

A Machine Learning-Based Approach for Discovering Correlations among Clinical Assessments in Substance Use Disorder

João Silva^{1*}, Pedro Costa¹, Ana Beatriz²

¹ Department of Psychiatry and Substance Use Research, Faculty of Medicine, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.

² Department of Machine Learning and Behavioral Sciences, Faculty of Medicine, University of Campinas, Campinas, Brazil.

*E-mail ✉ joao.silva@ufrj.br

Abstract

This research explores the condition of substance use disorder (SUD) and patterns of substance consumption by applying three distinct unsupervised machine learning approaches aligned with the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). The investigation relied on the 2019 National Survey on Drug Use and Health (NSDUH) dataset, drawn from a random sample of individuals who completed professional clinical evaluations. However, their official diagnoses remained undisclosed to the researchers. Through pattern recognition and relational insights from the models, the method supports accurate SUD classification. A licensed mental health expert thoroughly reviewed and confirmed the study's outcomes. Specifically, the algorithms used were k-means clustering, hierarchical clustering, and density-based spatial clustering of applications with noise (DBSCAN), each applied to dedicated datasets covering alcohol, marijuana, and cocaine. Cluster evaluation incorporated silhouette scores and 95% confidence intervals (CI). Findings from this work hold potential to enhance and support conventional psychiatric assessments.

Keywords: Diagnostic and statistical manual (DSM), Substance use disorder, Unsupervised machine learning, K-means clustering, Hierarchical clustering, DBSCAN

Introduction

Substance use disorder (SUD) represents a significant challenge for healthcare systems because of its wide-ranging detrimental impacts on those affected. Substance use strengthens harmful patterns such as drug-seeking and drug-taking behaviors [1], which substantially alter brain function and influence subsequent actions. Different substances produce unique physiological responses and are associated with specific behavioral outcomes in humans [2, 3]. Previous research has primarily relied on predetermined criteria to identify

SUD [4]. These criteria are formalized in the Diagnostic and Statistical Manual (DSM) of mental disorders [5], serving as a standard reference for evaluating clinical information related to SUD. Given the severe repercussions of substance use, it is essential to develop robust analytical strategies that support better management of personal behaviors [4, 6]. Machine learning has gained increasing traction as an analytical approach in psychiatric research, particularly when applied to clinical datasets involving drug-related SUD [7]. Using newly acquired data, the machine learning model—built from existing records—can generate diagnostic predictions for users.

The following sections review relevant literature on applying unsupervised machine learning to mental health conditions, the specific algorithms used to model such disorders, and the datasets used in those works. Notably, only one of the reviewed studies incorporated DSM criteria. These studies are outlined in the initial section of **Table 1**. For instance, Srividya *et al.* [8] developed an

Access this article online

<https://smerpub.com/>

Received: 11 September 2023; Accepted: 01 December 2023

Copyright CC BY-NC-SA 4.0

How to cite this article: Silva J, Costa P, Beatriz A. A Machine Learning-Based Approach for Discovering Correlations among Clinical Assessments in Substance Use Disorder. J Med Sci Interdiscip Res. 2023;3(2):171-83. <https://doi.org/10.51847/VPG03xOfTa>

approach that applied multiple machine learning techniques to assess mental health status within a specific population. The method modeled participants' questionnaire responses using unsupervised machine learning, generating meaningful clusters validated through mean opinion scores. Resulting classification models were then used to forecast individuals' mental health conditions. This research employed k-means, hierarchical, and k-medoids clustering methods. Recognizing the scarcity of reliable methods for assessing the predictive power of unsupervised clustering models, Tarekegn *et al.* [9] introduced an extended cross-validation framework to gauge clustering quality and

forecast cluster assignments for unseen data points. Their focus centered on evaluating model effectiveness and the accuracy of allocating new data points to appropriate clusters. The study used k-fold cross-validation to estimate label probabilities in the training and test sets. It supplemented these estimates with root-mean-square error (RMSE) and mean absolute percentage error (MAPE) metrics. The Calinski–Harabasz index was used to determine the optimal number of clusters, and the final model was evaluated using k-means. This produced three clusters for the Emotions dataset, four for the Chronic Disease dataset, and five for the Yeast dataset.

Table 1. Overview of previous studies investigating machine learning applications in substance use disorder and related mental health conditions.

Reference	Study title (first three words)	Target mental health condition	Machine learning approaches applied	Data source and sample characteristics
Srividya <i>et al.</i> [8]	Behavioral Modeling Mental	General mental health status	Combination of supervised and unsupervised learning methods, including Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbors (KNN), and Logistic Regression	Questionnaire-based survey containing 20 items; two study groups consisting of 300 and 356 participants
Tarekegn <i>et al.</i> [9]	Cross-Validation Approach	No specific psychiatric condition investigated	Unsupervised clustering using K-means with K-fold cross-validation; model performance assessed using RMSE and MAPE	Multiple benchmark datasets, including chronic disease records, emotion-related data, and yeast dataset
Stevens <i>et al.</i> [10]	Identification Analysis Behavioral	Autism Spectrum Disorder (ASD)	Unsupervised learning approach using hierarchical clustering techniques	Dataset collected from 2,400 children recruited through community centers throughout the United States
Jing <i>et al.</i> [11]	Analysis of Substance Use	Substance Use Disorder (SUD)	Supervised machine learning using Random Forest classification	Participants aged 10–22 years recruited through advertisements, public announcements, digital random dialing, and printed recruitment materials
Mannes <i>et al.</i> [12]	Correlates of Mild to Moderate	Alcohol Use Disorder (AUD)	Supervised statistical modeling using multinomial logistic regression methods	Two participant groups: 150 individuals from a suburban inpatient addiction treatment facility and 438 individuals from an urban medical center; participants aged ≥ 18 years
Mintz <i>et al.</i> [13]	A Cascade of Care	Alcohol Use Disorder (AUD)	Supervised learning and statistical analysis based on multinomial logistic regression	Same dataset as Mannes <i>et al.</i> [12]: 150 suburban inpatient treatment participants and 438 urban medical center participants, all aged ≥ 18 years
Hayley <i>et al.</i> [14]	DSM-5 Cannabis Use	Cannabis Use Disorder (CaUD)	Supervised learning approach using logistic regression	National Epidemiologic Survey on Alcohol and Related Conditions III (NESARC-III); 36,309 adults aged 18 years or older

