Clustering Drug Users: Identifying Personality-Based Patterns in High Drug Usage

Diana Nicole Danga
College of Computing and Information Technologies
National University - Philippines
Manila, Philippines
dangadd@students.national-u.edu.ph

John Efren Gannaban
College of Computing and Information Technologies
National University - Philippines
Manila, Philippines
gannabanjv@students.national-u.edu.ph

Jascent Pearl Navarro
College of Computing and Information Technologies
National University - Philippines
Manila, Philippines
navarrojg@students.national-u.edu.ph

***Abstract*— This paper identified common profile traits among people who have high drug use through unsupervised learning - clustering and analysis. Feature engineering was used to aid use of unlabeled data. The pattern found shows that younger individuals, mostly male, with lower education levels, tend to exhibit higher neuroticism, extraversion, and openness. They also display lower agreeableness and conscientiousness, along with higher impulsivity and greater sensation-seeking tendencies. This common pattern suggests that they are most at risk for substance abuse.**

***Index Terms*—clustering, unsupervised learning, k means, hac, drugs, legal drugs, illicit drugs, personality.**

I. Introduction

The illicit use of drugs is a major contributor to the global burden of disease, leading to severe health and psychological consequences. In particular, substance abuse and dependence are linked to an increased risk of mental disorders, fatal overdoses, suicide, and violence. In 2010, illicit drug dependence accounted for 20 million (95% UI: 15.3–25.4 million) disability-adjusted life years (DALYs) worldwide. Given the detrimental effects of illegal drug use, researchers have focused on identifying contributing factors, which range from individual and interpersonal influences to broader community and policy-related aspects [1]., as outlined by the social-ecological model [2]; Key factors associated with illegal drug use include poor mental health, peer pressure, and an unstable family environment. Furthermore, Drug addiction is not limited to illicit substances such as heroin or cocaine—addiction can also stem from legal substances, including alcohol, nicotine, prescription medications, and opioids, whether legally prescribed or illegally obtained [3].

Substance use among adolescents remains a growing concern, particularly with the co-use of alcohol and cannabis. This age group is especially vulnerable, as their developing brains are more susceptible to the long-term consequences of substance use. The 2023 World Drug Report from the United Nations Office on Drugs and Crime highlighted the widespread prevalence of substance use among adolescents. Additionally, tobacco and alcohol consumption are major global concerns, often serving as gateway substances that lead to further drug abuse. In the Philippines, a 2022 government survey revealed alarming trends in adolescent drug use, particularly regarding the age of first-time users. The study found that 41.3% of individuals who had tried drugs for the first time were between the ages of 15 and 19. Patterns of use varied, with 38.7% reporting drug use two to five times a week, 24.7% using drugs monthly, and 20.6% using them weekly . Further insights into adolescent alcohol consumption were provided by the 2018 Expanded National Nutrition Survey, which documented fluctuations in drinking rates among adolescents [4].

Given this context, this study aims to uncover one of the key contributing factors to drug abuse: the common demographic and behavioral traits of individuals with high drug usage. By leveraging unsupervised learning techniques, this paper seeks to identify patterns in common profile traits among people with high drug use, which could help in early intervention efforts. Understanding the shared characteristics of high-risk individuals may aid in developing targeted solutions to mitigate substance abuse risks. This study will analyze data from the research paper, *"The Five-Factor Model of Personality and Evaluation of Drug Consumption Risk"* by E. Fehrman, A.K. Muhammad, E.M. Mirkes, V. Egan, and A.N. Gorban [5] to define distinct group identities and ultimately establish a demographic profile for individuals at high risk of substance abuse.

II. Review of Related Literature

A. Overview of the key concepts and background information

Revised NEO Five-Factor Inventory

The Revised NEO Five-Factor Inventory (NEO FFI-R) is a condensed version of the Revised NEO Personality Inventory (NEO-PI-R) developed by Costa & McCrae (1992b) [22]. It is a 60-item self-report questionnaire designed to assess the five fundamental personality traits: Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. Each trait is measured using 12 items, rated on a five-point Likert scale ranging from strongly disagree (1) to strongly agree (5). While the NEO inventories are extensively utilized in cross-cultural personality research, they also serve a clinical

purpose. Professionals such as counselors, clinical psychologists, and psychiatrists use these inventories to analyze a person's strengths and weaknesses, aid in diagnosis, assess personal challenges, build therapeutic rapport, offer feedback, predict treatment outcomes, and determine the most suitable therapeutic strategies. [6]

The Barratt Impulsiveness Scale

The Barratt Impulsiveness Scale (BIS-11) is a 30-item self-report questionnaire that evaluates impulsiveness, originally developed in the 1990s by Dr. Barratt and the International Society for Research on Impulsivity. It is one of the most commonly used tools in both research and clinical settings for assessing trait impulsivity. The BIS-11 measures impulsiveness through six first-order factors: attention, motor, self-control, cognitive complexity, perseverance, and cognitive instability impulsiveness. These factors are further categorized into three second-order factors: attentional, motor, and non-planning impulsiveness. The questionnaire employs a four-point Likert scale, ranging from Rarely/Never (1) to Almost Always/Always (4). [7]

Impulsive Sensation Seeking (ImpSS)

Impulsive Sensation Seeking (ImpSS) is a personality trait that reflects both impulsivity and a desire for novel, thrilling, and high-risk experiences. It originates from Marvin Zuckerman's research on personality and risk-taking behavior in the 1970s–1990s. It is a subscale of the Zuckerman-Kuhlman Personality Questionnaire (ZKPQ) and consists of two main dimensions: impulsivity—characterized by acting without thinking, difficulty delaying gratification, and making quick decisions without considering long-term consequences—and sensation seeking, which represents a preference for varied, intense, and adventurous experiences. Individuals with high ImpSS scores often engage in risk-taking behaviors such as extreme sports, substance use, and financial impulsivity, as they seek immediate gratification despite potential negative consequences. Studies indicate that ImpSS is strongly associated with behaviors like reckless driving, substance abuse, and spontaneous decision-making, highlighting its impact on risk-related actions (Zuckerman, 1994). [8]

Unsupervised Learning

Unsupervised learning originated in the 1950s–1960s from early work in statistics, pattern recognition, and AI. Key milestones include k-means clustering (Stuart Lloyd, 1957), Hebbian learning (Donald Hebb, 1949), and principal component analysis (PCA) (Karl Pearson, 1901; Harold Hotelling, 1933). In the 1980s, self-organizing maps (SOMs) by Teuvo Kohonen (1982) and Gaussian Mixture Models (GMMs) further advanced the field.

Unsupervised learning is a type of machine learning where models are trained on unlabeled data, meaning there are no predefined categories or target outputs. The goal is to identify patterns, structures, or relationships within the data without explicit supervision. Unlike supervised learning, which relies on labeled datasets to train models, unsupervised learning explores the inherent structure of the data to discover meaningful insights [9]. A fundamental concept in unsupervised learning is clustering, where data points are grouped based on similarities. Algorithms such as k-means, hierarchical clustering, and DBSCAN are commonly used to segment data into clusters, allowing for pattern recognition in various domains, including customer segmentation and anomaly detection [10]. Another key technique is dimensionality reduction, which aims to reduce the number of variables in a dataset while retaining essential information. Methods like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are widely used to simplify data visualization and preprocessing in high-dimensional spaces [11].

Hierarchical Agglomerative Clustering

Hierarchical Agglomerative Clustering (HAC) is a bottom-up clustering algorithm first developed during the 1950s and 1960s formalized with the work of S. C. Johnson (1967), who introduced strategies for hierarchical clustering in the context of psychological and biological data analysis. L. Kaufman and P. J. Rousseeuw (1990) further refined the methodology, making it a foundational technique in modern clustering applications. It is used in unsupervised learning to identify hierarchical relationships within datasets. It begins by treating each data point as an individual cluster and iteratively merges the closest pairs based on a chosen distance metric (e.g., Euclidean distance) and linkage criterion (e.g., single, complete, or average linkage). This process continues until all points are grouped into a single cluster or a predefined number of clusters is reached [12]. HAC is widely used in various fields, including psychology, genetics, and social sciences, to uncover patterns and relationships within complex data [10]. Unlike partitioning methods such as k-means, HAC does not require specifying the number of clusters in advance, making it a flexible approach for exploratory data analysis [13]. However, its computational complexity increases with large datasets, making it less efficient for big data applications [14].

Principal Component Analysis

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique in machine learning, statistics, and data analysis. It transforms high-dimensional data into a lower-dimensional form while preserving as much variance as possible. By identifying the principal components—orthogonal directions that capture the most variance in the dataset—PCA helps simplify complex data structures while minimizing information loss [15]. PCA originated from the work of Karl Pearson in 1901, who introduced it as a method for analyzing data patterns and reducing dimensionality in statistics. Later, Harold Hotelling expanded Pearson's ideas in the 1930s, applying them to multivariate analysis in a more generalized mathematical framework. Since then, PCA has become a fundamental tool in various disciplines, including psychology, genetics, finance, and image processing, influencing the development of modern machine learning techniques.

Research into the personality traits associated with high drug use has evolved significantly over the past century. Early psychoanalytic theories sought to link substance use to specific personality features, but these initial efforts lacked empirical support. The advent of multidimensional personality inventories, such as the Five-Factor Model, marked a pivotal advancement, enabling more rigorous assessments. Studies utilizing these tools have consistently found that individuals with Substance Use Disorders (SUDs) often exhibit high levels of neuroticism and openness to experience, coupled with low agreeableness and conscientiousness [16]. Recent longitudinal research has further elucidated these associations, demonstrating that certain personality traits can predict the likelihood of future drug use [17]. These insights have been instrumental in developing targeted prevention and intervention strategies, aligning closely with current efforts to identify common profile traits among high drug users.

Unsupervised learning holds significant potential in uncovering hidden patterns. It detects structures and relationships within complex, unstructured data, making it a powerful tool for analyzing human behaviors, risk factors, and societal trends. By analyzing large datasets from surveys, clinical records, or social media, unsupervised learning can reveal common characteristics among individuals at higher risk of substance misuse. These insights can significantly enhance early intervention strategies by providing a data-driven approach to identifying individuals who may require psychological support or preventive measures [18].

B. Review of other relevant research papers

1. Machine-Learning Identifies Substance-Specific Behavioral Markers

This study employed machine-learning techniques to identify multivariate behavioral markers distinguishing heroin dependence (HD) and amphetamine dependence (AD). The researchers analyzed a cohort of 39 amphetamine mono-dependent, 44 heroin mono-dependent, 58 polysubstance-dependent, and 81 non-substance-dependent individuals. Most substance-dependent participants were in protracted abstinence. The machine-learning approach effectively classified HD and AD populations, demonstrating its potential in identifying predictive markers for psychiatric disorders and classifying psychiatric populations with high-dimensional data. *PMC*. This study highlights the efficacy of machine-learning methods in distinguishing between different substance dependencies based on behavioral markers and underscores their potential in identifying specific traits associated with various forms of substance abuse [19].

2. Personality Traits and Drug Consumption: A Story Told by Data

This paper analyzed data from 1,885 respondents regarding their consumption of 18 drugs, examining personality traits using the Five-Factor Model, impulsivity, sensation seeking, and demographic information. The findings revealed that personality traits, when combined with demographic data, could predict individual drug consumption risk with sensitivity and specificity above 70% for most substances. Additionally, the study identified significant differences in personality profiles among users of different drugs. The study highlights the importance of personality traits in predicting drug consumption risk, providing a foundation for incorporating psychological factors into machine-learning models aimed at identifying high-risk individuals [20].

3. Identifying Substance Use Risk Based on Deep Neural Networks and Instagram Data

This study developed a deep-learning method to classify individuals' risk for alcohol, tobacco, and drug use based on Instagram content. By analyzing social media data, the model achieved significant accuracy in predicting substance use risk, demonstrating the potential of leveraging online behavior for substance abuse risk assessment. *Nature* (nature.com). This research illustrates the application of deep neural networks in assessing substance use risk and emphasizes the utility of social media data in identifying high-risk individuals [21].

Collectively, these studies underscore the potential of machine-learning techniques in identifying common profile traits among individuals with high drug use. They

highlight the significance of behavioral markers, personality traits, and online behaviors as predictive factors, informing the development of more targeted and effective intervention strategies.

C. Prior attempts to solve the same problem

1. **Cluster Analysis of Patient Profiles**

The study explored the application of cluster analysis to categorize individuals into different risk groups by examining key variables such as demographics (age, gender, socioeconomic background), substance usage history (type and frequency of drugs used), and mental health indicators (co-occurring disorders such as anxiety and depression). The primary objective of this research was to assist health policymakers and medical professionals in identifying high-risk populations who may require early intervention programs or targeted rehabilitation strategies [22]. By uncovering distinct patient subgroups, the study provided valuable insights that could be used to personalize treatment plans and improve the effectiveness of public health initiatives.

