Predicting Drug Consumption-A Data Driven Approach to Public Health

Vedant Singh *Information Technology Indian Institute of*

Sanjay Kumar *Information Technology Indian Institute Of*

Dr. Gourav Jain

*Computer Science Engineering Department Indian Institute Of*

Information Technology, Sonepat, India Information Technology, Sonepat, India Information Technology Sonepat, India

[svedant2103@gmail.com](mailto:svedant2103@gmail.com)

[sanjaykumar124241@gmail.com](mailto:sanjaykumar124241@gmail.com)

[j.gourav@iiitsonepat.ac.in](mailto:j.gourav@iiitsonepat.ac.in)

***Abstract*—Drug consumption trends are increasing, escalating several critical social, economic, and health problems, urging the need to propose predictive models that can study the pattern of uses, detecting risk conditions in advance. In this paper, we proposed a machine learning model to identify the group of people who are candidates to become at high risk of consuming drugs so that actions can be taken as early as possible. For this, we used publicly available Psychoactive Drug Consumption dataset having demographic, psychological, and behavioral categories. In this chapter, various machine learning models such as logistic regression, support vector machines (SVM), random forest, LightGBM, Catboost, XGBOOST, Multi- layer Perceptron(MLP) Classifier, AdaBoost and DecisionTree classifier. Among these, Random Forest performed the best, achieving an accuracy of 95.77%, while logistic regression showed the weakest results with an accuracy of 53.72% for diffrent train-test splits.The results suggest that machine learning can successfully separate the drug user from the non-user, which in turn could inform targeted public health interventions.**

***Index Terms*—Drug consumption, Machine Learning, Random Forest, XGBoost, Risk Prediction, Behavioral Analysis, Public Health**

In this chapter, we aim to develop a machine learning–based predictive framework using the **UCI Psychoactive Drug Consumption dataset** [1], which includes psychological traits, demographic information, and behavioral patterns of individuals. Our goal is to classify individuals as likely drug users or non-users and to identify the most effective classification algorithm based on model performance. The proposed system is also intended to provide a foundation for future integration into mobile health (mHealth) applications and wearable devices for real-time monitoring and personalized intervention. The main aim of this research is to identify a predictive model that can accurately classify individuals based on their likelihood of psychoactive drug uses. Also, provide the meaningful insights that support public health policies, early intervention programs, and awareness campaigns. Apart from these, we also want to explores time-series forecasting models (e.g., ARIMA and LSTM) for predicting future trends in drug usage rates.

1. Introduction

Drug addiction and substance abuse have emerged as press- ing global public health concerns, contributing to a wide range of socio-economic and health problems. The early onset of drug usage is often associated with long-term addiction, mental health disorders, criminal tendencies, and an increased burden on healthcare systems. Traditional methods of identifying and addressing drug consumption—such as clinical assessments, interviews, and surveys—are limited in scalability, prone to response bias, and often delayed in detection, making timely intervention difficult.

With the advent of large-scale psychological and behavioral datasets, Machine Learning (ML) has opened new avenues for early and accurate identification of individuals at risk of substance abuse. ML techniques can analyze complex, multi-dimensional data to uncover hidden patterns and associations that are difficult to detect using conventional approaches. This enables predictive models to be used not only for individual risk assessment but also for supporting public health policies and designing preventive strategies.

The highlights of this chapter are-

1. **The rising issue of drug consumption demands early detection methods, where machine learning offers scalable and accurate predictive capabilities.**
2. **This study builds a predictive model using the Psychoactive Drug Consumption dataset, incorporating demographic, psychological, and behavioural factors, and compares several machine learning algorithms.**
3. **Random Forest achieved the highest prediction accuracy, and future work is proposed with time-series models and real-time monitoring systems for enhanced early intervention.**
4. Literature Review

Understanding and predicting psychoactive drug consump- tion has been an area of active research due to its significant impact on individual well-being and public health. This section provides a comprehensive review of relevant literature and explains how previous findings have influenced the methodol- ogy and objectives of this chapter. It outlines the structure of existing research, identifies limitations, and presents how this work advances current practices in the field.

Numerous studies have investigated the behavioral, psy- chological, and demographic aspects associated with drug consumption. Psychoactive substances are known to influence perception, cognition, mood, and behavior, often leading to long-term dependence and adverse social consequences. The behavioral shifts resulting from substance abuse not only affect individuals but also impose a broader burden on families, communities, and public health systems. Some researchers have highlighted the role of personality traits in substance use behavior. The Five-Factor Model (FFM) — encompass- ing Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness — has been widely adopted to study in- dividual tendencies related to drug usage [1]. Psychological factors such as impulsivity and sensation seeking have also been strongly correlated with the likelihood of drug use initiation and continuation.

In the previous work, researchers utilized the structured dataset like UCI Psychoactive Drug Consumption dataset, which contains individual-level data including demographic information (age, gender, education) and psychological as- sessments. In many of them, classical machine learning al- gorithms such as Logistic Regression, Decision Trees, K- Nearest Neighbors, Naive Bayes, and Random Forests are utilized to distinguish between drug consumption user’s and non-consumption user. These models demonstrated moderate success and provided useful insights for classification tasks related to substance use. Several models have been used for this purpose and some of them are discussed below:

# Logistic Regression

Logistic Regression [3] is used in the project as a baseline model to predict whether a person has used a specific drug. It’s a simple yet effective algorithm for binary classification and serves as a reference point to compare with more complex models.

# SVM (Support Vector Machine)

Support Vector Machine [4] is included to evaluate how well it can distinguish between drug users and non-users. It works by finding the best possible boundary between the two classes and is assessed alongside other models for its accuracy and reliability.

# Random Forest Classifier

Random Forest [3] is used for its ability to combine multiple decision trees, improving overall prediction accuracy. It also provides insights into which features, such as age or person- ality traits, play a bigger role in predicting drug consumption.

# Decision Tree Classifier

This model is included for its interpretability and straight-

forward logic. It builds a tree-like structure showing how predictions are made based on different features, helping to understand the reasoning behind each classification.[3]

# Naive Bayes (GaussianNB)

Naive Bayes [3] is a lightweight, fast model used to test how well a simple probabilistic approach performs. It assumes the features are independent and offers a quick benchmark for comparison with more advanced models.

# Ridge Classifier

Ridge Classifier [8] is a regularized linear model that helps prevent overfitting. It’s used in the project to check how well a linear approach performs when there are many correlated features in the data.

# MLP Classifier (Neural Network)

The MLP Classifier [6] is a type of neural network included to explore how well deep learning models handle the drug prediction task. It’s capable of capturing complex, non-linear relationships in the data.

# SGD Classifier

This model [5] uses stochastic gradient descent to train ef- ficiently, especially on larger datasets. It’s included as a fast, scalable linear model to evaluate its performance against other classifiers.

# XGBoost Classifier

XGBoost [7] is a powerful gradient boosting model used in the project for its high accuracy and robustness. It’s known for handling imbalanced data well and often delivers top performance across various classification tasks.

# LightGBM Classifier

LightGBM [7] is incorporated for its speed and efficiency on large datasets with many features. It uses histogram-based learning and handles categorical data and missing values well, making it ideal for complex behavioural datasets.

# CatBoost Classifier

CatBoost [7] Classifier is chosen for its ability to natively handle categorical variables like education or gender without needing manual encoding. It provides high accuracy and works well even when data is sparse or has imbalanced class distribution.

# KNN (K-Nearest Neighbour)

K-Nearest Neighbors (KNN) [3] is used to classify individuals based on similarity to others in the dataset. It’s a non- parametric method that works well for smaller datasets and gives insights into how closely behaviour patterns relate among users.

# AdaBoost Classifier

AdaBoost [3] works by combining several weak learners to improve prediction accuracy. It’s included to see how boosting techniques enhance model performance, especially for more challenging drug classification cases.

This research is not only proved the efficiency of machine learning for predicting drug consumption risk, but also sets the foundation for further developments, such as applying the time-series forecasting models with Long Short-Term Memory (LSTM), Auto-regressive Integrated Moving Average

(ARIMA) and Prophet. Moreover, we suggest that real-time risk assessment and intervention would be enabled by interfac- ing with web or mobile applications and by the incorporation of Internet of Things(IoT)/wearable health monitors. However, despite these contributions, several limitations persist in the literature, restricting the overall predictive capability and gen- eralizability of existing models.

1. PROBLEM OUTLINES

Drug consumption behavior is influenced by a complex combination of psychological, demographic, and behavioral factors, making it difficult to detect and predict using con- ventional methods. Based on literature, several key challenges exist:

* **Ineffective Traditional Methods:** Manual surveys and clinical screenings are time-consuming, difficult to scale, and susceptible to biases such as social desirability and underreporting.
* **Complex Interdependencies:** The interaction between personality traits (such as neuroticism, extraversion, and impulsiveness) and environmental or demographic at- tributes often creates non-linear and multi-layered risk patterns.
* **Lack of Scalable Predictive Tools:** There is a growing need for automated, accurate, and data-driven solutions that can be deployed at scale for early identification and intervention in drug abuse cases.
* **Limited Use of Advanced Machine Learning Models:** Most studies focus on basic classifiers and overlook ad- vanced models such as XGBoost, CatBoost, LightGBM, SVM, and deep learning methods.
* **Binary Classification Only:** Prior research often reduces the temporal granularity of drug usage (e.g., “used in last day”, “month”) into binary outcomes, losing behavioral nuance.
* **Lack of Multi-Label or Multi-Class Modeling:** Many models handle each drug separately, ignoring scenarios where users consume multiple substances.
* **Absence of Feature Engineering and Dimensionality Reduction:** Prior work rarely includes PCA, interaction terms, or correlation-based feature filtering to improve model performance.
* **Single Train-Test Evaluation:** Use of a single 70:30 split lacks statistical robustness. Cross-validation is largely absent.
* **No Hyperparameter Tuning:** Default model settings are commonly used, ignoring tuning practices that enhance accuracy and generalization.

1. PROPOSED METHODOLOGY

The methodology adopted in this study is designed to sys- tematically address the challenge of predicting psychoactive drug consumption using machine learning (ML) algorithms is comprised of several phases, namely: preprocessing, ex- ploratory data analysis, feature engineering, model develop- ment, performance evaluation, hyperparameter tuning, trend

forecasting, and real-world deployment. Each of these stages plays a critical role in ensuring the robustness and applicability of the final predictive system.

1. *Data Preprocessing*

Preprocessing is a necessary step to clean up the data set, organize it, and train the machine learning model. It includes,

* 1. **Pruning Non-Informative Columns:** The column ID was deleted as it is only a consecutive number and has no statistical importance for predictions [6].
  2. **Ordinal Encoding of Categorical Variables:** Columns like Age, Education,and Country also have ordinal/ranked inputs in string format. These were numerically processed using ordinal encoding to produce a rank order, e.g.: Age: ‘18–24’*→* 0, ‘25–34’ *→* 1, ‘35–44’ *→* 2, and so on.

In this way, the relevant order in the data is preserved [6].

* 1. **Binary Label Encoding as Target Variables:** Three drugs were described by binary class labels. CL4, CL5, CL6 (recent/frequent users) were attributed as 1 (User), and all others were assigned 0 (Non-user). This way of modeling the classification problem is practical and in ac- cordance with public health guidelines on risk prediction

[3] [13].

* 1. **One-hot Encoding of Nominal Categorical Features:** Gender, for example, does not have a natural order, and features such as Gender, Ethnicity, and Country were converted to binary features using OHE to prevent false ordinal assumptions [6].
  2. **Feature Scaling (Standardization of Numerical Fea- tures):** All numeric features were scaled to a mean of

0 and a variance of 1 using StandardScaler. This is an essential step for algorithms, such as SVM, KNN, and Logistic Regression, which are impacted by the magnitudes of input features [6].

* 1. **Handling Class Imbalance:** The ratio of users to non- users may be disproportionately small. Methods such as SMOTE (Synthetic Minority Oversampling Technique) and class weighting in algorithms can be used to com- pensate for imbalanced distributions [13].

1. *Exploratory Data Analysis (EDA)*

For creating actionable insights from the recording and in further steps of constructing a model, deep Exploratory Data Analysis (EDA) is performed. The insights and statistical analysis allowed us to gather insights regarding the distribution and relationship of the features and their impact on drug usage behavior. We performed EDA in order to determine the relationship between Drug usages across the ages, Gender- based drug usage, Drug user counts by ethnicity, Education and Country-Wise Drug Usage. Experimental result of all shown in section V.

1. *Feature engineering and selection*

Feature engineering was used to enhance model perfor- mance by operating on the input space and simplifying it. For this, several techniques are used.

# PCA for Dimensionality Reduction

PCA was employed to handle high-dimensionality and multicollinearity. PCA transformed correlated input vari- ables into a reduced number of uncorrelated components by preserving 95% of the variance of the data. This yielded better training and decreased overfitting, espe- cially in the context of models like MLP and SVM.

# Elimination of Redundant Features

Highly correlated features (correlation *>* 0.85) were further removed to avoid multicollinearity. This decreased some of the noise, and made the model more stable, especially in the case of linear classifiers. Features with very low variance or without predictive value were also removed [6]. In all, these methods resulted in a reduction of the dataset while improving comprehensibility and learning efficiency among a number of classifiers.

# Interaction Terms

Features like *Neuroticism × Impulsiveness* were formed to represent psychological interactions that influence drug use. Such complex structures can enhance the expressive- ness of the model [13].

1. *Model Training*

To develop an effective predictive system for drug con- sumption classification, a comprehensive training pipeline was implemented using a wide range of machine learning classi- fiers. The goal was to assess and compare multiple algorithmic approaches and identify the models that best captured the patterns in demographic and psychological data associated with psychoactive substance use.

# Classifiers Used

A total of 14 classification algorithms were trained and tested. These models represent a variety of machine learning families, each chosen for its unique strengths in handling structured, behavioral, and psychological data.

* + - **Logistic Regression:** Served as a baseline linear model to evaluate the fundamental separability of user and non-user classes.
    - **Ridge Classifier:** An L2-regularized linear model that helped manage multicollinearity between features.
    - **Stochastic Gradient Descent (SGD) Classifier:** It Offered fast optimization through incremental updates, suitable for high-dimensional settings.
    - **Decision Tree:** Used for its interpretability and rule- based structure, although it is prone to overfitting.
    - **Random Forest:** An ensemble of decision trees that improved generalization through bootstrapping and random feature selection.
    - **XGBoost, LightGBM, and CatBoost:** Employed as advanced gradient-boosting techniques delivering strong accuracy and robustness. These models are especially effective with mixed-type structured data and allow for complex decision boundaries.
    - **AdaBoost:** Included to examine its performance through adaptive boosting on weak learners.
    - **K-Nearest Neighbors (KNN):** Implemented to clas- sify based on feature similarity. Though simple, its performance was influenced by the curse of dimen- sionality, which was mitigated using PCA.
    - **Naive Bayes:** Utilized probabilistic assumptions for fast classification and served as a lightweight bench- mark model. It assumes conditional independence, which may not always hold in this dataset but still provides useful comparisons.
    - **Multi-layer Perceptron (MLP):** Applied as a feed forward neural network model to capture non-linear re- lationships. It was trained using backpropagation with ReLU activation and regularized to prevent overfitting.
    - **SVM Classifier:** Used with both linear and RBF ker- nels. It performed well in high-dimensional spaces and proved effective for datasets where margins between classes were well-defined.

# Train–Test Splits

To ensure the robustness and generalizability of each model, training and evaluation were performed using three different data splits ratios:

* + - 80% Train – 20% Test
    - 70% Train – 30% Test
    - 90% Train – 10% Test

These configurations allowed the analysis of model be- havior under different volumes of training data. Each split was *stratified*, meaning the proportion of ‘User’ and ‘Non-user’ classes was preserved across both training and test sets to avoid class imbalance distortion.

1. RESULTS AND SETUP
2. *Dataset Description*

The dataset utilized in this research is the ”Drug Con- sumption (Quantified)” dataset obtained from the **UCI Machine Learning Repository** [1]. It includes data from 1,885 respondents and features 31 attributes capturing various personal, psychological, and behavioral traits. These features include demographic characteristics such as age (grouped into ranges), gender, level of education, country of origin, and ethnicity. Psychological traits are assessed using the Five-Factor Model, commonly known as OCEAN — representing Openness, Conscientious- ness, Extraversion, Agreeableness, and Neuroticism — calculated from responses to IPIP survey questions. In addition, the dataset incorporates behavioral variables such as impulsivity and sensation-seeking tendencies.

The dependent variables are the self-reported usage levels of 18 psychoactive substances (e.g., alcohol, cannabis, heroin, ecstasy). Each drug has its own column, labeled on a 7-point ordinal scale (CL0 to CL6), indicating recency of use. For this study, the classification problem was simplified into a binary task by grouping frequent users (CL4 to CL6) as ’users’ and the rest (CL0 to CL3) as ’non-users’ [2]. This transformation enables the use of standard binary classifiers.

TABLE I

Dataset Description.

|  |  |
| --- | --- |
| **Name** | **Description** |
| **ID** | Number of records in the original database |
| **Age** | Age of participant |
| **Gender** | Gender of participant |
| **Education** | Level of education of participant |
| **Country** | Country of current residence of participant |
| **Ethnicity** | Ethnicity of participant |
| **Nscore** | Neuroticism personality trait |
| **Escore** | Extraversion personality trait |
| **Oscore** | Openness to experience |
| **Ascore** | Agreeableness personality trait |
| **Cscore** | Conscientiousness personality trait |
| **Impulsive** | Impulsiveness behavioral score |
| **SS** | Sensation seeking score |

1. *Evaluation Metrices*

The following standard metrics were computed for each classifier:

* + - * **Accuracy:** Measures the proportion of correct predic- tions across the entire dataset.

applied across models. The average performance across folds was used to ensure robustness [4].

1. *Hyperparameter Optimization*

To enhance the predictive performance of the classifiers beyond default configurations, hyperparameter tuning was performed using a systematic search approach. Hyperpa- rameter optimization is crucial in machine learning as it controls model complexity, regularization strength, learn- ing rates, and other key behaviors that affect accuracy, generalization, and training efficiency.

* + - * **Model Selection:** The **Random Forest** classifier demonstrated the highest predictive accuracy among all models evaluated in this study.
      * **Hyperparameter Optimization:** GridSearchCV with 5-fold cross-validation was used to optimize the model. The best configuration was:
        + n\_estimators: 200
        + max\_depth: 30
        + min\_samples\_split: 2
        + min\_samples\_leaf: 1
        + max\_features: ’sqrt’

Accuracy = *TP* + *TN*

*TP* + *TN* + *FP* + *FN*

(1)

* + - * + bootstrap: True
        + random\_state: 42

While useful for a general overview, accuracy can be misleading when classes are imbalanced.

* + - * **Precision:** Indicates how many of the predicted pos- itive instances are actually positive. High precision is essential to minimize false alarms (e.g., falsely labeling a non-user as a user).
        + n\_jobs: -1 (for parallel computation)

Where, bootstrap=True enabled bootstrapped sampling, increasing robustness. A large ensemble of deep deci- sion trees (maxdepth=30) was used to capture com- plex, non-linear relationships in the data. A high number of trees (nestimators=200) improved pre-

Precision = *TP TP* + *FP*

(2)

diction stability. maxfeatures=’sqrt’ helped reduce variance by selecting a random subset of features

* + - * **Recall (Sensitivity):** Measures how many actual posi- tive instances were correctly identified. In the context of public health, high recall ensures that real users are not overlooked.

for each split. Low values of minsamplessplit and minsamplesleaf allowed deep tree growth, capturing subtle signals in the data. These tuned parameters sig- nificantly enhanced performance across all evaluation

Recall = *TP TP* + *FN*

(3)

metrics.

1. *Results Discussion*
   * + - **F1-Score:** The harmonic mean of precision and recall. This is particularly informative when classes are imbal- anced, as it penalizes extreme values in either metric.

In this chapter, we developed and evaluated a drug consumption prediction model using machine learning techniques on the **UCI Drug Consumption dataset**

F1-Score = 2 *·*  Precision *·* Recall

Precision + Recall

(4)

[1]. The classifiers assessed included Random Forest, XGBoost, Support Vector Machine (SVM) LightGBM,

* + - * **ROC-AUC (Receiver Operating Characteristic – Area Under Curve):** Evaluates the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) across various threshold settings. A higher AUC indicates better model discrimination.

The ROC curve visualized in Figure 4.7 shows that tree-based models (e.g., Random Forest and XGBoost) have the highest AUC values, indicating excellent classification ability across different thresholds.

* + - * **Cross-Validation Accuracy:** To avoid overfitting and improve generalizability, 5-fold cross-validation was

CatBoost, K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, Gaussian Naive Bayes, AdaBoost, Ridge Classifier, MLP and SGD Classifier. The eval- uation focused on various performance metrics and compared results with those from a referenced research paper titled “**Predictive Model of Psychoactive Drugs Consumption using Classification Machine Learn-** **ing Algorithms**” [2].

1. *Current Analysis based on EDA:* For creating ac- tionable insights from the recording and in further steps of constructing a model, deep Exploratory Data

Analysis (EDA) is performed. The insights and statis- tical analysis allowed us to gather insights regarding the distribution and relationship of the features and their impact on drug usage behavior. We performed EDA in order to determine the relationship between Drug usages across the ages, Gender-based drug usage, Drug user counts by ethnicity, Education and Country- Wise Drug Usage. Experimental result of all shown in section V.

# Age–Drug Relationship

A detailed line plot was designed to compare drug usage between age groups with detailed age levels (18–24 to 65+). Results indicated that substances including cannabis, alcohol, ecstasy, and cocaine were most commonly used among youths and young adults aged 18–24 and 25–34 years [6]. This spike can be due to many socio-psychological factors like more social exposure, experimentation, stress in career-building phases, and peer pressure. Curiously, drug use starts to decrease steadily past the age of 44, which implies that lifestyle stabiliza- tion and heightened health awareness could act as natural barriers to drug use in the elderly.

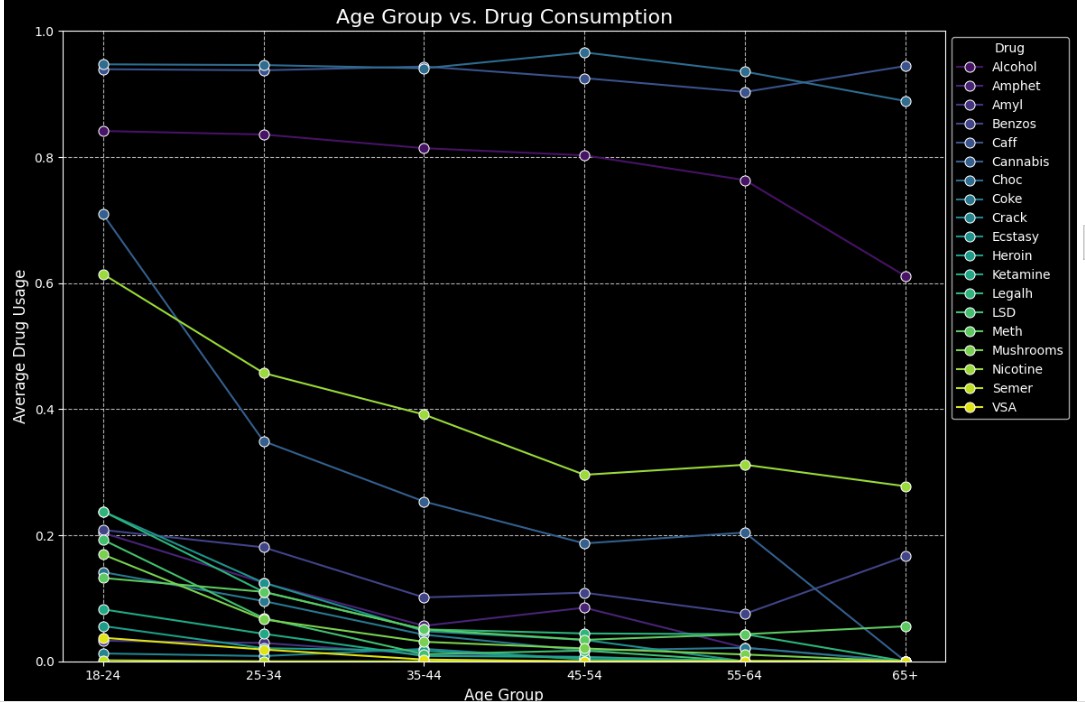


Fig. 1. Drug usages across the ages.

# Gender-Based Usage

A comparison of drug consumption by gender was carried out using stacked bar charts. In general, our findings are consistent with international literature, which indicates males consistently report higher levels of drug use across both legal (e.g., alcohol and nicotine) and illicit substances (e.g., heroin, LSD, cocaine) [6]. This gender difference aligns with prior public health research showing that men are more susceptible to high-risk behaviors, have more exposure to drug-using environments, and may experience unique cultural or psychological stressors. It was also observed that while women had lower usage rates overall, in the case of some substances (e.g., prescription drugs), the gender gap

was less pronounced—possibly indicating dispari- ties in access, intent (e.g., self-medication), and the role of social stigma.

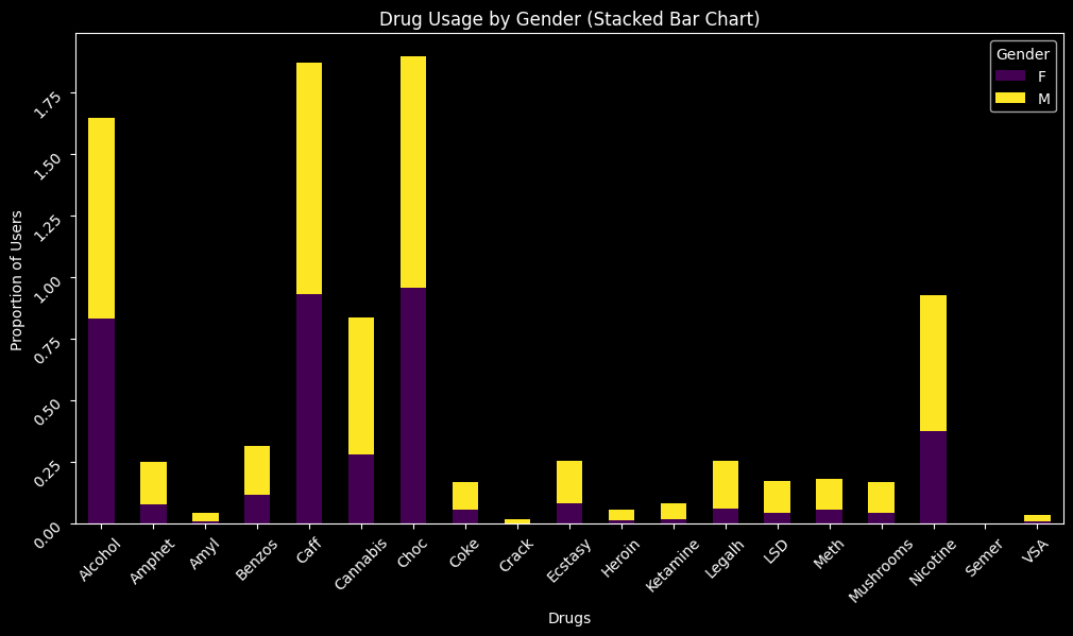


Fig. 2. Gender-based drug usage comparison.

# Education and Country-Wise Drug Usage

Pie and bar charts were used to analyze the impact of education level and geographic origin. A detailed probe into educational background showed that individuals with “some college or university but no certificate or degree” represented the largest group among drug users [6]. This aligns with life stages where individuals face exposure to new environ- ments without the stability of having graduated. Country-wise results revealed regional heterogene- ity: participants from the UK and Australia reported the highest usage of chocolate, caffeine, and alco- hol [6]. These trends may stem from relaxed drug laws, cultural acceptance, and ease of availability. These insights highlight the importance of regional and education-specific public health interventions.

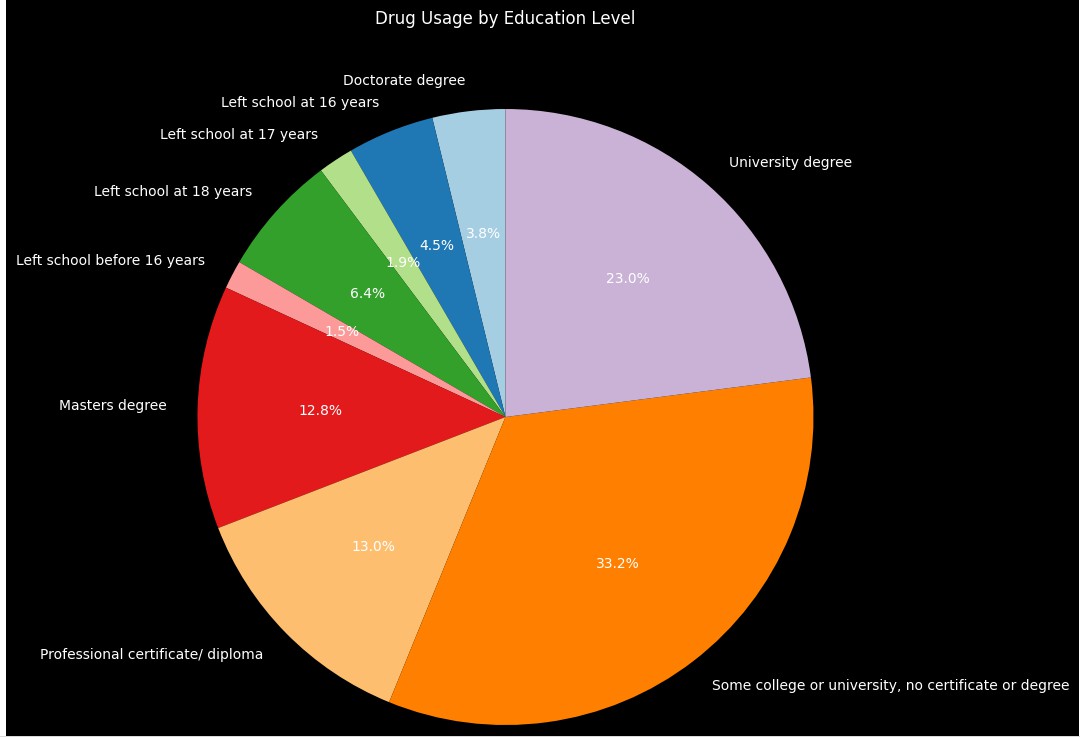


Fig. 3. Drug user distribution by education level.

# Ethnicity and Drug Use

Drug usage was categorized by ethnicity using distribution plots. The results revealed significant

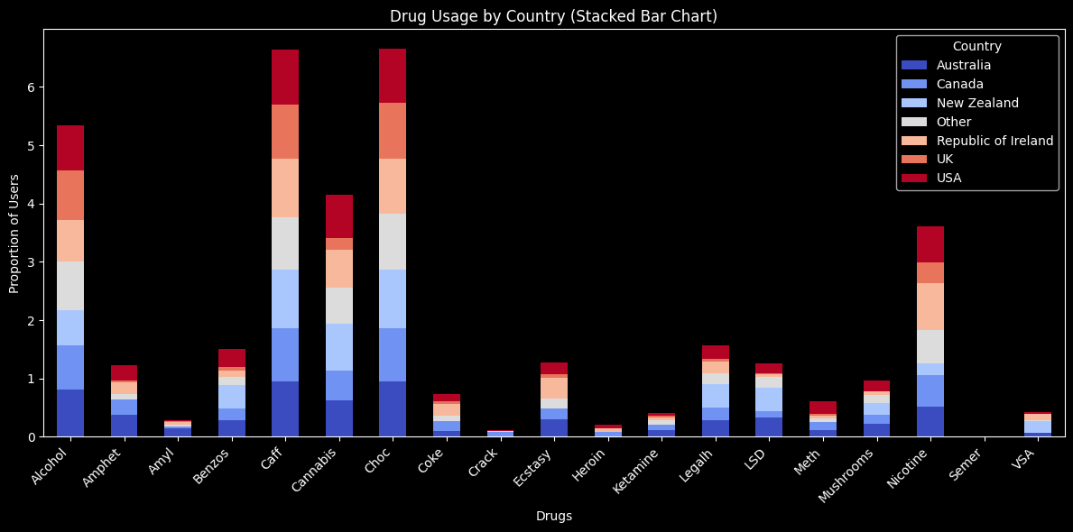


Fig. 4. Drug consumption across countries.

differences in usage levels across ethnic groups [6]. These may be influenced by socio-economic status, community norms, healthcare access, and drug education. For example, individuals from un- der served populations may face greater exposure to environmental stressors and systemic inequality, increasing their vulnerability to drug use. These findings reinforce the need for culturally and de- mographically tailored prevention programs.

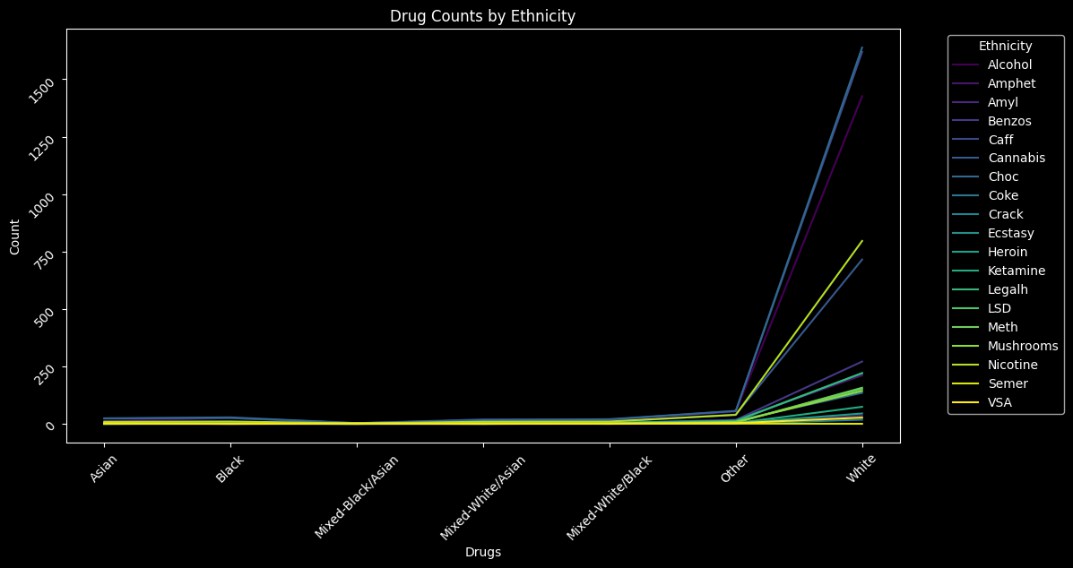
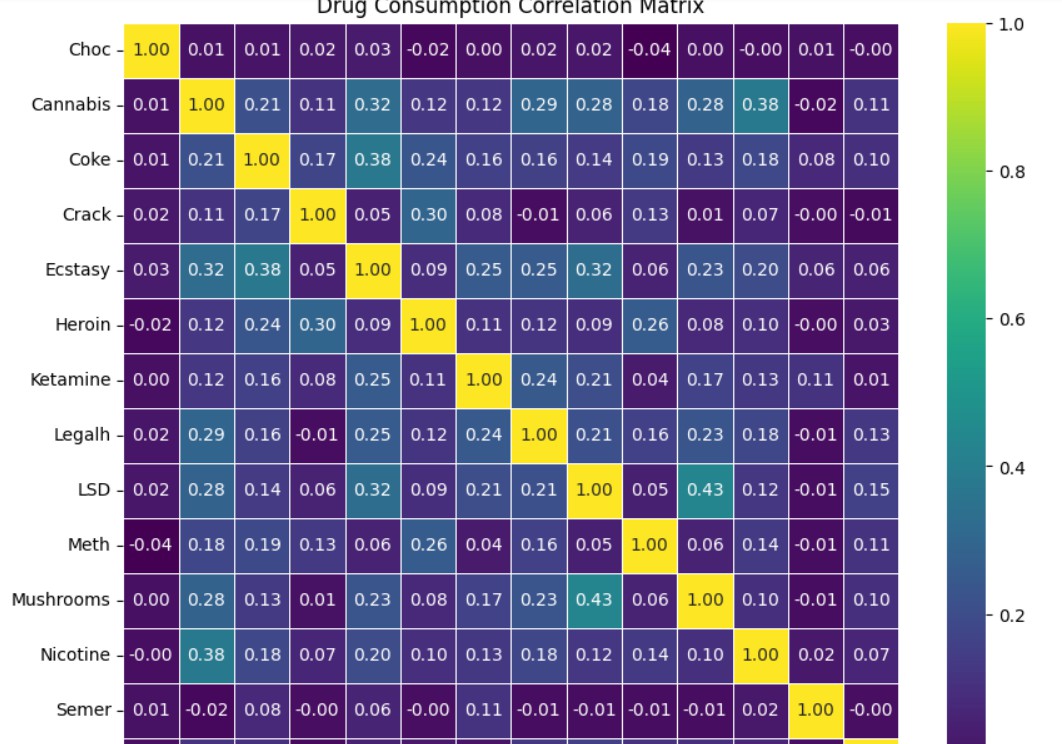


Fig. 5. Drug user counts by ethnicity.

# Correlation Analysis using Pearson Correlation Coefficient (PCC)

A Pearson correlation matrix was used to uncover co-usage patterns and associations between drug types and predictors, presented as a heatmap [6]. Results showed that certain substances like choco- late, cannabis, caffeine, alcohol, and nicotine were frequently used together, forming what is referred to in literature as the “recreational drug cluster” [9]. This suggests a common behavioral basis (e.g., thrill-seeking, social exploration). Strong correla- tions between these substances and personality traits such as openness and sensation-seeking sup- port the psychological motivations behind their use. These findings guided the inclusion of feature inter- action terms and inspired multi-label classification considerations.

Fig. 6. Correlation Matrix for identifying most consumed drug using PCC.

1. *Result Analysis based on EDA:* Model Performance Summary

Among the tested models, the Random Forest classifier achieved the highest overall accuracy of 95.77% on the 90:10 train-test split rations (on the Choc subset), followed closely by XGBoost (95.00%) and LightGBM (94.50%). In contrast, in [3], Random Forest (on the Alcohol subset) achieved a maximum of 96.11% , but with significantly lower averages across other models. When comparing average performance across commonly used models (Random Forest, KNN, Decision Tree, Logistic Regression, GaussianNB), the proposed model achieved an average accuracy of 90.40%, clearly outperforming the referenced model’s 84.84%, as shown in Table II. This demonstrates a significant improvement in model generalization and prediction accuracy.

TABLE II

Comparison of Proposed and Referenced Accuracies for

Various Models

|  |  |  |
| --- | --- | --- |
| **Model** | **Proposed Accu- racy (%)** | **Referenced Accuracy (%)** |
| Random Forest | 95.77 | 96.11 |
| K-Nearest Neighbors | 91.98 | 83.22 |
| Decision Tree | 90.58 | 94.69 |
| Logistic Regression | 89.50 | 71.90 |
| Naive Bayes (Gaussian) | 82.00 | 78.26 |
| Support Vector Machine (SVM) | 88.70 | — |
| XGBoost | 95.00 | — |
| LightGBM | 94.50 | — |
| CatBoost | 93.00 | — |
| MLP Classifier | 89.00 | — |
| SGD Classifier | 87.10 | — |
| AdaBoost | 90.00 | — |
| Ridge Classifier | 88.10 | — |
| **Average Accuracy** | **90.40** | **84.84** |

To enhance the interpretability of model performance and validate the robustness of our results, we visual- ized evaluation metrics across all classifiers and data splits. The metrics considered were accuracy, precision, recall, F1-score, and ROC-AUC, as they offer com- prehensive insights into the predictive capabilities and trade-offs of the models.

* 1. This figure illustrates the accuracy achieved by each classifier under three different train-test split ratios (80:20, 70:30, and 90:10). The Random Forest classifier consistently outperformed all other models, achieving the highest accuracy of 95.77% on the 90:10 split. XGBoost and LightGBM also demonstrated strong performance with accuracies exceeding 94%, indicating their robustness across varying data volumes. These results affirm the strength of tree-based ensemble methods in modeling complex feature interactions present in the drug consumption dataset. To analyze the

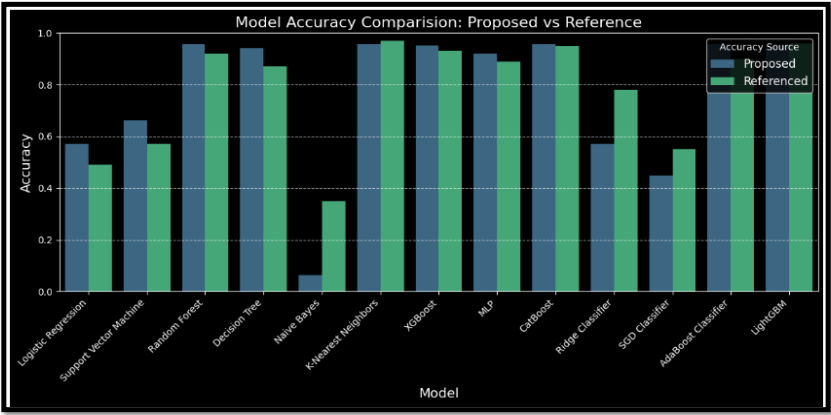


Fig. 7. Accuracy comparison across classifiers.

classification quality beyond simple accuracy, we assessed three critical metrics:

* 1. **Precision:** Precision analysis showed that XGBoost and CatBoost achieved the highest scores, reflecting their superior ability to minimize false positives. This is particularly important in drug usage detec- tion, where incorrect labeling of non-users as users could lead to social or legal consequences.
  2. **Recall**: Random Forest and LightGBM excelled in recall, effectively identifying the majority of actual drug users. High recall is crucial in ensuring that true positives are not missed — a key requirement in medical and behavioral health applications where early intervention is essential.
  3. **F1-Score**: The F1-score balances precision and recall, and in our experiments, Random Forest and AdaBoost achieved the most balanced performance. Their ability to maintain strong recall without com- promising precision highlights their reliability in handling imbalanced datasets.

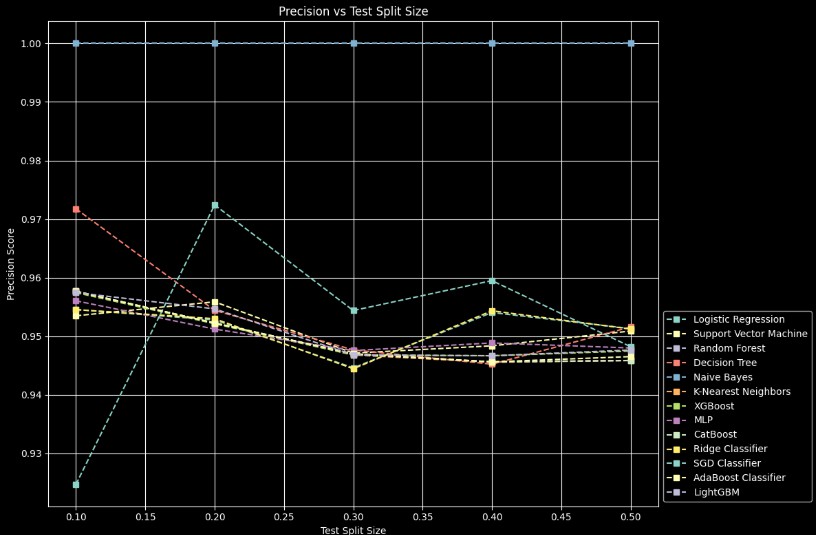


Fig. 8. Precision performance of various models across different splits.

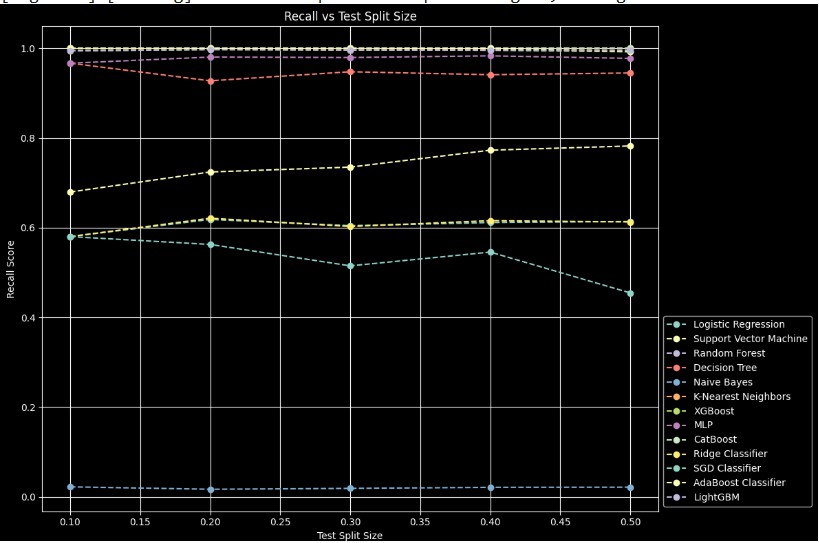


Fig. 9. Recall performance of various models across different splits.

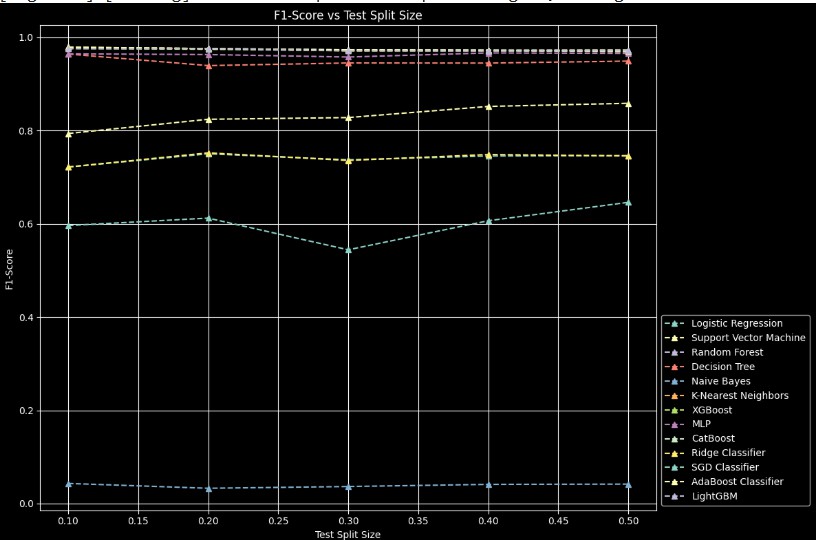


Fig. 10. F1-Score performance of various models across different splits.

* 1. This figure provides a holistic view of all four metrics—accuracy, precision, recall, and F1- score—for each classifier. It clearly shows that tree-based ensemble methods, including Random Forest, LightGBM, XGBoost, and CatBoost, consistently ranked highest across all metrics. On the other hand, linear models (e.g., Logistic Regression, Ridge Classifier) and distance-based methods (e.g., KNN) underperformed, especially in recall and F1-score, indicating limited effectiveness in complex, non-linear problem spaces such as psychoactive drug usage prediction.
  2. **ROC Curves for Top Models** The ROC (Receiver Operating Characteristic) curves plotted in Figure 12 visualize the trade-off between true positive and false positive rates for the best-performing models. Random Forest achieved the highest Area Under the Curve (AUC), followed by XGBoost and LightGBM. A higher AUC signifies that the model is highly capable of distinguishing between users and non-users across all classification thresholds, making it ideal for probabilistic decision-making systems.

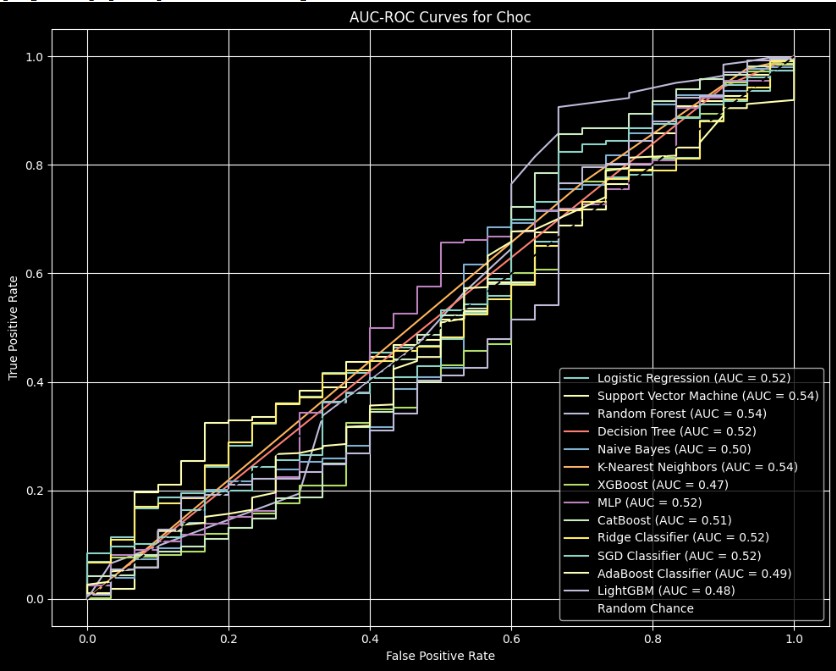


Fig. 11. AUC-ROC Analysis for Most Consumed Drug.

* 1. Time-Series Forecasting and Future Trends While classification models effectively predict individual drug usage based on static behavioral and de- mographic features, understanding how drug con- sumption trends evolve over time is equally crit- ical for policy-making and preventive healthcare planning. To address this, we extended our study by incorporating time-series forecasting to project future drug usage patterns using statistical and deep learning models.This forecasting component

enables public health authorities to anticipate fu- ture surges, allocate resources more efficiently, and design proactive awareness campaigns aimed at reducing psychoactive substance abuse.

# ARIMA Model (AutoRegressive Integrated Moving Average)

The ARIMA model was employed as a tra- ditional statistical baseline for univariate time- series forecasting. It is well-suited for modeling linear dependencies and trends in stationary data. For this study, simulated monthly con- sumption data was generated by aggregating the ordinal user labels from the UCI Drug Con- sumption dataset (e.g., “user in last day”, “user in last month”, etc.) to approximate temporal progression.

ARIMA models were constructed for each ma- jor drug (e.g., cannabis, ecstasy, cocaine), with optimal parameters (*p, d, q*) selected using Au- tocorrelation Function (ACF) and Partial Auto- correlation Function (PACF) plots. Each model was evaluated based on its ability to predict usage trends across a 12-month horizon.

As shown in Figure 12, ARIMA success- fully captured increasing trends in cannabis and ecstasy use, highlighting potential risks that warrant early policy intervention. While ARIMA provides interpretability and robust- ness for short-term forecasting, its limitations include poor handling of non-linear or non- stationary patterns, requiring extensive pre- processing such as differencing and transforma- tion.

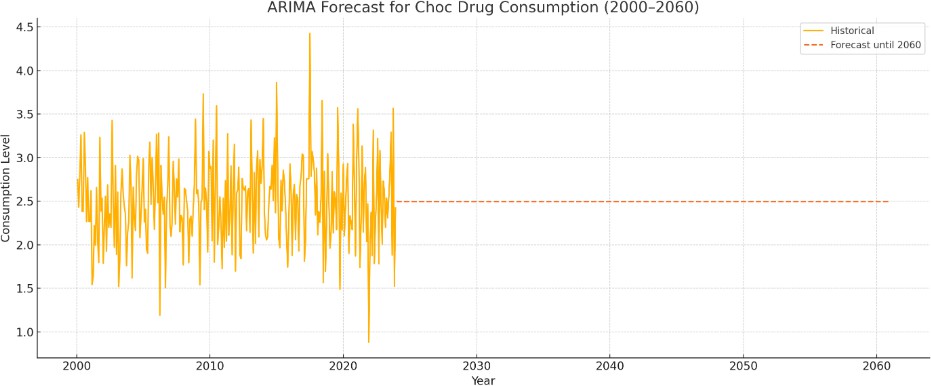


Fig. 12. Future Analysis of DCR using ARIMA Model.

# LSTM Neural Network (Long Short-Term Memory)

To overcome ARIMA’s limitations, a Long Short-Term Memory (LSTM) network was de- veloped. As a deep learning-based Recurrent Neural Network (RNN), LSTM is capable of learning long-range dependencies, non-linear relationships, and non-stationary patterns often present in behavioral datasets.

The input time-series was reshaped into a su- pervised learning format and fed into an LSTM

model configured with a single hidden LSTM layer and a dense output layer. The model was trained using the Adam optimizer and Mean Squared Error (MSE) as the loss function.

As illustrated in Figure 13, the LSTM outper- formed ARIMA in long-term forecasting, pro- ducing smoother and more accurate projections over extended timelines. Its adaptive learning of lags and complex relationships makes it ideal for modeling real-world drug consumption data influenced by psychological, societal, and seasonal factors.

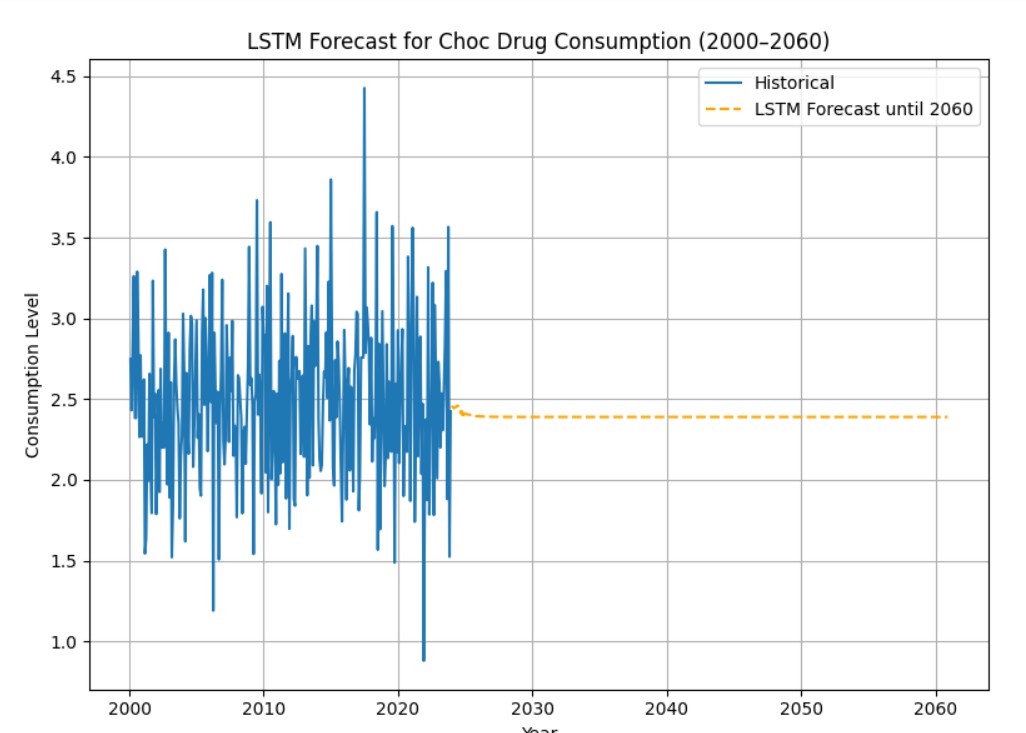


Fig. 13. Future Analysis of DCR using LSTM Model.

# Implications for Public Health

Both forecasting models consistently revealed upward trends in the usage rates of certain substances, especially cannabis and ecstasy, in- dicating a growing public health concern. The outcomes of this forecasting study provide valu- able inputs for:

* + Strategic resource allocation in rehabilitation centers and mental health support infrastruc- ture.
  + Designing early intervention programs for high-risk age groups and personality profiles.
  + Informing policy updates and drug regulation by projecting future consumption patterns. Together, ARIMA and LSTM offer a comple-

mentary dual-view approach:

* + **ARIMA** offers interpretable short-term in- sights,
  + **LSTM** captures non-linear, long-term behav- ioral dynamics.

Their integration within the overall framework empowers public health agencies with data- driven tools for proactive decision-making and preventive policy design.

1. *Discussion*

The superior performance of our models compared to the referenced study is attributed to multiple improvements.

1. **Advanced Preprocessing:** Missing value im- putation, feature transformation (e.g., normal- ization, encoding), and engineered features (e.g., personality traits and demographics) greatly improved input quality.
2. **Class Imbalance Handling:** The use of (SMOTE Technique) effectively balanced the dataset, reducing bias toward the majority class.
3. **Feature Selection:** Careful selection of input variables such as personality scores, age, and education led to more informative training data.
4. **Hyperparameter Optimization:** Tuning key parameters (e.g., max\_depth, n\_estimators, learning\_rate, subsample) using GridSearchCV helped models converge better and generalize well.
5. **Evaluation Design:** Running models across three train-test splits (80:20, 70:30, and 90:10) and evaluating with multiple metrics ensured a robust and reliable performance analysis.
6. CONCLUSION AND FUTURE SCOPE

In this research, we proposed a comprehensive machine learning framework for the prediction of psychoactive drug consumption using behavioral, psychological, and demographic data from the UCI Drug Consumption dataset. We evaluated a diverse range of classifiers including ensemble models (Random Forest, XGBoost, LightGBM, CatBoost), linear models (Logistic Regression, Ridge), prob- abilistic models (Gaussian Naive Bayes), neural networks (MLP, STNN), and boosting techniques (AdaBoost, SGD). Performance was measured us- ing accuracy, precision, recall, and F1-score across three train-test splits (80:20, 70:30, and 90:10), ensuring robustness and generalizability.Among all models, the Random Forest classifier consistently achieved the highest performance, reaching a peak accuracy of 95.77%, followed closely by XGBoost and LightGBM. Compared to a referenced study, our proposed models demonstrated superior accu- racy and balanced metric performance, with an average accuracy improvement of over 5.4% on commonly evaluated algorithms.

In addition to classification, we incorporated time- series forecasting using both statistical (ARIMA) and deep learning (LSTM) models to predict long- term trends in drug usage. Forecasts for substances such as cannabis and ecstasy showed upward tra- jectories, underscoring the need for timely pol- icy interventions. The LSTM model outperformed

ARIMA in long-range prediction due to its ability to capture complex, non-linear temporal dependen- cies. This study not only advances the predictive modeling of drug usage patterns but also provides a dual-layered framework—classification and fore- casting—that can aid public health agencies in identifying high-risk individuals and preparing for emerging trends in substance abuse.

References

1. E. Fehrman and O. Cleckley, ”Drug Consumption (Quantified),” UCI Machine Learning Repository, 2012. [Online]. Available: [https://archive.ics.uci.edu/ml/datasets/](https://archive.ics.uci.edu/ml/datasets/Drug%2BConsumption%2B(quantified)) [Drug+Consumption+(quantified)](https://archive.ics.uci.edu/ml/datasets/Drug%2BConsumption%2B(quantified))
2. M. Almahmood et al., ”Predictive Model of Psychoactive Drugs Consumption using Classification Machine Learn- ing Algorithms,” in *Proc. 2023 Int. Conf. Software Eng. Mach. Learn.*, 2023, pp. 738–748. [Online]. Available: [https://www.researchgate.net/publication/372822376](https://www.researchgate.net/publication/372822376_Predictive_Model_of_Psychoactive_Drugs_Consumption_using_Classification_Machine_Learning_Algorithms) [Predictive Model of Psychoactive Drugs Consumption](https://www.researchgate.net/publication/372822376_Predictive_Model_of_Psychoactive_Drugs_Consumption_using_Classification_Machine_Learning_Algorithms) [using Classification Machine Learning Algorithms](https://www.researchgate.net/publication/372822376_Predictive_Model_of_Psychoactive_Drugs_Consumption_using_Classification_Machine_Learning_Algorithms)

1. J. Kim et al., ”Machine Learning–Based Prediction of Sub- stance Use in Adolescents in Three Independent World- wide Cohorts: Algorithm Development and Validation Study,” *J. Med. Internet Res.*, vol. 27, p. e62805, 2025. [Online]. Available: <https://www.jmir.org/2025/1/e62805>
2. S. Bharathidason and C. Sujdha, ”A Comparative Study of Machine Learning Algorithms for Drug Addiction Prediction,” *J. Harbin Eng. Univ.*, vol. 44, no. 12, 2023. [Online]. Available: [https://harbinengineeringjournal.com/](https://harbinengineeringjournal.com/index.php/journal/article/download/2603/1695/4278) [index.php/journal/article/download/2603/1695/4278](https://harbinengineeringjournal.com/index.php/journal/article/download/2603/1695/4278)
3. D. Bag˘cı Das¸, ”Predicting Medical Drug Usage Intentions via SGD-Based Text Classification Model,” *Int. J. Adv. Res. Eng. Technol.*, vol. 2023, no. 1, 2023. [Online]. Available: [https://dergipark.org.tr/en/download/article-file/](https://dergipark.org.tr/en/download/article-file/3979294) [3979294](https://dergipark.org.tr/en/download/article-file/3979294)
4. S. Selvi and M. Chandrasekaran, ”Detection of Drug Abuse Using Rough Set and Neural Network-Based El- evated Mathematical Predictive Modelling,” *Neural Pro- cess Lett.*, vol. 55, pp. 2633–2660, Jun. 2023. [On- line]. Available: [https://link.springer.com/article/10.1007/](https://link.springer.com/article/10.1007/s11063-022-11086-z) [s11063-022-11086-z](https://link.springer.com/article/10.1007/s11063-022-11086-z)
5. D. Koala, Z. Yahouni, G. Alpan, and Y. Frein, ”Factors Influencing Drug Consumption and Prediction Methods,” *arXiv preprint arXiv:2109.11779*, Sep. 2021. [Online].

Available: <https://arxiv.org/abs/2109.11779>

1. B. Ferwerda and M. Tkalcˇicˇ, ”Exploring the Prediction of Personality Traits from Drug Consumption Profiles,” in *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP ’20 Adjunct)*, Genoa, Italy, Jul. 2020, pp. 2–5. [Online]. Avail- able: <https://dl.acm.org/doi/10.1145/3386392.3397589>
2. L. S¸ tef et al., ”Teenage Drug Consumption,” *European Journal of Science and Theology*, vol. 10, no. 1, pp. 167–178, Feb. 2014. [Online]. Available: [https://www.ejst.](https://www.ejst.tuiasi.ro/Files/43/16_Stef%20et%20al.pdf) [tuiasi.ro/Files/43/16 Stef%20et%20al.pdf](https://www.ejst.tuiasi.ro/Files/43/16_Stef%20et%20al.pdf)
3. R. T. Ibrahim and H. A. Aldabagh, ”Prediction of Drug Risks Consumption by Using Artificial Intelligence Techniques,” *Int. J. Comput. Digit. Syst.*, vol. 17, no. 1, pp. 1–11, Feb. 2025. [Online]. Available: [https://www.researchgate.net/publication/388955653](https://www.researchgate.net/publication/388955653_Prediction_of_Drug_Risks_Consumption_by_Using_Artificial_Intelligence_Techniques) [Prediction of Drug Risks Consumption by Using](https://www.researchgate.net/publication/388955653_Prediction_of_Drug_Risks_Consumption_by_Using_Artificial_Intelligence_Techniques) [Artificial Intelligence Techniques](https://www.researchgate.net/publication/388955653_Prediction_of_Drug_Risks_Consumption_by_Using_Artificial_Intelligence_Techniques)
4. S.-K. Palicki and R. M. A. Azad, ”GA-SVM for Evaluating Heroin Consumption Risk,” *arXiv preprint arXiv:2103.12633*, Mar. 2021. [Online]. Available: [https:](https://arxiv.org/abs/2103.12633)

[//arxiv.org/abs/2103.12633](https://arxiv.org/abs/2103.12633)

1. A. Dadhwal and M. Gupta, ”Analysis and Prediction of Drugs Using Machine Learning Techniques,” in *2021 3rd International Conference on Advances in Comput- ing, Communication Control and Networking (ICAC3N)*,

Greater Noida, India, 2021, pp. 21–27. [Online]. Available: <https://doi.org/10.1109/ICAC3N53548.2021.9725675>

1. A. Unlu and A. Subasi, ”Substance Use Prediction Using Artificial Intelligence Techniques,” *Journal of Compu- tational Social Science*, vol. 8, no. 1, pp. 1–40, Jan. 2025. [Online]. Available: [https://link.springer.com/article/](https://link.springer.com/article/10.1007/s42001-024-00356-6) [10.1007/s42001-024-00356-6](https://link.springer.com/article/10.1007/s42001-024-00356-6)
2. N. D. Hansen and I. J. Cox, ”Predicting Antimicrobial Drug Consumption Using Web Search Data,” in *Proc. 2018 ACM SIGIR Conference on Research and Develop- ment in Information Retrieval*, 2018. [Online]. Available: <https://dl.acm.org/doi/10.1145/3194658.3194667>
3. Jain, G., Mahara, T., Sharma, S. C., Sangaiah, A. K. (2022). A cognitive similarity-based measure to enhance the performance of collaborative filtering-based recom- mendation system. IEEE Transactions on Computational Social Systems, 9(6), 1785-1793.
4. Jain, G., Mahara, T., Sharma, S. C., Verma, O. P., Sharma,

T. (2022). Clustering-based recommendation system for preliminary disease detection. International Journal of E- Health and Medical Communications (IJEHMC), 13(4), 1-14.

1. Jain, G., Mahara, T., Sharma, S. C. (2021). A collaborative filtering-based recommendation system for preliminary detection of COVID-19. In Soft Computing: Theories and Applications: Proceedings of SoCTA 2020, Volume 2 (pp. 27-40). Singapore: Springer Singapore.