The research yielded encouraging outcomes, achieving an overall RMSE below 0.075 and a MAPE under 12.5%

across the three datasets. Another investigation by Stevens *et al.* [10] focused on identifying and examining

behavioral phenotypes associated with autism spectrum disorder (ASD) through unsupervised machine learning applied to various ASD subgroups and their interrelationships. Employing a Gaussian mixture model, the analysis uncovered 16 distinct subgroups from a sample of $n = 1034$ (drawn from a larger pool of $n = 2400$ children with ASD). This subset served as retrospective treatment data, with responses analyzed via linear regression to assess how different subgroups responded to varying levels of treatment intensity. Hierarchical agglomerative clustering was first conducted before fitting the regression model. The findings highlighted two overarching behavioral phenotypes, within which certain subgroups displayed notably low profiles that varied in severity. High variance in the regression models suggested that clustering captured similarities effectively, making it well suited for tailoring treatment based on cluster membership.

In the concluding part of the literature review, we examine prior works that applied supervised machine learning to substance use disorders (SUDs) along with their respective datasets. All of these studies incorporated DSM-5 criteria, with a summary provided in the latter portion of **Table 1**. According to Jing *et al.* [11], machine learning techniques helped pinpoint key health, psychiatric, psychological, and contextual factors predictive of SUD. These indicators were also leveraged to forecast individuals at elevated risk of developing the disorder. The study recruited 600 participants (494 males and 206 females) who completed questionnaires across multiple assessment periods at ages 10–12, 12–14, 16, 19, and 22. We used a random forest classifier to identify the most influential variables associated with SUD, with 30 variables emerging as the strongest predictors. Prediction accuracy for SUD was 74% among youths aged 10–12 and rose to 86% by age 22. The random forest classifier outperformed other machine learning algorithms in identifying at-risk individuals aged 10–22. The study by Mannes *et al.* [12] investigated clinical and functional characteristics as predictors of alcohol use disorder (AUD) severity. It assessed the relationships between alcohol consumption levels, comorbidities, and functional impairment across mild, moderate, and severe AUD categories. A total of 588 participants aged 18 and older, recruited between 2016 and 2019, underwent DSM-5 AUD assessments for current substance use. We used multinomial logistic regression to examine associations among potential validators, psychiatric conditions, functional impairment, and the three AUD

severity levels. Results indicated that craving, binge-drinking frequency, problematic use, and harmful drinking showed stronger associations with all three AUD levels relative to the no-AUD reference group. All potential validators showed stronger links to severe AUD effects than to mild or moderate ones.

Using 2015–2019 NSDUH data, Mintz *et al.* [13] introduced a framework to identify gaps in care for individuals with AUD. The model evaluated adults based on several metrics: AUD prevalence, the share who received healthcare in the past year and were screened for alcohol use, those receiving brief interventions for alcohol misuse, individuals informed about treatment options, and those who ultimately received treatment. Outcomes were stratified by AUD severity, revealing that while most people with AUD accessed healthcare, only a small minority received interventions, referrals, or actual treatment. Similar care cascade patterns appeared across severity levels, underscoring that many individuals with AUD do not receive appropriate care despite screening opportunities.

In Hayley *et al.* [14], researchers examined connections between older and newer generations of illicit and prescription drugs. The analysis included 36,309 adults aged 18 and above from Wave 3 of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC-III). It evaluated comorbidity between DSM-5 cocaine use disorder (CoUD), DSM-5 SUD, and general substance use. Multivariable logistic regression showed that individuals with current DSM-5 CoUD had significantly higher lifetime use of all examined substances, including several newer classes of illicit and prescription stimulants. Current DSM-5 CoUD was also strongly linked to various other DSM-5 SUDs, indicating substantial comorbidity.

Overall, while numerous studies have applied machine learning to SUDs, they vary considerably in the types of SUD examined, the algorithms employed, and the datasets utilized. These existing works leave room for new approaches that integrate key elements from prior research. The present study aims to address this gap. The core objective of the method proposed in this paper is to automatically distinguish individuals with SUD from those without, using clinical assessment data from the NSDUH. This database captures behaviors that mental health professionals would typically evaluate when determining a DSM diagnosis. A mental health professional validated the proposed approach, providing

a rapid, accurate classification tool that minimizes human error and bias.

The key contributions of this work include applying unsupervised machine learning to the NSDUH 2019 dataset, guided by DSM-5 criteria. To the best of our knowledge, this is the first study to use unsupervised machine learning techniques to process NSDUH 2019 data for this purpose.

- Unsupervised machine learning is particularly suitable here as it enables the model to derive classifications organically from inherent data correlations rather than relying on pre-assigned labels.
- At the time of the study, the NSDUH 2019 dataset was among the most recent available, containing 56,137 records, of which approximately half were utilized. DSM-5 remains the prevailing standard for SUD classification among mental health professionals.
- The study delivers an automated SUD classification system grounded in DSM-5 criteria, with final validation provided by a mental health professional.

Unsupervised learning was selected over supervised methods because the dataset lacks explicit class labels or target variables indicating SUD versus non-SUD status. This approach allows classifications to emerge directly from data patterns without requiring prior labeling. Supervised learning would have necessitated accurate pre-labeling of the data, which was not feasible given the available information. Section 2 details the proposed methodology; Section 3 covers the materials and methods; Section 4 presents the results and discussion; and Section 5 provides analysis and interpretation of the machine learning outcomes. The final section summarizes the key conclusions and offers recommendations for future work.

Methodology

The flowchart in **Figure 1** illustrates the overall workflow of the proposed methodology, comprising six main stages: raw data acquisition, data mapping, data encoding, data curation, clustering, and validation. The process begins with raw data from the NSDUH. It culminates in an unsupervised machine learning model whose outputs are evaluated through statistical techniques and expert review by a mental health professional.

