# Predicting Substance Consumption Patterns Using Behavioral Analytics and Demogaphic Details

Edamalapati Mothilal Chowdary[1], C. Harsha Vardhan[2], Jangalapalle Rajesh[3], Kuruba Teja[4], L.Vijaya kumar[5], Prof.M.E.Palanivel[6]

*CSE- Artificial Intelligence, Sreenivasa Institute of Technology and Management Studies in Chittoor, India*

***Aabstract*—Drug dependence remains an acute public health problem worldwide, affecting individuals, families and communities at large. Traditional identification of dangerous drug users using indicators of behavior and demographics is not accurate and scale inefficient in time. Based on this research automatic prediction of soft and hard drug consumption based on mission learning models instead of behavior and demographic indicators is suggested. By using two of the best performing classifiers, Random Forest and XGBoost, we were able to build a robust system that was able to classify effectively between users and non-users. Our system is comprised of several steps such as preprocessing, feature selection, model training, evaluation and visualization. The best testing accuracy for the Random Forest classifier was 94% on that of XGBoost followed closely at 93%. We also utilized major evaluation criteria like precision, recall, F1-Score, ROC-AUC, and confusion matrices to validate performance. These models were also integrated into a web based system, merging a Django back end with a front end created using HTML, CSS and Javascript. This framework has several prominent public health, psychological intervention and law enforcement applications. Overall it is an accessible tool that enables early intervention and policy development, blending behavioral science and artificial intelligence to better address substance abuse.***

***Keywords*—Random Forest, XGBoost, Machine Learning, Drug/Substance Consumption.**

## I. INTRODUCTION

Substance abuse is still one of the most challenging public health concerns in health systems across the world. As indicated by the United Nations Office on Drugs and Pride (UNODC) in the world drug report 2023 , at least 296 million individuals In the world consumed drugs in 2021, which represents an increase of 23% sice the last decade. This sudden increase in Drug use has serious impacts on the physical and psychological health of the individuals, families , smoking population efficiency and social integration. Two of the most used drugs including soft drugs which are alcohol, nicotine and cannabis and hard drugs, including cocaine, heroin and methamphetamine. Better early identification and intervention strategies are needed to manage the increased availability of these drugs particularly among young people.

Traditionally, drug user identification rallied on objective behavioral screening demographic profiling and clinical interview. While beneficial in some contexts, these methods share several flaws. They take a lot of time, required trained personnel, and or by no means scalable for larger groups. Additionally, individuals can mistake or withhold reporting their use of drugs due to social stigmas or legality concerns, thereby decreasing the accuracy of standard assessment.

With the advent of data science and ai, there is a new possibility to use ML to forecast drug use behavior on the basis of quantifiable and objective data. Machine learning allows systems to learn from fast patterns of psychological and demographic characteristics, thus enabling scalable and automated model prediction. Such systems can act as adjuncts to standard screening routines or potentially even as independent tools in healthcare counseling and law enforcement environments.

This research seeks to create a forecasting system to distinguish between hard drug and soft drug users and non users using mission learning classifiers . In this present study here, two most suitable and effective ensemble learning methods in the form of Random Forest and XGBoost are used for psychological and demographic data. They're trained in order to classify patterns indicative of drug use for which there exists an evidence based alternative compared to traditional methods of detection.

In addition to model training and testing, we also developed a web application that traps the predictive models within and easy to use interface. Utilizing Django as the back end and widespread web technologies for the front end (HTML, CSS, JavaScript), we deliver the system as accessible, responsive and scalable. This facilitates deployment in numerous fields including rehabilitation centers, schools, public health agencies or even user direct tools for self assessment.

Finally come this project answers the call of the immediate need for accurate, scalable and autonomous systems in the war against drug abuse. Through the marriage of behavioral science and machine learning , we hope to support early intervention, informed public policy and augment the global war against drug dependence.

