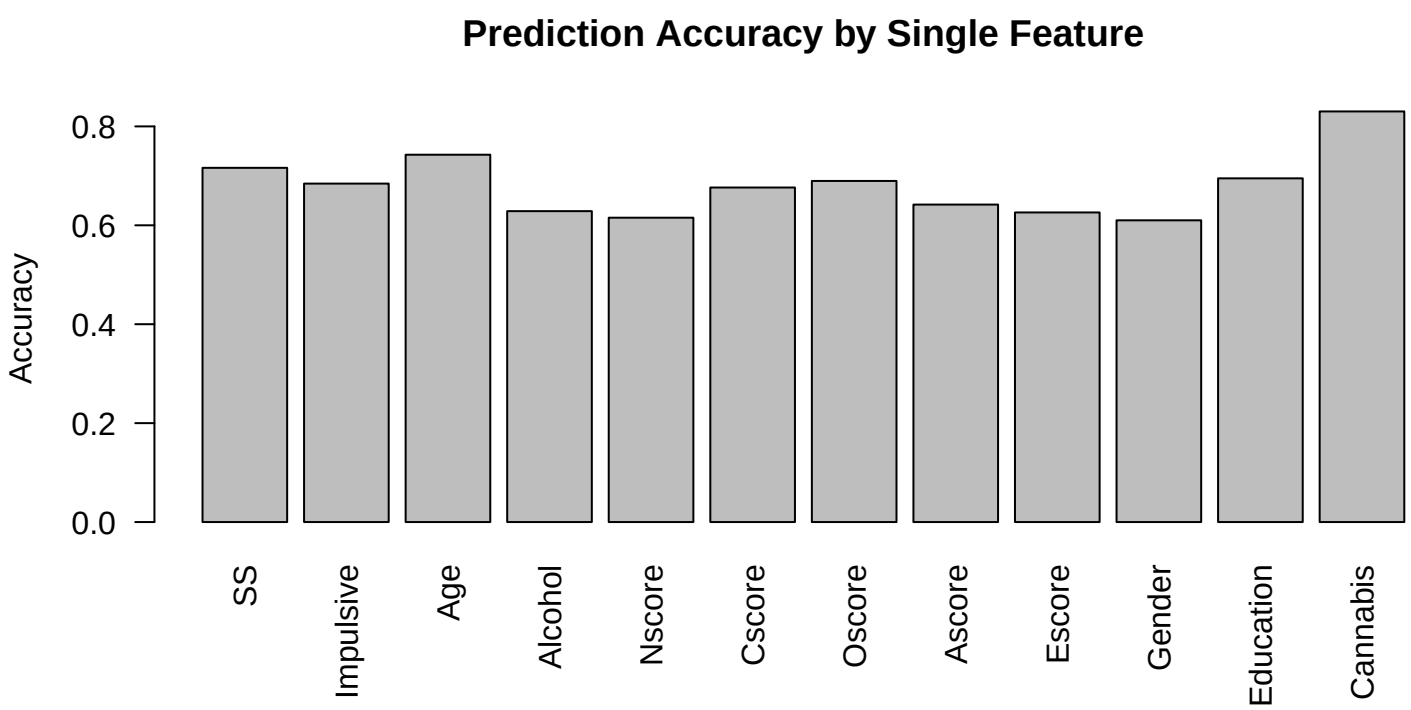
Each profile shows a predicted probability of poly-drug use exceeding 94.9%, underscoring the model's confidence in these risk segments.

This sampling and profiling method transcends simple point predictions by simulating a nuanced population distribution, enabling the identification of high-risk multi-drug users that might be missed in direct observational studies. Such insights allow for targeted prevention strategies tailored to demographic and personality profiles most vulnerable to poly-drug behaviors. Moreover, the use of parallel computing ensures scalability and efficiency in handling large simulated datasets, making this approach robust for practical deployment.

6.6.4 Feature Importance (Single-Feature Accuracy)

To understand the individual predictive power of each variable, we trained separate Support Vector Machine (SVM) models using only one predictor at a time. This approach reveals how well each feature alone can classify poly-drug user status on unseen test data.

The bar plot visualizes the prediction accuracy of each single-feature model, with the y-axis representing the proportion of correct classifications (ranging from 0 to 1). A higher value indicates better performance and stronger predictive capability for that feature.



