MIT 209 Project Proposal Submission

Submitted by:

Ceed Jennelle B. Lorenzo

Submitted to:

Dr. John Ed Augustus Escorial

Table of Contents

Introduction

Background of the Study

Illegal drug usage is a complicated issue in contemporary culture that impacts people of all demographics and is a major public health concern [1]. Understanding the behavioral and psychological elements that lead to drug use is crucial for addressing these problems. Strong correlations between personality factors and drug use patterns have been shown by research.

Research has demonstrated strong associations between personality traits and drug consumption patterns. The Five-Factor Model (NEO-FFI-R), impulsivity (BIS-11), and sensation-seeking (ImpSS) have been shown to predict individual drug use significantly. For example, studies indicate that higher impulsivity and sensation-seeking are commonly associated with the consumption of drugs like heroin and cocaine. At the same time, traits like agreeableness and conscientiousness are negatively correlated with substance abuse. The correlation between certain traits and specific drugs highlights the potential for machine-learning models to classify drug users based on personality profiles [2].

Recent advances in technology provide a good opportunity to utilize machine learning techniques to analyze large datasets about the individual's information on their personality traits, demographic attributes, and history of drug use. This study explores the underlying patterns of illegal drug consumption and develops a predictive risk model that can identify at-risk individuals based on their personality traits.

The dataset utilized in this study is from https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified and contains records for 1885 respondents, each characterized by 12 attributes. These attributes include widely used psychological measures such as NEO-FFI-R (measuring neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking). Additionally, demographic factors such as education level, age, gender, country of residence, and ethnicity are also recorded. All the input attributes are originally categorical but have been quantified to allow for real-valued input features, providing a robust foundation for statistical and machine-learning analysis.

This study uses machine learning techniques to create models that can forecast illegal drug use trends utilizing demographic data and personality traits. Additionally, the classification

classes can be made simpler by binarizing classes, which makes it easier to spot more general trends in user versus non-user behavior. Healthcare professionals, legislators, and social workers may find it easier to target preventive initiatives and identify high-risk individuals for early intervention if they have access to predictive models that can evaluate the likelihood of drug use.

Problem Statement

Illegal drug consumption remains a significant global challenge, with adverse effects on public health, social stability, and individual well-being [1]. Understanding the underlying factors that drive drug use behaviors can provide valuable insights for prevention and treatment programs. Past research has predominantly focused on environmental and social influences, and yet there is increasing evidence that an individual's personality traits and demographic factors also come into play in predicting drug consumption patterns.

Personality traits such as impulsivity, neuroticism, and sensation-seeking are known to be associated with risky behaviors, but their specific role in drug use is often nuanced. The interactions between these traits and demographic variables such as age, gender, education, and ethnicity remain underexplored. With this, the study aims to identify specific at-risk groups based on their personality traits.

Drug consumption patterns in the dataset were classified into multiple levels, from "Never Used" to "Used in the Last Day", which makes this a multi-class classification problem. In some areas, it may be more useful to transform these classes into binary categories (e.g., user vs non-user).

This study aims to address these challenges by leveraging machine learning techniques to predict illegal drug usage patterns using a dataset of 1885 respondents with personality and demographic information. The study will focus on building a predictive model to identify at-risk individuals and evaluate the risk factors associated with illegal drug substance abuse.

Objectives

The primary objective of this study is to predict illegal drug usage patterns using personality traits and socio-demographic factors. By applying machine learning algorithms, the research aims to improve the understanding of how psychological and demographic variables influence the likelihood of consuming illegal drugs. This ultimately provides insights for prevention, intervention, and policy-making.

This study seeks to answer the following objectives:

1. To examine the relationship between personality traits (NEO-FFI-R, BIS-11, ImpSS) and socio-demographic factors (age, gender, education level, ethnicity, and country of residence) in predicting illegal drug usage patterns.
2. To identify the most significant personality and demographic predictors for illegal drug consumption, providing insights for targeted interventions and public health strategies.
3. To develop and evaluate machine learning models for predicting illegal drug usage, focusing on multi-class classification (e.g., frequency of use) and binary classification (user vs. non-user).
4. To identify and analyze existing gaps and recommendations in related studies and the current research.