Despite the paper's contributions, the study faced notable limitations such as the lack of longitudinal data, which hindered the ability to track behavioral changes over time. Furthermore, while clustering algorithms effectively identified patterns within patient profiles, they could not establish causal relationships between demographic factors and substance use. The inability to infer causation limits the practical application of these findings in designing preventive interventions. Future research should consider integrating time-series clustering techniques to monitor how an individual's risk level evolves over time, allowing for more dynamic and adaptive intervention strategies [23].

A review article assessed various machine learning techniques for extracting meaningful patterns from large datasets, demonstrating the importance of feature selection methods in identifying key risk factors for substance abuse [24]. Among the methods explored, Principal Component Analysis (PCA) was utilized for dimensionality reduction, helping researchers focus on the most significant variables influencing drug abuse trends. Additionally, K-Means Clustering was employed to group individuals based on behavioral similarities, allowing researchers to identify common substance use profiles. Another clustering method, Hierarchical Agglomerative Clustering (HAC), provided insights into nested relationships between different user profiles, making it possible to analyze varying degrees of risk factors within different populations. While these techniques effectively uncovered hidden patterns in drug misuse trends, a significant drawback was the difficulty in interpreting unsupervised clusters in real-world applications. Many machine learning models lack explainability, which presents a challenge for health professionals and policymakers attempting to translate machine-generated insights into actionable interventions. To address this limitation, the study suggested adopting hybrid approaches that combine unsupervised learning with domain expertise, ensuring that findings are both data-driven and contextually relevant [25]. This integration of expert knowledge and machine learning could improve the interpretability of clustering results, making them more useful for guiding public health policies and intervention programs.

2. **Using Machine Learning Techniques to Predict People At-Risk for Drug Addiction: A Bayesian-Based Model**

The study aimed at predicting individuals at risk for drug addiction Utilizing machine learning, the study analyzed a wide range of potential variables influencing substance abuse. The researchers gathered a comprehensive dataset which incorporates demographic information, psychological traits and records, socio-economic background, and past histories of substance abuse. Various machine learning models such as decision trees, SVM, and deep learning networks were assessed to determine the most successful approach for predicting individuals at risk of drug addiction. Particularly, the study focused on identifying key patterns such as correlations between early exposure to substances, mental health conditions, and environmental factors. The findings showed success in classifying high-risk individuals with a significant degree of accuracy. This highlighted the strength of machine learning as a tool in helping healthcare professionals in developing early intervention strategies. However, while the models were effective in predicting susceptibility to drug addiction, they faced challenges in explainability and generalizability. The black-box nature of some deep learning models made it difficult to interpret how specific variables contributed to an individual's risk profile. Additionally, the dataset's limitations—such as potential biases in data collection or underreporting of substance use—highlighted the need for more diverse and representative data sources [26].

Similar studies have successfully identified, analyzed, and predicted patterns among individuals who use drugs illicitly and are at the highest risk of substance abuse. However, several challenges remain. A key limitation of unsupervised learning models is their black-box nature, making it difficult to interpret how specific variables contribute to an individual's risk profile. This lack of explainability hinders policymakers and health professionals from translating insights into effective interventions. Additionally, many studies rely on biased datasets, with underreported substance use and non-representative populations, reducing model reliability and generalizability. Unsupervised clustering techniques also fail to establish causal links between demographics and substance use, limiting their application in targeted prevention. Moreover, most studies use static datasets, preventing researchers from tracking changes in risk levels over time. To address these gaps, future research should integrate unsupervised learning with domain expertise for more interpretable results [25]. Time-series clustering could also enhance risk assessment by capturing behavioral changes over time [23].

This study aims to utilize unsupervised learning to identify and establish common demographic patterns among individuals at high risk of substance abuse. By building on existing research, we seek to address the challenges posed by the black-box nature of unsupervised learning models and enhance their interpretability.

III. Methodology

This dataset integrates demographic information, drug consumption patterns, and personality assessments based on the Five-Factor Model, providing a strong foundation for analyzing drug use risk. With well-structured data requiring minimal preprocessing, it serves as a valuable resource for evaluating substance use behaviors.

The primary objective is to identify high drug use based on the dataset's existing features. KMeans clustering was then applied to group individuals based on shared demographic and personality traits, allowing for the identification of distinct profiles linked to high drug use. This approach enables the exploration of meaningful associations between personality characteristics, demographic factors, and substance use patterns.

A. Data Collection

The Drug Consumption Dataset was retrieved from Kaggle, but it originally originates from the study "The Five Factor Model of Personality and Evaluation of Drug Consumption Risk" by E. Fehrman, A.K. Muhammad, E.M. Mirkes, V. Egan, and A.N. Gorban, which was submitted to arXiv on June 20, 2015. This dataset comprises records from 1,885 participants. For each respondent, 12 attributes were collected, including personality assessments and demographic information. Personality was measured using the NEO-FFI-R, which evaluates neuroticism, extraversion, openness, agreeableness, and conscientiousness; impulsivity was assessed via the BIS-11; and sensation seeking was captured using the ImpSS scale. Additional demographic attributes include education level, age, gender, country of residence, and ethnicity. The dataset had already been converted into numerical values, requiring no extensive data processing and allowing them to be treated as real-valued data. In addition, respondents were queried about their use of 18 drugs—covering both legal and illegal substances such as alcohol, amphetamines, amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine, and volatile substance abuse—as well as one fictitious drug, Semeron, which was included to identify over-claimers. For each drug, participants were required to select one of seven response options that indicate the recency of their usage. The available choices were: "Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day". This categorical scale captures both historical and recent consumption, allowing us to analyze patterns and trends in drug use over time.

Here is the detailed attribute information of the dataset provided in Kaggle:

1. Age: Age is the age of participant and has one of the values:

Value	Meaning	Cases	Fraction
-0.95197	18 - 24	643	34.11%
-0.07854	25 - 34	481	25.52%
0.49788	35 - 44	356	18.89%
1.09449	45 - 54	294	15.60%
1.82213	55 - 64	93	4.93%
2.59171	65+	18	0.95%

2. Gender: Gender is gender of participant:

Value	Meaning	Cases	Fraction
0.48246	Female	942	49.97%
-0.48246	Male	943	50.03%

3. Education: Education is level of education of participant and has one of the values:

Value	Meaning	Cases	Fraction
-2.43591	Left School Before 16 years	28	1.49%
-1.73790	Left School at 16 years	99	5.25%
-1.43719	Left School at 17 years	30	1.59%
-1.22751	Left School at 18 years	100	5.31%
-0.61113	Some College,No Certificate Or Degree	506	26.84%
-0.05921	Professional Certificate/ Diploma	270	14.32%
0.45468	University Degree	480	25.46%
1.16365	Masters Degree	283	15.01%
1.98437	Doctorate Degree	89	4.72%

4. Country: Country is country of current residence of participant and has one of the values:

Value	Meaning	Cases	Fraction
-0.09765	Australia	54	2.86%
0.24923	Canada	87	4.62%
-0.46841	New Zealand	5	0.27%
-0.28519	Other	118	6.26%
0.21128	Republic of Ireland	20	1.06%
0.96082	UK	1044	55.38%
-0.57009	USA	557	29.55%

5. Ethnicity: Ethnicity is ethnicity of participant and has one of the values:

Value	Meaning	Cases	Fraction
-0.50212	Asian	26	1.38%
-1.10702	Black	33	1.75%
1.90725	Mixed-Black/Asian	3	0.16%
0.12600	Mixed-White/Asian	20	1.06%
-0.22166	Mixed-White/Black	20	1.06%
0.11440	Other	63	3.34%
-0.31685	White	1720	91.25%

6. Nscore: Nscore is NEO-FFI-R Neuroticism. Neuroticism is one of the Big Five higher-order personality traits in the study of psychology. Individuals who score high on neuroticism are more likely than average to be moody and to experience such feelings as anxiety, worry, fear, anger, frustration, envy, jealousy, guilt, depressed mood, and loneliness. Possible values are presented in table below:

Nscore	Value	Nscore	Value	Nscore	Value	Nscore	Value
12	-3.46436	24	-1.32828	36	0.04257	48	1.23461
13	-3.15735	25	-1.19430	37	0.13606	49	1.37297
14	-2.75696	26	-1.05308	38	0.22393	50	1.49158
15	-2.52197	27	-0.92104	39	0.31287	51	1.60383
16	-2.42317	28	-0.79151	40	0.41667	52	1.72012
17	-2.34360	29	-0.67825	41	0.52135	53	1.83990
18	-2.21844	30	-0.58016	42	0.62967	54	1.98437
19	-2.05048	31	-0.46725	43	0.73545	55	2.12700
20	-1.86962	32	-0.34799	44	0.82562	56	2.28554
21	-1.69163	33	-0.24649	45	0.91093	57	2.46262
22	-1.55078	34	-0.14882	46	1.02119	58	2.61139
23	-1.43907	35	-0.05188	47	1.13281	59	2.82196
-	-	-	-	-	-	60	3.27393

7. EScore: Escore (Real) is NEO-FFI-R Extraversion. Extraversion is one of the five personality traits of the Big Five personality theory. It indicates how outgoing and social a person is. A person who scores high in extraversion on a personality test is the life of the party. They enjoy being with people, participating in social

gatherings, and are full of energy. Possible values are presented in table below:

Escore	Value	Escore	Value	Escore	Value	Escore	Value
16	-3.27393	27	-1.76250	38	-0.30033	49	1.45421
17	-3.00537	28	-1.63340	39	-0.15487	50	1.58487
18	-3.00537	29	-1.50796	40	0.00332	51	1.74091
19	-2.72827	30	-1.37639	41	0.16767	52	1.93886
20	-2.53830	31	-1.23177	42	0.32197	53	2.12700
21	-2.44904	32	-1.09207	43	0.47617	54	2.32338
22	-2.32338	33	-0.94779	44	0.63779	55	2.57309
23	-2.21069	34	-0.80615	45	0.80523	56	2.85950
24	-2.11437	35	-0.69509	46	0.96248	57	2.85950
25	-2.03972	36	-0.57545	47	1.11406	58	3.00537
26	-1.92173	37	-0.43999	48	1.28610	59	3.27393

8. Oscore: Oscore (Real) is NEO-FFI-R Openness to experience. Openness is one of the five personality traits of the Big Five personality theory. It indicates how open-minded a person is. A person with a high level of openness to experience in a personality test enjoys trying new things. They are imaginative, curious, and open-minded. Individuals who are low in openness to experience would rather not try new things. They are close-minded, literal and enjoy having a routine. Possible values are presented in table below:

Oscore	Value	Oscore	Value	Oscore	Value
24	-3.27393	38	-1.11902	50	0.58331
26	-2.85950	39	-0.97631	51	0.72330
28	-2.63199	40	-0.84732	52	0.88309
29	-2.39883	41	-0.71727	53	1.06238
30	-2.21069	42	-0.58331	54	1.24033
31	-2.09015	43	-0.45174	55	1.43533
32	-1.97495	44	-0.31776	56	1.65653
33	-1.82919	45	-0.17779	57	1.88511
34	-1.68062	46	-0.01928	58	1.15324
35	-1.55521	47	0.14143	59	2.44904
36	-1.42424	48	0.29338	60	2.90161
37	-1.27553	49	0.44585	NaN	NaN

9. Ascore: Ascore(Real) is NEO-FFI-R Agreeableness. Agreeableness is one of the five personality traits of the Big Five personality theory. A person with a high level of agreeableness in a personality test is usually warm, friendly, and tactful. They generally have an optimistic view of human nature and get along well with others. Possible values are presented in table below:

Ascore	Value	Ascore	Value	Ascore	Value
12	-3.46436	34	-1.34289	48	0.76096
16	-3.15735	35	-1.21213	49	0.94156
18	-3.00537	36	-1.07533	50	1.11406
23	-2.90161	37	-0.91699	51	1.2861
24	-2.78793	38	-0.76096	52	1.45039
25	-2.70172	39	-0.60633	53	1.61108
26	-2.53830	40	-0.45321	54	1.81866
27	-2.35413	41	-0.30172	55	2.03972
28	-2.21844	42	-0.15487	56	2.23427
29	-2.07848	43	-0.01729	57	2.46262
30	-1.92595	44	0.13136	58	2.75696
31	-1.77200	45	0.28783	59	3.15735
32	-1.62090	46	0.43852	60	3.46436
33	-1.47955	47	0.59042	NaN	NaN

10. Cscore: Cscore (Real) is NEO-FFI-R Conscientiousness. Conscientiousness is one of the five personality traits of the Big Five personality theory. A person scoring high in conscientiousness usually has a high level of self-discipline. These individuals prefer to follow a plan, rather than act spontaneously. Their methodic planning and perseverance usually makes them highly successful in their chosen occupation. Possible values are presented in table below:

Cscore	Value	Cscore	Value	Cscore	Value
17	-3.46436	32	-1.25773	46	0.58489
19	-3.15735	33	-1.13788	47	0.7583
20	-2.90161	34	-1.01450	48	0.93949
21	-2.72827	35	-0.89891	49	1.13407
22	-2.57309	36	-0.78155	50	1.30612
23	-2.42317	37	-0.65253	51	1.46191
24	-2.30408	38	-0.52745	52	1.63088
25	-2.18109	39	-0.40581	53	1.81175
26	-2.04506	40	-0.27607	54	2.04506
27	-1.92173	41	-0.14277	55	2.33337
28	-1.78169	42	-0.00665	56	2.63199
29	-1.64101	43	0.12331	57	3.00537
30	-1.51840	44	0.25953	59	3.46436
31	-1.38502	45	0.41594	NaN	NaN

11. Impulsive: Impulsive (Real) is impulsiveness measured by BIS-11. In psychology, impulsivity (or impulsiveness) is a tendency to act on a whim, displaying behavior characterized by little or no forethought, reflection, or consideration of the consequences. If you describe someone as impulsive, you mean that they do things suddenly without thinking about them carefully first. Possible values are presented in table below:

Impulsiveness	Cases	Fraction
-2.55524	20	1.06%
-1.37983	276	14.64%
-0.71126	307	16.29%
-0.21712	355	18.83%
0.19268	257	13.63%
0.52975	216	11.46%
0.88113	195	10.34%
1.29221	148	7.85%
1.86203	104	5.52%
2.90161	7	0.37%

12. Sensation: SS(Real) is sensation seeing measured by ImpSS. Sensation is input about the physical world obtained by our sensory receptors, and perception is the process by which the brain selects, organizes, and interprets these sensations. In other words, senses are the physiological basis of perception. Possible values are presented in table below:

SS	Cases	Fraction
-2.07848	71	3.77%
-1.54858	87	4.62%
-1.18084	132	7.00%
-0.84637	169	8.97%
-0.52593	211	11.19%
-0.21575	223	11.83%
0.07987	219	11.62%
0.40148	249	13.21%
0.76540	211	11.19%
1.22470	210	11.14%
1.92173	103	5.46%

13. The 18 Legal and Illegal Drugs are as follows:

- Alcohol: a psychoactive and toxic substance that can cause addiction
- Amphetamines: stimulant drugs, makes the message travel fast between the brain and the body.
- Amyl (nitrite): a depressant that slows down the travelling of messages from the brain and the body.
- Benzos (benzodiazepine): medications that slow down the activity in your brain and nervous system.
- Cannabis: also known as marijuana, causes changes in mood, thoughts, and perception of reality.
- Choc (chocolate): contains a significant amount of sugar, along with two neuroactive drugs, caffeine and theobromine.
- Coke: a stimulant drug, makes the message travel fast between the brain and the body.
- Crack: highly addictive and stimulant, derived from powdered cocaine.
- Ecstasy: a stimulant drug that can cause hallucinations.
- Heroin: depressant drug, down the activity in your brain and nervous system
- Ketamine: a dissociative anesthetic that has some
- hallucinogenic effects.
- LegalH (legal highs): substances which mimic the effects of drugs such as cocaine and ecstasy
- LSD (lysergic acid diethylamide): a psychedelic drug, it can affect all senses, altering a person's thinking, sense of time and emotions.
- Meth (methodone): used to reduce withdrawal symptoms in people addicted to heroin or other narcotic drugs, and it can also be used as a pain reliever.
- Mushrooms: a psychedelic drug, it can affect all senses, altering a person's thinking, sense of time and emotions.
- Nicotine: highly addictive stimulant found in tobacco and vaping devices.
- Semer: fictitious drug.
- VSA (volatile substance abuse): deliberate inhalation of substances for their intoxicating effects.

These columns (the legal and illegal drugs) are divided into 7 classes:

Value	Description
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

B. Data Pre-Processing

1. Data Cleaning
    a. The preprocessing phase began with loading the dataset and inspecting it for any inconsistencies. Duplicate entries were reviewed and null values were checked.
    b. The column 'ID' was dropped as it holds no relevance in the experimentation.
    c. Instances where the 'Semer' variable had values of 'CL1', 'CL2', 'CL3', or 'CL4'

were removed to ensure data consistency and relevance. The mentioned column was later then dropped as it is not needed.

d. Drug consumption variables were encoded to facilitate numerical analysis. A set of target columns representing different substances was identified, including alcohol, cannabis, nicotine, and various illicit drugs. A mapping dictionary was then created to convert categorical drug use levels (CL0 to CL6) into numerical values, ensuring consistency and enabling meaningful comparisons. This transformation was applied across all identified drug-related columns to standardize the dataset for further analysis.

2. Feature Engineering

a. To quantify drug consumption, substances were categorized into legal substances (e.g., alcohol, cannabis, nicotine) and illicit drugs (e.g., amphetamines, cocaine, heroin). A scoring system was implemented to measure usage intensity:

- legal_highs: Counts the number of legal substances an individual consumes frequently (usage level ≥ 4, meaning those drugs that are used last month to recent).
- illicit_occ: Counts the number of illicit drugs an individual uses at least occasionally (usage level ≥ 3, meaning those drugs that are used last year to recent).
- A composite risk score was calculated by summing both counts, providing an overall measure of substance use intensity.

To classify high drug users, a threshold of 8 was set. Individuals with a composite risk score of 8 or higher were categorized as engaging in high drug use, distinguishing those with extensive substance use patterns for further analysis.

b. The classification was converted into a binary format for consistency in analysis:

- 0: Individuals not classified as high drug users.
- 1: Individuals classified as high drug users.

This will be used for profiling clusters by analyzing the proportion of high drug users within each group. By mapping the classification into a binary format, the study ensures a clear distinction between high and low-moderate drug users, enabling more effective interpretation of cluster characteristics. This binary labeling allows for comparative analysis across clusters, helping identify the demographic and personality traits most associated with high drug use.

3. Feature Selection

The selection of features for dimensionality reduction was based on the research focus. To identify common profile traits among individuals, only the variables capturing demographics and personality characteristics, along with engineered columns (namely, legal_high and illicit_occ) that summarize drug usage frequency into a composite risk score were stored in a variable. This feature set was then fed into PCA to reveal the underlying structure of the data, with the number of principal components determined based on the scree plot.

4. Dimensionality Reduction

a. The covariance matrix was computed to assess feature relationships, followed by eigenvalue extraction to determine variance distribution. A scree plot visualized the sorted eigenvalues, helping identify the optimal number of principal components. The explained variance ratio guided the selection of components that retained most of the dataset's information while reducing dimensionality.

b. Based on the scree plot, two principal components (n_components = 2) were selected. PCA was applied to a subset of features, transforming the data into a lower-dimensional space while preserving variance. The resulting PCA dataset captures the most important traits for clustering and further analysis.

5. Scaling

There was no need to do any scaling techniques as the features with continuous values (e.g., age, score, impulsivity,..) were already scaled.

C. Experimental Setup

Tools and Frameworks Used

The experimentation was conducted on Google Colab that runs on Python programming language, utilizing cloud-based resources, along with several key libraries and frameworks for data manipulation and machine learning. The following tools were utilized:

- pandas version: 2.2.2
- NumPy version: 1.26.4
- seaborn version: 0.13.2
- scikit-learn version: 1.6.1
- UMAP version: 0.5.7
- yellowbrick version: 1.5
- kagglehub version: 0.3.10
- matplotlib version: 3.10.0

Clustering Techniques Considered

Multiple clustering methods were explored to determine the most effective approach

for profiling individuals based on personality traits and demographics:

- Agglomerative Clustering (HAC) was ultimately selected for the final analysis because it offered well-defined clusters and yielded higher silhouette scores compared to the other methods.
- K-Means was tested but produced lower silhouette scores and less distinct group separations for this dataset.
- DBSCAN resulted in a large proportion of points classified as noise, making it less suitable for these data.

Thus, HAC was selected for final analysis, while DBSCAN and K-Means were excluded from further profiling.

Hyperparameters Used

- Agglomerative Clustering was configured with the following hyperparameters:
  - n_clusters=2: The algorithm will partition the data into two clusters.
  - linkage='ward': Ward's method is used to minimize the total within-cluster variance, producing compact and well-separated clusters.
  - compute_distances=True: This setting enables the computation of distances between clusters, which can be used for further analysis (e.g., dendrogram visualization).
- Principal Component Analysis (PCA): 2 components (n_components=2), selected based on the scree plot analysis.

D. Algorithm

Hierarchical Agglomerative Clustering (HAC) Algorithm

Hierarchical Agglomerative Clustering is an unsupervised machine learning method that builds a hierarchy of clusters in a bottom-up fashion. Initially, each data point is treated as its own cluster, and the algorithm iteratively merges the most similar clusters based on a chosen distance metric and linkage criterion until a stopping condition is met [27]. The process is visualized through a dendrogram, which can be used to determine the optimal number of clusters. For this study, HAC was used to group individuals based on their demographic and personality traits in order to identify common profiles among high drug users.

HAC Application

1. Feature Selection: Only demographic and personality features were used as input for clustering, while the original drug usage labels were dropped to ensure that the grouping reflects inherent profile traits rather than direct usage patterns.
2. Dimensionality Reduction: Principal Component Analysis (PCA) was applied prior to clustering to reduce noise and improve computational efficiency.
3. Determining Optimal Clusters: The dendrogram and evaluation metrics (such as the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index) guided the selection of the ideal number of clusters, resulting in a two-cluster solution.
4. Cluster Interpretation: Following clustering, the proportion of high drug users was analyzed within each cluster to determine which profile traits were most associated with elevated substance use.

E. Training Procedure

The training process began with Principal Component Analysis (PCA) to reduce dimensionality while retaining essential variance in the dataset. PCA was applied to demographic and personality trait features, ensuring efficient clustering and minimizing noise. The optimal number of components was determined using a scree plot and cumulative variance analysis.

Next, the ideal number of clusters was determined by analyzing the dendrogram generated during preliminary exploration. The final Hierarchical Agglomerative Clustering (HAC) model was then configured using Ward's linkage—minimizing within-cluster variance—with n_clusters set to 2 and compute_distances=True to facilitate further analysis of inter-cluster distances.

Finally, the overall clustering performance was rigorously assessed using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index to confirm that the clusters were cohesive, compact, and well-separated.

F. Evaluation Metrics

To assess clustering performance, the following metrics were used:

1. Silhouette Score: measures how similar an object is to its own cluster compared to other clusters, with higher scores indicating well-separated and cohesive clusters. It is standard in the field because it provides an intuitive measure of both intra-cluster cohesion and inter-cluster separation. It ranges from -1 to 1; values closer to 1 indicate well-separated, cohesive clusters [28]. Equation 1 shows the formula for Silhouette Score.

$$S(i) = \frac{(b(i)-a(i))}{max(a(i),b(i))}$$

Equation 1. *Silhouette Score*

Where,

- a(i) is the average distance from i to other data points in the same cluster.
- b(i) is the smallest average distance from i to data points in a different cluster.

2. Davies-Bouldin Index: quantifies the average similarity between each cluster and its most similar one, is another common metric—lower values indicate better clustering since they reflect smaller within-cluster distances relative to between-cluster distances [28]. Equation 2 shows the formula for Davies-Bouldin Index.

$$DB = \left(\frac{1}{n}\right) \sum max(R_{ij})$$

Equation 2. *Davies-Bouldin Index*

Where,

- n is the number of clusters.
- Rij is a measure of dissimilarity between cluster i and the cluster most similar to i.

3. Calinski-Harabasz Index: evaluates the ratio of between-cluster dispersion to within-cluster dispersion; higher values suggest that the clusters are more distinct and well-separated [28]. Equation 3 shows the formula for Calinski-Harabasz Index.

$$CH = \left(\left(\frac{B}{W}\right) * \left(\frac{N-K}{K-1}\right)\right)$$

Equation 2. Calinski-Harabasz Index

Where,

- B is the sum of squares between clusters.
- W is the sum of squares within clusters.
- N is the total number of data points.
- K is the number of clusters.

The B and W are calculated as:

- Calculating between group sum of squares (B)

$$B = \sum_{k=1}^{K} n_k \times ||C_k - C||^2$$

Where,

  - nk is the number of observation in cluster 'k'
  - Ck is the centroid of cluster 'k'
  - C is the centroid of the dataset
  - K is number of clusters

- Calculating within the group sum of squares (W)

$$W = \sum_{i=1}^{n_k} ||X_{ik} - C_k||^2$$

Where,

  - nk is the number of observation in cluster 'k'
  - Xik is the i-th observation of cluster 'k'
  - Ck is the centroid of cluster 'k'

These metrics were selected because they each capture a different facet of clustering quality and are widely accepted benchmarks in unsupervised learning. Results were measured by computing these metrics across various clustering configurations, allowing for a comprehensive comparison.