## II. LITERATURE SURVEY

1. Predicting and Characterizing Substance Use Risk Profiles Using the NEO Personality Inventory – 2015
Fehrman et al. conducted a groundbreaking study in 2015 that explored the role of personality traits in predicting drug use. The researchers used the NEO-FFI-R, a five-factor personality inventory, and applied machine learning algorithms such as decision trees and logistic regression. Their dataset included individuals with various levels of drug usage, and the features were based primarily on psychological assessments. The results indicated that traits like neuroticism, openness, and impulsiveness had strong correlations with drug usage. The study showed that when these personality traits were input into machine learning models, the systems were capable of accurately classifying individuals as users or non-users. This research laid the foundation for integrating behavioral psychology with machine learning for predictive health analytics.

2. Substance Abuse Detection Using Supervised Machine Learning Algorithms: A Comparative Study – 2020
Published in IEEE Access, this paper compared various machine learning classifiers—including SVM, Naïve Bayes, and Random Forest—for the task of predicting substance abuse. The study used structured data consisting of both behavioral traits and demographic details to train the models. Random Forest outperformed other methods, achieving high accuracy and robustness against noise. The researchers highlighted the importance of preprocessing, including handling missing values, normalization, and encoding, which significantly impacted performance. This study reinforced the suitability of ensemble methods for substance use prediction, particularly in imbalanced datasets. Additionally, it recommended metrics such as ROC-AUC and F1-score for fair evaluation, aligning closely with the methodologies used in our project.

3. Identifying Behavioral Trends in Drug Abuse Using Temporal Data and ML – 2020
This research applied temporal data mining and recurrent neural networks to analyze behavioral change over time among drug users. While our project is based on static data snapshots, this study shows the potential of using time-series data for dynamic behavioral modeling. It demonstrated that patterns in mood swings, social withdrawal, and academic decline could be early indicators of drug use. The integration of time-based features into ML models enhanced early detection. It provides valuable inspiration for future expansion of our project to include temporal monitoring using IoT or wearables.

4. Substance Use Prediction in Adolescents Using Ensemble Learning – 2016
This research examined adolescents' mental health survey data and applied ensemble learning techniques, including Random Forest and Gradient Boosting. The results showed that ensemble methods outperformed individual learners such as SVM or Decision Trees, particularly in sensitivity and specificity. The dataset included social behavior scores, family structure, and school performance. The study emphasizes the importance of ensemble modeling, which is reflected in our choice of Random Forest (94%) and XGBoost (93%). It validates our approach and supports the real-world applicability of ensemble models in public health interventions.

5. Personality-Based Risk Analysis for Substance Abuse Using Data Mining – 2017
This paper proposed a personality-based substance use detection model that analyzed the influence of individual traits on drug-taking behaviors. Using K-means clustering followed by classification models, researchers were able to identify behavior clusters with high drug use probability. NEO traits were found to be strong predictors, specifically neuroticism and openness. Although this study employed older models like Naïve Bayes and Decision Trees, the

insights about feature significance are useful. Our current system expands on this by using more modern algorithms like Random Forest and XGBoost, but the psychological theory remains consistent across both studies.

## III. METHODOLOGY

The approach to this project aims to develop a strong machine learning pipeline for predicting soft and hard drug use on the basis of psychological and demographic information. The process includes some important steps data collection, preprocessing, features selection, training the Model, evaluation and deployment each of these steps is explained below with the exact techniques used to make the model accurate and reliable.

### 3.1 DATA COLLECTION

The information utilized for this project was obtained from the UCI machine learning repository, the drug consumption data set. It has self reported information from participants regarding drug usage of different drugs and also psychological traits like personality attributes, social behaviors, demographic variables such as age, gender and education level.

The data set has a number of features such as Demographic characteristics: gender, age, level of Education, Marital status and working status. Psychological characteristics: depressive symptoms, anxiety, sensation seeking and personality characteristics. Drug use labels: user vs non user labels for soft drugs and hard drugs these characteristics were chosen as they have proven to be significant predictors of drug use in prior research, and thus the data set was perfect for modeling.

