Analysis and Prediction of Relapse Likelihood for Selected Addictive Substances Following Treatment

Thomas Kidu[1], Erica King[2], Yeninda Tjoa[2], Cathy Wang[2]

[1]Data Science, Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, United States
[2]Applied Mathematics and Statistics, Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, United States

***Abstract*—Relapse following treatment for substance use disorders remains a critical public health concern, with relapse rates often comparable to those of other chronic diseases. In this study, we leveraged large-scale survey data from 2022 and 2023 (over 100,000 respondents) and derived a binary relapse outcome based on recent substance use following past-year treatment. We then built predictive models to estimate relapse probability. Multiple supervised learning models were trained, including Random Forest, Extra Trees, XGBoost, Gradient Boosting, and a neural network. Next, we proposed an ensemble model that combined the five classifiers, which achieved the best performance (Accuracy = 0.953, Recall = 0.890, F1-score = 0.864, AUC = 0.989) compared to each single model. Notably, high recall was prioritized to ensure accurate identification of high-risk relapsers. In addition, we analyzed feature importance and applied explainable AI techniques to identify key relapse predictors. The analysis revealed that frequent marijuana use was associated with a lower relapse risk, potentially indicating a substitution effect. In contrast, high-risk substances such as opioids, stimulants, and nonmedical pain relievers consistently contributed to higher relapse predictions. These findings demonstrate the feasibility of relapse risk modeling using population survey data and provide actionable insights for the early identification of individuals at risk of relapse.**

***Keywords*— relapse, rehabilitation, neural network, classification, random forest, xgboost, ensemble model, feature importance, explainable artificial intelligence**

I. Introduction

A. Motivation

Substance use disorders (SUDs) continue to pose a serious public health challenge, with rising rates of drug addiction and overdose in recent years. A critical concern in addiction recovery is the high likelihood of relapse after treatment completion. Research by the National Institute on Drug Abuse (NIDA) has shown that 40–60% of individuals treated for substance use disorders experience relapse, a rate on par with other chronic diseases like hypertension or asthma. [1-2]. This high relapse rate underscores the need for a better understanding of relapse risk factors and the development of tools to predict and prevent relapse. Early identification of patients at higher risk of relapse could enable targeted interventions and support during recovery.

In recent years, machine learning (ML) algorithms have been applied to model addiction treatment outcomes, including relapse. A 2024 systematic review by de Mattos et al. surveyed 28 studies using ML for SUD treatment outcomes and found outcomes like treatment adherence and relapse were commonly predicted with algorithms ranging from logistic regression to tree ensembles and neural networks. These studies span various substances (opioids, alcohol, cocaine, etc.) and demonstrate the potential of ML to enhance predictive accuracy in SUD treatment. However, the review also highlighted methodological gaps, such as inconsistent feature definitions and limited external validation [3].

Cavicchioli et al.[4], applied regularized logistic regression to predict alcohol relapse in patients undergoing therapy, achieving an AUC of about 0.76 in identifying those who dropped out or relapsed. Their model found that higher Alcohol Severity Index scores were the strongest predictors of relapse risk. In another study, Davis et al. [5] used machine learning to identify individual and environmental predictors of post-treatment opioid and stimulant use in youth. They observed that factors such as gender and peer-group involvement influenced relapse; notably, greater engagement in support groups significantly reduced the likelihood of relapse. These efforts indicate that ML models can capture complex, multifactorial influences on relapse. Still, many prior works used relatively small clinical datasets (on the order of a few hundred patients) or focused on specific substances, limiting generalizability.

In this study, we performed a comprehensive analysis of relapse risk using an extensive and nationally representative dataset: the 2022 and 2023 waves of the National Survey on Drug Use and Health (NSDUH). Administered annually by the Substance Abuse and Mental Health Services Administration (SAMHSA), the NSDUH provides self-reported data on substance use, treatment history, mental health, and sociodemographic characteristics from a broad cross-section of the U.S. population. Each wave contains approximately 2,400 variables per respondent, offering a rich foundation for predictive modeling. To ensure data quality and model interpretability, we implemented an extensive preprocessing pipeline including handling missing data, outlier detection, variance filtering, and correlation analysis, which reduced the feature set from over 2,400 variables to 54 highly informative features. These features span key domains such as substance use history, treatment exposure, mental health status, and early use behaviors. We then defined a binary relapse outcome based on self-reported substance use in the past month following past-year treatment, enabling us to model relapse likelihood using supervised machine learning. We trained several classification models and ultimately proposed an ensemble method that combines the strengths of the top-performing approaches. To address class imbalance and improve sensitivity to relapse events, we incorporated synthetic oversampling techniques. Finally, we applied explainable AI (XAI) methods to interpret model outputs and highlight key predictors of relapse.

II. Data collection and preprocessing

A. Data Collection

We obtained data for this study from the publicly accessible National Survey on Drug Use and Health (NSDUH) conducted by the U.S. Substance Abuse and Mental Health Services Administration (SAMHSA) for the years 2022 and 2023. NSDUH is an annual cross-sectional survey that provides nationally representative data on substance use behaviors, mental health conditions, and related health outcomes among the civilian, non-institutionalized U.S. population aged 12 years and older [6].

The 2022 survey comprised responses from 59,069 individuals, while the 2023 dataset contained 56,705 respondents. Each survey record represented one participant and included over 2,600 features covering several domains. These variables encompassed demographic factors such as age, sex, race/ethnicity, education, and criminal justice involvement; detailed substance use information including frequency, age of initiation, and recent usage; mental health assessments like depressive episodes and psychological distress; and clinical diagnoses aligned with DSM-5 criteria (e.g., opioid use disorder, stimulant use disorder). Additional variables captured treatment utilization, criminal justice engagement, and co-occurring conditions. Data were retrieved from SAMHSA's public use file repository along with supporting documentation, including official codebooks, methodology reports, and guidance notes for interpreting survey responses and managing special response codes such as "Don't know," "Legitimate skip," or "Blank" [6]. These resources facilitated accurate preprocessing and analysis.