Significance

The significance of this study lies in the potential to understand and predict illegal drug usage patterns through a multitude of data from an individual's personality traits and demographic factors. This study also seeks to integrate machine learning models to analyze the psychological and demographic factors in the field of illegal drug usage. The insight generated by the model provides contributions to public health, behavioral psychology, and data science.

The ability to drug usage based on personality and demographic factors allows early targeted intervention programs for these individuals who display characteristics of more prone to substance abuse. This research also advances machine learning methodologies by addressing a complex multi-class classification problem. The study explores how drug consumption data can be efficiently modeled and classified, providing insights into the best approaches for binarizing multi-class problems and evaluating risk. This can contribute to the broader field of

predictive modeling and data science, offering techniques that can be applied to other areas of behavioral prediction.

Review of Related Literature

Substance abuse remains an unsolved issue in the world. The Public Health and the World Drug Problem reported the global impact of drug substance abuse in 2019. It points out that 1.3% of the total global disease burden is attributed to drug use, with 500,000 deaths annually linked to drug-related disorders, infectious diseases, and accidents. There is also a stark disparity in access to pain management medications, with 5.5 billion people, or 83% of the world's population, living in regions where access to moderate or severe pain treatment is inadequate. Drug use also contributes significantly to the spread of HIV, with 25% of new HIV infections outside sub-Saharan Africa occurring among people who inject drugs [1].

Furthermore, 23% of global hepatitis C cases and 33% of related deaths are tied to injection drug use. Despite these significant challenges, 31 million people globally suffer from drug use disorders, but only one in six has access to effective treatment. The disparity in treatment access is especially notable in WHO African Region countries, where 90% of people do not receive adequate opioid analgesics. Furthermore, 20 million people require palliative care annually, but only 3 million (15%) receive it. These numbers underscore the severe public health challenges posed by drug use worldwide and the need for greater healthcare access and intervention efforts [1].

The statistics presented by the Public Health and the World Drug Problem highlight the global impact of drug use and how it can be detrimental to the well-being of an individual. Hence, it gives room for more explanation of the various behavioral and demographic aspects of the individual and their likelihood of consuming illegal drugs. Interventions can be made earlier to help these individuals.

The relationship between personality traits, demographic characteristics, and the propensity to use illegal drugs has been explored in various studies. The Big Five personality traits—Neuroticism, Openness, Extraversion, Agreeableness, and Conscientiousness—play a significant role in this context. Research indicates that higher levels of Neuroticism, Openness, and Extraversion are positively associated with illegal drug use, while higher levels of Agreeableness and Conscientiousness are negatively associated with such behaviors. These

traits are important predictors of both the likelihood to use illegal drugs and the frequency of use among young individuals [3].

Demographic factors also influence drug use patterns. For instance, younger age, male gender, and urban residence are often correlated with higher rates of illegal drug use. Additionally, socioeconomic factors such as lower educational attainment and economic status can contribute to a higher likelihood of drug use. This multidimensional approach, combining personality assessments with demographic profiling, helps identify at-risk populations, which can be crucial for developing targeted preventive interventions [3].

These findings are supported by a broader body of research that consistently highlights the importance of personality traits as predictors of various behaviors, including substance use and abuse. By understanding these associations, psychologists and public health professionals can better tailor their approaches to prevention and intervention, potentially reducing the incidence of illegal drug use among vulnerable populations [4].

The Five-Factor Model (NEO-FFI-R)

The NEO Five-Factor Inventory-Revised (NEO-FFI-R) is a psychological assessment tool to measure the five major personalities: neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. These traits are often linked to various behaviors, including risk-taking and drug consumption. The NEO-FFI-R is widely used in personality research and is backed by extensive cross-cultural validations [5]. These personality traits are neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. They serve as predictors for various behaviors, including risky and addictive behaviors such as substance abuse. Neuroticism, for instance, is associated with emotional instability, often linked to higher susceptibility to drug use, while conscientiousness is negatively associated with substance abuse due to its correlation with self-discipline and impulse control [6]. Several studies also demonstrate that individuals with high neuroticism and low conscientiousness are more likely to resort to substance abuse [7].