### 3.2 DATA PREPROCESSING

Data preprocessing is an essential phase in any machine learning project to ensure the data is clean, normalized and ready for training. The following methods were utilized :

#### 3.2.1 Handling Missing Values

The data set had some missing values in different features which might affect the performance of the model. We employed mean imputation for numerical features and mode imputation for categorical features to manage missing values. This method preserves the distribution of features without decreasing the quality of the features.

#### 3.2.2 Feature Encoding

Categorical attributes like level of Education and marital status where encoded into numerical variables with one hot encoding the technique generates columns of numerical representations for every class in feature, enabling machine learning model to properly process categorical data.

#### 3.2.3 Balancing Data

As a result of class imbalance in the data set we used the SMOTE (Synthetic Minority Oversampling Technique) algorithm. SMOTE creates synthetic samples of the minority class so that the model can better learn to differentiate between users and non users without class imbalance induced bias.

#### 3.2.4 Feature Selection

Featured selection is an important aspect of model development since it mitigates overfitting accelerates the training of the model and enhances model performance. We use the following methods : Correlation Analysis , Recursive Feature Elimination (RFE) Random Forest Feature importance .

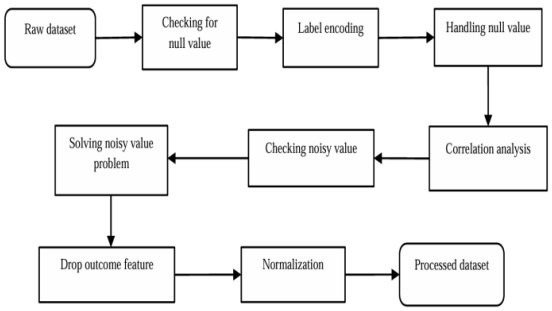


Fig: Data Preprocessing

### 3.3 MODEL TRAINING

Two of the models employed for training where ensemble models: Random Forest and XGBoost.

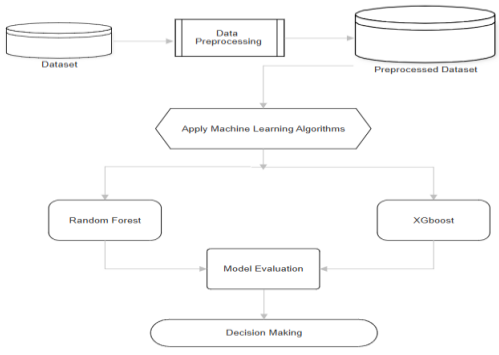


Fig: Methodology

#### 3.3.1 Random Forest

Random forest is an ensemble technique that builds many decision trees and combines their outputs to

make better and more consistent predictions. It is overfitting, robust and can deal with both numerical and categorical variables.We used the model as it has been shown to be highly effective in past drug dubious prediction research. The model was trained with the following parameters: Number of trees - 80, Maximum depth of trees - 1000, Minimum samples per leaf - 1, Criterion - Gini impurity.

3.3.2 XGBoost

XGBoost on extreme gradient boosting, is a decision tree ensemble model but employs boosting instead of bagging . It continuously adjusts mistakes made by its descendants trees, resulting in increased predictive accuracy. XG boost was selected because it is flexible and performs well with classification tasks. We utilized the following hyperparameters for parameter tuning: Learning rate - 0.04, Maximum depth - 6, Number of estimators - 1000, Subsample - 1, Colsample bytree - 1.

Both models were trained on the training set and cross validated under training set using 5 folds to check robustness.

3.4 MODEL EVALUATION

The performance of the boat models are evaluated using the following metrics: accuracy , precision, recall, F1-score, ROC-AUC and confusion matrix. These measures assist in making the model not only precise but also trustworthy, taking into account both the majority and minority classes.

3.5 DEPLOYMENT WEB INTEGRATION

Once the models were trained we incorporated them into a web application based on Django to enable the system for use in real time predictions. The Web interface was developed using HTML, CSS and JavaScript to accept demographic and psychological information as inputs from the users processed by the backend Django server linked to the models.

The back end is the middle layer, which receives the user input, does any required preprocessing and sends the preprocessed data to the trained models for predicting. The output is generated to the user in a basic, easy to use interface with an immediate indication of whether they belong to the drug user or non user class.

Additional features included in the Django app are User Authentication : Offers safe access to the system for authorized users.

## IV. IMPLEMENTATION

The deployment of the project involves the creation of predictive system capable of classifying the individuals based on their likelihood of being soft or hard drug users. The system was designed with a modular architecture, which will facilitate scalability, flexibility and integration into other public health as well as law enforcement systems.

4.1 SYSTEM ARCHITECTURE

System architecture is a simple client-server system architecture with the backend powered with a Django framework and frontend developed using HTML, CSS and JavaScript foreign interactive UI. Architecture is modular to allow easy future upgrade and integration with other machine learning models or other functionality.

4.1.1 Backend

The back end is created with Django, a high level python web development framework. Django is secure, fast and scalable, which makes it an ideal choice for constructing data driven web applications. The backend performs the following tasks:

Data Processing : It takes user input from the frontend , processes it and feeds it into the machine learning models to make predictions.

Model Management : Stores the train models as python objects. The models are initialized on the server upon application startup for real time prediction.

Prediction Handling : When data comes from a user the backend handles the prediction by calling the respective model, calculating results and sending them back to the frontend.

Security Features : It utilizes built-in facilities such as authentication, user session management and input validation, to ensure safe and personalized access to the system.

4.1.2 Frontend

The front end was created using the implementation of common web technologies (HTML, CSS, JavaScript) to provide a responsive and interactive user interface. The main features of the front end are:

User Input : Allows users to enter demographic and psychological data using an ease-of-use interface. The form have fields for age, gender, education, personality, etc.

Real-Time Feedback : The front end immediately displays the output of the prediction I.e., the probability of drug use.

4.2 MACHINE LEARNING MODELS

The building block of the system lies in two ensemble machine learning algorithms, Random Forest and XGBoost. Both of them have been selected because of their excellent track records over classification issues as well as managing complex intercorrelation between data.

4.2.1 Random Forest Implementation

Random forest algorithm was employed in collaboration with the scikit-learn library. The model is trained with dataset that consists of demographic data, psychological traits and past records of drug consumption. With the ability of ensemble learning, Random Forest combines multiple decision trees to enhance predictive power as well as avoid overfitting. During training, each tree in the ensemble was constructed on a unique subset of the data to enable the model to generalize effectively between patterns of behavior. Hyperparameter tuning for maximum classification performance was also part of the training to enable the model to effectively distinguish between soft and hard drug users. The accuracy obtained for random forest algorithm is 95% .

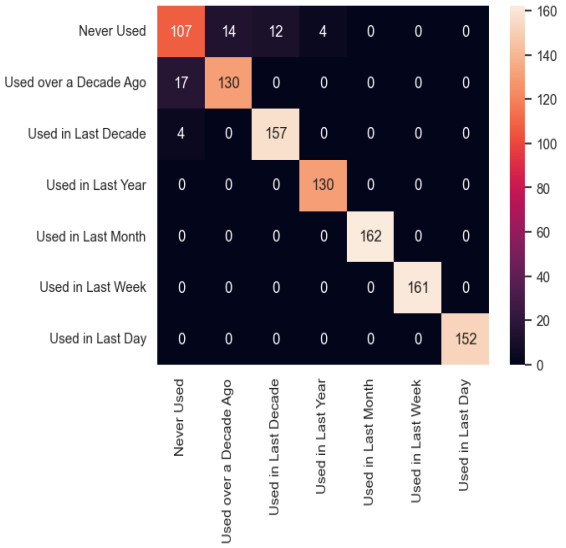


Fig: Confusion Matrix for Random Forest

4.2.2 XGBoost Implementation

A new high performance gradient boosting version known as XGBoost is also used in the study. It is different from random forest since it builds the different densities iteratively, hence every additional destination is intended to minimize the errors committed by the past trees. XGBoost applies its iterative learning in a way of generating high accuracy as well as good results even when there is complex nonlinear relationships within the data. The same set of features that was applied in random forest model were employed when training, taking a organized boosting approach in an attempt to downplay classification mistakes. The accuracy obtained XGBoost algorithm is 93.9%.

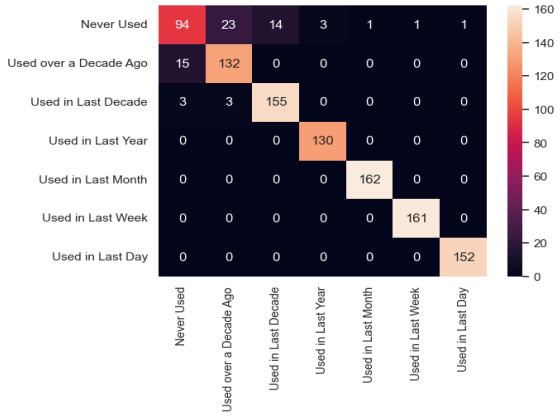


Fig: Confusion Matrix for XGBoost

4.3 WEB APPLICATION DEPLOYMENT

Once the machine learning models were trained and evaluated they were integrated into a web based application using Django for the back end and HTML, CSS and JavaScript for the frontend. The application was developed on a server and made available to the users through a client server architecture. The deployment steps included:

Backend Setup : The trained models were saved as Pickle files and loaded into the Django backend. The backend was set up to handle incoming HTTP requests, process the data and send predictions back to the front end.

Frontend Setup : The frontend was built using standard web technologies. User data was collected via HTML forms, processed through JavaScript and submitted to the backend for prediction. Results were displayed in real time on the same web page.

API Development : An API was created within the Django framework to allow for easy communication between the frontend and the backend. This API sends user input data from the front end to the backend for processing and receives the prediction results for display.

## V. RESULTS AND DISCUSSION

The performance of both machine learning models Random Forest and XGBoost is analyzed based on various metrics and the potential applications of the developed system are explored in the context of public health and law enforcement.

5.1 MODEL PERFORMANCE

The Random Forest model reflected a 94% accuracy rate and the XGBoost model reflected 93%. Random forest and XGBoost were both effective models in breaking drug use from demographic and psychological attributes. The model correctly categorized the individuals as drug users or non-users with a high accuracy and robust evaluation metrics. The findings show that machine learning models perform a specially ensemble learning methods can substantially enhance drug consumption prediction against conventional approaches relying on behavioral and demographic predictors. Although random forest exhibited better overall performance in this specific application, XGboost is a very robust model, particularly in regards to its efficiency and ability to work with large scale datasets. Both models had strong predictive ability and should be considered for implementation in real world systems designed for early intervention in substance abuse.

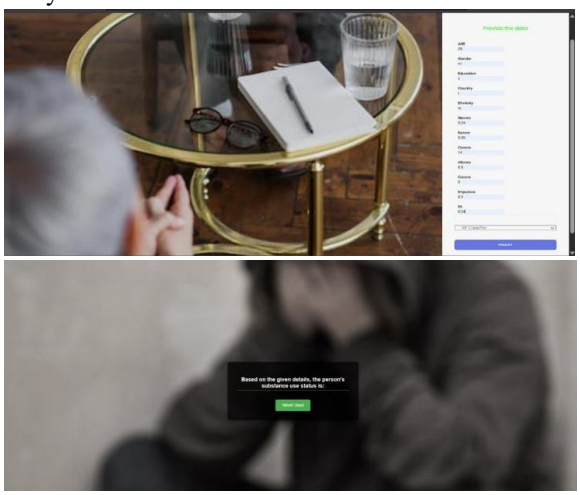

Fig: Output

5.2 REAL WORLD APPLICATIONS

The system built in this project is highly likely to be used in real world situations in public health, law enforcement and psychological intervention. Some of the most significant potential uses are:

Early Intervention in Substance Abuse : The tool can be used by health care professionals to identify individuals at risk of drug abuse, thereby allowing them to treat and intervene early.