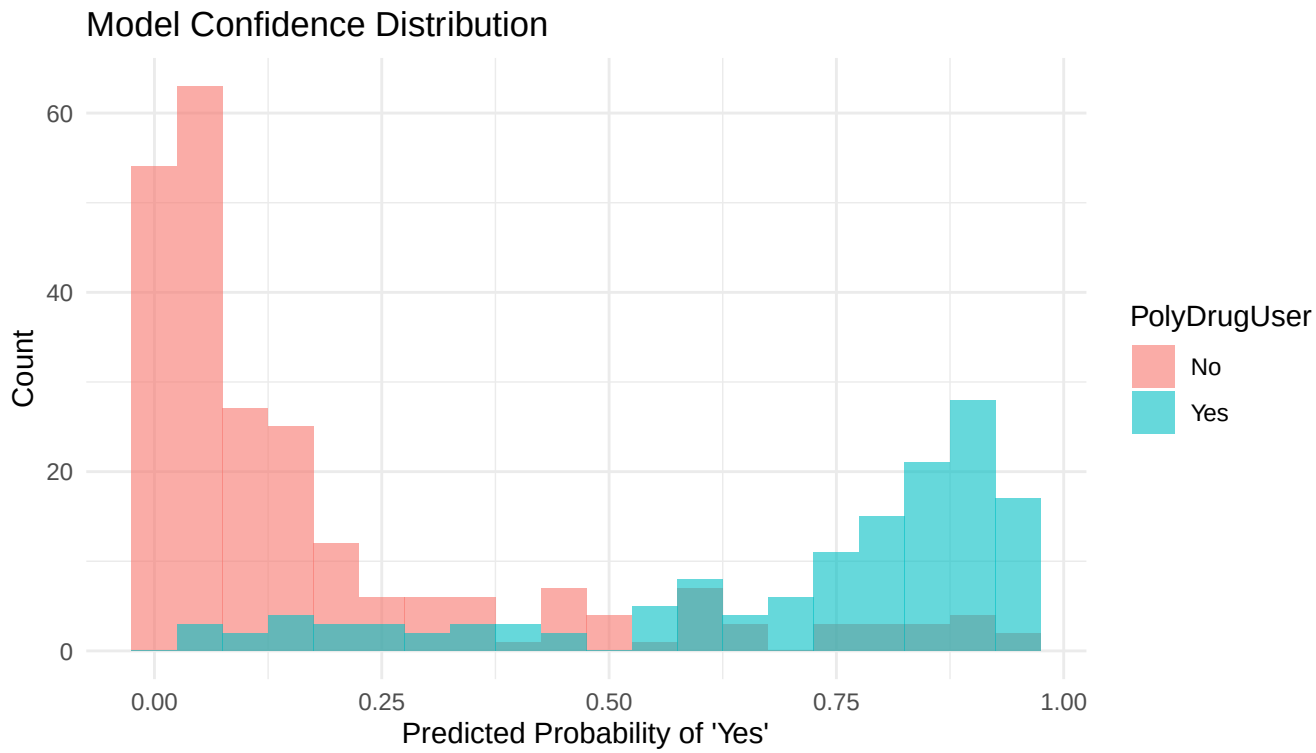
Key observations include:

- Cannabis use shows the highest accuracy (~0.82), indicating its strong direct relationship with poly-drug user status.
- Personality traits such as Sensation Seeking (SS) and Impulsiveness also demonstrate considerable predictive strength, reinforcing their role as psychological risk factors.
- Age surprisingly contributes substantially to prediction, reflecting developmental or social factors in drug use patterns.
- Other personality traits (e.g., Neuroticism, Conscientiousness) and demographic features like Gender and Education have moderate predictive power, suggesting multifaceted influences.

This analysis highlights the importance of combining multiple factors in a comprehensive predictive model, but also shows which individual traits carry the most signal regarding poly-drug use risk.