Personality traits measured by NEO-FFI-R provide valuable insights into exploring the risk of an individual resorting to risky behaviors such as substance abuse. As noted by Fehrman et al. (2021), users of substances like heroin and ecstasy often exhibit distinct personality profiles, such as high neuroticism and impulsivity. These traits contribute to the psychological

understanding of substance abuse and support the use of personality assessments in predicting drug-related behaviors [2].

Impulsivity (BIS-11)

Impulsivity is another significant factor in predicting drug use and is commonly measured through the Barratt Impulsiveness Scale (BIS-11). Impulsivity reflects a tendency toward unplanned actions, often driven by a lack of forethought. It is a multi-dimensional construct encompassing attentional, motor, and non-planning components, each of which correlates with behaviors like substance abuse [8].

Impulsion-Seeking (ImpSS)

Sensation-seeking, measured using the ImpSS scale, refers to a person's predisposition to seek out new and intense experiences, often with little regard for the risks involved. Zuckerman (1994) describes sensation-seeking as a trait linked to a variety of high-risk activities, including drug use. Individuals who score high on the ImpSS scale are more prone to experimentation with substances, driven by a desire for novel experiences [9].

Several studies have established a connection between sensation-seeking and substance abuse. High sensation seekers are more likely to use substances such as marijuana, ecstasy, and cocaine, as these drugs provide the intense experiences they seek (Ahn et al., 2016) [10]. Additionally, individuals with high sensation-seeking tendencies tend to exhibit a greater frequency of drug use and are more likely to progress from experimentation to habitual use (Fehrman et al., 2021).

The statistics presented by the Public Health and the World Drug Problem highlight that 31 million people suffer from drug use disorders worldwide, and yet only one in six receive effective treatment [1]. This lack of access to treatment resonates with findings that socio-demographic factors like age, gender, education level, ethnicity, and country of residence significantly influence patterns of drug use and the availability of healthcare interventions [11], [12].

For example, results show that regions like sub-Saharan Africa experience a higher percentage of new HIV infections among people who inject drugs. This is consistent with research indicating that ethnicity and country of residence play crucial roles in access to care

and treatment. In many low-resource settings, drug use is compounded by a lack of access to medical care, as reflected in the statistic that 83% of the world's population lives in areas with inadequate access to medication for severe pain [13].

Furthermore, the age group most impacted by drug use—mainly teenagers and young adults—corresponds with previous research showing that younger people are more likely to experiment with drugs. Because people with lower levels of education are more prone to abuse substances and have less access to appropriate treatment alternatives, treating drug use disorders is made more difficult by socioeconomic and educational characteristics. [13].

This global public health crisis reported by the Public Health and the World Drug Problem reinforces the necessity of integrating socio-demographic data into the predictive modeling of drug consumption, to develop more targeted and effective interventions based on these risk factors. Hence, it is also necessary to explore the various machine learning techniques to predict the likelihood of an individual resorting to illegal drug use according to their personality traits and demographic factors.

Logistic regression has been applied to model binary outcomes determining whether an individual is a drug user or not. This study [2] employed logistic regression to analyze personality traits using the Five-Factor Model (NEO-FFI-R) and demographic factors like age and gender to predict drug use. Their findings highlighted that neuroticism, impulsivity, and low conscientiousness are strong predictors of drug consumption. A cross-validation method was used in this study to evaluate the model's performance, including metrics such as sensitivity and specificity to measure its ability to correctly identify drug users. Logistic regression can handle classifications in predicting if an individual is a drug user or not by calculating the odd ratio for each variable and making it a binary classification
[2], [10].

Decision trees and random forests are used to predict substance use behavior by combining demographic factors like education and impulsivity, with studies showing younger age, male gender, and low education as associated factors. [10], [11]. In these studies, both of these models were tested using precision, recall, and F1 scores, indicating the model's effectiveness in predicting drug use across various demographic groups.

In contrast, neural networks offer a more complex model for drug use prediction, capable of capturing non-linear patterns between multiple predictors. Neural networks have been

effectively used in studies to predict drug consumption by incorporating large datasets of psychological and demographic attributes. While these models achieve higher accuracy than traditional models, they are less interpretable compared to logistic regression and decision trees [10]. Studies using neural networks often emphasize sensation-seeking and impulsivity as leading factors in predicting drug use [2], [14].