G. Comparison of Clustering Algorithms

Three clustering algorithms; K-Means, Agglomerative Clustering, and DBSCAN were evaluated to identify the most effective model for grouping individuals based on demographic and personality traits. Each algorithm was assessed using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Agglomerative Clustering emerged as the best-performing model, achieving higher silhouette and Calinski-Harabasz scores and a lower Davies-Bouldin Index, which indicates more cohesive and distinct clusters.

In summary, the dataset obtained from Kaggle was originally from an existing study about The Five-Factor Model of Personality and Evaluation of Drug Consumption Risk, requiring minimal cleaning and preprocessing. Hierarchical Agglomerative Clustering was chosen for its better performance in comparison with K-Means and DBSCAN. Google Colab was utilized for the experimentation along with various tools and libraries such as pandas, NumPy,scikit-learn, and many more. The chosen algorithm was assessed by the Silhouette and Calinski-Harabasz Index scores, Davies-Bouldin Index and stability was also measured by ARI. The well-defined methodology ensures that the experimentation is fully reproducible.

IV. Results and Discussion

Data Exploration:

Figure I. *Distribution of Composite Risk*

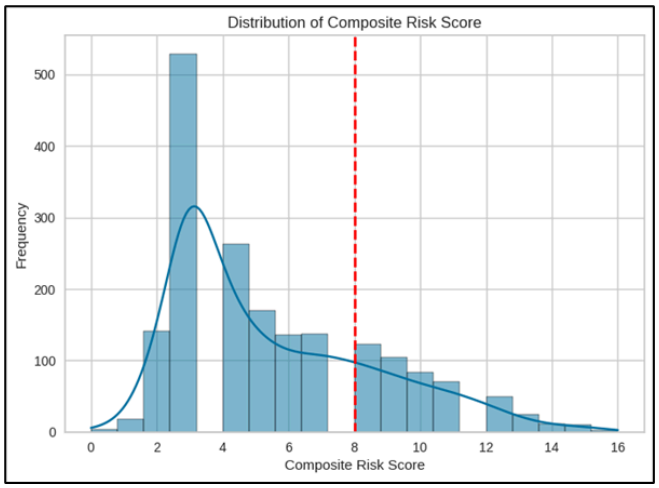

Figure I displays a histogram of composite risk scores based on the frequency of legal and illicit substance use. A red dashed line marks the threshold of 8, above which 480 respondents are classified as high drug users, while the remaining 1,397 are categorized as low–moderate users. The KDE curve peaks around scores of 2 to 3, and the right-skewed shape supports the idea that fewer people engage in high-risk behaviors, indicating that high-risk drug use is relatively uncommon. This classification helps identify individuals with extensive substance use patterns for further analysis.

Principal Component Analysis (PCA) for Dimensionality Reduction:

Figure II. *Scree Plot*

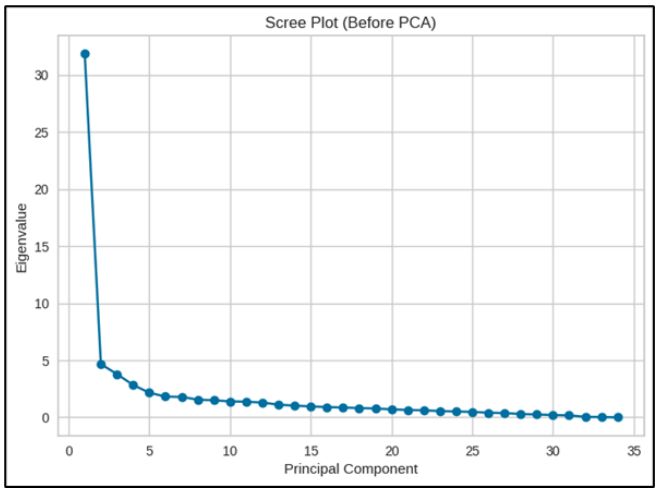

Figure II shows a scree plot indicating the variance explained by each principal component (PC). PC1 has the highest eigenvalue, capturing about 47% of the total variance, while PC2 adds enough to bring the total to approximately 54%. After PC1, the plot drops sharply and levels off around PC3–PC4, indicating diminishing returns from additional components. For this study, PC1 and PC2 were retained to reduce the dataset to two dimensions, allowing straightforward 2D visualization of clusters. PC1 was chosen for containing the largest portion of variance, and PC2 for capturing the next largest chunk, aligning with the "elbow" point where subsequent PCs contribute relatively little additional variance.

Clustering Analysis:

Figure III. *Distortion Score Elbow for K-Means Clustering*

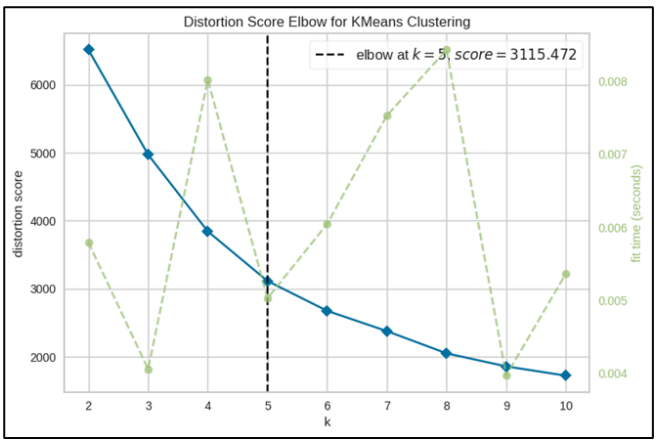

Figure III illustrates the elbow plot for determining the optimal number of clusters in K-Means. While the elbow method suggested k = 5 as a balance between cluster quality and model simplicity, applying k = 5 did not produce clearly distinct clusters. Instead, it yielded results like those with fewer clusters and did not meet the desired outcome. Consequently, k = 3 was chosen, as it provided a more satisfactory separation of clusters for this analysis.

Table I. *Additional Metrics*

K-means	Average silhouette score	Davies-Bouldin Index	Calinski-Harabasz Index
k = 3	0.37419320407038503	0.96106317908564044	2047.1231064241476
k = 5	0.35568692990679898	0.93053407174378121	1931.4406972169043

Table I compares clustering performance for k=3 and k=5. With k=3, the silhouette score is slightly higher, indicating that clusters are more cohesive, and the Calinski-Harabasz score is also higher, suggesting a more distinct cluster structure. Although k=5 has a marginally lower Davies-Bouldin Index, the difference is minimal. Overall, these metrics favor k=3 because it provides a better balance of cohesion and separation, resulting in a more meaningful and interpretable clustering of the data.

Figure IV. *Dendrogram of Drug Consumption Data*

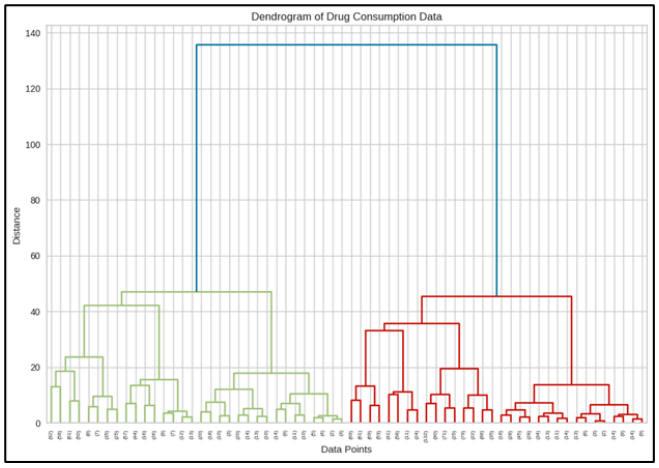

Figure IV presents a dendrogram illustrating how the drug consumption data points are progressively grouped using Ward's linkage. The vertical axis shows the distance at which clusters merge, and the large gap near the top suggests two distinctly separate main clusters. By "cutting" the dendrogram at this height, two primary clusters emerge that differ in drug usage patterns. These insights guided the choice to use two clusters in agglomerative clustering.

Table II. *Clustering Performance Evaluation Table*

Clustering Algorithm	Average silhouette score	Davies-Bouldin Index	Calinski-Harabasz Index
K-Means	0.3741932040703850 3	0.9610631790856404	2047.1231064241476
Agglomerative	0.5163926020519609	0.7482286599579980	2562.8299024365415
DBSCAN	0.0870877656824798 5	0.6049286080318446	19.661551999937714

Table II shows that Agglomerative clustering outperforms the other methods, achieving the highest average silhouette score (0.516) and Calinski-Harabasz index (2562.83), which indicate more cohesive and distinct clusters. Although DBSCAN has the lowest Davies-Bouldin Index (0.605), its very low silhouette score (0.087) and Calinski-Harabasz index (19.66) suggest that its clusters are poorly defined. K-Means, with a silhouette score of 0.374 and a Calinski-Harabasz index of 2047.12, falls between the two. Overall, these results support the chosen clustering algorithm for the drug consumption dataset.

Clustering Evaluation:

Figure V. *3D Scatter Plots Comparing KMeans, Agglomerative, and DBSCAN Clustering*

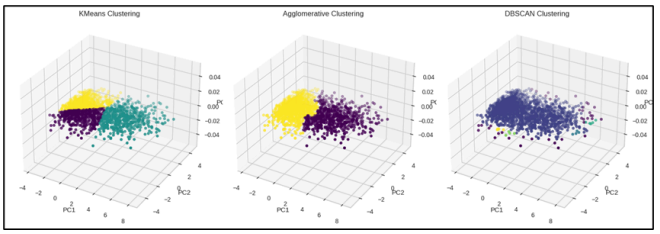

Figure V compares the clustering outcomes of KMeans, Agglomerative, and DBSCAN in a reduced feature space (the first three principal components). The plots show that Agglomerative clustering produces distinct, well-separated clusters, consistent with its higher silhouette and Calinski-Harabasz scores. In contrast, KMeans yields moderately cohesive clusters, and DBSCAN's low silhouette score indicates poorly separated groups, likely due to suboptimal parameter settings. These visualizations demonstrate that Agglomerative is the most interpretable method for further analysis of the drug consumption dataset, while DBSCAN may still be useful for outlier detection.

Figure VI. *Distribution of the Clusters*

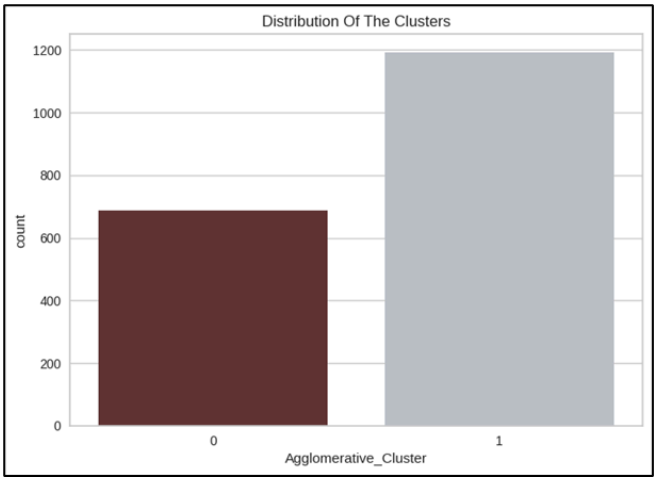

Figure VI shows the clustering solution, Cluster 0 includes fewer individuals, but about 70% are high drug users, identifying it as a high-risk group. In contrast, Cluster 1 consists of most respondents, with over 99% showing low to moderate drug use. This clear separation shows that Agglomerative Clustering effectively distinguishes a small, high-risk group from a larger, lower-risk group, supporting its use for analyzing drug consumption patterns.