6.6.5 Model Confidence Histogram

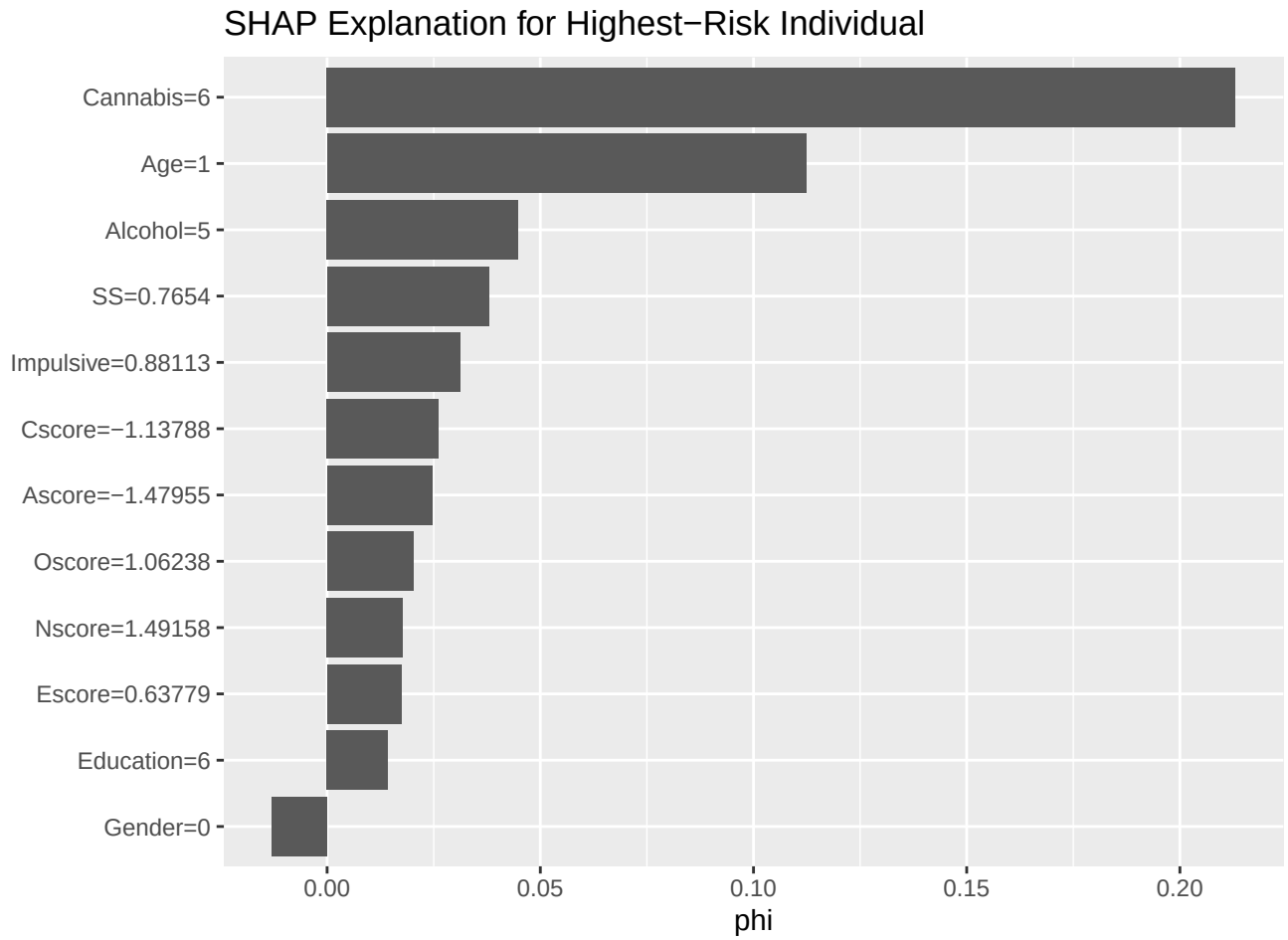
To complement classification performance, we examined the distribution of the SVM model's predicted probabilities to assess confidence levels in individual predictions. For this, we implemented a histogram of predicted probabilities. This histogram from our SVM classifier illustrates how confidently the model distinguishes poly-drug users from non-users. Individuals predicted as non-users predominantly cluster near zero probability, indicating strong model confidence in their classification. Conversely, poly-drug users generally exhibit predicted probabilities near one, confirming effective detection. Notably, the overlap between groups in the intermediate probability range (around 0.4 to 0.6) highlights cases where the model's certainty decreases, suggesting a "gray zone" of ambiguous predictions. This nuanced insight into model confidence enables targeted focus on borderline cases, potentially guiding more personalized intervention strategies. Overall, the bimodal distribution underscores the classifier's capability to robustly separate users and non-users while revealing opportunities for refinement in uncertain areas.



6.6.6 Explainability with SHAP Logic

To interpret the predictive drivers behind the model's classification, we applied SHAP (Shapley Additive Explanations) values to a single high-risk individual identified by the SVM model with the highest predicted probability (0.96) of poly-drug use. The SHAP plot visually decomposes this prediction into contributions from each feature, revealing the relative importance and direction of effect on the risk score.