Studies show that traits like neuroticism, impulsivity, and sensation-seeking, alongside demographic factors like age, gender, and education level, are crucial predictors of drug consumption. Moreover, risk assessment models based on these predictors allow for more targeted public health interventions, offering valuable tools for early detection and prevention of drug abuse. This body of research provides a strong foundation for the development of more sophisticated, data-driven approaches to addressing the global drug problem.

Methodology

This study employs a quantitative, observational research design. Machine learning models will be developed and evaluated for predictive performance, utilizing both multi-class and binary classification approaches.

Dataset

The dataset consists of records for 1,885 respondents, with 12 attributes known for each respondent. These attributes include personality traits (NEO-FFI-R, BIS-11 for impulsivity, and ImpSS for sensation-seeking), as well as demographic factors like age, gender, education level, ethnicity, and country of residence. Additionally, respondents provided information about their usage of 18 legal and illegal drugs, categorized into seven classes ranging from "Never Used" to "Used in Last Day." This data structure allows for both multi-class and binary classification, where the latter groups drug use into User vs. Non-user categories [2].

In the dataset, it collected both legal and illegal substances. However, this study will only cover the predictive analysis of illegal substances consumed by individuals and their personality traits and demographic features. Several drugs were used in the questionnaire such as cannabis, cocaine, heroin, and among others. Semeron, a fictitious drug was introduced to identify individuals who were over-claimers. Each drug participant can select from one of seven possible responses, ranging from “Never Used” to “Used in Last Day” providing a nuanced view of their drug usage history. This classification scheme offers 18 separate seven-class classification problems—one for each drug. The dataset presents additional opportunities, such as transforming the problem into binary classification tasks, where drug users and non-users can be more broadly categorized, and evaluating the risk of drug consumption for specific substances.

The dataset was initially loaded and inspected using basic Pandas functions to understand its composition and structure, which assisted in planning further cleaning and manipulation steps. An assessment of missing data was conducted to identify and quantify null values, ensuring the integrity and completeness of the dataset for robust analysis. The data types of each column were reviewed and adjusted as necessary to align with the analytical requirements of subsequent processing steps.

To uncover underlying patterns and distributions within the data, exploratory data analysis (EDA) was performed using both statistical summaries and visualization techniques: Descriptive statistics provided a foundational understanding of the central tendencies and variability within the sociodemographic and personality variables. Correlation matrices were constructed to detect and illustrate the relationships between personality traits and drug usage, informing the potential predictive power of certain traits.

Multiple machine learning models were developed to predict drug usage based on identified significant predictors such as XGBoost, Random Forest Classifier, and Support Vector Machine. The performance of each model was critically evaluated using accuracy, precision, recall, and F1 score metrics.

Approximately 70% of the original dataset were used to train the models, allowing them to learn the underlying patterns and relationships between the predictors and the response variable (drug usage). The selection of 70% for training is a balance between having enough data to effectively train the models and leaving a substantial portion for independent testing. The remaining 30% of the data formed the testing set. This subset was not used during the model training phase and served as new, unseen data for evaluating the models. This evaluation helps in assessing the generalization ability of the models to new data, which is crucial for ensuring the models' practical applicability.

A fixed random state (seed) was used during the splitting process to ensure the reproducibility of the research. This setting allows the partitioning to be exactly replicable, ensuring that the same training and testing sets can be regenerated for future verification or comparison studies.

Software and Tools

The following tools was employed for data analysis:

- Python (Scikit-learn): For implementing logistic regression, decision trees, and random forests.
- Pandas and NumPy: For data preprocessing and handling.
- Matplotlib and Seaborn: For visualizing model performance and data insights.

- Jupyter Notebook: This will serve as the main environment for developing, testing, and visualizing models. This interactive platform allows for step-by-step model development and performance tracking [2].

Results and Discussion

In this section, a comprehensive analysis of the data visualized on the patterns of illegal drug usage across various demographics are presented. The following results elucidate the prevalence and intensity of drug consumption, distinguished by demographic categories such as age, gender, and geographic location. This discussion integrates these findings with existing research to interpret the potential drivers behind these patterns and their implications for public health strategies and policy formulation.

The primary objective of this analysis is to examine the relationship between personality traits and socio-demographic factors. Specifically, the study explored how traits measured by NEO-FFI-R, BIS-11, and ImpSS scales interact with variables such as age, gender, education level, ethnicity, and country of residence to influence drug consumption behaviors.