B. Data Preprocessing Stages

Given the complex and high-dimensional nature of the NSDUH datasets, we implemented a structured data preprocessing stage aimed at ensuring data integrity, reducing redundancy, and maintaining interpretability. The detailed preprocessing steps are as follows:

Initially, we had a dataset consisting of 2,636 variables, many NSDUH variables contain conditional responses, resulting in numerous missing or uncertain values. Utilizing the official NSDUH codebook, we systematically mapped each variable to its meaningful categories (e.g., 0 = "No," 1 = "Yes," or specific numeric codes for unknown values). Responses labeled as uncertain or refused ("Don't know," "Refused") were treated as missing. Variables containing more than 50% missing values were excluded at this stage. This step significantly reduced the feature count to approximately 2,101. In addition, respondents lacking critical outcome measures or sampling weights were removed, leaving a sample size of 54,244 and 52,036 for 2022 and 2024, respectively.

Next, we applied a variance threshold procedure to eliminate features with negligible variability (threshold set at ε = 1e-4). Low-variance variables typically offered minimal analytical value due to uniform or nearly uniform responses across the sample [7]. Removing these features reduced dataset noise and computational complexity, further refining the dataset down to 1,594 features.

Subsequently, we performed outlier detection using Isolation Forest, a robust method well-suited to high-dimensional datasets [8]. Isolation Forest identifies anomalies by isolating individual observations through randomly partitioned decision trees. Observations deviating significantly from typical response patterns yield shorter isolation paths and, consequently, higher anomaly scores. Mathematically, the anomaly score $s(x, n)$ for a point x is calculated as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \text{ ,where } c(n) \approx 2\ln(n-1) + -\frac{2(n-1)}{n} \quad (1)$$

$E(h(x))$ is is the average path length of sample x, and $c(n)$is a normalization factor based on sample size n. This approach is well-suited for high-dimensional data and identifies records that are isolated more easily (i.e., potential outliers). Using this method, we removed approximately 6,000 anomalous samples, reducing noise.

To address potential redundancy and multicollinearity among variables, we applied the Pearson correlation filtering across all continuous features [9]. The Pearson correlation r measures linear between two features X and Y , and is calculated as:

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}} \quad (2)$$

Where: $\bar{x}$ and $\bar{y}$ are the means of X and Y . We used a correlation threshold of 0.85 to identify and remove highly correlated feature pairs. When two features exceeded this threshold, we retained the one with greater interpretability or predictive relevance, depending on our modeling goals. This step significantly reduced redundancy in the data and helped prevent overfitting. As a result, the feature space was reduced from 1,594 to 467 features, keeping a balance between information retention and model generalizability. The exact threshold was selected based on empirical evaluation to ensure that a sufficient number of diverse and informative variables were preserved.

Finally, guided by the objectives of this study and to enhance interpretability, we manually selected a focused set of substance-specific and clinically relevant features for relapse prediction. We identified a final set of 53 variables categorized into five core domains: (1) demographic factors (age group, sex, race/ethnicity, education), (2) age of first substance use (e.g., marijuana, heroin, stimulants), (3) frequency of substance use within the past year, (4) DSM-5 clinical diagnoses for substance use disorders (e.g., opioid use disorder, stimulant use disorder), and (5) mental health indicators (e.g., depressive episodes, psychological distress). These selected variables provided comprehensive yet focused coverage of the key constructs pertinent to relapse prediction modeling. (Details of the input feature names are listed in the appendix).

C. Relapse Feature Definition

An important step in our study was defining a meaningful relapse outcome using cross-sectional survey data. Because NSDUH data do not track individuals longitudinally after treatment, we defined relapse based on respondents' self-reported substance use behaviors within the past month, contextualized by their recent (past-year) treatment experiences.

First, respondents who reported receiving any form of substance use treatment within the past year (as determined by composite variables such as inpatient or outpatient treatment indicators) were identified as potentially at-risk for relapse. Next, we examined their reported substance use in the past 30 days, including substances like alcohol, marijuana, cocaine, opioids, stimulants, hallucinogens etc.. We aggregated these substance use indicators into a continuous "relapse score," reflecting both the diversity and intensity of recent substance use following treatment.

For respondents without recent treatment, relapse was assigned a score of zero by definition. This acknowledges that relapse inherently requires previous treatment or recovery. To clearly differentiate relapse cases, we converted this continuous relapse score into a binary outcome: respondents with a relapse score greater than zero were classified as relapse cases (relapse = 1), whereas those with a score of zero were categorized as non-relapse cases (relapse = 0). Importantly, all variables used in constructing this relapse outcome (treatment indicators, recent substance use variables) were excluded from model inputs to prevent data leakage, ensuring realistic and unbiased predictive modeling.

Following the completion of all preprocessing steps, the cleaned datasets from the 2022 and 2023 NSDUH surveys were merged, resulting in a combined analytic sample of approximately 102,497 records with 53 selected input features, comprising 81% non-relapse and 19% relapse cases. This final dataset served as the input for model training and evaluation in subsequent analysis.

III. Methodology

A. Baseline Models for Relapse Prediction: Tree-Based Classifiers and Neural Network

We explored multiple machine learning methodologies, including tree-based classifiers and neural network architectures that accurately predict relapse likelihood. Our primary objective was to evaluate and compare the predictive performance of various models, identifying the most effective approaches for capturing complex relationships and patterns in substance use data.

We initially implemented a Random Forest (RF) classifier, a robust ensemble method that combines predictions from multiple decision trees [10]. Each tree in a random forest is trained on distinct bootstrap samples of the training dataset, using random subsets of features at each decision split. This randomness helps reduce the correlation between trees, mitigating overfitting and improving generalization. Our random forest model was configured with 500 trees to enhance stability and provide reliable estimates of feature importance based on Gini impurity. As discussed later in Section IVB, this approach exhibited strong predictive performance, emphasizing its suitability for structured survey data.