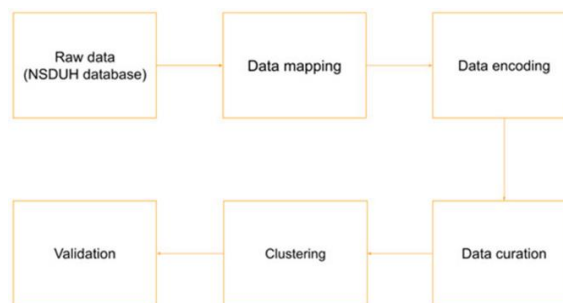


Figure 1. Overview of the proposed model architecture and procedural workflow, illustrating the successive application of data mapping, data encoding, data curation, clustering, and validation procedures to the alcohol, marijuana, and cocaine datasets individually.

Materials and Methods

This section outlines the resources and procedures used to achieve the desired outcomes described in the methodology.

Dataset

The study utilized information drawn from a comprehensive publicly accessible online repository that contains extensive records on the consumption of illicit drugs, tobacco, and alcohol. The National Survey on Drug Use and Health (NSDUH) provides statistical data [15] and includes structured questionnaires aligned with the DSM for evaluating mental health conditions linked to addictive substances [16]. The specific questionnaire follows the most recent edition, DSM-5, released in 2013. Clinical assessment data were gathered from $n = 56,137$ respondents, each of whom completed a psychiatric evaluation to determine their level of SUD. Survey participants included adolescents (aged 12–17 years) and adults (aged 18 and older) residing in the United States [17]. Data collection followed a random selection process, with questionnaires distributed to households irrespective of any known history of substance exposure. Consequently, whether participants belonged to an addicted or non-addicted category was unknown beforehand. The analysis focused on variables such as self-reported substance use, perceived risk and availability of the substance, and indicators of substance abuse and dependence. The latter measures were particularly important for diagnosing the severity of SUD. All variables are identified by their respective code names in the dataset.

Data preprocessing

Examination of the extensive dataset revealed records covering ten categories of drugs along with alcohol and tobacco. However, the present study concentrated exclusively on SUD related to alcohol, marijuana, and cocaine. The selected features consisted of responses to questions grounded in DSM-5 diagnostic standards for substance use disorder.

Data cleaning

Data cleaning is a crucial preprocessing step that addresses missing values to minimize bias and prevent inflated estimates of predictive accuracy in psychiatric studies [18]. Many features contained missing entries, resulting from unanswered questions during psychiatric assessments or participant dropout [19]. The missing data pattern was determined to be missing completely at random (MCAR), as the likelihood of missingness did not depend on observed or unobserved values [20]. Given the MCAR nature of the data, mean imputation was applied to avoid introducing bias while preserving variability across input variables. Although this approach can slightly reduce the spread of extreme values, it was selected for its simplicity and suitability, especially given the minimal proportion of missing data (approximately 3.16%). Prior research [21, 22] supports mean imputation as a reliable technique when missing values constitute less than 5%, rendering the impact negligible. Similarly, Bonfiglio *et al.* [23] successfully employed mean imputation in a comparable scenario where missing data ranged from 1.94% to 3.15%, confirming its appropriateness for datasets with low levels of incompleteness.

Feature correlation

Feature correlation analysis was conducted on the fully populated raw data as the subsequent essential step. This process identifies linear relationships among variables, enabling the removal of irrelevant or redundant features to enhance computational efficiency [24, 25].

Feature scaling

Feature scaling normalizes the data so that all variables are on a comparable scale, ensuring no single feature disproportionately influences the analysis due to its original magnitude. This third preprocessing step was performed on the cleaned dataset after eliminating redundant or irrelevant information. The technique standardizes the range of participant responses collected

during the survey [26], transforming minimum and maximum values to a uniform scale.

Dimensionality reduction

Dimensionality reduction addresses the challenges posed by high-dimensional, multivariate datasets [27] by projecting the data into a lower-dimensional space, such as two dimensions (2d), for improved visualization. This final preprocessing stage retained only the two most informative principal components. The alcohol and cocaine datasets originally contained 40 features each, while the marijuana dataset contained 35. **Figures 2-4** display the explained variance ratios for the principal components derived from the alcohol, marijuana, and cocaine datasets, respectively. We retained two principal components to clearly visualize clustering outcomes in a two-dimensional feature space.

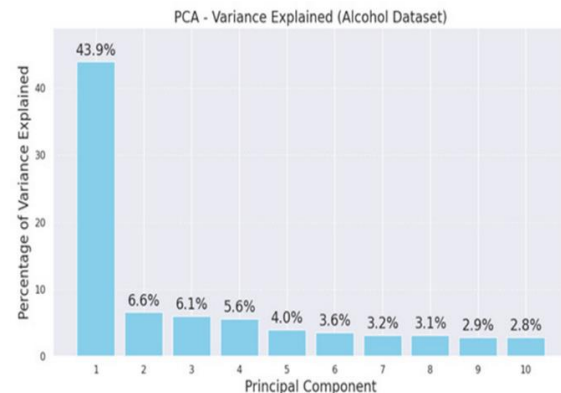


Figure 2. Proportion of variance accounted for by the principal components in the alcohol dataset.

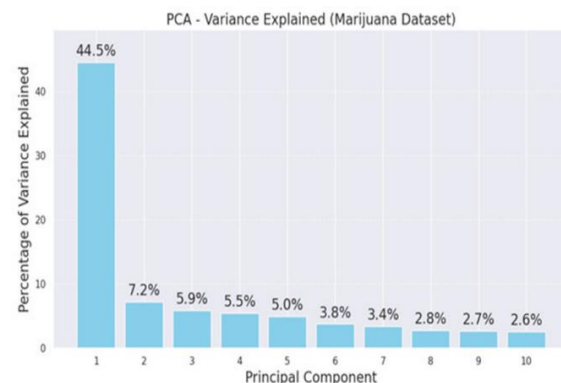


Figure 3. Proportion of variance explained by the leading principal components from the marijuana dataset.