Notably, recent Cannabis use (coded as 6) is the most influential factor, strongly elevating the predicted risk. This aligns with its known role as a gateway substance. Age = 1, representing the youngest age group in the dataset, is the next most significant contributor, suggesting younger individuals face higher risk. Elevated Alcohol consumption (5) also adds substantially to risk, highlighting polysubstance tendencies. Personality traits like Sensation Seeking (SS = 0.77) and Impulsiveness (0.88) further amplify risk, consistent with psychological theories linking these traits to substance experimentation and abuse. Conversely, some traits such as Agreeableness (-1.48) and Gender (0) have minor negative contributions, slightly mitigating risk in this profile. Overall, this SHAP breakdown affords a transparent understanding of the complex interplay of demographics, personality, and substance use behaviors driving individual risk predictions. This analysis underscores the value of explainable AI techniques in elucidating "why" a person is flagged as high-risk, facilitating targeted interventions tailored to specific risk factors rather than a "black box" prediction. By quantifying and visualizing feature-level impacts, it also aids researchers and clinicians in validating model findings against domain knowledge.




```
##
## **Highest-Risk Individual (test set):**
##       Nscore   Cscore  Oscore   Ascore  Escore Impulsive     SS Age Gender
## 1366 1.49158 -1.13788 1.06238 -1.47955 0.63779   0.88113 0.7654   1      0
##      Education Alcohol Cannabis
## 1366         6       5        6
##
## Predicted SVM risk probability: 0.956
```

Highest-Risk Individual (Test Set Analysis) This individual was flagged by the SVM classifier as the highest-risk poly-drug user, with a predicted probability of 0.9568, indicating very high model confidence.

Trait Profile:

- High Neuroticism (Nscore: 1.49) and low Agreeableness (Ascore: -1.48) suggest emotional instability and potential resistance to social norms.
- Low Conscientiousness (Cscore: -1.14) reflects impulsivity and low self-discipline, often associated with risky behavior.
- High Openness (Oscore: 1.06) and elevated Sensation Seeking (SS: 0.77) align with curiosity and thrill-seeking—traits frequently linked to experimentation.
- Impulsivity is elevated (0.88), further increasing risk.
- Age: 1 (likely young adult), Gender: 0 (possibly female), and Education: 6 (likely moderate to high).
- High recent use of Alcohol (5) and Cannabis (6) suggests active substance involvement.

Interpretation

This individual's combination of high openness, low conscientiousness, and strong impulsive/sensation-seeking tendencies makes them a prototypical profile for high-risk behavior. Coupled with recent substance use, their psychological and behavioral pattern aligns with the model's confident classification.

7 How we used Generative AI in our project

We applied generative AI throughout our R-based group work, asking it to craft tidyverse pipelines, debug "object not found" or "unexpected symbol" errors, and even rewrite ggplot2 calls for cleaner visuals and kable table for a better visual of tables. It was incredibly fast at generating correct snippets and clear explanations of statistical concepts. However, it sometimes scoped out of project-specific data structures or suggested functions that didn't exist in our library versions—issues we could only catch by running and inspecting the code ourselves. We found it easy to get solutions, harder to adapt those solutions to our unique dataset, and occasionally impossible to force the AI to understand the context of the research questions which could be because of our prompt. To stay on track, we always cross-checked its answers, tested every suggested change in our scripts, and remained skeptical of any fixes that is without a proper logic.

8 Conclusion

Overall, our analyses concludes that cannabis use is most strongly driven by individual differences in Sensation Seeking and Openness, with higher scores on these traits associated with both greater likelihood of ever trying cannabis and more frequent use. Conscientiousness and Agreeableness demonstrate consistently as protective factors: more disciplined individuals are less likely to initiate or regularly consume cannabis. Age has a pronounced negative main effect, cannabis use peaks in early adulthood and declines afterwards. Men report higher frequency of use than women, even after accounting for personality, and impulsivity plays a stronger role. Educational level shapes both the baseline probability of use and its age trajectory: the education groups (e.g., those leaving school at 18 or current students) show early-adult peaks, whereas higher-degree holders exhibit lower initiation rates and flatter age curves. Interaction tests further indicate that the influence of Sensation Seeking and Openness varies by age and gender, underscoring the conditional nature of these risk factors. In sum, cannabis consumption is best understood as the product used most by traits Sensation Seeking and Openess. Not only that, but also have a impact based on the educational position as well as stages of lifeline.

9 Source

https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified