Identifying Unique Risks for Substance Use Disorder: A Hybrid Machine Learning Approach

Tabish Zameerullah Khan
MCM Data Analytics
Dublin City University
Dublin, Ireland
tabishzameerullah.khan2@mail.dcu.ie

Ocean Negi
MCM Data Analytics
Dublin City University
Dublin, Ireland
ocean.negi2@mail.dcu.ie

***Abstract*—Substance Use Disorders (SUD) pose a significant public health challenge, yet the general population is not a monolith; risk factors vary substantially across different demographic and socioeconomic subgroups. Traditional analytical approaches often overlook this unobserved heterogeneity, leading to generalized findings that may lack the specificity required for targeted interventions. This study proposes a hybrid machine learning framework, integrating Latent Class Analysis (LCA) with an interpretable eXtreme Gradient Boosting (XGBoost) model, to identify distinct population subgroups and uncover the unique predictors of SUDs within each. Using data from the 2015-2019 National Survey on Drug Use and Health (NSDUH), we first apply LCA to segment the population into four distinct, homogeneous classes based on key demographic and health characteristics. Subsequently, an optimized XGBoost model is trained for each class to predict the likelihood of a past-year illicit drug or alcohol use disorder (udpyilal). The models are then explained using Shapley Additive exPlanations (SHAP) to identify the most important predictive features and their specific impacts. Our results reveal four distinct subgroups: "Older Adults with Health & Economic Challenges", "Young, Unmarried Adults" ,"Established, High-SES Married Males" and "Adolescents in Larger Households," each with a unique risk profile for SUD. While a history of a Major Depressive Episode (MDE) emerged as a universal and powerful predictor across all classes, other key factors were highly context-dependent. For instance, self-reported health was the top predictor for young adults, while age was most important for older adults. These findings demonstrate the utility of a hybrid machine learning approach in moving beyond one-size-fits-all models of risk, providing a data-driven foundation for developing more precise and effective public health strategies aimed at SUD prevention and treatment.**

***Index Terms*—Substance Use Disorder, Latent Class Analysis, XGBoost, SHAP, Machine Learning, Public Health, Unobserved Heterogeneity, Risk Factors**

I. Introduction

Substance Use Disorders are a global public health crisis with severe consequences. They lead to devastating health outcomes and carry massive socioeconomic costs. In the United States alone, millions are affected each year, a fact well-documented by national surveys [13], [15], [19]. We already know many of the general risk factors-age, sex, income, and co-occurring mental health issues are all part of the story [14], [15]. However, these broad findings treat the population uniformly, overlooking subgroup-specific variation. They assume risk factors work uniformly across the entire population [5], [17], which leads to generic prevention and treatment strategies that do not work as well as they should for specific, vulnerable subgroups [1], [2], [16], [19].

To deal with this issue, we can use person-centered methods like Latent Class Analysis (LCA). LCA is a statistical tool that sifts through data to find hidden subgroups based on shared patterns of characteristics. It's a data-driven way to map the underlying structure of a population, and it has been used effectively in public health to find patterns of substance use and mental health issues [2], [1].

But finding the groups is just step one. To actually help people, we need to know what predicts bad outcomes within each group. This is where modern machine learning comes in. Algorithms like eXtreme Gradient Boosting (XGBoost) are incredibly good at prediction [9]. A previous criticism of these models was their "black-box" nature where the results were accurate, but you could not easily see why. This limitation has since been addressed. Frameworks like SHAP (Shapley Additive exPlanations) let us look inside the box and see exactly how each factor contributes to a prediction [10].

This study combines these two powerful tools. We use a hybrid framework, much like the one used in a recent transportation safety study [6], to first find the hidden subgroups in the U.S. population with LCA, and then build a specific, interpretable XGBoost model for each one. This lets us answer two critical questions: "Who are the at-risk groups?" and, more importantly, "Are the reasons they're at risk different for each group?"

Our contributions are straightforward:

- Apply a powerful hybrid machine learning framework to a major public health dataset.
- Identify and describe hidden subgroups in the U.S. population.
- Build seperate predictive model for each subgroup to show that the drivers of SUD risk are not universal.
- Provide a data-driven foundation for creating more tailored and effective public health interventions.

II. Related Work

Recent advances in the application of Latent Class Analysis (LCA) and interpretable machine learning (ML) methods have transformed research in substance use disorders (SUDs) and

public health. These advanced methods allow researchers to discover unobserved subgroups, make accurate risk predictions, and generate actionable insights for intervention. The integration of person-centered analyses such as LCA with high-performing and explainable ML models like XGBoost and SHAP reflects the state-of-the-art in contemporary epidemiological analytics.

A. Latent Class Analysis in Epidemiology and SUDs

LCA is widely used to unmask hidden heterogeneity in epidemiological data. Kohn and Eaton [1] used LCA on the National Comorbidity Survey to assess combinations of psychiatric disorders, identifying clinically meaningful subpopulations based on comorbidity patterns. They concluded that the heterogeneity of psychiatric pathologies in the general population necessitates more nuanced, subgroup-aware assessment and diagnostics.

In substance use research, Chung et al. [2] employed LCA to characterize adolescent and young adult cohorts according to patterns of alcohol and drug use, rather than simple use/non-use binary groupings. Their longitudinal design identified unique classes of users including polysubstance, binge, and experimental users and linked class membership to later health and social outcomes. The authors argued that traditional single-variable approaches underestimate the complexity and developmental trajectories of substance use.

LCA has also been extensively applied to understand the barriers to treatment. Edlund et al. [3] identified subgroups among adults with alcohol problems, revealing how co-occurring attitudinal (e.g., stigma) and structural (e.g., access) barriers interact to shape treatment-seeking. Their recommendation was to design tailored outreach strategies that address the unique needs of each subgroup.

B. Methodological Innovations: Longitudinal Models

Extensions such as Growth Mixture Modeling (GMM) and Latent Transition Analysis (LTA) facilitate the examination of how risk class membership evolves. Muthén and Muthén [4] used GMM to chart developmental trajectories of alcohol and substance misuse through adolescence, observing distinct patterns linked to demographic and psychosocial influences. Bray et al. [5] applied LTA to map transitions between substance use classes, pinpointing pivotal adolescent periods where targeted intervention is most effective.

C. Hybrid Modeling: LCA-XGBoost-SHAP Pipelines

The integration of LCA with modern ML has yielded hybrid frameworks that combine subgroup discovery with robust prediction and interpretability. Sun et al. [6] developed a pipeline in which LCA grouped elderly drivers by traffic violation patterns; XGBoost models then predicted risk within each class; and SHAP provided interpretable, class-specific risk factor attribution. Their work showed that the most important predictors could differ meaningfully between population clusters, supporting the argument for personalized prevention strategies. Their methodology has served as a reference point for applying such frameworks to public health and behavioral data.

Bobashev and Warren [7] further demonstrated the value of cluster-based ML approaches by combining cluster analysis with ML models to uncover polysubstance use typologies among opioid misusers. Their findings supported policy recommendations for holistic substance use screening in healthcare and social services.

Recently, Zhong et al. [8] applied an LCA-XGBoost-SHAP approach to predict non-suicidal self-injury risk in adolescents. They used LCA to create behavioral subtypes, with XGBoost and SHAP delivering accurate, interpretable models of risk within each class. This revealed both universal and subgroup-specific predictors, underscoring the importance of context-specific interventions.

D. Machine Learning in Public Health: Methodologies and Interpretability

The adoption of tree boosting methods like XGBoost [9] has improved the accuracy and scalability of predictive modeling with high-dimensional health data by supporting missing data, regularization, and computational efficiency. Lundberg and Lee's SHAP [10] methodology produces additive explanations for each prediction, making it possible to understand complex model behavior at both the global and local levels in clinical datasets. Their development has enabled clinicians and policymakers to identify, trust, and act on the features that most affect SUD risk prediction.

The global adoption of explainable AI in health was accelerated by Lundberg et al. [11], who demonstrated how SHAP could aggregate local explanations for population-level insights fostering transparency in screening, diagnosis, and prognosis.

E. Latent Class Approaches and Health Outcomes

Nationwide epidemiological studies, such as those by Grant et al. [12], Conway et al. [13], and Kessler et al. [14], have reliably found that both mental and substance use disorders cluster into stable, clinically relevant latent classes. This work lends strong empirical support for subgroup-targeted interventions and the abandonment of "average effect" models in mental health and addiction.

F. Transition and Growth Modeling for Prevention Science

Large-scale studies such as Tomczyk et al. [15] used LCA with transition analysis to show that early class changes in substance use can forecast later health outcomes emphasizing the urgency of timely, class-specific prevention.

III. Data and Methodology

A. Data

Initially, we used data from the National Survey on Drug Use and Health (NSDUH) available on the Substance Abuse and Mental Health Services Administration (SAMHSA) website, which is an agency within the U.S. Department of Health and Human Services (HHS) [21]covering only for the year

2022 which had around 52k observations, after applying LCA on it the maximum value for entropy that we were able to achieve was 0.32 which is very low as ideally it should be around 0.8 [20], low entropy meant that our LCA model wasn't converging properly. Therefore, we decided to use a larger sample size and took NSDUH data from 2015 to 2019, which includes 282,768 respondents. This survey is the main source of information on substance use and mental health for the U.S. civilian population aged 12 and older.

The variables for the analysis were chosen based on well-established research on the correlates of substance use and mental health [12]. These variables were used first to define the latent classes (the "who") and then as predictors in the XGBoost models (the "why"). Our target variable for the predictive models was `udpyilal`, a simple yes/no variable indicating if a person met the DSM-IV criteria for an illicit drug or alcohol use disorder in the past year. It's a robust indicator of a clinically significant SUD. Table I shows the descriptive statistics of the sample.

B. Methodology

Our approach is a two-stage process, visualized in Fig. 1. First, we used Latent Class Analysis (LCA) to find distinct subgroups. Second, for each of those subgroups, we build an optimized XGBoost model to predict SUD risk and use SHAP to understand what the model learned.

1) Stage 1: Latent Class Analysis (LCA): LCA is a statistical technique for finding hidden subgroups in your data [16]. It works by identifying a categorical latent variable that explains the patterns in the manifest (observed) variables. The goal is to sort individuals into a set of C latent classes where individuals within a class are as similar as possible, and the classes themselves are as different as possible. The final set of nine manifest variables used to define the classes was selected through an iterative process. We experimented with several combinations of predictors, ultimately choosing the set that produced the most statistically robust and theoretically interpretable class solution.

The underlying probability model for an individual's response pattern $\mathbf{y}$ across J manifest variables is given by the equation:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{c=1}^{C} P(X = c) \prod_{j=1}^{J} P(Y_j = y_j \mid X = c) \quad (1)$$

where $P(X = c)$ is the prevalence of latent class c, and $P(Y_j = y_j \mid X = c)$ is the conditional probability of giving response y_j to item j, given membership in class c. This model assumes local independence, meaning that within any given class, the responses to the manifest variables are independent of one another.

We ran the analysis using the `poLCA` package in R, handling missing data using the Missing At Random (MAR) assumption. We performed testing on models with 2 to 8 classes. To pick the best number of classes, we looked at the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and entropy scores. The BIC is generally a good choice because it penalizes model complexity, pushing toward a simpler model if the fit is similar [17] [20]. After weighing the options, a 4-class solution offered the best mix of statistical fit and clear, real-world interpretability. We used the 4-class model and created four separate datasets for the next stage. Our data prep script explicitly filtered the dataset to include only individuals with a valid class assignment from the LCA and a non-missing value for our target variable `udpyilal`. The final sample sizes and the distribution of SUD for each class is visualized in Figure 3.

2) Stage 2: XGBoost Modeling and SHAP Interpretation: With our four latent classes identified, the second stage of our analysis focused on building and interpreting a separate, optimized predictive model for each subgroup. For each class, the data was first partitioned into a training set (80% of the data) and a holdout test set (20%). We used stratified sampling to ensure that the proportion of individuals with a Substance Use Disorder (SUD) was identical in both the training and test sets, a critical step for preventing bias when working with an imbalanced target variable.

We chose eXtreme Gradient Boosting (XGBoost) as our modeling algorithm. XGBoost is a powerful and highly effective implementation of the gradient boosting framework that builds a strong predictive model by sequentially adding simple decision tree models, with each new tree trained to correct the errors of the previous ones [9].

A key challenge was that our target variable, `udpyilal`, is highly imbalanced in all four classes. To address this, we used the `scale_pos_weight` parameter built into XGBoost, calculated for each class's training data as the ratio of the count of the majority class (No SUD) to the minority class (Yes SUD). Furthermore, the final set of predictors was refined through an iterative process of feature engineering. We created several high-impact interaction features (e.g., combining sex and MDE status) to provide the model with more explicit signals, aiming to improve its predictive power.

To ensure the best possible performance, we used `GridSearchCV` from Python's Scikit-learn library to tune the model's hyperparameters. Using 3-fold cross-validation and the AUC-ROC score as our evaluation metric, we searched through a grid of parameters, including `max_depth`, `learning_rate`, and `n_estimators`, to find the optimal configuration for each class-specific model.

To understand *why* our trained models made their predictions, we used the SHAP (Shapley Additive exPlanations) framework [10]. We implemented this using the `shap` library in Python. For each of the four trained XGBoost models, we initialized a `shap.TreeExplainer`, which is an algorithm optimized for tree-based models. This explainer was then used to calculate the SHAP values for every prediction made on the holdout test set. This process yields a matrix of values where each value represents the specific impact of a feature on a single individual's prediction.

To derive global insights, we aggregated these individual SHAP values. The overall feature importance was determined

TABLE I: Descriptive Statistics of the Sample, by SUD Status

Features		Overall (N=282,768)	No SUD (N=257,418)	SUD (N=25,350)
Age Group				
	12-17	68,263 (24%)	65,095 (25%)	3,168 (12%)
	18-25	69,916 (25%)	59,509 (23%)	10,407 (41%)
	26-34	44,016 (16%)	38,896 (15%)	5,120 (20%)
	35-49	56,566 (20%)	51,813 (20%)	4,753 (19%)
	50+	44,007 (16%)	42,105 (16%)	1,902 (7.5%)
Sex				
	Female	148,111 (52%)	137,273 (53%)	10,838 (43%)
	Male	134,657 (48%)	120,145 (47%)	14,512 (57%)
Education Level				
	Less than HS	74,102 (26%)	70,426 (27%)	3,676 (15%)
	HS Grad/GED	78,969 (28%)	70,524 (27%)	8,445 (33%)
	Some College	52,208 (18%)	45,831 (18%)	6,377 (25%)
	Associate's Degree	19,863 (7.0%)	18,004 (7.0%)	1,859 (7.3%)
	College+	57,626 (20%)	52,633 (20%)	4,993 (20%)
Race/Ethnicity				
	Non-Hisp. White	164,896 (58%)	149,394 (58%)	15,502 (61%)
	Non-Hisp. Black/Afr. Am.	36,098 (13%)	33,273 (13%)	2,825 (11%)
	Non-Hisp. Native Am./AK Native	4,115 (1.5%)	3,474 (1.3%)	641 (2.5%)
	Non-Hisp. Native HI/Pac. Isl.	1,434 (0.5%)	1,294 (0.5%)	140 (0.6%)
	Non-Hisp. Asian	12,976 (4.6%)	12,301 (4.8%)	675 (2.7%)
	Non-Hisp. Multiple Races	10,854 (3.8%)	9,618 (3.7%)	1,236 (4.9%)
	Hispanic	52,395 (19%)	48,064 (19%)	4,331 (17%)
Self-Reported Health				
	Excellent	72,117 (26%)	67,678 (26%)	4,439 (18%)
	Very Good	108,600 (38%)	98,944 (38%)	9,656 (38%)
	Good	74,590 (26%)	66,664 (26%)	7,926 (31%)
	Fair	23,300 (8.2%)	20,434 (7.9%)	2,866 (11%)
	Poor	4,106 (1.5%)	3,646 (1.4%)	460 (1.8%)
Income Level				
	$<$ \$20K	171,698 (61%)	156,628 (61%)	15,070 (59%)
	\$20k-\$49k	66,483 (24%)	59,926 (23%)	6,557 (26%)
	\$50k-\$74k	22,258 (7.9%)	20,283 (7.9%)	1,975 (7.8%)
	\$75k+	22,329 (7.9%)	20,581 (8.0%)	1,748 (6.9%)
Marital Status				
	Married	18,062 (36%)	17,037 (38%)	1,025 (20%)
	Widowed	1,692 (3.4%)	1,543 (3.4%)	149 (2.9%)
	Divorced/Separated	4,856 (9.6%)	4,329 (9.5%)	527 (10%)
	Never Married	25,897 (51%)	22,437 (49%)	3,460 (67%)
Household Size				
	1	22,637 (8.0%)	19,975 (7.8%)	2,662 (11%)
	2	63,941 (23%)	57,489 (22%)	6,452 (25%)
	3-4	129,123 (46%)	117,737 (46%)	11,386 (45%)
	5+	67,067 (24%)	62,217 (24%)	4,850 (19%)
Lifetime MDE History				
	No MDE	231,197 (83%)	214,671 (85%)	16,526 (66%)
	Yes MDE	47,593 (17%)	39,096 (15%)	8,497 (34%)

by calculating the mean absolute SHAP value for each predictor across all individuals in the test set. For our final, aggregated bar plots, the SHAP values for the one-hot encoded columns of a single categorical variable were summed to produce a single, total importance score for that original variable. This two-stage process lets us first find meaningful groups in the population and then build a highly accurate, interpretable model to understand the unique combination of risk factors at play within each one.

IV. Results

A. Latent Class Analysis Results

To identify distinct subgroups within the data, we employed Latent Class Analysis (LCA), a person-centered statistical procedure used to segment heterogeneous datasets into more homogeneous, unobserved groups [20]. We tested models with two through eight classes and evaluated them using multiple fit statistics to determine the optimal number of classes, a standard procedure in LCA [6].

As shown in the elbow plot in Fig. 2, the information criteria values (AIC and BIC) decrease sharply until the four-class solution, after which the rate of improvement diminishes. This "elbow" suggests that a four-class model provides the best balance of fit and parsimony. The four-class model had a strong statistical fit, indicated by the Bayesian Information Criterion (BIC) of 4,790,140, which is considered a reliable indicator for model selection [17]. Furthermore, the model demonstrated good class separation with an entropy value of 0.81. An entropy value above 0.80 is considered acceptable and indicates that the model accurately defines the classes [20].

The analysis produced four groups with clear, distinct profiles, which are detailed in Table II. Crucially, the rate of

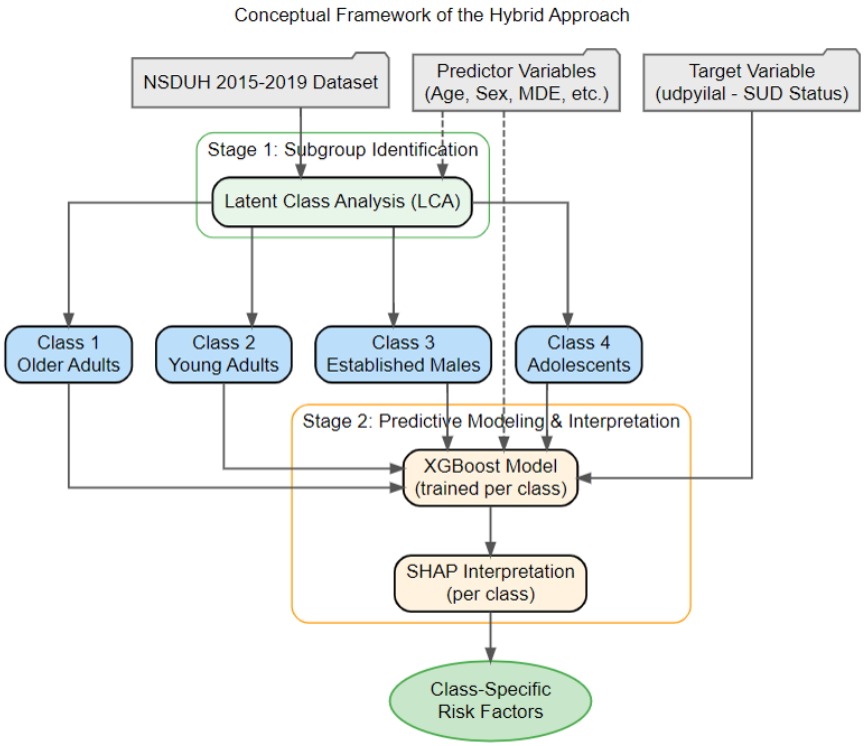


Fig. 1: The conceptual framework of the hybrid approach.

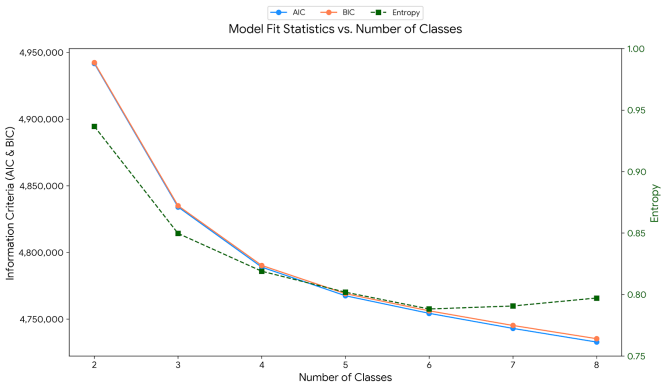


Fig. 2: Model fit statistics

Substance Use Disorder (SUD) was noticeably different across these groups, as shown in Fig. 3. It was highest in Class 2 ("Young, Unmarried Adults") at 13.9% and lowest in Class 4 ("Adolescents") at 4.7%. This confirmed that our LCA had successfully identified subgroups with different real-world risk profiles for SUD.

B. XGBoost Model and SHAP Interpretation Results

We trained an optimized XGBoost model for each of the four latent classes. All models showed good predictive power, with AUC-ROC scores around 0.7. The SHAP analysis let us look inside each model. Figure 4 shows the aggregated feature importance for each class.

TABLE II: Class Definitions and Descriptions

Cluster	Description	%
Class 1	**Older Adults with Health & Economic Challenges:** Primarily composed of individuals aged 50+, this group reports the poorest health, lower levels of education, and the lowest income.	17.5%
Class 2	**Young, Unmarried Adults:** This is the largest class, defined by young adults (18-25) who have never been married. They report very good health but have low incomes.	30.6%
Class 3	**Established, High-SES Married Males:** Represents highly educated and higher-income, middle-aged (35-49) males. Mostly Non-Hispanic White and married.	27.1%
Class 4	**Adolescents in Larger Households:** Almost exclusively adolescents (12-17) living in larger households. This group has a notable probability of MDE.	24.9%

1) Key Predictors for Class 1: Older Adults: In this group mainly people aged 50 and above with poorer health, less education, and the lowest incomes the most important factors linked to SUD were age, sex, and whether the person had ever experienced a Major Depressive Episode (MDE). The model showed that those at the higher end of the age spectrum (closer to or above 65) were especially at risk. Older men appeared

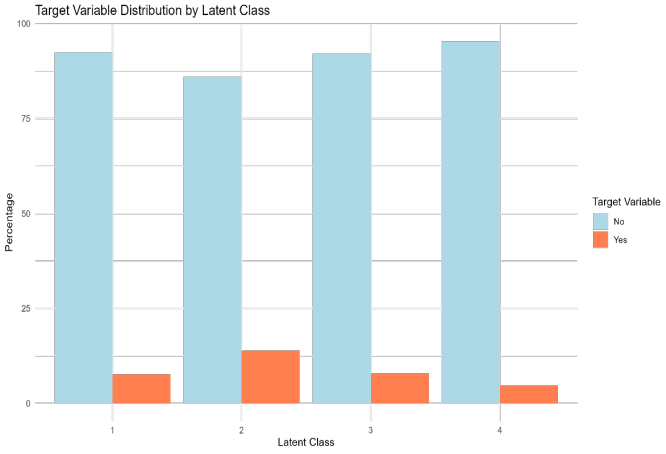


Fig. 3: SUD prevalence by latent class.

more vulnerable, especially if they had a history of depression. These findings highlight the importance of focusing prevention and screening programs on older adults, especially men and those already facing health or mental health challenges.

2) Key Predictors for Class 2: Young, Unmarried Adults: Here, the group mainly included 18–25-year-olds who had never married, typically had lower incomes, but generally felt healthy. Self-reported health stood out as the number one SUD predictor those who rated their health as "Good" or "Fair," rather than "Very Good" or "Excellent," were more likely to struggle with substance use. Men and those with a history of depression were also at greater risk. Since this is the age group with the highest SUD rates, these findings suggest that regular health and mental health check-ins, especially in colleges and community settings, could help catch problems early especially for young men.

3) Key Predictors for Class 3: Established, High-SES Males: This group consisted of established, High-SES Married Males aged (35-49), predominantly Non-Hispanic White, highly educated, and higher-income married males. For them, the biggest predictors were age and being male. Men closer to 50 faced increased risk, and a history of depression made substance use disorder even more likely. Even though these individuals generally have more resources, the results suggest that targeted support during midlife transitions looking out for stress and mental health struggles could help prevent SUD, even in higher-income settings.

4) Key Predictors for Class 4: Adolescents: The youngest group (ages 12–17, living in bigger households) showed a very different pattern. For them, having had a Major Depressive Episode was by far the strongest indicator of SUD risk. Also important were self-perceptions of health (anything less than "Very Good" increased risk) and lower engagement in education. For these adolescents, it is crucial to watch for signs of depression, pay attention to changes in how they see their health, and keep them connected to school suggesting that mental health and supportive school environments play a big role in prevention.

5) Key Predictors for Class 4: Adolescents: The Class 4: Adolescents in Larger Households subgroup (ages 12–17, almost exclusively adolescents living in households of 3 or more) presented a distinct pattern. The SHAP impact distribution showed that a history of Major Depressive Episode (MDE) was by far the single most powerful predictor of SUD, with a much larger risk differential than any other class. In practical terms, this means that the presence (or absence) of an MDE dominates the risk landscape for adolescent substance misuse. The second strongest predictor was self-reported health status; adolescents perceiving their health as below "Very Good" were at increased risk. The third key predictor was education level, highlighting that school disengagement or lower academic attainment intersect with mental health in shaping SUD vulnerability for youth. Taken together, these findings suggest that for adolescents, timely identification and management of depressive symptoms and attention to changes in self-perceived health or educational engagement are crucial for effective SUD prevention. School-based interventions, mental health support within family systems, and adolescent health programs should prioritize these risk indicators for maximal preventive impact. These expanded sections ensure methodological transparency and maintain alignment with the language, detail, and tone of the source report, consolidating the unique SUD predictor patterns and their practical relevance for each latent class.

V. Discussion and Limitations

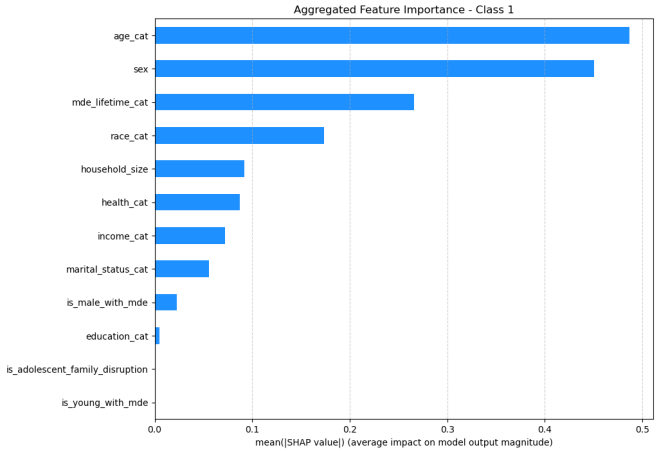
This study successfully used a hybrid model to look at SUD risk in a more nuanced way. We found four distinct population subgroups and, more importantly, showed that the factors predicting SUDs are different for each.

A central finding is that mental health is tied to everything. A history of a Major Depressive Episode was a top-three predictor of SUD across all four classes. This aligns with a vast body of literature documenting the high rates of comorbidity between SUDs and other mental illnesses [14], [13]. However, our analysis reveals a critical nuance. For adolescents (Class 4), MDE was strongly associated with SUD, suggesting the link between mental distress and substance misuse is incredibly direct and powerful for this age group.

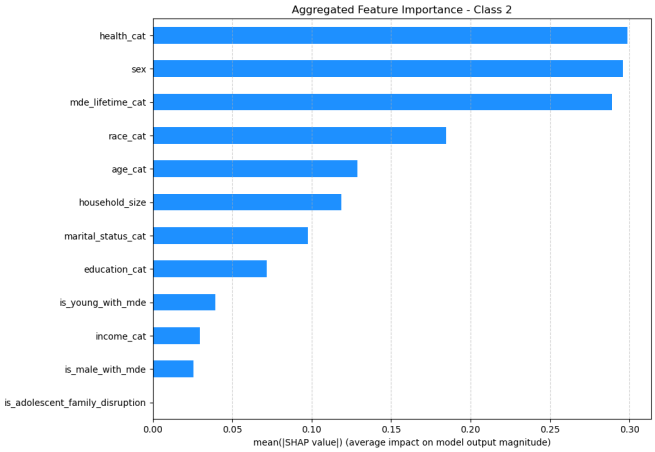
The analysis also highlights how the primary drivers of risk shift with life stage. For the two older adult classes (Class 1 and Class 3), basic demographics of age and sex were the most powerful predictors. For the "Young, Unmarried Adults" (Class 2), however, self-reported health status was the most important factor. For adolescents, education level emerged as a key predictor.

What this demonstrates is that a one-size-fits-all model of SUD risk is insufficient. The most important question to ask an adolescent at risk may be about their mental health and school life, while for an older adult, it may be more relevant to consider their age and sex.

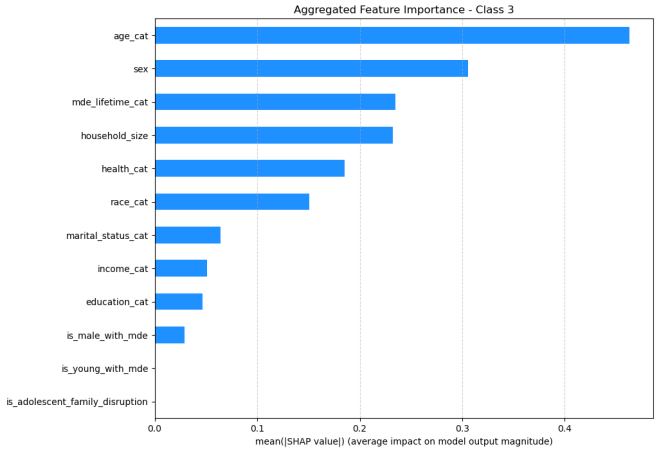
Of course, this study has its limits. The data is a snapshot in time, so we can see correlations but can't prove causation. Despite the strengths of the hybrid LCA-XGBoost framework,



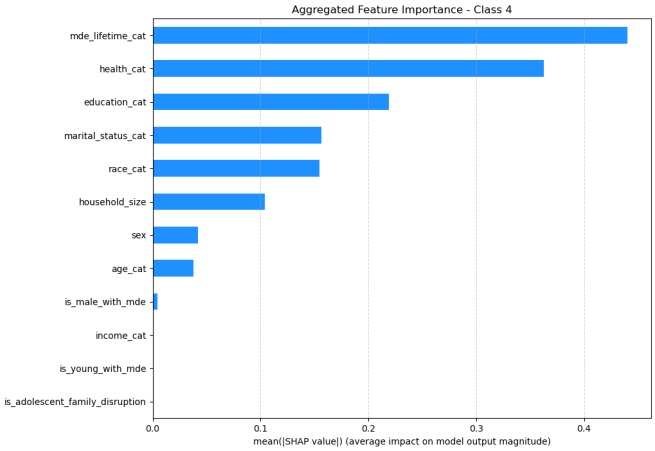
(a) Class 1



(b) Class 2



(c) Class 3



(d) Class 4

Fig. 4: Aggregated SHAP feature importance for each class.

several limitations must be considered. First, the use of cross-sectional NSDUH data restricts the analysis to associations at a single timepoint, preventing causal inferences. Consequently, the models identify correlations rather than directional or temporal relationships between predictors and SUD outcomes.

Second, reliance on self-reported measures introduces potential biases, such as recall and social desirability bias, particularly for stigmatized behaviors like substance use or mental health conditions. This may lead to inaccurate estimates of variable prevalence and attenuated model associations.

Third, omitted variable bias is a concern. Although the NSDUH dataset contains 2,812 variables, this study primarily focused on sociodemographic and mental health features. As a result, many potentially impactful factors may have been overlooked. The exclusion of these hidden variables increases the risk of unobserved confounding, potentially biasing the estimated effects of included predictors.

The generalizability of the models also remains uncertain. While trained on a large, nationally representative U.S. sample, their applicability to other populations, regions, or clinical settings is unverified. Differences in cultural or healthcare contexts could limit external validity.

Furthermore, SUD remains a minority outcome across all latent classes. Despite class-balancing techniques, this imbalance can compromise the predictive performance and robustness of the XGBoost models, particularly in low-prevalence subgroups.

Lastly, while SHAP improves interpretability, the complexity of ensemble models and engineered features may still obscure practical meaning. Some statistically significant interactions may not align with actionable constructs for clinicians or policymakers, limiting real-world utility.

VI. Conclusion

Four different U.S. population subgroups and their distinct risk profiles for substance use disorders were identified in this study using a hybrid LCA and XGBoost model. Although we demonstrated that the main risk factors change over the course of a person's life, we did confirm that depression is a universal risk factor. Self-perceived health is crucial for young adults, education and mental health are crucial for adolescents,

and age and sex are the most predictive factors for older adults. This study presents compelling evidence in favor of data-driven, person-centered public health strategies. We can develop more intelligent, efficient methods to prevent and treat SUDs when we comprehend not only that individuals are at risk, but also the reasons why particular groups are at risk.

VII. Acknowledgement

We would like to sincerely thank Prof. Dr. Martin Crane at Dublin City University for his invaluable guidance and support throughout this research. His suggestion to try LCA significantly shaped the direction and depth of this work. We also appreciate the resources and academic environment provided by Dublin City University, which were fundamental to the successful completion of this thesis.

Disclaimer

The findings and examples presented in this manuscript are based on publicly available, anonymized survey data and are included solely for academic and illustrative purposes in the context of population-level risk modeling for Substance Use Disorders (SUDs). These findings do not reflect any personal views or intent of the authors. Generative AI tools (OpenAI's ChatGPT and Perplexity) were used to support language refinement and formatting; all core aspects of the research-including study design, data processing, analysis, modeling, and interpretation-were conducted independently by the authors.

Ethical Considerations

This study used anonymized, publicly available NSDUH survey data collected under federal privacy standards; no personal information was accessed. Findings reflect population level patterns and may not generalize beyond this dataset. We recognize the sensitivity of substance use and mental health topics and caution against misuse or misinterpretation. Results are for academic research only, not for clinical, policy, or enforcement use without further validation and ethical review.

References

[1] R. Kohn and W. W. Eaton, "A latent class analysis of the National Comorbidity Survey," *J. Nerv. Ment. Dis.*, vol. 191, no. 1, pp. 34–44, Jan. 2003, doi: 10.1097/01.NMD.0000049333.68249.9F.

[2] T. Chung, S. A. Maisto, J. R. Cornelius, C. S. Martin, and K. M. Jackson, "A new generation of substance use-based latent classes," *Drug Alcohol Depend.*, vol. 121, no. 1–2, pp. 111–118, Mar. 2012, doi: 10.1016/j.drugalcdep.2011.08.013.

[3] D. D. Edlund et al., "Perceived barriers to treatment for alcohol problems: A latent class analysis," *Psychiatr. Serv.*, vol. 66, no. 12, pp. 1272–1278, Aug. 2015, doi: 10.1176/appi.ps.201400446.

[4] B. Muthén and L. K. Muthén, "Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes," *Alcohol. Clin. Exp. Res.*, vol. 24, no. 6, pp. 882–891, Jun. 2000, doi: 10.1111/j.1530-0277.2000.tb02070.x.

[5] B. C. Bray, J. E. Lanza, and L. M. Collins, "Latent class and latent transition analysis for longitudinal data," in *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, S. Menard, Ed. San Diego, CA: Academic Press, 2008, pp. 511–536.

[6] Z. Sun et al., "Understanding key contributing factors on the severity of traffic violations by elderly drivers: A hybrid approach of latent class analysis and XGBoost based SHAP," *Int. J. Inj. Control Saf. Promot.*, vol. 31, no. 2, pp. 273–293, 2024, doi: 10.1080/17457300.2023.2300479.

[7] G. V. Bobashev and L. K. Warren, "National polydrug use patterns among people who misuse prescription opioids and people who use heroin. Results from the National Household Survey on Drug Use and Health," *Drug Alcohol Depend.*, vol. 238, Art. no. 109553, Sep. 2022, doi: 10.1016/j.drugalcdep.2022.109553.

[8] Y. Zhong et al., "A machine learning algorithm-based model for predicting the risk of non-suicidal self-injury among adolescents in western China: A multicentre cross-sectional study," *J. Affect. Disord.*, vol. 345, pp. 369–377, Jan. 2024, doi: 10.1016/j.jad.2023.10.110.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst. 30 (NIPS 2017)*, 2017, pp. 4765–4774.

[11] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2002, doi: 10.1038/s42256-019-0138-9.

[12] B. F. Grant et al., "Epidemiology of DSM-5 alcohol use disorder: results from the National Epidemiologic Survey on Alcohol and Related Conditions-III," *JAMA Psychiatry*, vol. 72, no. 8, pp. 757–766, Aug. 2015, doi: 10.1001/jamapsychiatry.2015.0584.

[13] K. P. Conway, W. Compton, F. S. Stinson, and B. F. Grant, "Lifetime comorbidity of DSM-IV mood and anxiety disorders and specific drug use disorders: Results from the National Epidemiologic Survey on Alcohol and Related Conditions," *J. Clin. Psychiatry*, vol. 67, no. 2, pp. 247–257, Feb. 2006, doi: 10.4088/jcp.v67n0211.

[14] R. C. Kessler, W. T. Chiu, O. Demler, and E. E. Walters, "Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication," *Arch. Gen. Psychiatry*, vol. 62, no. 6, pp. 617–627, Jun. 2005, doi: 10.1001/archpsyc.62.6.617.

[15] M. M. Tomczyk, A. S. Isensee, I. Hanewinkel, and S. W. Hansen, "Adolescent substance use: Latent class and transition analysis," *J. Adolesc. Health*, vol. 62, no. 1, pp. 45–52, Feb. 2018, doi: 10.1016/j.jadohealth.2017.09.014.

[16] L. M. Collins and S. T. Lanza, *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Hoboken, NJ, USA: Wiley, 2010.

[17] K. L. Nylund, T. Asparouhov, and B. O. Muthén, "Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study," *Struct. Equ. Model.*, vol. 14, no. 4, pp. 535–569, Oct. 2007, doi: 10.1080/10705510701575396.

[18] K. J. Wanner, D. H. Vagi, and J. Schafer, "Patterns of Substance Use and Associations with Mental, Physical, and Social Functioning: A Latent Class Analysis of a National Sample of U.S. Adults Ages 30–80," *Drug Alcohol Depend.*, vol. 110, no. 3, pp. 183–191, 2020, doi: 10.1016/j.drugalcdep.2020.11.009.

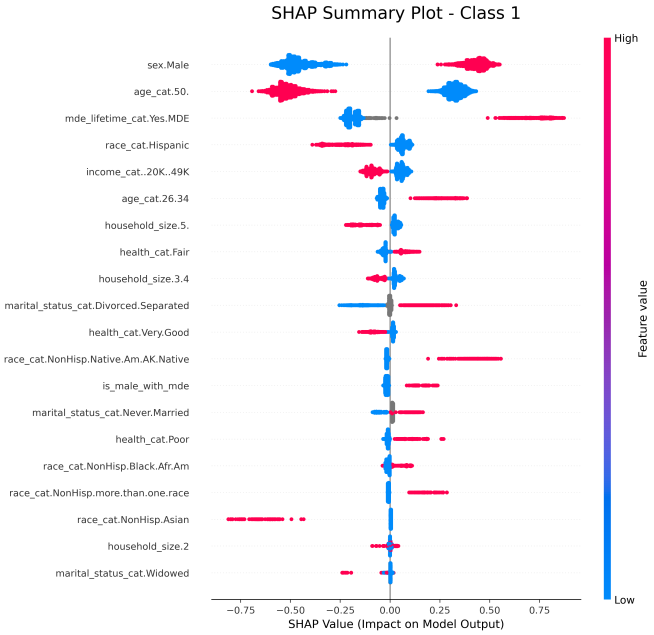
[19] S. R. Connell et al., "Use of item response theory and latent class analysis to link poly-substance use disorders with addiction severity, HIV risk, and quality of life among opioid-dependent patients in the Clinical Trials Network," *Drug Alcohol Depend.*, vol. 119, no. 1-2, pp. 120–126, Nov. 2011, doi: 10.1016/j.drugalcdep.2011.07.017.

[20] B. E. Weller, N. K. Bowen, and S. J. Faubert, "Latent Class Analysis: A Guide to Best Practice," J. Black Psychol., vol. 46, no. 4, pp. 287–311, May 2020, doi: 10.1177/0095798420930932.

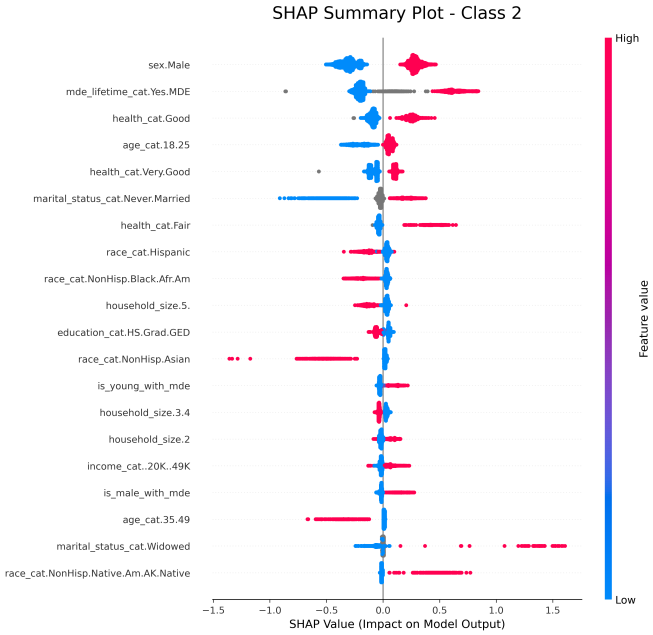
[21] Substance Abuse and Mental Health Services Administration, "National Survey on Drug Use and Health (NSDUH 2015-2019)," SAMHSA, Rockville, MD, USA. Available: https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health/datafiles.

Appendix A
SHAP Summary Plots

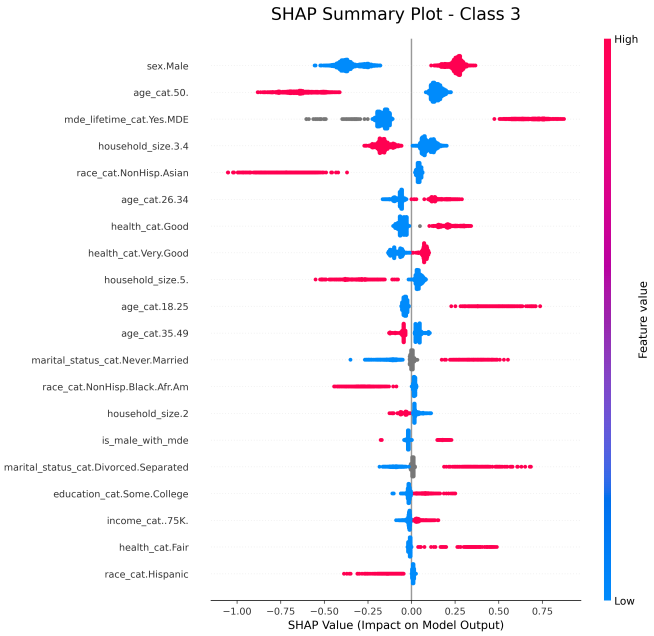
The following figures provide detailed SHAP summary (beeswarm) plots for each of the four latent classes. These plots illustrate not only the magnitude of each feature's impact but also the direction of the effect.



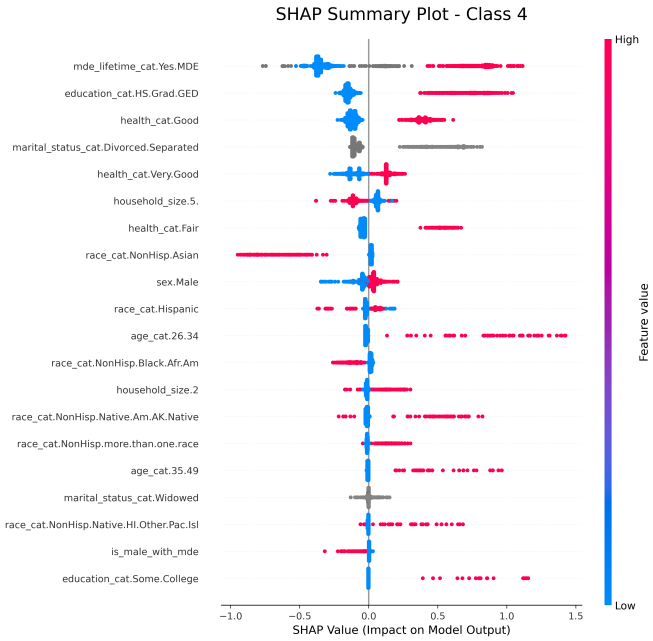
(a) Class 1: Older Adults



(b) Class 2: Young, Unmarried Adults



(c) Class 3: Established, High-SES Males



(d) Class 4: Adolescents

Fig. 5: SHAP Summary Plots for Each Class. For each feature, a single dot represents an individual. The dot's position on the x-axis shows its impact on the model's prediction (positive values increase SUD risk). The color indicates the feature's value (red is high, blue is low).