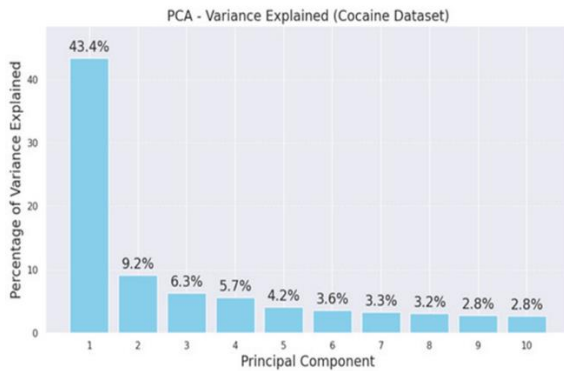


Figure 4. Proportion of variance explained by the primary principal components in the cocaine dataset.

Cluster analysis

This study employed three clustering methods: k-means, hierarchical, and DBSCAN. After dimensionality reduction, the processed data were fed into these models to produce two-dimensional (2d) visualizations of the clustering results. As noted by Sarker [28], clustering approaches generally fall into three categories: partitioning, density-based, and hierarchical.

K-means clustering

K-means is a partitioning-based clustering method favored for its straightforward implementation and strong performance among distance-based algorithms. It minimizes the sum of squared Euclidean distances between each data point and its assigned cluster centroid [28, 29].

Hierarchical clustering

Agglomerative hierarchical clustering was applied, beginning with each data point as its own separate cluster and progressively merging the most similar clusters until a single comprehensive cluster is formed. Like k-means, this method requires specifying the desired number of clusters in advance. A dendrogram was utilized to guide this bottom-up merging process [30].

DBSCAN

The DBSCAN algorithm identifies clusters based on the density of data point distributions. It relies on two key parameters—epsilon (eps) and minimum points (minPts)—to determine data density and categorize points as core, border, or outlier points [31, 32].

Validation

Validation aims to evaluate the overall quality of the identified clustering structure and confirm that data points have been appropriately assigned to their respective groups [9, 33]. Silhouette scores and 95% confidence intervals (CIs) served as internal validation metrics to assess clustering effectiveness and to determine how the resulting clusters relate to one another across algorithms.

Results and Discussion

This section presents visual representations of the outcomes obtained from the three clustering algorithms. Results are discussed starting with k-means, followed by hierarchical clustering, and finally DBSCAN. Across all three methods, the clustering patterns for SUD classification were highly consistent for alcohol, marijuana, and cocaine, with validation supported by silhouette scores and 95% confidence intervals.

Alcohol dataset

K-means clustering was performed with two clusters, as illustrated in **Figure 5**. The visualization reveals tightly grouped data points within each cluster, with only a few points farther away, reflecting greater Euclidean distance from the cluster center. The green cluster extends further along the PCA1 axis than the red cluster, whereas the red cluster shows a broader spread along the PCA2 axis. The red cluster appears notably denser in its core, suggesting a higher likelihood of accurate classification. The concentration of points in a single area reinforces the model's reliability and consistency, reflecting shared response patterns among many participants. Although the green cluster is less dense overall, its main concentration remains clearly discernible in the plot. This distribution aligns with the expectation that SUD can be broadly divided into two primary categories, with one encompassing most respondents. The green cluster occupies roughly two-thirds of the red cluster's spatial extent. This disparity likely stems from the standardized nature of the survey questions, which yield variability primarily from individual experiences. Cluster boundaries are thus shaped by feature patterns derived from participant responses. The smaller size of the green region can be attributed to fewer respondents whose answers were less diverse than those of the larger group. In the hierarchical clustering results shown in **Figure 6**, data points within the orange cluster are more closely clustered, with shorter intra-cluster distances. This

cluster is dense and evenly distributed across its range, and it contains substantially more respondents than the blue cluster. Within the blue cluster, points near the inter-cluster boundary are relatively dense but spread more widely along the x-axis. This pattern shows greater distances between points in the blue cluster than in the orange cluster, further confirming the greater density of the orange group relative to the green cluster observed earlier.

For DBSCAN, the overall visualization produced outcomes comparable to those of k-means and hierarchical clustering. Cluster centroids and boundaries aligned closely across the three techniques, as shown in **Figure 7**. However, DBSCAN identified four clusters in total, including two dominant ones that closely resembled the structures from the other methods. When restricted to two clusters, the DBSCAN results closely aligned with those from k-means and hierarchical clustering.

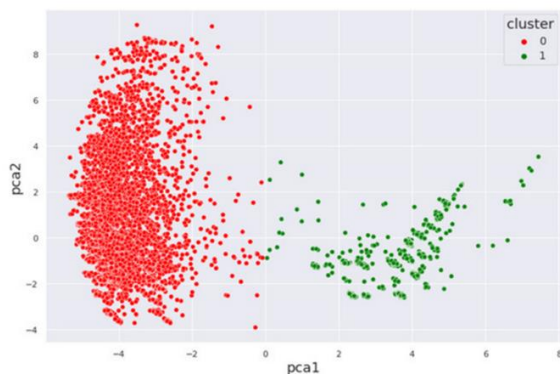


Figure 5. Visualization of k-means clustering applied to the alcohol dataset, displaying two clearly separated groups of data points determined via the elbow method.

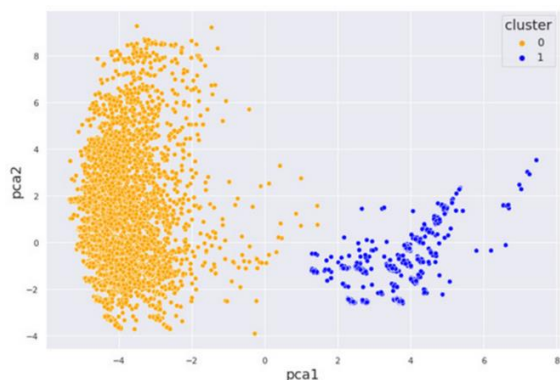


Figure 6. Hierarchical clustering results for the alcohol dataset, highlighting a greater number of data

points in the orange cluster compared to the blue cluster.



Figure 7. DBSCAN output on the alcohol dataset, displaying three primary clusters accompanied by noise points. The general layout and positioning of data points stayed largely consistent throughout the three clustering approaches.