Descriptive statistics were done to compute the summary of the dataset. This is to understand its distribution and then calculate the average usage rates for the various illegal drugs, roviding a clear picture of which drugs were most and least commonly used.

Table 1 Summary of Usage Rates for the various illegal drugs

Illegal Drugs	Drug Usage Rates
Mushroom	0.479299 (≈47.9% of participants have used mushrooms)
Ecstasy	0.458599 (≈45.9% have used ecstasy)
LSD	0.433121 (≈43.3% have used LSD)
Meth	0.242038 (≈24.2% have used meth)
Volatile substance abuse (VSA)	0.228238 (≈22.9% have used Volatile substance Abuse (VSA))
Ketamine	0.209660 (≈21% have used Ketamine)
Heroin	0.148620(≈15% have used heroin)

These values represent the relative prevalence of each drug's usage in the dataset. The top three (mushrooms, ecstasy, and LSD) have notably higher usage rates compared to meth, VSA, ketamine, and heroin. This suggests that, among the participants studied, mushrooms were the

most commonly used illegal drug, followed closely by ecstasy and LSD. Heroin, being the lowest on the list, indicates a lower proportion of participants reporting heroin use.

As noted in the data analysis, the relative prevalence rates for each drug's usage show distinctive patterns among the study participants. The most commonly used drugs are mushrooms, ecstasy, and LSD, indicating a propensity among the sample toward using psychedelic and recreational substances. This could suggest that participants are more inclined to experiment with drugs that are perceived as enhancing sensory experiences and social interactions, especially within youth and festival cultures.

Mushrooms are the most prevalently reported drug with approximately 47.93% of participants having used them, mushrooms lead in usage rates. This high prevalence might be linked to the cultural and possibly social environments of the participants, where such substances might be more readily available or accepted.

Ecstasy and LSD both follows closely, with usage rates of 45.86% and 43.31%, respectively, both drugs are similarly associated with social and recreational contexts. Ecstasy, often associated with dance music scenes and social gatherings, and LSD, known for its strong hallucinogenic effects, appear to attract users interested in profound sensory and emotional experiences.

On the other end of the spectrum, heroin has the lowest usage rate at about 14.86%. Its low prevalence within the dataset could reflect the general societal attitude toward heroin as a high-risk substance with severe addiction potential and social stigmatization. This stark contrast in the usage rates between psychedelics and opiates like heroin could be indicative of the differences in perceived safety, legal consequences, and social acceptance of these drugs.

The personality profiles, represented through scores on neuroticism, extraversion, openness, agreeableness, and conscientiousness, offer further insights into potential behavioral patterns and drug usage. For example, higher openness might correlate with increased likelihood of trying substances like LSD and mushrooms, which are often sought for their mind-expanding qualities. Conversely, lower agreeableness and conscientiousness could be associated with riskier behavior patterns, including higher drug use.

While one might expect higher neuroticism scores to correlate with substance use as a coping mechanism, the data does not explicitly confirm this without deeper statistical analysis to

understand correlations and causations. Higher extraversion might be associated with substances like ecstasy, which are popular in social settings.

To address the first objective of this research, the proponent examined the complex interplay between personality traits, socio-demographic factors, and their influence on illegal drug usage patterns. Utilizing a comprehensive dataset, a heatmap was used to visually represent correlation coefficients between various dimensions such as NEO-FFI-R personality traits (including Openness and Impulsivity), demographic characteristics (age, gender, education level, ethnicity, and country of residence), and patterns of drug consumption. This analysis aims to uncover significant predictors within these dimensions, which helps us understand how individual differences and socio-cultural contexts contribute to drug-related behaviors. This exploration not only aligns with established psychological theories and prior empirical findings but also seeks to provide a nuanced understanding of the predictors of drug use, which is crucial for developing targeted interventions.