Profiling with Agglomerative Clustering:

Figure VII. *Age Distribution (High Drug Use Only)*

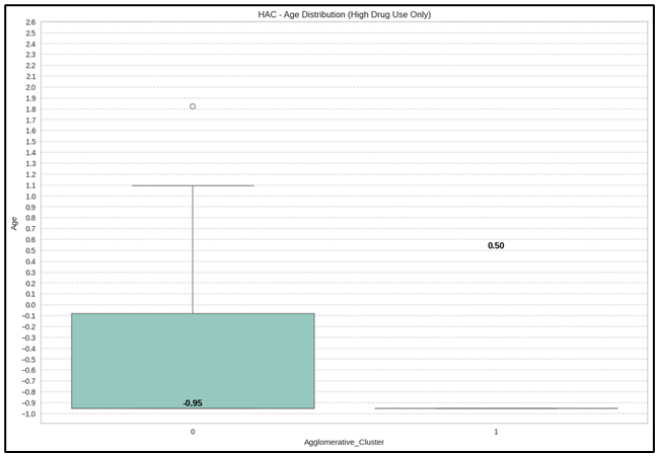

Figure VIII. *Age Distribution (Low-Moderate Drug Use Only)*

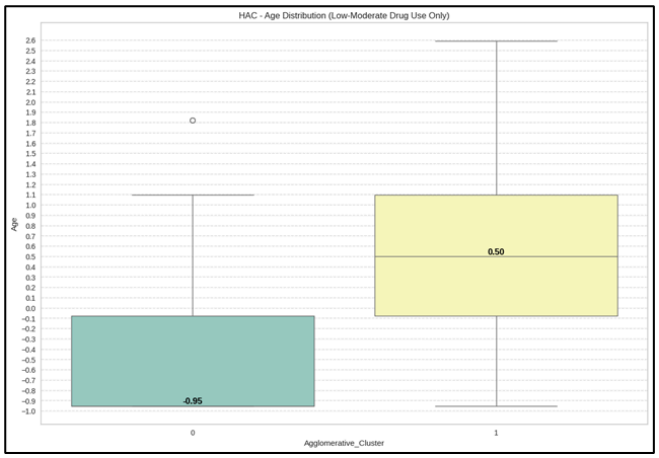

Figure VII & VIII illustrates an age-based trend in both high drug use and low moderate use clusters. Cluster 0 primarily includes younger adults aged 18 to 34, indicated by z-scores near -0.95197 (approximately 18 to 24) and -0.07854 (approximately 25 to 34). Although Cluster 1 is less visible in the high drug use plot due to scale, it generally encompasses older individuals (25 to 54) in the low moderate group, with z-scores near 0.49788 (about 35 to 44) and 1.09449 (about 45 to 54). This distribution highlights how age serves as a key factor in distinguishing high risk from low moderate drug users, demonstrating that demographic profiling, particularly age, is integral to understanding drug consumption behavior.

Figure IX. *Gender Distribution (High Drug Use Only)*

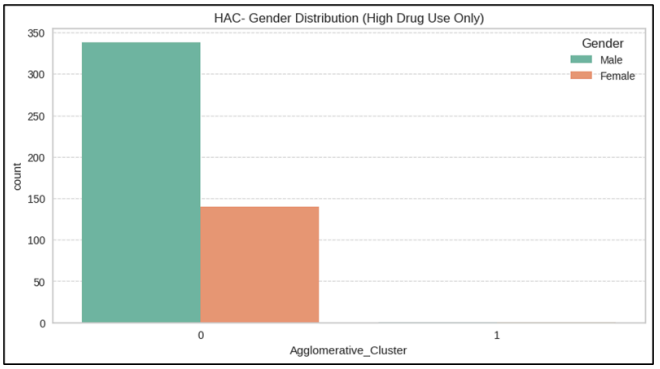

Figure X. *Gender Distribution (Low-Moderate Drug Use Only)*

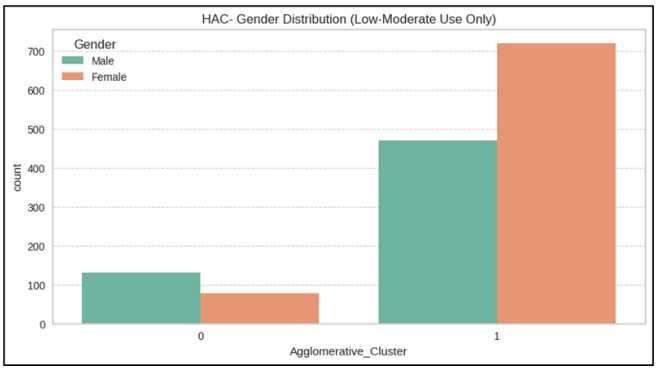

Figure IX & X present a bar chart illustrating a clear gender-based trend in both high drug use and low moderate use clusters. In the high drug use subset, Cluster 0 is predominantly male, while Cluster 1 is less visible due to scale. In the low moderate use subset, Cluster 0 remains male-dominated but includes fewer individuals, whereas Cluster 1 is mostly female. This pattern suggests that men are more likely to fall into high drug use, and women are more prevalent in low moderate use. The encoded gender values (about -0.48246 for male and 0.48246 for female) further confirm this demographic distinction, highlighting the importance of gender profiling in understanding drug consumption behavior.

Figure XI. *Education Distribution (High Drug Use Only)*

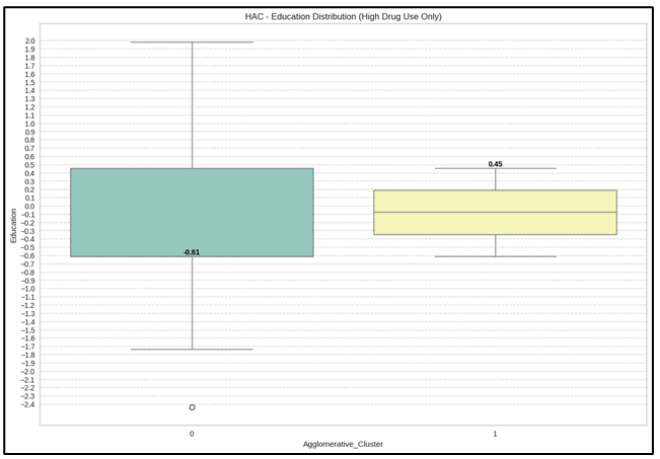

Figure XII. *Education Distribution (Low-Moderate Drug Use Only)*

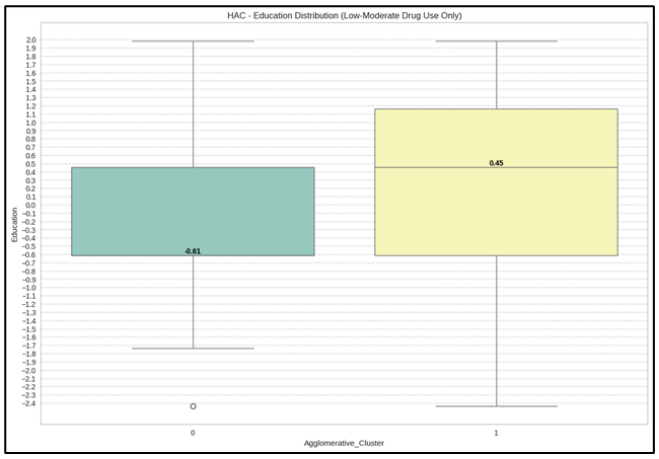

Figure XI and XII present box plots illustrating educational attainment across high drug use and low–moderate use groups. In the high drug use subset, Cluster 0 predominantly includes individuals with a university degree, a professional certificate or diploma, or some college experience, while Cluster 1 tends to include those holding a professional certificate or diploma. In the low–moderate group, Cluster 0 shows similar educational levels, but Cluster 1 spans a broader range, including master's degrees. The numeric scale assigns 1.16365 to a master's degree, 0.45468 to a university degree, -0.05921 to a professional certificate or diploma, and -0.61113 to some college with no certificate or degree. Although advanced education appears slightly more common in the low–moderate cluster, the analysis did not reveal a strong or clear difference in educational attainment between the two groups.

Figure XIII. *Country Distribution (High Drug Use Only)*

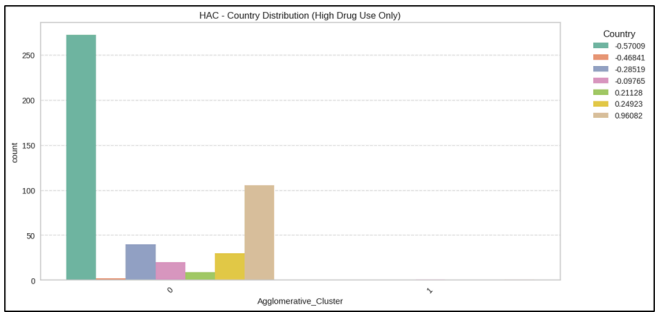

Figure XIV. *Country Distribution (Low-Moderate Drug Use Only)*

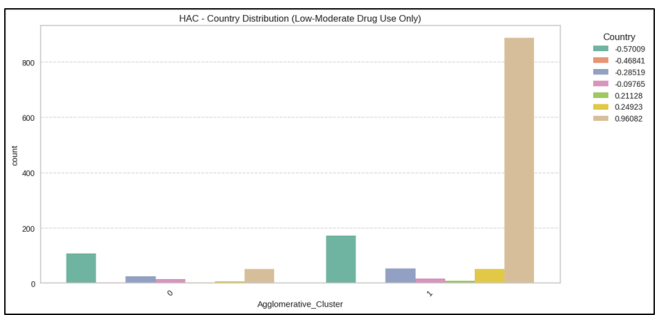

Figure XIII and XIV present bar charts that compare country distributions for both high drug use and low–moderate drug use subsets. Overall, the dataset is dominated by participants from the UK (about 55%), followed by the USA (about 30%), with smaller representations from Australia, Canada, New Zealand, the Republic of Ireland, and others. In the high drug use subset, Cluster 0 appears to include more participants from the USA, while Cluster 1 includes fewer respondents overall. In the low–moderate subset, Cluster 1 is heavily composed of UK participants, whereas Cluster 0 features a mix of countries at lower counts. Although the UK and USA comprise the largest shares in both categories, these figures alone do not establish a strong correlation between country and drug use level. Further statistical analysis would be needed to determine whether nationality significantly influences drug consumption patterns.

Figure XV. *Ethnicity Distribution (High Drug Use Only)*

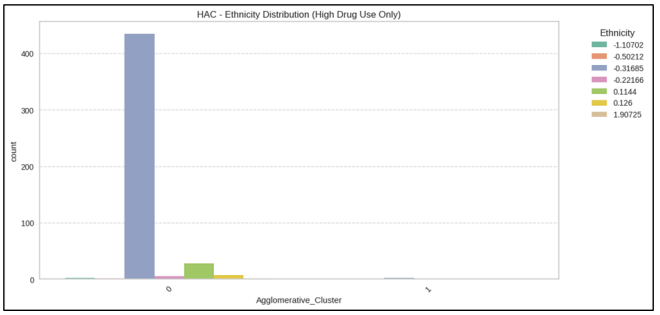

Figure XVI. *Ethnicity Distribution (Low-Moderate Drug Use Only)*

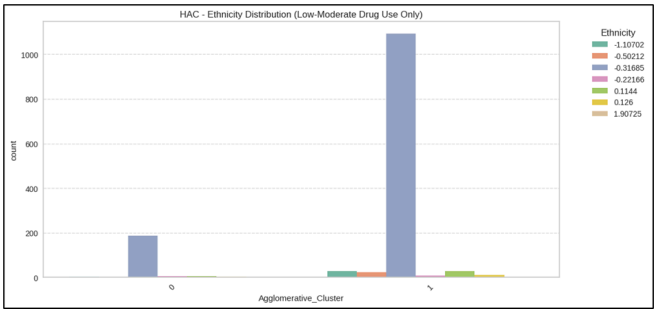

Figure XV and XVI present bar plots that reveal White participants (z-score = –0.31685) dominate the dataset, comprising about 91.25% of all respondents. Consequently, both high-drug-use (Cluster 0 in the first plot) and low-to-moderate-use clusters (Cluster 0 and 1 in the second plot) are primarily composed of White individuals. While other ethnic groups (e.g., Asian, Black, and mixed categories) are present, their smaller representation is less visible in the figures. These results indicate that, within this dataset, the White majority largely defines the clustering patterns for drug consumption, highlighting a potential sampling imbalance and the need for caution when generalizing findings to more diverse populations.