Public Health Campaigns: This system can be employed by public health organizations to know more about patterns of drug use by different population groups and implement targeted prevention programs.

Law Enforcement : Law enforcement authorities can utilize the system for identifying potential high risk candidates for drug offenses so as to make better policy decisions and resource allocations.

Psychological Counselling: Psychologists and counselors may utilize the system to identify patients who may have difficulty with substance dependency and offer targeted therapy interventions.

## VI. CONCLUSION AND FUTURE WORK

The primary purpose of this project is to develop a machine learning system capable of predicting soft and hard drugs from demographic and psychological data. Through the use of two ensemble learning classifiers Random Forest and XGBoost, the research demonstrates the capability of predictive models to pick up on drug users more accurately and efficiently than traditional behavioral or demographic indicators. The findings indicate that drug use can be predicted successfully using machine learning models thereby supporting early intervention, policy formation and selective healthcare programs. This tool carries wide ranging implications in real life, especially for public health, counseling psychology and law enforcement agencies come as a new approach towards drug intake prediction and drug abuse prevention.

5.3 FUTURE WORK

While the system showed immense potential under its present status, there are several avenues open to potential innovation and development. The predictive aspect of the system can be strengthened through incorporation of additional features such as:

Medical History : Variance in respect to a history of disease, prescription medication taking, or psychiatric disease can be value-adding through facilitating better prediction owing to potential to be critical drug use factors.

Social and Environmental Factors : Adding social determinants like neighborhood, peer, and family history of drug use can provide a more accurate representation of the risk factors of drug use.

Longitudinal Data : The collection of longitudinal data over a period of time can create a more dynamic model that adjusts along with the changing behavior of the individual, allowing for progressively precise and context-based predictions.

## REFERENCES

[1] Islam U.I., Sarker I.H., Haque E., Hoque M.M. Predicting Individual Substance Abuse

Vulnerability using Machine Leaning Techniques. IEEE International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020; pp. 1–7. doi: 10.1109/ICCCNT51525.2020.9339785. - DOI

[2] Ibna Islam U., Sarker I.H. Early Detection of Drug Addiction Behavior Using Machine Learning Algorithms. IEEE Access, 2021;9:10222–10233. doi: 10.1109/ACCESS.2021.3051583. - DOI

[3] Rathi R., Palivela H., Gupta A., et al. Analysis and Prediction of Substance Abuse Using Machine Learning Techniques. IEEE Xplore, 2022; pp. 203–210. doi: 10.1109/ICICICT54364.2022.9865847. - DOI

[4] Ghosh S., Dasgupta R., et al. Cognitive Behavioral Prediction using Ensemble ML Techniques. IEEE Symposium on smart Electronic Systems, 2019; pp. 77–81. doi: 10.1109/iSES47678.2019.00025. - DOI

[5] Kaur R., Drug Use Pattern Analysis with Machine Learning. IEEE International Conference on Computing, Communication and Security (ICCCS),2021; pp. 1–6. doi: 10.1109/ICCCS51487.2021.9433726. - DOI

[6] Pal S., et al. Predicting Substance Use Disorders Using Supervised Learning. *IEEE I2CNER.* 2020; pp. 122–126. doi: 10.1109/I2CNER51239.2020.9330045. - DOI

[7] Mehta H., Verma A. Analysis of Drug Consumption Dataset Using Machine Learning. *IEEE PuneCon.* 2022; pp. 1–5. doi: 10.1109/PuneCon56963.2022.10063090. - DOI