Subsequently, we employed Gradient Boosted Trees using Gradient, an optimized library designed specifically for efficient training and high performance with structured tabular datasets [11]. Gradient boosting iteratively builds decision trees, each one targeting residual errors from previous ensemble iterations. Gradient integrates regularization terms, efficient handling of missing and sparse data, and sophisticated optimization algorithms, significantly boosting predictive accuracy. Hyperparameters, including the learning rate, tree depth, and the number of estimators, were systematically tuned using cross-validation. This model demonstrated excellent accuracy and robust identification of relapse predictors (results detailed in Section IVB).

We also evaluated the Extra Trees (Extremely Randomized Trees) model, closely related to Random Forest but introducing even more randomization by selecting split thresholds at random rather than optimizing for the best threshold at each node [12]. The increased randomness introduced by Extra Trees typically reduces variance further, enhancing generalization, particularly in datasets that exhibit noise or complex interactions among features. Configured similarly to the Random Forest with 500 trees, this model provided a complementary perspective and strong predictive capability, as shown in Section IVB.

In addition to traditional gradient boosting, we implemented Extreme Gradient Boosting (XGBoost), a scalable and highly efficient gradient boosting framework known for its performance on structured data [13]. XGBoost enhances the standard boosting algorithm by incorporating advanced regularization (L1 and L2), parallelized tree construction, and intelligent handling of sparse data and missing values. These optimizations significantly reduce overfitting and training time while improving predictive accuracy.

In addition to tree-based models, we explored a simple feedforward Neural Network (NN) architecture to investigate whether nonlinear relationships beyond the capacity of trees could be detected. Our neural network consisted of two dense hidden layers; the first layer included either 64 or 128 neurons with ReLU (Rectified Linear Unit) activation, followed by a smaller second hidden layer, and finally an output layer with sigmoid activation optimized for binary classification. The model was trained using binary cross-entropy loss. Although neural networks can theoretically capture intricate nonlinear patterns, their effectiveness on tabular datasets often requires larger training sets or more complex architectures. Nevertheless, this neural network offered valuable insight into whether non-tree models might detect unique predictive signals missed by ensemble tree methods.

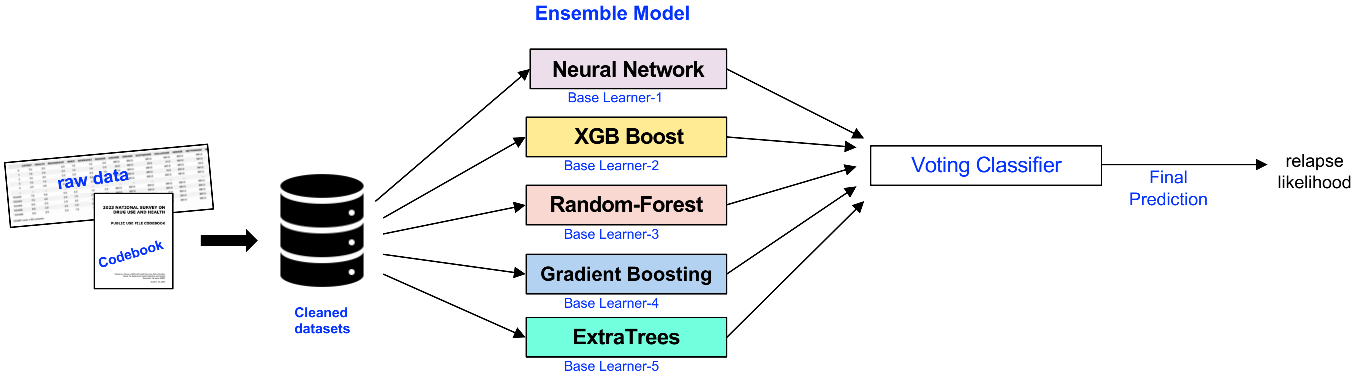


Fig.1 Overview of the Proposed Ensemble Framework for Relapse Prediction

B. Final Proposed Model: Ensemble Learning Via Voting Classifier

After extensively evaluating individual models and guided by well-established ensemble learning principles [14], we constructed a final Ensemble Voting Classifier that combined the predictions from the top five individual models: Random Forest, XGBoost, Extra Trees, Gradient Boosting, and the Neural Network. Ensemble learning approaches aggregate diverse individual classifiers, capitalizing on their complementary strengths to achieve enhanced prediction stability and accuracy.

In our ensemble implementation, we utilized a soft-voting strategy, averaging the predicted probabilities from each component model. Ensemble methods, as described by Dietterich [14], typically lead to reduced variance, improved generalization to unseen data, and more robust predictive performance by offsetting individual model weaknesses. The resulting ensemble provided superior stability and consistency in relapse prediction compared to any single model, as detailed further in Section IV.B (see also Figure 1 for a schematic representation of the ensemble methodology).

The final ensemble model represents a powerful predictive approach, combining the interpretability and strong performance of tree-based classifiers with the flexibility and potential nonlinear capturing abilities of neural networks. This integration provided the optimal balance for accurately predicting relapse, guiding targeted interventions, and informing clinical decision-making. As shown in Equation -4, the ensemble computes the average predicted probability for each class c, across all models:

$$P_{\text{ensemble}}(c) = \frac{1}{N}\sum_{i=1}^{N} P_i\,(c) \qquad (4)$$

Where N=5 is the number of models and $P_i(c)$ is the predicted probability for class c from the i-th model. The final predicted class $\hat{y}$ is determined by selecting the class with the highest averaged probability:

$$\hat{y} = \underset{c}{\text{argmax}} P_{\text{ensemble}}(c) \qquad (5)$$

This balances the strengths of tree-based models and neural networks, improving generalization, reducing variance, and enhancing relapse prediction performance.

IV. Evaluation and Results

A. Evaluation Matrix

To assess model performance, we employed several standard classification metrics: accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). Given the imbalanced nature of the dataset, where relapse cases represent a minority class, we placed particular emphasis on recall and F1-score, as these metrics are more informative for evaluating predictive performance on rare but clinically significant outcomes. Accuracy represents the overall proportion of correct predictions and is computed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (6)$$

where TP (true positives) denotes correctly predicted relapse cases, TN (true negatives) refers to correctly predicted non-relapse cases, FP (false positives) are incorrect relapse predictions, and FN (false negatives) are missed relapse cases. While accuracy provides a general overview, it can be misleading in imbalanced settings. For instance, a model that predicts only the majority class (non-relapse) may still achieve high accuracy without identifying true relapse cases. Precision, or positive predictive value, measures the proportion of predicted positive cases that are actually positive:

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (7)$$

It reflects the model's ability to minimize false positives, essential for preventing over-identification of relapse cases.

Recall, or sensitivity, quantifies the proportion of actual relapse cases that were correctly predicted:

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (8)$$

This metric is especially critical in our study context. Failing to identify individuals at high risk of relapse can lead to missed opportunities for timely intervention. Therefore, our evaluation prioritized high recall to ensure that the model reliably captures true relapse cases. F1-score serves as a composite metric that balances precision and recall. It is calculated as the harmonic mean:

Table 1: Performance Metrics of Individual Models and Final Ensemble Model for Relapse Prediction

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	0.939	0.835	0.852	0.843	0.983
Extra Trees	0.936	0.828	0.846	0.837	0.978
XGBoost	0.942	0.844	0.858	0.851	0.983
Decision Tree	0.920	0.802	0.781	0.791	0.932
Gaussian NB	0.847	0.652	0.464	0.542	0.864
Gradient Boosting	0.942	0.846	0.854	0.850	0.984
Neural Network	0.939	0.848	0.837	0.842	0.982
Ensemble Model (5-Combined model)	**0.953**	**0.8364**	**0.890**	**0.864**	**0.989**

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (9)$$

F1-score is particularly suited for imbalanced classification tasks, as it penalizes extreme trade-offs and only yields high values when both precision and recall are strong. Finally, we reported the Area Under the ROC Curve (AUC) to evaluate the model's ability to distinguish between relapse and non-relapse classes across varying decision thresholds. AUC provides a threshold-independent summary of model discriminability.

B. *Model Performance Results*

The comparative performance of all trained models including individual classifiers and the final ensemble modelis summarized in Table 1. Across the board, ensemble tree-based models and neural networks consistently outperformed simpler classifiers such as Decision Trees and Gaussian Naive Bayes, particularly in the clinically important metrics of recall, F1-score, and AUC.Among individual models, XGBoost exhibited the strongest performance, achieving an accuracy of 0.942, a recall of 0.858, and an F1-score of 0.851. These results suggest that XGBoost was highly effective at identifying relapse cases without substantially compromising precision. Closely following XGBoost was Gradient Boosting, which matched its accuracy (0.942) and achieved a comparable recall (0.854) and F1-score (0.850). These high scores reinforce the capability of boosting techniques to optimize classification performance by sequentially correcting model errors.

Random Forest and Extra Trees also yielded strong and consistent results, with Random Forest achieving an accuracy of 0.939, recall of 0.852, and F1-score of 0.843. Extra Trees showed a slightly lower accuracy (0.936) and F1-score (0.837), but maintained strong predictive utility, benefiting from the high variance-reducing nature of randomized split thresholds. The Neural Network model also performed competitively, with an accuracy of 0.939, recall of 0.837, and F1-score of 0.842 suggesting that even a relatively shallow feedforward architecture can effectively capture nonlinear relationships in high-dimensional tabular survey data.

In contrast, Gaussian Naive Bayes performed substantially worse, with accuracy dropping to 0.847, and more critically, recall falling to 0.464 and F1-score to 0.542. This underperformance likely results from the naive independence assumption inherent in the algorithm, which is inappropriate for our dataset, where survey features are highly correlated. Similarly, while the single Decision Tree model offered interpretability, it underperformed across all metrics, most notably, with an F1-score of only 0.791 and recall of 0.781, likely due to its tendency to overfit and lack of ensemble stability.

The final implemented ensemble model aggregated predictions from the five top-performing classifiers: Random Forest, Extra Trees, XGBoost, Gradient Boosting, and a Neural Network. achieved the best overall performance across most evaluation metrics, with an accuracy of 0.9530, a recall of 0.890, F1-score of 0.864, and an AUC of 0.989. While precision (0.8364) remained comparable to the top individual models, the substantial improvement in recall and F1-score indicates the ensemble's superior ability to correctly identify high-risk relapse cases. These results confirm that combining multiple learners enhances sensitivity and robustness, yielding a more balanced and reliable predictive model than any single classifier.

C. *Confusion Matrix Analysis of Individual Models and Ensemble Approach*

Figure 2 presents the confusion matrices for the five top-performing individual models Random Forest, Extra Trees, XGBoost, Gradient Boosting, and Neural Network as well as the final proposed ensemble model. These matrices offer a detailed view of model performance by reporting the number of true positives (high-risk relapsers correctly identified), true negatives (low-risk relapsers correctly identified), false positives (low-risk cases incorrectly flagged as high-risk), and false negatives (missed high-risk cases).

Among the individual models, XGBoost achieved the highest true positive count, correctly identifying 3,401 high-risk relapsers, followed closely by Gradient Boosting (3,385) and Random Forest (3,374). The Neural Network and Extra Trees classifiers trailed slightly, with 3,316 and 3,353 true positives, respectively. False negatives (i.e., missed relapse cases) ranged from 559 in XGBoost to 644 in the Neural Network, indicating variation in sensitivity across models. Similarly, false-positive cases incorrectly classified as high-risk were highest for Extra Trees (694) and lowest for XGBoost (627), reflecting trade-offs between recall and precision.

The ensemble model, which aggregated predictions from all classifiers. It correctly identified 3,525 high-risk relapsers, reducing false negatives to 435, the lowest across all models. Simultaneously, it maintained a competitive false positive count (618), comparable to the best individual models. This balance between true positives and false positives reflects the ensemble model's superior recall (0.890) and F1-score (0.864),

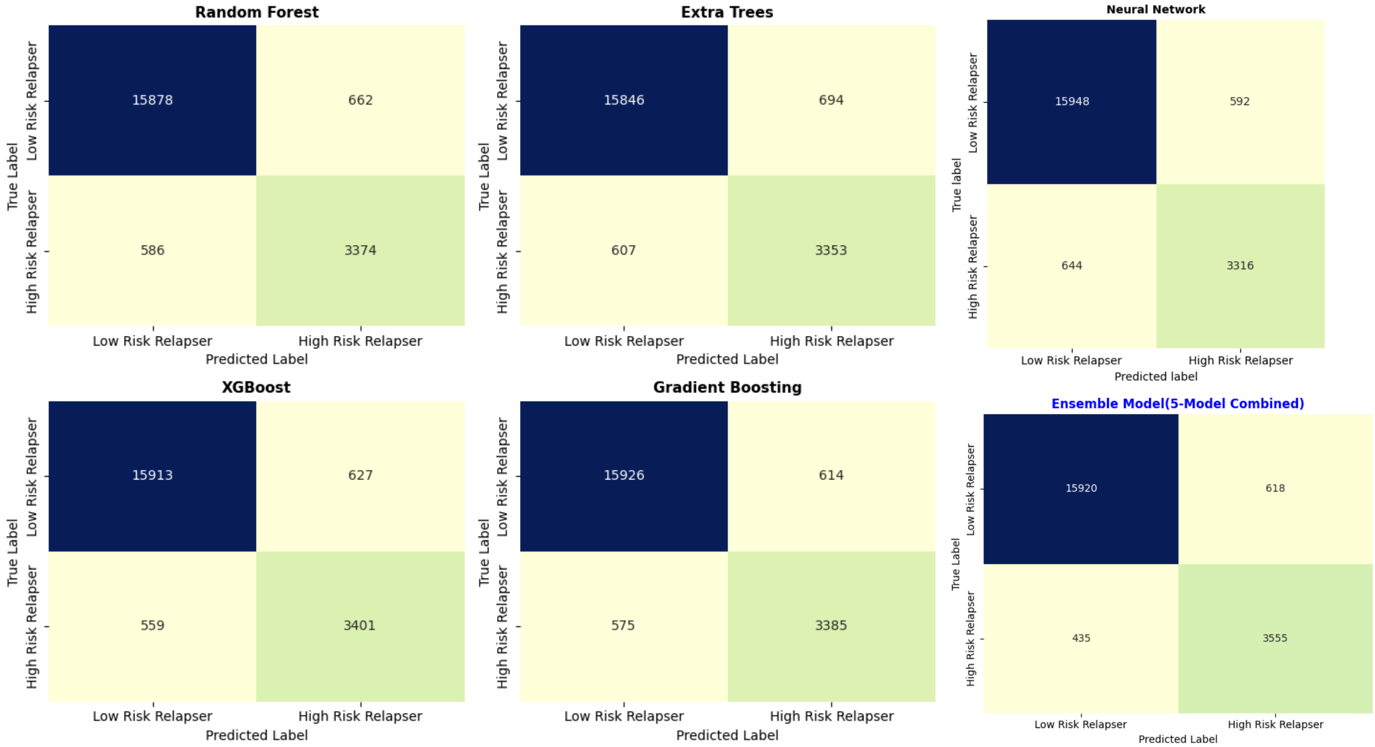


Fig. 2 Confusion Matrices of Individual Models and the Ensemble Model for Relapse Prediction

as previously shown in Table 1. These confusion matrix outcomes reaffirm the ensemble model's ability to generalize more effectively and detect relapse-prone individuals with higher sensitivity, while keeping false alarms at a manageable level. This is particularly important in clinical risk prediction, where failing to identify high-risk individuals could result in missed opportunities for intervention.

D. Interpretability and Key Predictors of Relapse: Feature Importance and Explainable AI Insights

An important aspect of this work was not only to predict relapse but also to understand which features drive the predictions. We employed two main approaches for model interpretability: feature importance analysis-based on eachmodels and explainable artificial intelligence (SHapley Additive Explanations) values for model-agnostic explanation of predictions:

Feature Importance: Across the five best-performing models Random Forest, Extra Trees, Gradient Boosting, XGBoost, and Neural Network we observed substantial consistency in feature importance rankings. As illustrated in Figure 3, marijuana-related predictors consistently emerged as the strongest factors influencing relapse likelihood. Notably, frequency of marijuana use in the past year (MRJYDAYS), overall marijuana use (MJYRTOT), and age at first marijuana use (MJAGE) were among the top three variables in all tree-based models, and were also highly ranked by permutation-based importance in the neural network model.

In addition, the use of other substances, including hallucinogens (HALLNDAYYR, frequency of use; and HALLUCAGE, age of first use), nonmedical pain relievers (PNRNMYR), and stimulants (STMNMYR, METHNDAYYR), consistently demonstrated moderate predictive importance. These findings underscore the relevance of polysubstance use patterns in predicting relapse. Clinical and behavioral indicators, such as self-rated mental health status (HEALTH), presence of any past-year substance use disorder diagnosis (UD5ILLANY), and criminal history or tranquilizer misuse (BOOKED, TRQNMYR), contributed moderately to predictive performance, especially within the ensemble approaches.

In comparison, demographic factors like age group (CATAG7) and race (NEWRACE2) exhibited relatively lower predictive strength. This suggests that dynamic clinical and behavioral characteristics offer more reliable insights into relapse likelihood compared to static sociodemographic attributes. Thus, the models emphasize the clinical importance of recent substance use behaviors and mental health status in relapse risk prediction.

Explainable AI (XAI) Interpretation: To further clarify the factors driving relapse predictions, we implemented SHapley Additive exPlanations (SHAP), an advanced interpretability framework rooted in cooperative game theory [15]. SHAP quantifies each feature's contribution to individual model predictions, where positive SHAP values reflect an increased likelihood of relapse (high-risk prediction), and negative SHAP values indicate lower relapse likelihood.

The SHAP summary plot (Figure 4) confirmed the influential role of marijuana-related variables. Interestingly, increased marijuana use days (MRJYDAYS) exhibited negative SHAP

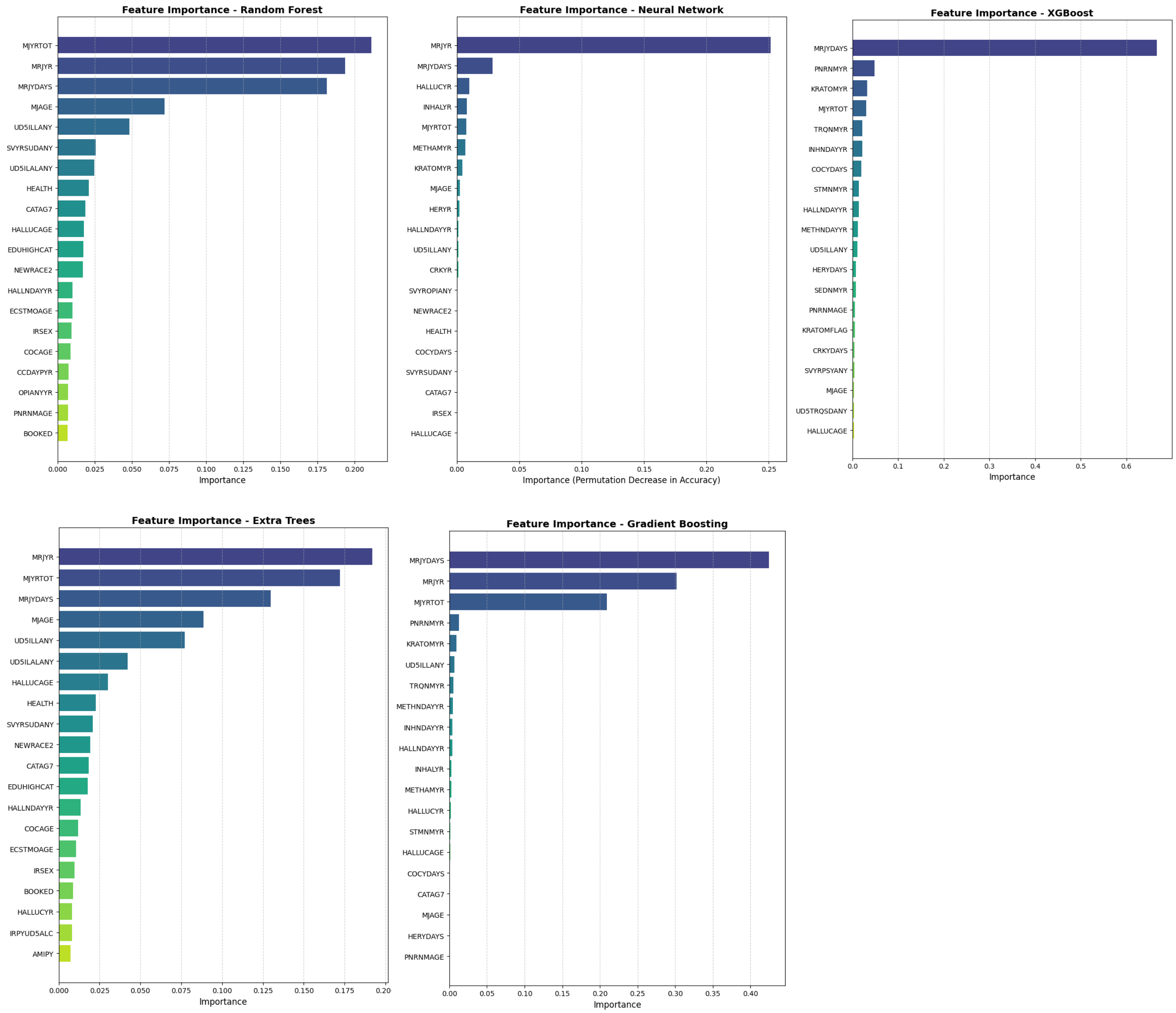
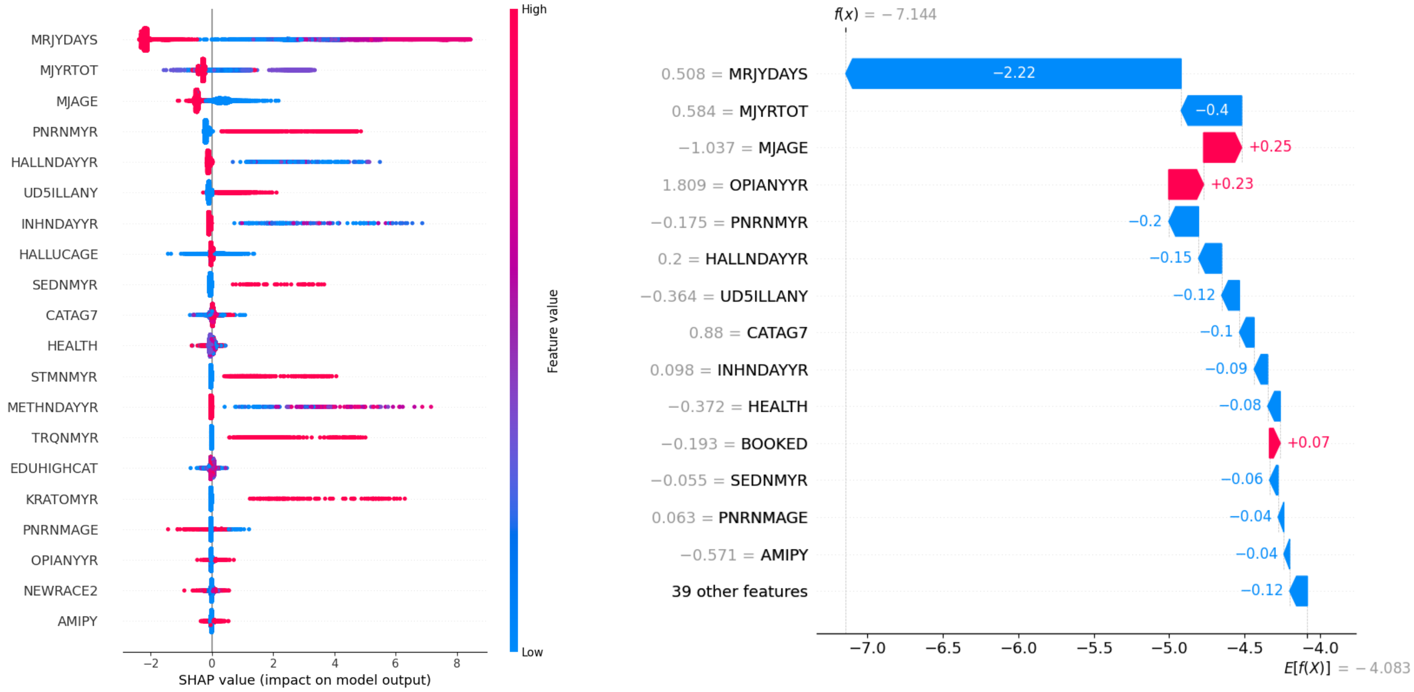


Fig. 3 Feature Importance Across Five Predictive Models (Random Forest, Neural Network, XGBoost, Extra Trees, and Gradient Boosting)

values, indicating lower predicted relapse risk with higher marijuana usage. This aligns with existing research suggesting marijuana's potential as a harm-reduction alternative or substitute substance during recovery, reducing relapse into more harmful drug use Lucas et al.,[16]. On the other hand, increased use of several other substances consistently elevated relapse risk. For example, frequent nonmedical pain reliever use (PNRNMYR), stimulants (STMNMYR), methamphetamine (METHNDAYYR), tranquilizers (TRQNMYR), opioids (OPIANYR), and kratom (KRATOMYR) showed pronounced positive SHAP values, strongly suggesting heightened relapse vulnerability with higher usage. These findings align with clinical evidence highlighting that misuse of high-risk substances significantly compromises recovery and elevates relapse risks.

Mental health and developmental factors were also influential. Specifically, the presence of any substance use disorder diagnosis (UD5ILLANY) and frequent inhalant use (INHNDAYYR) were strongly linked to increased relapse risk. Additionally, poorer perceived mental health status (HEALTH) was moderately predictive of relapse. Importantly, the age of first hallucinogen use (HALLUCAGE) showed a clear trend, indicating that younger age at initial substance exposure contributed to greater relapse likelihood, consistent with literature documenting adverse impacts of early substance initiation on addiction severity.

Demographic features offered further insights. Younger age categories (CATAG7) tended to increase relapse predictions, highlighting greater vulnerability among younger individuals. While these demographic indicators had lower overall predictive strength than clinical features, their inclusion enhanced the nuanced understanding of relapse dynamics.



(a)

(b)

Fig. 4 SHAP-Based Global and Local Interpretability of Relapse Risk Predictions

V. Discussion and Conclusion

A. Discussion

This study highlights significant predictors of relapse following substance use treatment and demonstrates the applicability of machine learning (ML) methods using large-scale survey data. Substance use patterns, particularly frequency and early age of initiation, emerged as robust predictors across all models. These findings align with clinical evidence indicating individuals who start substance use at younger ages typically experience greater dependency, complicating sustained recovery efforts, Grant & Dawson et.al [17]. Notably, treatment history and severity measures, such as prior substance use disorder diagnoses, showed high predictive relevance. This indicates that clinical assessments embedded within survey data effectively identify individuals at elevated relapse risk, underscoring the importance of tailored relapse-prevention strategies based on clinical severity [18].

The ensemble model outperformed individual models by achieving the highest recall (approximately 89%) and F1-score (around 86%), significantly improving the identification of relapse cases while maintaining reasonable precision. The superior performance of ensemble methods corroborates prior studies that advocate combining multiple classifiers to enhance generalization and robustness [14].

The explainable AI (SHAP) analysis provided interpretability, demonstrating that high marijuana use frequency negatively correlated with relapse, possibly reflecting a substitution effect during recovery, as supported by Lucas et al. [16]. In contrast, frequent use of substances such as opioids, stimulants, sedatives, and hallucinogens consistently predicted higher relapse likelihood, aligning well with clinical observations of substance-related relapse vulnerability, as studied by Darke et al. [19].

B. Conclusion and Future Works

This research presented a robust predictive framework for substance relapse using nationally representative NSDUH data, combining detailed feature selection, multiple ML algorithms, and interpretability methods. The final ensemble model effectively predicted relapse with high recall, ensuring critical relapse cases were accurately identified. Interpretability via SHAP values confirmed clinically relevant predictors, enhancing the model’s potential utility in clinical and public health settings by highlighting modifiable risk factors for targeted interventions.

Future research could enhance model validation through prospective studies tracking individuals post-treatment, providing longitudinal relapse data. Additionally, integrating external data sources (e.g., electronic health records, treatment specifics, social determinants of health) could enrich predictive accuracy. Refining relapse definitions to distinguish between relapse severity levels or employing continuous risk scoring could further enhance predictive nuance. Lastly, practical implementation should emphasize integrating ML predictions with clinical judgment, creating comprehensive, proactive relapse prevention strategies.

References

[1] A. T. McLellan, C. O. Kleber, and R. C. O'Brien, "Drug dependence, a chronic medical illness: implications for treatment, insurance, and outcomes evaluation," Journal of the American Medical Association, vol. 284, no. 13, pp. 1689–1695, 2000.

[2] Substance Abuse and Mental Health Services Administration, "Key Substance Use and Mental Health Indicators in the United States: Results from the 2022 National Survey on Drug Use and Health," Center for Behavioral Health Statistics and Quality, Rockville, MD, 2023. [Online]. Available: SAMHSA.gov/data (accessed Jan. 2025).

[3] B. P. de Mattos, C. Mattjie, R. Ravazio, R. C. Barros, and R. Grassi-Oliveira, "Craving for a robust methodology: A systematic review of machine learning algorithms on substance-use disorders treatment outcomes," International Journal of Mental Health and Addiction, Oct. 2024 (online). DOI: 10.1007/s11469-024-01403-z

[4] M. Cavicchioli, F. Calesella, S. Cazzetta et al., "Investigating predictive factors of dialectical behavior therapy skills training efficacy for alcohol and concurrent substance use disorders: A machine learning study," Drug and Alcohol Dependence, vol. 224, p. 108723, 2021.