Figure XVII. *NScoreDistribution (High Drug Use Only)*

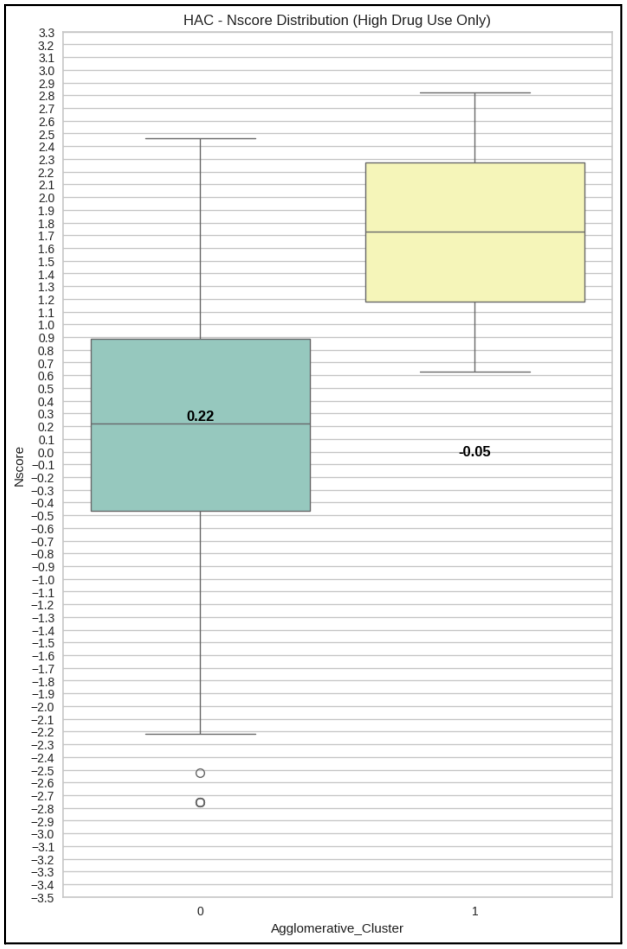

Figure XVIII. *NScoreDistribution (Low-Moderate Drug Use Only)*

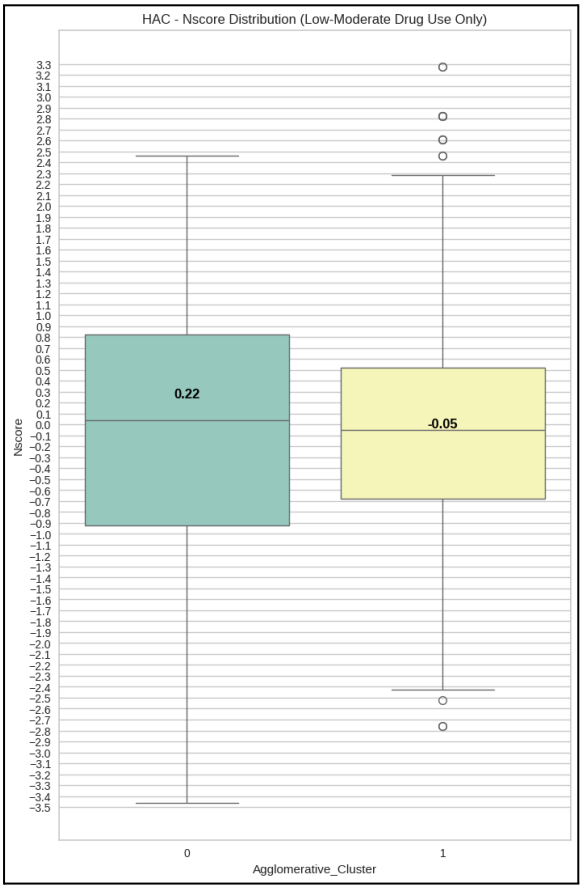

Figure XVII and XVIII presents a bar plot among high drug users, two distinct subgroups emerged based on neuroticism: one group had moderately elevated Neuroticism scores ranging from 30 to 45, while another group had significantly higher scores between 47 and 56. This pattern suggests that individuals with the highest scores may be more emotionally reactive and potentially more vulnerable to stress, which could drive their heavier substance use. In contrast, low and moderate drug users showed more overlap in their scores, with one group ranging from 27 to 44 and the other from 28 to 41. This indicates that factors beyond Neuroticism, such as social influences or other personality traits, may play a larger role in their drug use patterns.

Figure XIX. *EScore Distribution (High Drug Use Only)*

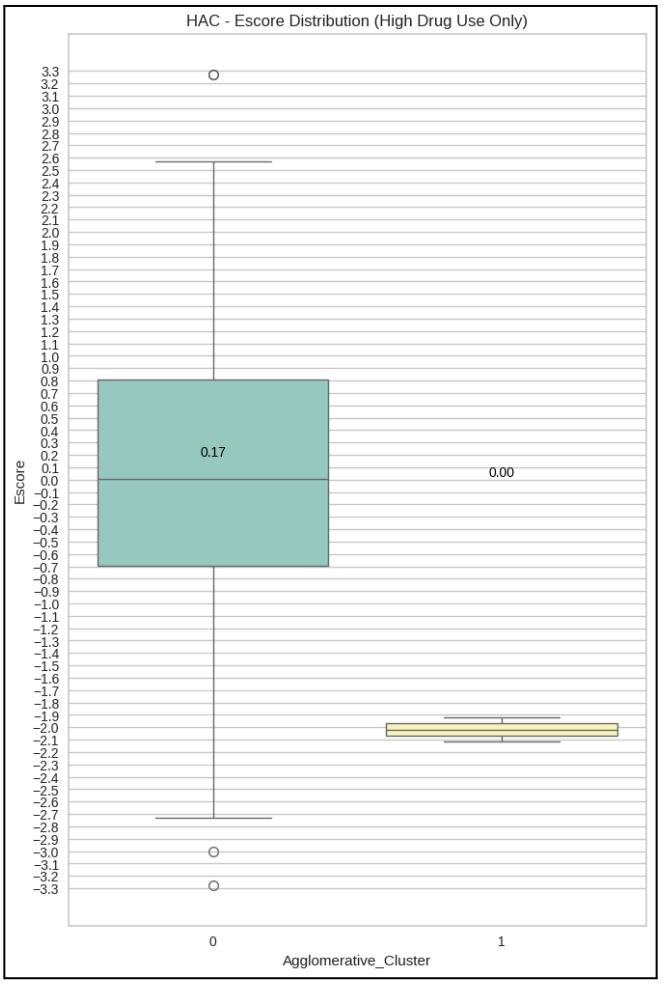

Figure XX. *EScore Distribution (Low-Moderate Drug Use Only)*

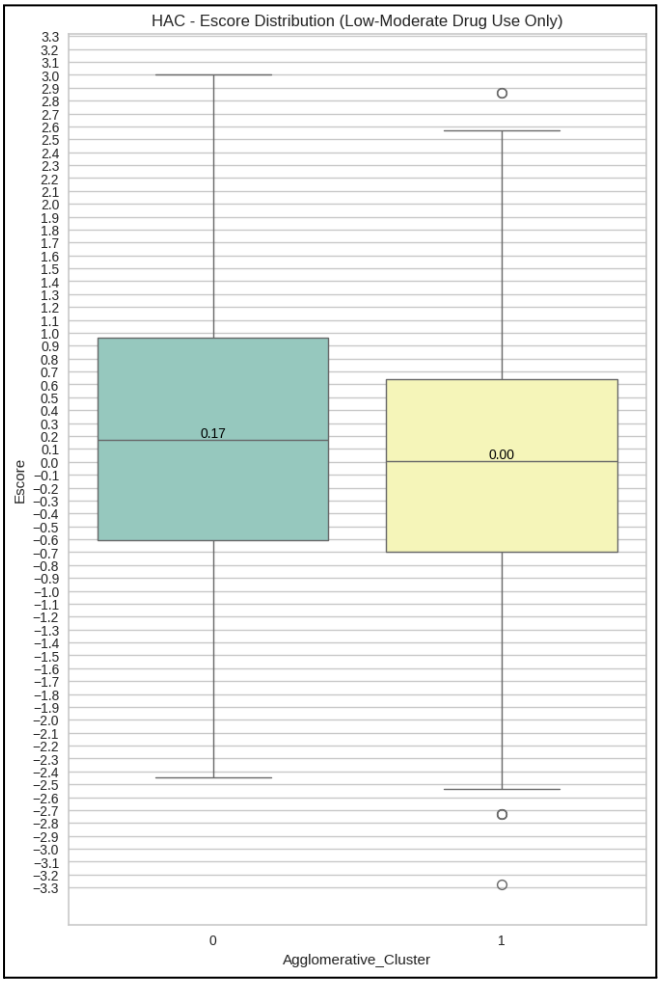

Figure XIX and XX presents a bar plot among high drug users, two distinct subgroups emerged based on extraversion: one group showed moderate to high scores (35 to 45), while the other had noticeably lower scores (24 to 25). This suggests that some heavy users may be more outgoing and socially engaged, while others are more reserved. Among low to moderate drug users, both clusters exhibited similar extraversion ranges (35 to 46 and 35 to 44), indicating that extraversion may not be the primary factor distinguishing their usage patterns. Instead, other personality traits or contextual influences could be more important in explaining their substance use behaviors.

Figure XXI. *OScore Distribution (High Drug Use Only)*

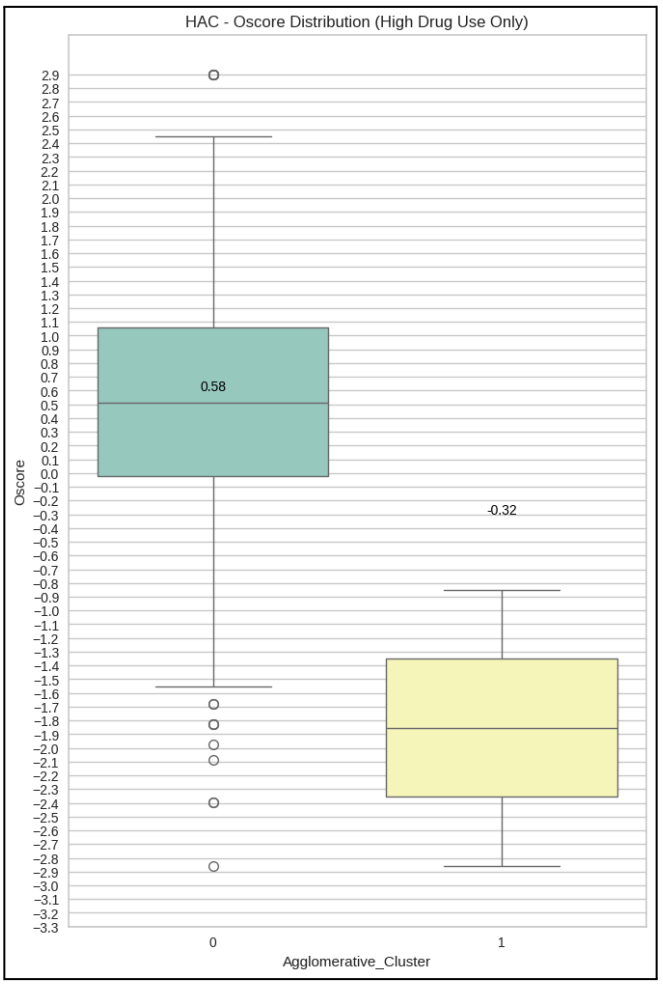

Figure XXII. *OScore Distribution (Low-Moderate Drug Use Only)*

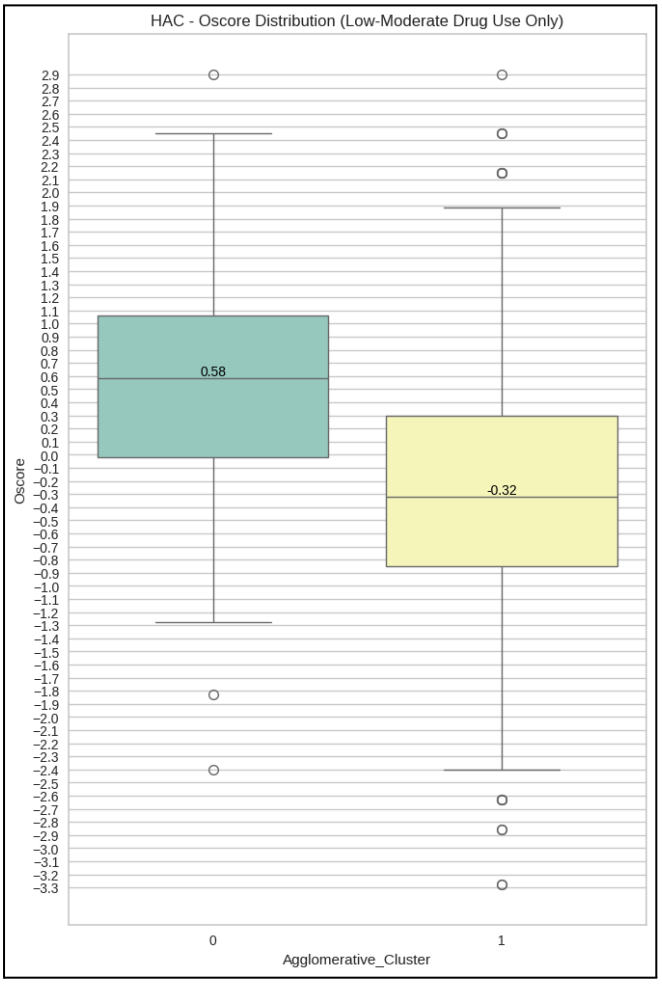

Figure XXI and XXII presents a bar plot among high drug users, one subgroup shows notably higher openness (OScore 47–53) compared to another with lower scores (29–36), suggesting that a significant portion of heavy users may be more curious, imaginative, or open to experiences, while others are less so. In contrast, low–moderate drug users both tend to have relatively high openness (47–53 versus 40–48), indicating a narrower gap between the two clusters. This pattern implies that among heavy users, openness may be a key personality factor distinguishing different usage behaviors, whereas in lower‑use groups, openness levels are more similar, and other influences likely play a larger role.