Marijuana dataset

Figure 8 shows the results of k-means clustering on the marijuana dataset. The red cluster covers a larger area and has a higher concentration of tightly packed points than the green cluster. Because the algorithm has limited ability to isolate outliers, these points are typically absorbed into nearby groups. Several observations located far from their centroids are still assigned to their respective clusters. The orange cluster features relatively large distances between its members and a lower overall density than the green cluster. Early points in the hierarchy show tight associations, but separation grows steadily at higher levels. The green cluster stands out for both broader coverage and stronger internal density across a wider range. **Figure 9** displays the hierarchical clustering results for the marijuana dataset.

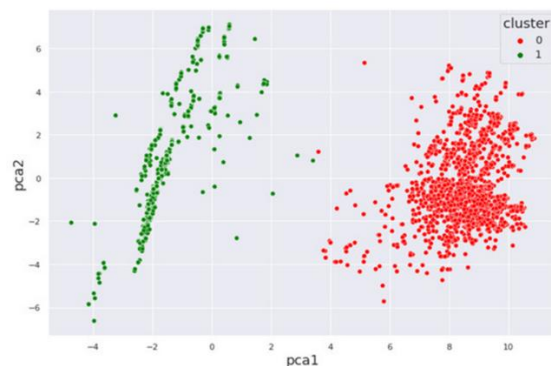


Figure 8. K-means clustering outcomes for the marijuana dataset. A small number of points located

between the clusters exhibit notably larger Euclidean distances.

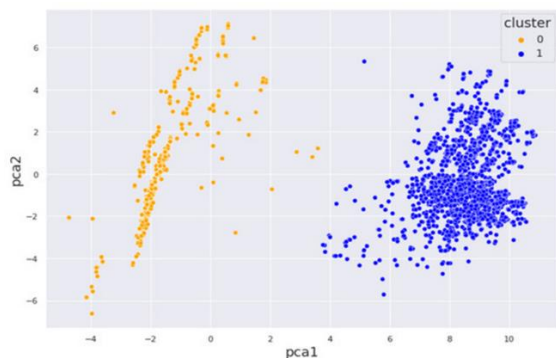


Figure 9. Hierarchical clustering for the marijuana dataset. Unlike the k-means approach, the two clusters are not positioned around the origin.

In the visualization, the two groups are sharply separated. Points in the orange cluster are more spread out, with greater distances between them, while only a handful are positioned close together. This group has lower density than the blue cluster, which covers a broader area with tighter grouping and reduced spread. The data clearly separate into two categories, with the larger containing most respondents in the blue cluster. These hierarchical clustering results closely mirror those from k-means.

Figure 10 shows the DBSCAN results for the marijuana dataset, with two dominant clusters and four smaller ones. Broadly, the grouping structure matches the patterns seen with k-means and hierarchical methods. The main variation is that the four smaller clusters, located farther away from the primary ones, could be incorporated into the larger clusters, effectively reducing the solution to two main groups.

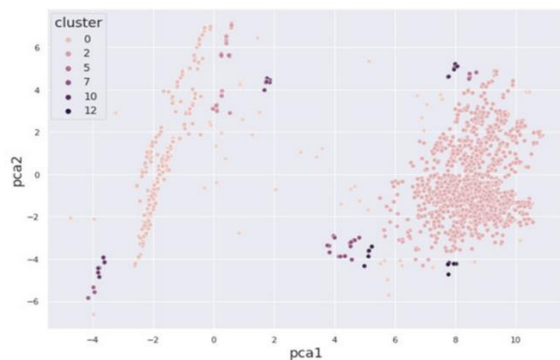


Figure 10. DBSCAN clustering visualization for the marijuana dataset. No noise cluster is present, indicating that the data are free of outliers.

Cocaine dataset

Figure 11 shows the K-means clustering results for the cocaine dataset. A substantial Euclidean distance sharply separates the two clusters. The red cluster demonstrates strong internal cohesion, with most points positioned near the centroid at relatively short distances. In contrast, points within the green cluster are more widely spaced, though they still exhibit shared characteristics. A notable distinction from the earlier alcohol and marijuana results is that these clusters span a much wider range of values along both PCA1 and PCA2.

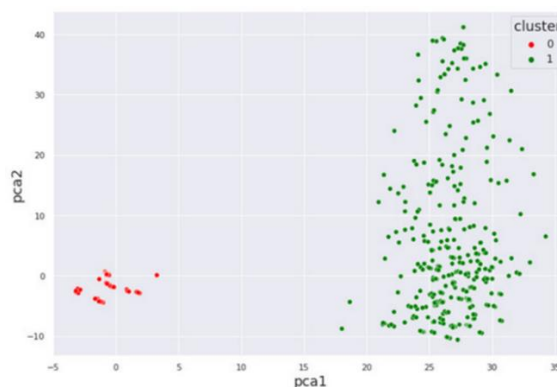


Figure 11. K-means clustering outcomes for the cocaine dataset, featuring distinctly separated clusters. Intra-cluster distances are notably larger within the green cluster.

Figure 12 presents hierarchical clustering results for the cocaine dataset. This method yields cluster densities per group that are comparable to those observed with k-means. **Figure 13** shows DBSCAN results for the cocaine dataset, with two primary clusters and four smaller ones. While the overall structure is similar to k-means and hierarchical clustering, DBSCAN allows overlapping points and groups. In total, the plot shows six clusters. The smaller clusters are situated inside or near the main ones, with their centroids positioned relatively close to those of the larger clusters.

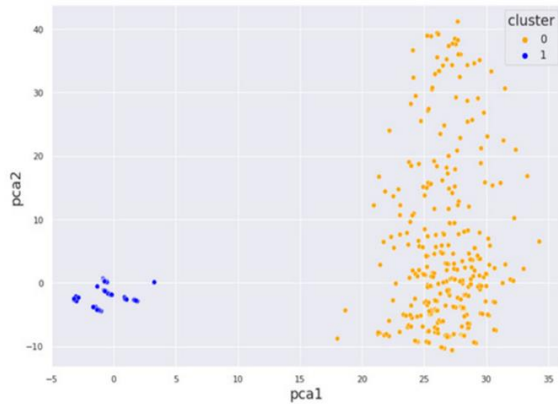


Figure 12. Hierarchical clustering results for the cocaine dataset, showing a high concentration of points in the primary cluster.