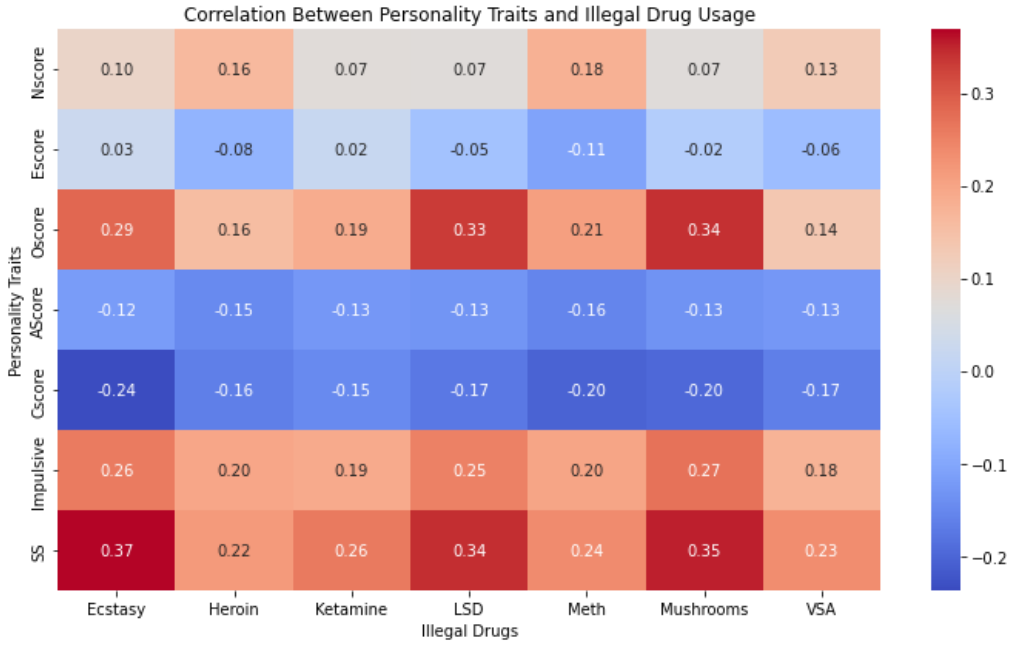

Figure 1 Correlation Between Personality Traits and Illegal Drug Usage

These coefficients likely relate to traits such as Openness (from NEO-FFI-R) and Impulsivity (from BIS-11) with certain drug use categories, particularly psychedelics like LSD, mushrooms, and ecstasy, which show correlations in the 0.25 to 0.37 range. This suggests a significant positive relationship, indicating that higher openness and impulsivity are associated with higher usage of these substances. This aligns with previous research indicating that openness is associated with

psychedelic use due to curiosity and the desire for new experiences (MacLean, Johnson, & Griffiths, 2011). Impulsivity is often linked to riskier behaviors, including experimenting with drugs (de Wit, 2009).

There is a moderate positive correlation between 0.19 – 0.24. These could correspond to traits such as Neuroticism or Extraversion with substances like cannabis and amphetamines. These drugs might appeal to individuals with higher extraversion due to their social usage contexts, while those with higher neuroticism might use these substances to manage stress or emotional turmoil. The research by Fairbairn et al. provides relevant insights where they found that individuals high in extraversion experience more mood enhancement from alcohol in social settings, potentially due to greater reward sensitivity. This suggests that extraversion could influence the use of substances like alcohol in socially rewarding situations, which might help explain patterns of substance use among those with high extraversion [15].

Negative correlations in this range might involve conscientiousness, typically associated with lower drug usage due to its links with self-discipline and carefulness (Kotov, Gamez, Schmidt, & Watson, 2010) [16]. Drugs that are less socially acceptable or perceived as riskier (like heroin or meth) might show stronger negative correlations with conscientiousness.

The heatmap shows correlations around 0.10 to 0.22 with gender, possibly indicating that males are slightly more likely than females to use drugs, which is consistent with literature reporting higher substance use rates among men [17]. The correlations between age and drug use likely vary, with younger adults showing higher drug use rates, particularly for substances like ecstasy and cannabis, which are popular in younger demographics [18]. Education might show mixed effects. Lower educational attainment may correlate with higher usage of certain illegal drugs due to socioeconomic factors and reduced access to education on substance abuse risks [6]. When it comes to ethnicity and country of residence, these factors likely have specific correlations depending on the cultural context and the legality of drugs in various countries. For example, countries with stricter drug laws might see lower drug usage rates, or certain ethnic groups might have varying tendencies toward drug use based on cultural norms.