[5] [5] J. Davis, P. Rao, B. Dilkina *et al*., "Identifying individual and environmental predictors of opioid and psychostimulant use among adolescents and young adults following outpatient treatment," *Drug and Alcohol Dependence*, vol. 233, p. 109359, 2022.

[6] Substance Abuse and Mental Health Services Administration (SAMHSA). Public-Use Data Files and Codebooks. [Online]. Available: https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health

[7] I. Guyon and A. Elisseeff, "An Introduction to Feature Extraction," in Feature Extraction: Foundations and Applications, Springer, 2006, pp. 1-25.

[8] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proceedings of the 8th IEEE International Conference on Data Mining, 2008, pp. 413–422.

[9] K. Pearson, “Note on Regression and Inheritance in the Case of Two Parents,” *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[10] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[11] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[12] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely Randomized Trees," Machine Learning, vol. 63, no. 1, pp. 3–42, 2006.

[13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 785–794, 2016.

[14] T. G. Dietterich, "Ensemble Methods in Machine Learning," in Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 1857, Springer, , pp. 1–15.

[15] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[16] Lucas, P., Walsh, Z., Crosby, K., Callaway, R., Belle-Isle, L., Kay, R., ... & Holtzman, S. (2017). Substituting cannabis for prescription drugs, alcohol and other substances among medical cannabis patients: The impact of contextual factors. Drug and Alcohol Review, 36(3), 326-333.

[17] Grant, B. F., & Dawson, D. A. (1998). Age at onset of drug use and its association with DSM-IV drug abuse and dependence: results from the national longitudinal alcohol epidemiologic survey. Journal of Substance Abuse, 10(2), 163-173.

[18] Miller, W. R., Walters, S. T., & Bennett, M. E. (2001). How effective is alcoholism treatment in the United States? Journal of Studies on Alcohol, 62(2), 211-220.

[19] Darke, S., Ross, J., & Teesson, M. (2005). The Australian Treatment Outcome Study (ATOS): What have we learnt about treatment for heroin dependence? Drug and Alcohol Review, 24(1), 3-10.

APPENDIX

Table A Summary of All Input Features Used for Relapse Prediction Modeling

FEATURE_NAME	FULL_DESCRIPTION	CATEGORY	Remark
CATAG7	Age category (6 levels)	Demographics	Input feature
HEALTH	Health conditions	Demographics	
EDUHIGHCAT	Highest education level	Demographics	
IRSEX	Sex assigned at birth	Demographics	
NEWRACE2	Race/ethnicity	Demographics	
BOOKED	Ever arrested and booked	Demographics	
COCAGE	Age when first used cocaine	Age of First Use	
CRKAGE	Age when first used crack cocaine	Age of First Use	
ECSTMOAGE	Age of first MDMA (Ecstasy/Molly) use	Age of First Use	
HALLUCAGE	Age when first used hallucinogens	Age of First Use	
HERAGE	Age when first used heroin	Age of First Use	
METHAMAGE	Age when first used methamphetamine	Age of First Use	
MJAGE	Age when first used marijuana	Age of First Use	
PNRNMAGE	Age first misused prescription pain reliever	Age of First Use	
SEDNMAGE	Age first misused sedatives	Age of First Use	
STMNMAGE	Age first misused stimulants	Age of First Use	
TRQNMAGE	Age first misused tranquilizers	Age of First Use	
AMIPY	Any mental illness past year	Mental Health	
IRAMDEYR	Major depressive episode past year (adult)	Mental Health	
SPDPSTMON	Serious psychological distress past month	Mental Health	
KRATOMFLAG	Ever used kratom	Other Drug Use	
KRATOMYR	Used kratom in the past year	Other Drug Use	
CRKYR	Used crack cocaine in past year	Substance Use	
UD5CNSANY	CNS stimulant use disorder	Substance Use Disorder	
UD5ILALANY	Any drug or alcohol use disorder past year (DSM-5)	Substance Use Disorder	
UD5ILLANY	Any illicit drug use disorder past year (DSM-5)	Substance Use Disorder	
UD5OPIANY	Prescription opioid use disorder	Substance Use Disorder	
UD5TRQSDANY	Tranquilizer or sedative use disorder	Substance Use Disorder	
SVYROPIANY	Opioid use disorder severity past year	Substance Use Disorder Severity	
SVYRPSYANY	Prescription psychotherapeutic disorder severity	Substance Use Disorder Severity	
SVYRSUDANY	Past-year substance use disorder severity	Substance Use Disorder Severity	
CCDAYPYR	Days used cocaine in past 12 months	Substance Use Frequency	
CRKYDAYS	Days used crack cocaine in past year	Substance Use Frequency	
METHNDAYYR	Days used methamphetamine in past 12 months	Substance Use Frequency	
INHNDAYYR	Days used inhalants in past 12 months	Substance Use Frequency	
HALLNDAYYR	Days used hallucinogens in past 12 months	Substance Use Frequency	
HERYDAYS	Days used heroin in past 12 months	Substance Use Frequency	
MRJYDAYS	Days used marijuana in past 12 months	Substance Use Frequency	
COCYDAYS	Days used cocaine in past 12 months	Substance Use Frequency	
IRPYUD5ALC	Alcohol use disorder (DSM-5) past year	Substance Use	
MJYRTOT	Marijuana use frequency past year (days)	Substance Use	
COCYR	Cocaine use past year	Substance Use	
HERYR	Heroin use past year	Substance Use	
METHAMYR	Methamphetamine use past year	Substance Use	
INHALYR	Inhalants use past year	Substance Use	
PNRNMYR	Pain reliever misuse past year	Substance Use	
MRJYR	Marijuana use past year	Substance Use	
HALLUCYR	Hallucinogens use past year	Substance Use	
TRQNMYR	Tranquilizers misuse past year	Substance Use	
STMNMYR	Stimulants misuse past year	Substance Use	
SEDNMYR	Sedatives misuse past year	Substance Use	
LSDYR	LSD use past year	Substance Use	
OPIANYYR	Opioids use past year	Substance Use	