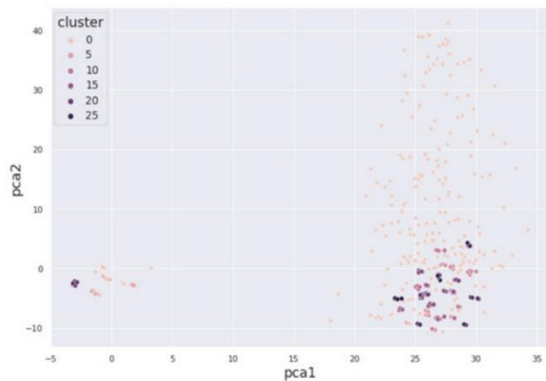


Figure 13. DBSCAN clustering visualization for the cocaine dataset. The outcomes reveal numerous closely spaced clusters.

Validation results

Table 2 displays the validation metrics computed for each algorithm. The effectiveness of the three clustering methods was evaluated across all datasets, with performance expressed as percentages. These measures were used to examine inter-cluster relationships, predictive reliability, and result deviations.

Table 2. Internal validation scores: Classical algorithms.

Algorithm	Dataset		
	Alcohol (%)	Marijuana (%)	Cocaine (%)
K-Means			
Sil_score	72.5	78.7	77.9
95% CI	± 0.002	± 0.002	± 0.002
Hierarchical			
Sil_score	83.4	84.7	78.9

95% CI	± 0.001	± 0.001	± 0.001
DBSCAN			
Sil_score	62.6	66.3	63
95% CI	± 0.002	± 0.002	± 0.002

The subsequent sections examine the internal validation outcomes derived from standard evaluation techniques. For the alcohol dataset, the silhouette score for k-means was 72.5% (95% confidence interval: ±0.001). Hierarchical clustering produced a higher silhouette score of 83.4%, with a narrower confidence interval of ±0.001, as detailed in **Table 2**. Overall, hierarchical clustering delivered superior validation scores compared to the other two conventional algorithms. Similarly, among the more advanced algorithms, k-means++ yielded better validation metrics than its counterparts. Regarding DBSCAN within the classical approaches, the silhouette scores were 62.6% for alcohol, 66.3% for marijuana, and 63% for cocaine — representing the lowest values recorded across all three techniques. The 95% confidence intervals remained stable across datasets for both k-means and DBSCAN, ranging from ±0.001 to ±0.003 for both conventional and advanced algorithms. This narrow range reflects strong stability and reliability in the clustering process, underscoring the models' robustness. The next paragraph compares the internal validation performance between classical and state-of-the-art methods. As illustrated in **Tables 2** and **3**, the results from the advanced algorithms are quite close to those of the classical ones. However, classical methods generally outperformed the state-of-the-art approaches in most cases. This suggests that newer algorithms do not automatically guarantee superior outcomes and that effectiveness can vary depending on the dataset's specific characteristics. In this study, the survey data collected from respondents regarding substance use disorder (SUD) was handled more effectively by the classical clustering techniques.

Table 3. Internal validation scores: State-of-the-art algorithms.

Algorithm	Dataset		
	Alcohol (%)	Marijuana (%)	Cocaine (%)
K-Means++			
Sil_score	73.1	78.8	78.4
95% CI	± 0.002	± 0.002	± 0.002
BIRCH			
Sil_score	68.7	70.7	67.1

95% CI	± 0.003	± 0.003	± 0.003
HDBSCAN			
Sil_score	40.8	58.6	81
95% CI	± 0.001	± 0.001	± 0.001

Interpretation of results

This section analyzes the modeled and processed data using machine learning techniques. Two distinct forms of interpretation are presented: one derived directly from the machine learning outputs and another based on expert evaluation by a mental health professional.

Machine learning-based interpretation

The machine learning interpretations draw on the visual patterns observed in the cluster plots, focusing on trends in cluster density, spatial coverage, and separation boundaries. Across all clustering outcomes, two primary groups consistently emerged: one large, dense cluster and one smaller, less dense cluster. The dense cluster is interpreted as representing individuals exhibiting minimal or no notable symptoms of substance use disorder (SUD). These behaviors show considerable variety yet remain closely aligned, reflecting the routine daily activities typical of people without SUD, which tend to show strong similarity across this group.

The less-dense cluster corresponds to individuals who show symptoms of SUD. In this group, each person's response to the disorder is more individualized, often limiting their ability to engage in normal activities in ways that differ markedly from others in the same category. This individuality reduces density within the cluster.

Notably, the major and minor clusters for the alcohol dataset are positioned closer together than those for marijuana or cocaine, with the smallest centroid distance observed in alcohol. This suggests that transitioning between non-AUD and AUD states is relatively easier compared to shifts between non-dependence and dependence for marijuana or cocaine (and vice versa). Consequently, rehabilitation efforts required to overcome AUD appear to demand the least intensity relative to cannabis use disorder (CaUD) or cocaine use disorder (CoUD).

A greater distance separates the clusters for the marijuana dataset than those for alcohol. This indicates that moving between non-CaUD and CaUD is more challenging than for alcohol, implying a slower progression toward or away from dependence on marijuana. Additionally, behaviors associated with marijuana dependence tend to

deviate more noticeably from normal functioning than those linked to alcohol.

In the cocaine dataset, the major and minor clusters are the most widely separated, showing the largest centroid distance among the three substances. This implies that transitioning to or from CoUD is the most difficult. Dependence on or recovery from cocaine thus occurs at the slowest rate. Rehabilitation for CoUD would likely require the greatest effort, followed by CaUD, and then AUD.

In the alcohol dataset, the major cluster spans a broader range of values than the minor cluster. This wider coverage suggests it encompasses a diverse set of everyday behaviors typical of non-users, moderate users, or low-risk users of alcohol — individuals whose daily lives show no substantial disruption from alcohol consumption and thus exhibit few AUD symptoms. The minor cluster, with its narrower range, reflects the restricted activities of those experiencing alcohol dependence, such as extreme binge or heavy drinkers whose behaviors are significantly constrained by the effects of alcohol.