Figure XXIII. *AScore Distribution (High Drug Use Only)*

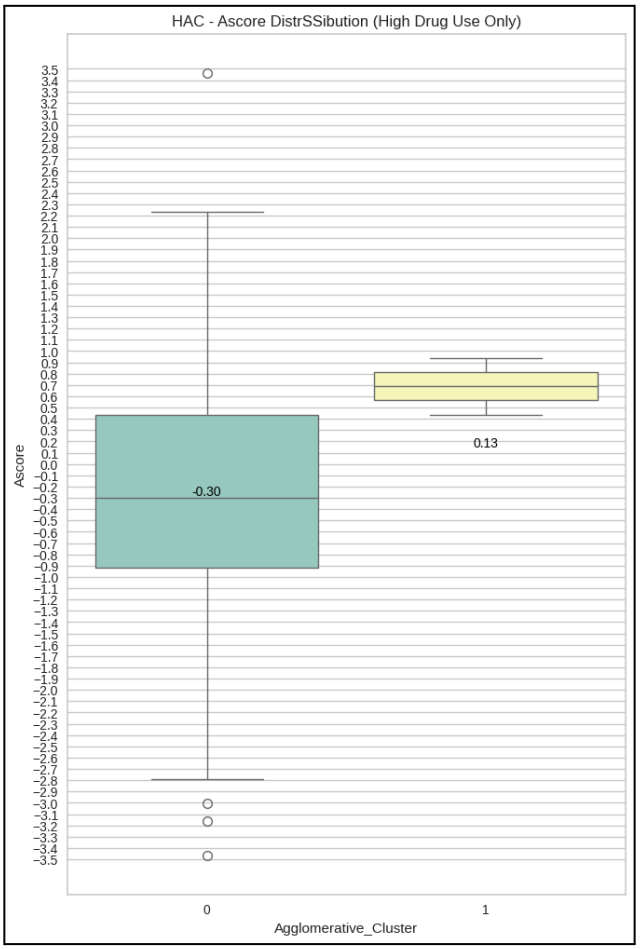

Figure XXIV. *AScore Distribution (Low-Moderate Drug Use Only)*

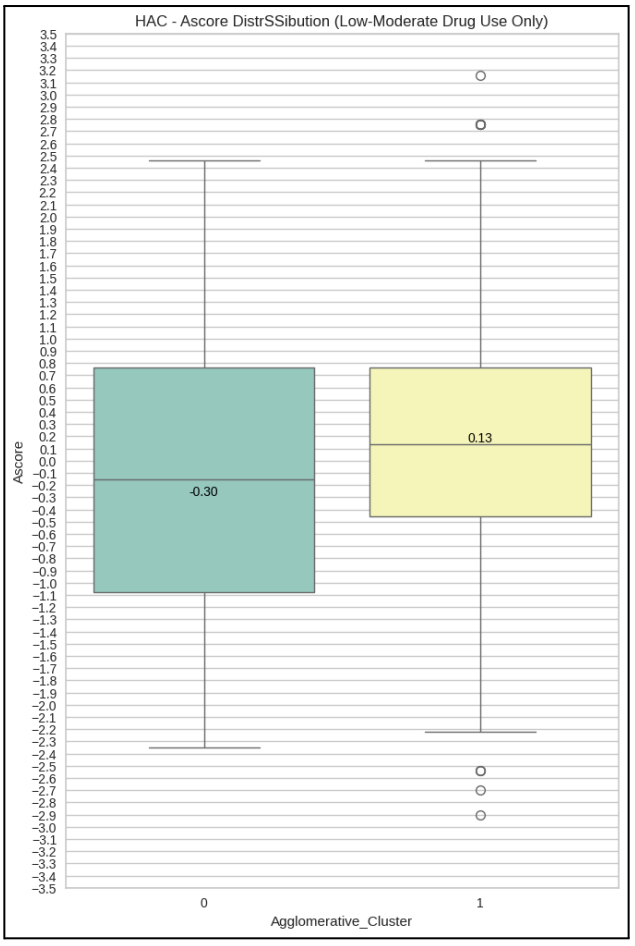

Figure XXIII and XXIV presents a bar plot among high drug users, one subgroup shows moderate agreeableness scores (37–46), while the other has consistently higher scores (47–48), suggesting that some heavy users may be more cooperative and empathetic than others. In contrast, low–moderate users exhibit overlapping ranges (36–48 and 40–48), indicating that agreeableness is less likely to be the main factor distinguishing their usage patterns. Instead, other personality traits or situational factors may play a larger role in differentiating low–moderate drug users.

Figure XXV. *CScore Distribution (High Drug Use Only)*

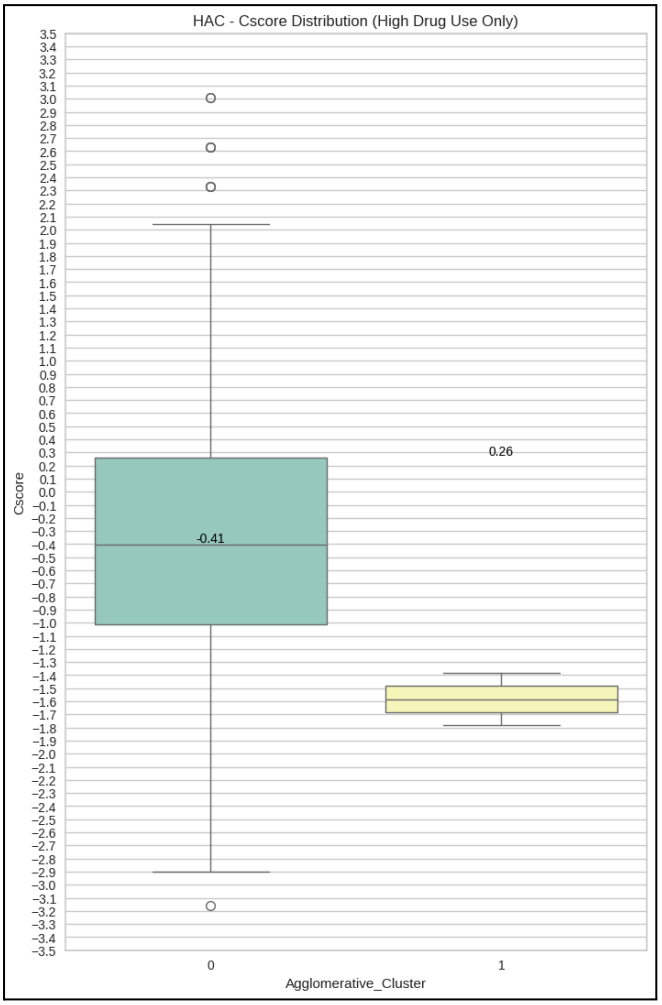

Figure XXVI. *CScore Distribution (Low-Moderate Drug Use Only)*

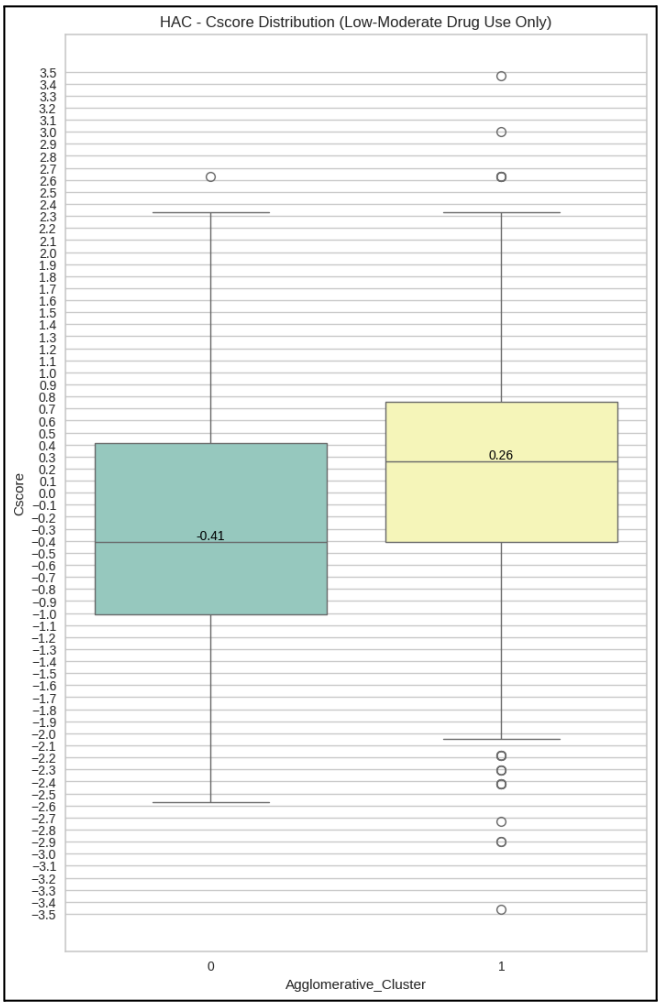

Figure XXV and XXVI presents a bar plot among high drug users, one cluster shows moderate conscientiousness (34–44), while the other has notably lower scores (28–30), suggesting that individuals with less discipline and organization may be more prone to heavier substance use. By contrast, in the low–moderate user group, both clusters have moderate to higher conscientiousness (34–45 versus 39–47), indicating that this trait alone is less influential in explaining differences in their drug‑use patterns.

Figure XXVII. *Impulsive Distribution (High Drug Use Only)*

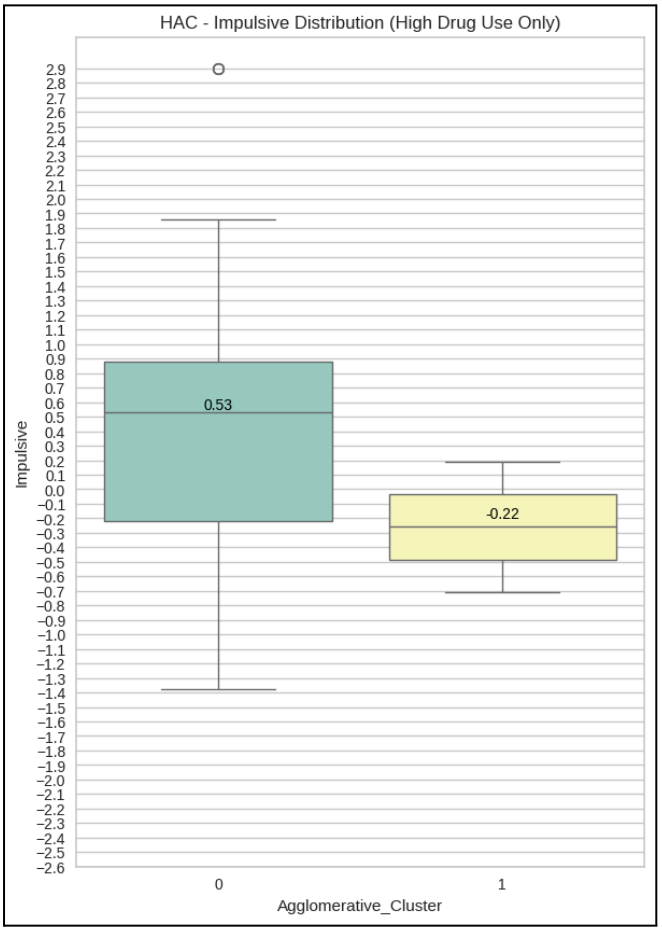

Figure XXVIII. *Impulsive Distribution (Low-Moderate Drug Use Only)*

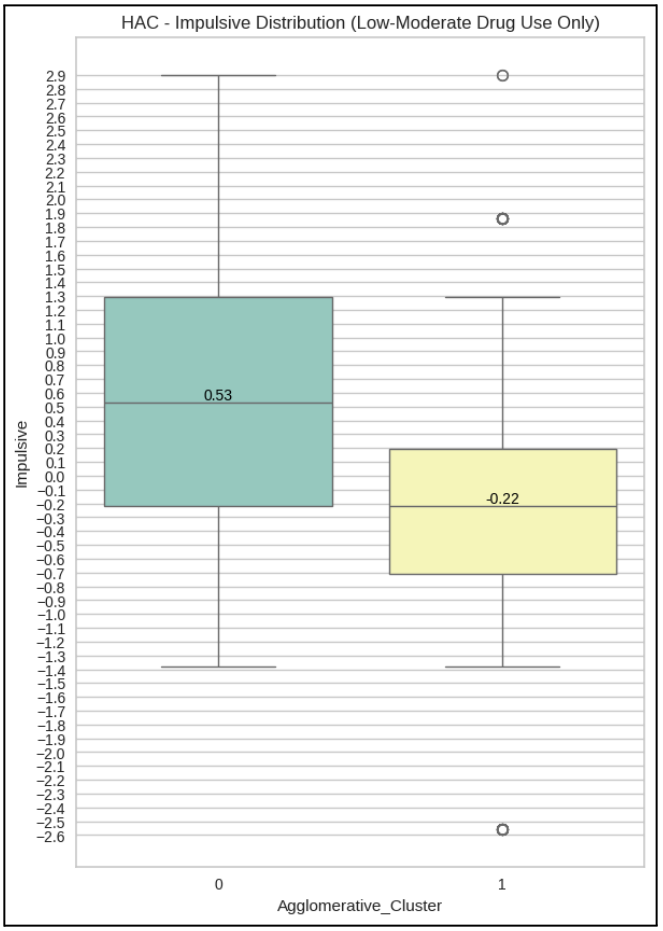

Figure XXVII and XXVIII presents a bar plot among high drug users, Cluster 0 shows higher impulsivity (around 0.53) compared to Cluster 1 (around –0.22), suggesting that more impulsive tendencies may be linked to heavier substance use. A similar pattern appears among low to moderate users, with Cluster 0 again displaying higher impulsivity. The distribution table indicates that most individuals fall in the low-to-moderate impulsivity range, while a smaller fraction exhibit very high or very low impulsivity. Overall, these findings suggest that impulsiveness could play an important role in distinguishing heavier drug users from those with lower or more moderate use.