To address the second objective of this research, the most significant personality and demographic predictors for illegal drug usage were looked into. Ecstasy usage is linked to traits such as Openness to Experience, Conscientiousness, and Sensation Seeking, with prevalent use among individuals in the 45-64 age range, also noting gender differences. Similarly, Heroin is

associated with Neuroticism, Openness, and Agreeableness, primarily among ages 25-34, 35-44, and 55-64, suggesting a base effect in consumption patterns. Ketamine shares predictors with Heroin but emphasizes age groups 25-34 and 55-64. LSD's use correlates with Extraversion, Openness, and Sensation-seeking across three distinct age categories from 25 to 64 years. Methamphetamine (Meth) shows a correlation with Neuroticism, Openness, and Agreeableness, significantly among males with a mid-level education. Mushrooms attract users with a wide range of personality traits from Neuroticism to Sensation Seeking, especially within the 25-44 age demographic. Lastly, Volatile Substance Abuse (VSA) highlights Openness and Sensation Seeking as key traits, with significant usage noted in the 25-44 age group, also influenced by geographical factors.

These findings underscore the role of personality traits such as Openness and Sensation Seeking in the propensity for drug experimentation, alongside marked demographic trends that provide crucial insights for targeted preventive and intervention strategies. The consistent appearance of certain age groups as more likely to engage in drug use emphasizes the need for focused health strategies to address these specific populations effectively.

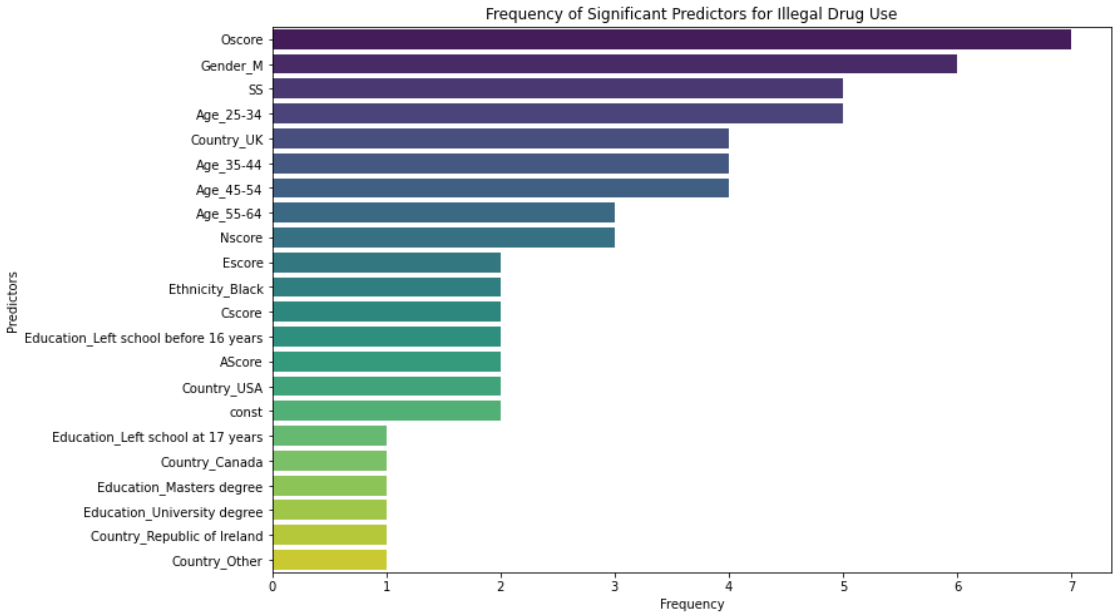

Figure 2 Frequency of Significant Predictors for Illegal Drug Usage

Figure 2 highlights the most frequent demographic and personality predictors associated with illegal drug usage. Among the top predictors, gender (Male) emerges as the most significant factor, followed by personality traits such as Nscore (Neuroticism) and Escore (Extraversion). Age

also plays a critical role, with age groups 25-34, 35-44, 45-54, and 55-64 appearing prominently. Education levels, including "Left school before 16 years," "Left school at 17 years," "Masters degree," and "University degree," also show strong associations with drug usage. Geographic factors like Country (UK, USA, Canada) and Ethnicity (Black) further contribute to the analysis. The chart suggests that illegal drug use is influenced by a combination of demographic variables (age, gender, education, and location) and personality traits, with gender and age being particularly dominant predictors.

```
=== SVM Binary Classification ===
              precision    recall  f1-score   support

           0       0.74      0.49      0.59       203
           1       0.76      0.90      0.83       363

    accuracy                           0.76       566
   macro avg       0.75      0.70      0.71       566
weighted avg       0.75      0.76      0.74       566
```