Similar patterns of wider coverage in major clusters and narrower coverage in minor clusters were observed for marijuana and cocaine. The major clusters represent normal behavioral ranges associated with non-users or low-risk users showing no significant CaUD or CoUD symptoms. The minor clusters indicate more limited activities among those affected by dependence, whose daily functioning is noticeably restricted.

When comparing the three substances, the CaUD region shows the broadest span, followed by AUD, and then CoUD. This suggests that behaviors associated with CaUD are the least restrictive, allowing individuals to remain more functional and closer to non-SUD patterns than those with AUD or CoUD. Consequently, the visible effects of SUD are least apparent in CaUD, moderately apparent in AUD, and most pronounced in CoUD. These findings align with existing research indicating increasingly permissive risk perceptions regarding marijuana use across various age groups, particularly among adults, emerging adults, and middle-aged adults [34].

Additionally, the distance between cluster boundaries is smallest for alcohol, followed by marijuana, and largest for cocaine. This means that behaviors at the AUD/non-AUD boundary are the most similar and difficult to differentiate. In contrast, CaUD/non-CaUD boundaries

show clearer distinctions, and CoUD/non-CoUD boundaries exhibit the most pronounced differences. In the DBSCAN results for alcohol, the smallest cluster likely corresponds to individuals with severe AUD symptoms, such as heavy drinkers. For marijuana and cocaine, the additional smaller clusters may indicate heterogeneous symptoms suggestive of potential poly-substance use disorder. This interpretation is supported by Dell *et al.* [35], who link marijuana and cocaine dependence to higher likelihoods of binge drinking. The model may also have identified potential endophenotypes for CaUD and CoUD, given that both substances show at least 50% heritability according to Wetherill *et al.* [36]. Overall, this study advances the field of mental health by introducing a computational model capable of automatically classifying individuals as having or not having SUD, with validation by a mental health professional. The tool can serve as a valuable supplement to psychiatric assessments, enabling faster diagnostics and timely interventions. The analysis also offers insights into the relative severity of SUD across substances and potential strategies for managing treatment. Finally, the model successfully uncovers underlying patterns in mental health disorders that may correlate with SUD severity or poly-substance involvement.

Interpretation by a mental health professional

The machine learning outputs reveal distinct behavioral patterns in the classification of substance use disorders (particularly within the group identified as addicted) when analyzing alcohol, marijuana, and cocaine data from the NSDUH. The larger cluster contains behaviors that are highly similar across individuals, corresponding to people who do not fulfill the DSM-5 diagnostic criteria for SUD. This may occur because they have no history of substance use, or because they have used substances but do not meet the full threshold for a disorder. In contrast, the smaller cluster shows greater behavioral diversity. Individuals with SUD tend to display a wider array of behaviors as captured in the NSDUH, consistent with meeting diagnostic criteria for a specific substance use disorder.

Alcohol stands out as a widely used substance with substantial participation in drinking-related behaviors. There is notable variability in patterns such as tolerance, withdrawal, and adverse consequences among both those who meet and do not meet criteria for alcohol use disorder (AUD). As a result, people can more easily adjust their alcohol consumption patterns (compared to

other substances), which influences whether they cross the diagnostic threshold for AUD. Consumption levels alone do not necessarily determine an AUD diagnosis.

For marijuana, a higher percentage of users who report consumption tend to meet the criteria for cannabis use disorder (CaUD) than is the case for alcohol and AUD. NSDUH assessments show stronger correspondence between reported behaviors and CaUD diagnosis. Individuals using marijuana at higher intensities are more likely to qualify for CaUD than comparable high-level alcohol users are for AUD.

Compared to alcohol and marijuana, cocaine users show an even higher likelihood of meeting DSM-5 criteria for stimulant use disorder. Cocaine is linked to a greater probability of fulfilling these criteria, potentially due to its distinct psychopharmacological effects relative to the other two substances.

In summary, people can engage in a broad spectrum of alcohol-related behaviors, with some meeting DSM-5 criteria for AUD and others not. In comparison, higher-frequency marijuana users are more prone to meeting CaUD criteria. Cocaine users, in turn, demonstrate the highest likelihood of satisfying the diagnostic standards for stimulant use disorder.

Clinical integration and diagnostic implications

Traditional diagnostic processes for substance use disorders typically depend on established approaches such as structured interviews (e.g., NSDUH), self-reported assessments, and DSM-5 criteria. Although these are clinically established, they can be lengthy, subject to subjective bias, and heavily reliant on clinician judgment.

This study contributes by offering visual and quantitative insights into SUD related to alcohol, marijuana, and cocaine. The machine learning findings support the following practical applications. First, the cluster visualizations serve as a decision-support tool for mental health professionals to distinguish SUD from non-SUD cases. Importantly, this approach does not require a separate control group. A new individual completing the NSDUH survey can be automatically assigned to the appropriate category using unsupervised learning-generated plots. Second, the quantified clustering results provide clear indicators of SUD severity, enabling more personalized treatment plans tailored to the disorder's specific severity.

The visualizations show that the SUD and non-SUD clusters for alcohol are positioned closest together,

followed by marijuana, with cocaine clusters being the most distant. This suggests that treatment strategies for cocaine-related SUD may need to be the most comprehensive, with moderately extensive approaches for marijuana and the least extensive for alcohol.

Conclusion

This study introduced a modeling approach to evaluate the severity of substance use disorder by leveraging data from the National Survey on Drug Use and Health (NSDUH) database, aligned with DSM-5 criteria. Three unsupervised machine learning algorithms — k-means, hierarchical clustering, and DBSCAN — were applied, consistently identifying two primary groups: respondents with and without substance use disorder.