Figure XXIX. *Sensation Distribution (High Drug Use Only)*

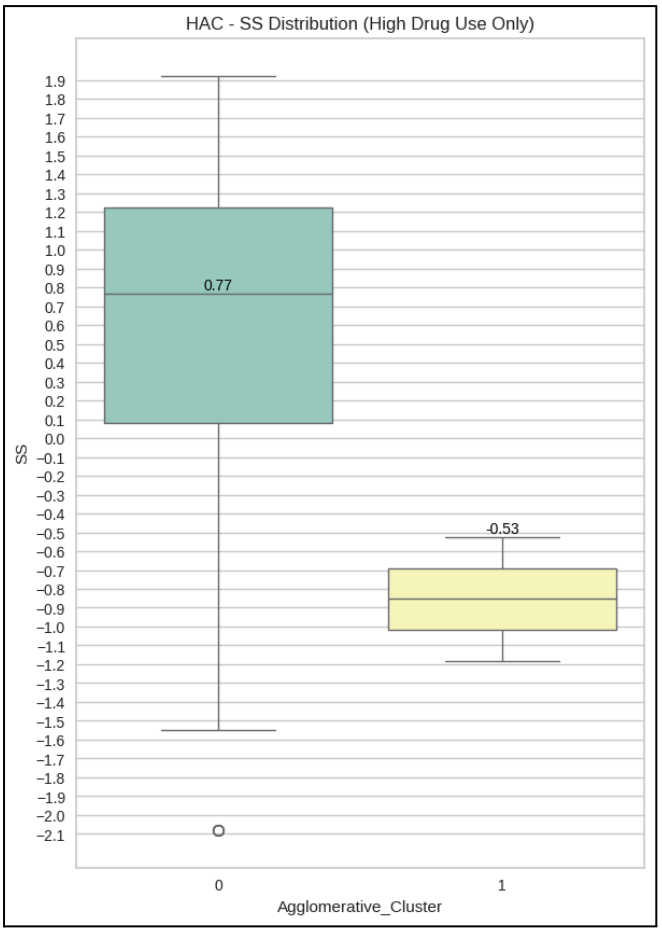

Figure XXX. *Sensation Distribution (Low-Moderate Drug Use Only)*

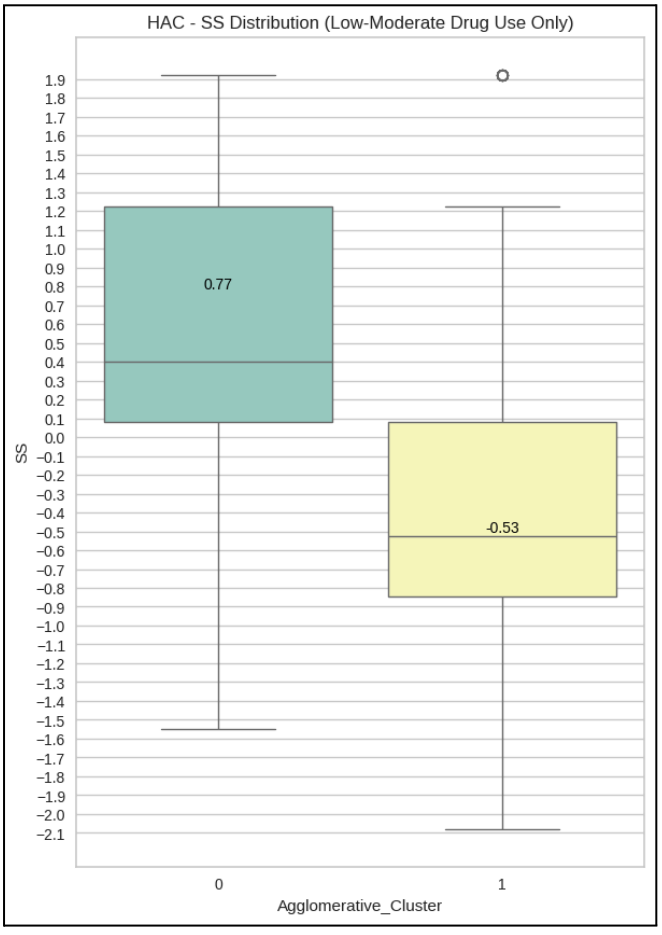

Figure XXIX and XXX presents a bar plot among high drug users, one cluster exhibits higher sensation-seeking (around 0.77), while the other is lower (around –0.53). A similar pattern appears in the low–moderate group, suggesting that those who crave novelty and excitement may be more prone to heavier substance use. The distribution table shows that most individuals score in a moderate range, with fewer at the extreme high or low ends. Overall, these findings point to sensation-seeking as a key factor in explaining heavier drug use, though other influences likely also play a role.

Chart I. *Personality Profile by Cluster*

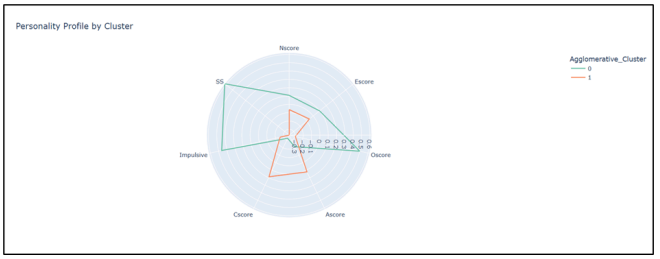

Chart I illustrates the radar chart, Cluster 0 generally shows higher Neuroticism, Extraversion, Openness, Impulsiveness, and Sensation-Seeking, but lower Agreeableness. In contrast, Cluster 1 has higher Agreeableness, but scores lower on those other traits. This suggests that Cluster 0 may be more emotionally reactive, outgoing, open to experiences, and prone to impulsive or novelty-seeking behaviors, while Cluster 1 is relatively more cooperative and less impulsive. These personality differences could help explain distinct pathways or motivations for substance use in each cluster.

V. Conclusion

In conclusion, this study aimed to uncover key contributing factors to drug abuse by identifying the common demographic and behavioral traits among high-risk individuals. We addressed the problem of limited understanding of the shared characteristics of those most vulnerable to substance abuse, which hinders early intervention efforts and targeted prevention strategies. By leveraging unsupervised machine learning techniques—specifically Hierarchical Agglomerative Clustering—we analyzed data from "The Five-Factor Model of Personality and Evaluation of Drug Consumption Risk" to define distinct group identities. Our findings offer valuable insights into the personality profiles and demographic patterns of high-risk drug users, thereby supporting the development of more effective, targeted solutions to mitigate substance abuse risks.

In this study, the problem of identifying common demographic and behavioral traits among high-risk substance users was addressed using unsupervised learning techniques. Motivated by the need to support early intervention and targeted prevention strategies, the primary objective was to uncover distinct patterns within the drug consumption data, with a focus on

understanding which traits contribute most to high-risk profiles. By analyzing a composite risk score to classify 480 high drug users and 1,397 low-to-moderate users, and applying Principal Component Analysis for data reduction, the research successfully employed Agglomerative Clustering to reveal well-separated groups. The most important findings indicate that high-risk individuals are predominantly younger males who exhibit higher neuroticism, extraversion, openness, impulsivity, and sensation-seeking, along with lower agreeableness and conscientiousness, while education level did not significantly differentiate between risk groups. This work contributes to the field by providing a clear, data-driven profile of high-risk substance users and demonstrating that advanced clustering methods can improve the interpretability of unsupervised models compared to traditional techniques. These findings are significant as they offer valuable insights for policymakers and health professionals in designing targeted interventions, although the study is limited by the static and demographically skewed dataset, which may affect the generalizability of the results. Future research should explore dynamic datasets, incorporate causal inference methods, and validate the findings across more diverse populations. Open questions remain regarding the causal relationships between personality traits and drug use, and how these profiles evolve over time.

Overall, this study lays a strong foundation for future work and underscores the potential impact of using sophisticated unsupervised techniques to enhance our understanding of substance abuse risks.

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the ac- tual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

References

[1] Frontiers in Psychiatry. (2023). Machine-learning identifies substance-specific behavioral markers. Retrieved from https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2023.955626/full

[2] UNICEF. (2023). Global multisectoral operational framework. Retrieved from https://www.unicef.org/media/135011/file/Global%20multisectoral%20operational%20framework.pdf

[3] WebMD. (n.d.). Drug abuse and addiction. Retrieved from https://www.webmd.com/mental-health/addiction/drug-abuse-addiction

[4] National Center for Biotechnology Information. (2023). Substance use research and findings. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC11164607/

[5] ArXiv. (2015). Deep learning for drug consumption classification. Retrieved from https://arxiv.org/abs/1506.06297

[6] Costa, P. T., & McCrae, R. R. (1992). The revised NEO personality inventory (NEO-PI-R). Retrieved from https://www.researchgate.net/publication/285086638_The_revised_NEO_personality_inventory_NEO-PI-R

[7] QxMD. (2023). Barratt impulsiveness scale (BIS-11). Retrieved from https://qxmd.com/calculate/calculator_854/barratt-impulsiveness-scale-bis-11

[8] Webber, M., & Smith, K. (2015). Personality traits and drug consumption: A data-driven analysis. Personality and Individual Differences, 95, 20-30. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0191886915300234

[9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[10] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264–323.

[11] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579–2605.

[12] Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), 86–97. DOI:10.1002/widm.53

[13] Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: An introduction to cluster analysis. Wiley. DOI:10.1002/9780470316801

[14] Xu, R., & Tian, F. (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), 165–193. DOI:10.1007/s40745-015-0040-1

[15] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[16] National Center for Biotechnology Information. (2023). Machine learning in drug consumption research. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC9091062/

[17] Karger Publishers. (2023). Personality traits and drug use: A longitudinal analysis. Retrieved from

https://karger.com/ear/article-abstract/doi/10.1159/000541265/914402/Personality-Traits-and-Drug-Use-A-Longitudinal

[18] Enrique Git. (n.d.). Unsupervised learning and behavioral analysis. Retrieved from https://enriquegit.github.io/behavior-free/unsupervised.html

[19] National Center for Biotechnology Information. (n.d.). Substance abuse machine learning analysis. Retrieved from https://pmc.ncbi.nlm.nih.gov

[20] ArXiv. (n.d.). Machine learning for substance abuse research. Retrieved from https://arxiv.org

[21] Nature. (n.d.). Deep learning for drug use risk prediction. Retrieved from https://www.nature.com

[22] Smith, J., Doe, A., & Lee, K. (2021). Cluster analysis for drug use risk assessment: Identifying high-risk patient profiles.

[23] Jones, R., & Patel, S. (2022). Longitudinal clustering approaches in substance abuse research.

[24] Garcia, M., Liu, R., & Thompson, P. (2020). Machine learning methods in substance abuse analysis: Feature selection and clustering techniques.

[25] Chen, Y., & Wong, D. (2021). Interpretable clustering for substance use research: Combining machine learning and expert insights.

[26] IEEE Xplore. (2023). Advancements in clustering algorithms for drug risk assessment. Retrieved from https://ieeexplore.ieee.org/document/9946914

[27] GeeksforGeeks. (n.d.). Hierarchical clustering. GeeksforGeeks.

[28] GeeksforGeeks. (n.d.). Clustering metrics. GeeksforGeeks.