Interpretations were derived from data visualizations and numerical outputs from a machine learning perspective, with final validation provided by a mental health professional. For future research, the proposed methodology will be extended to explore the cognitive and neurobiological characteristics associated with potential SUD in individuals.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Robbins TW, Clark L. Behavioral addictions. *Curr Opin Neurobiol.* 2015;30:66-72.
2. Ahn WY, Vassileva J. Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence. *Drug Alcohol Depend.* 2016;161:247-57.
3. Al Sukar M, Sleit A, Abu-Dalhoum A, Al-Kasasbeh B. Identifying a drug addict person using artificial neural networks. *Int J Comput Inf Eng.* 2016;10(4):611-6.
4. Saha TD, Chou SP, Grant BF. The performance of DSM-5 alcohol use disorder and quantity-frequency of alcohol consumption criteria: An item response theory analysis. *Drug Alcohol Depend.* 2020;216:108299.
5. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. 5th ed. Washington (DC): American Psychiatric Association Publishing; 2013.
6. Alcohol Research: Current Reviews Editorial Staff. Drinking patterns and their definitions. *Alcohol Res Curr Rev.* 2018;39(1):17-8.
7. Mak KK, Lee K, Park C. Applications of machine learning in addiction studies: A systematic review. *Psychiatry Res.* 2019;275:53-60.
8. Srividya M, Mohanavalli S, Bhalaji N. Behavioral modeling for mental health using machine learning algorithms. *J Med Syst.* 2018;42(5):88.
9. Tarekegn AN, Michalak K, Giacobini M. Cross-validation approach to evaluate clustering algorithms: An experimental study using multi-label datasets. *SN Comput Sci.* 2020;1(5):263.
10. Stevens E, Dixon DR, Novack MN, Granpeesheh D, Smith T, Linstead E, et al. Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. *Int J Med Inform.* 2019;129:29-36.
11. Jing Y, Hu Z, Fan P, Xue Y, Wang L, Tarter RE, et al. Analysis of substance use and its outcomes by machine learning I. Childhood evaluation of liability to substance use disorder. *Drug Alcohol Depend.* 2020;206:107605.
12. Mannes ZL, Shmulewitz D, Livne O, Stohl M, Hasin DS. Correlates of mild, moderate, and severe Alcohol Use Disorder among adults with problem substance use: Validity implications for DSM-5. *Alcohol Clin Exp Res.* 2021;45(10):2118-29.
13. Mintz CM, Hartz SM, Fisher SL, Ramsey AT, Geng EH, Grucza RA, et al. A cascade of care for alcohol use disorder: Using 2015-2019 National Survey on Drug Use and Health data to identify gaps in past 12-month care. *Alcohol Clin Exp Res.* 2021;45(6):1276-86.
14. Hayley AC, Stough C, Downey LA. DSM-5 cannabis use disorder, substance use and DSM-5 specific substance-use disorders: Evaluating comorbidity in a population-based sample. *Eur Neuropsychopharmacol.* 2017;27(7):732-43.
15. Substance Abuse and Mental Health Services Administration. National Survey on Drug Use and Health [Internet]. 2019 [cited 2023 Jan 1]. Available from: <https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health/national-releases/2019>

16. Hasin DS, O'Brien CP, Auriacombe M, Borges G, Bucholz K, Budney A, et al. DSM-5 criteria for substance use disorders: Recommendations and rationale. *Am J Psychiatry*. 2013;170(8):834-51.
17. Substance Abuse and Mental Health Services Administration. Key Substance Use and Mental Health Indicators in the United States: Results from the 2019 National Survey on Drug Use and Health [Internet]. 2020 [cited 2023 Jan 1]. Available from: <https://www.samhsa.gov/data/report/2019-nsduh-annual-national-report>
18. Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychol Med*. 2016;46(12):2455-65.
19. Chowdhry AK, Gondi V, Pugh SL. Missing data in clinical studies. *Int J Radiat Oncol Biol Phys*. 2021;110(5):1267-71.
20. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol*. 2017;9:157-65.
21. Valenti GD, Craparo G, Faraci P. The development of a short version of the internet addiction test: The IAT-7. *Int J Ment Health Addict*. 2023;23(2):1028-53.
22. Dluzniewski A, Casanova MP, Ullrich-French S, Brush CJ, Larkins LW, Baker RT. Psychological readiness for injury recovery: Evaluating psychometric properties of the IPRRS and assessing group differences in injured physically active individuals. *BMJ Open Sport Exerc Med*. 2024;10(2):e001869.
23. Bonfiglio AY, Munniksma A, Volman M, van Rooij F, Gaikhorst L. Teachers' attention to students' funds of identity in Dutch primary school classrooms. *Teach Teach Educ*. 2024;144:104584.
24. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing*. 2018;300:70-9.
25. Hauke J, Kossowski T. Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. *Quaest Geogr*. 2011;30(2):87-93.
26. Ali PJM, Faraj RH. Data Normalization and Standardization: A Technical Report. Koy Sanjaq (Iraq): Koya University .Machine Learning Technical Reports. 2014;1(1):1-6.
27. Kherif F, Latypova A. Principal component analysis. In: *Machine Learning*. Amsterdam: Elsevier; 2020. p. 209-25.
28. Sarker IH. Machine learning: Algorithms, real-world applications and research directions. *SN Comput Sci*. 2021;2(3):160.
29. Chakraborty S, Paul D, Das S, Xu J. Entropy regularized power k-means clustering. *arXiv [Preprint]*. 2020. Available from: <https://arxiv.org/abs/2001.03452>
30. Jiang X, Ma J, Jiang J, Guo X. Robust feature matching using spatial clustering with heavy outliers. *IEEE Trans Image Process*. 2020;29:736-46.
31. Zhang M. Use density-based spatial clustering of applications with noise (DBSCAN) algorithm to identify galaxy cluster members. *IOP Conf Ser Earth Environ Sci*. 2019;252(4):042033.
32. Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data Knowl Eng*. 2007;60(1):208-21.
33. Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: A clinical example. *J Clin Epidemiol*. 2003;56(9):826-32.
34. Waddell JT. Age-varying time trends in cannabis- and alcohol-related risk perceptions 2002-2019. *Addict Behav*. 2022;124:107091.
35. Dell NA, Srivastava SP, Vaughn MG, Salas-Wright C, Hai AH, Qian Z. Binge drinking in early adulthood: A machine learning approach. *Addict Behav*. 2022;124:107122.
36. Wetherill L, Agrawal A, Kapoor M, Bertelsen S, Bierut LJ, Brooks A, et al. Association of substance dependence phenotypes in the COGA sample. *Addict Biol*. 2015;20(3):617-27.