Figure 3 Support Vector Machine (SVM) for Binary Classification of User vs Non-User of Illegal Drugs

A support vector machine was used for the classification of models. It evaluates the performance of the model in predicting 2 classes: 0 for non-users of illegal drugs and 1 for user of illegal drugs. The model demonstrates a precision of 74% for non-users and 76% for users, indicating that it is slightly better at correctly identifying users than non-users. However, the recall for non-users is low at 49%, suggesting the model struggles to correctly classify this group, while it excels for users with a recall of 90%. This disparity results in an F1-score of 59% for non-users and a much higher 83% for users, reflecting the model's stronger performance in identifying users due to the better balance of precision and recall. The support numbers—203 non-users and 363 users—highlight that the dataset contains more users, which may influence the model's predictive tendencies. Overall, the model achieves an accuracy of 76%, with macro and weighted averages across metrics showing moderate to good performance, calculated at 75% and 76% respectively. This suggests while the model is effective at identifying drug users, improvements are needed to enhance its ability to detect non-users, potentially through model parameter tuning, more balanced data, or additional discriminative features.

```
=== SVM Multi-Class Classification ===
              precision    recall  f1-score   support

    Frequent       0.71      0.90      0.79       307
    Non-user       0.72      0.63      0.67       203
  Occasional       0.00      0.00      0.00        56

    accuracy                           0.71       566
   macro avg       0.48      0.51      0.49       566
weighted avg       0.64      0.71      0.67       566
```

Figure 4 Support Vector Machine (SVM) Multi-Classification of Frequency of Illegal Drug Usage

The SVM (Support Vector Machine) Multi-Class Classification report provides an insightful breakdown of the model's ability to classify individuals into three categories based on their drug usage: Frequent, Non-user, and Occasional. The model shows a strong performance for Frequent users, achieving a precision of 71%, recall of 90%, and an F1-score of 79% from a substantial support of 307 cases. This indicates the model's robust capability in accurately identifying individuals who frequently use drugs, capturing 90% of actual frequent users correctly.

For Non-users, the model demonstrates reasonable effectiveness with a precision of 72%, recall of 63%, and an F1-score of 67% based on 203 cases. While the precision is slightly higher than for Frequent users, the recall is notably lower, suggesting that the model is somewhat less adept at correctly identifying non-users, missing a significant proportion of actual non-users.

However, the model's performance drastically drops for Occasional users, for whom it fails to identify any correctly, resulting in zero precision, recall, and F1-score across 56 cases. This suggests a significant deficiency in the model's ability to distinguish occasional drug users from others, possibly due to insufficient data or inadequate feature differentiation within this category.

Overall, the model achieves an accuracy of 71%, which, while seemingly high, is primarily driven by its success in identifying Frequent users. The macro averages across all classes for precision (48%), recall (51%), and F1-score (49%) are significantly lower than the weighted averages of precision (64%), recall (71%), and F1-score (67%). This discrepancy indicates that the Frequent user class, due to its larger sample size, disproportionately influences the model's performance metrics. The notably poor performance in classifying Occasional users highlights a critical area for improvement, suggesting that enhancing the model's training with more balanced

data or more distinct features could help improve its overall classification ability, particularly for less frequent drug users.

```
Accuracy: 0.765017667844523
Precision: 0.7875
Recall: 0.8677685950413223
F1 Score: 0.8256880733944953
Classification Report:
              precision    recall  f1-score   support

    Non-User       0.71      0.58      0.64       203
        User       0.79      0.87      0.83       363

    accuracy                           0.77       566
   macro avg       0.75      0.72      0.73       566
weighted avg       0.76      0.77      0.76       566
```

Figure 5 XG Boost Classification

The results displayed show the performance metrics of a classification model used to predict whether individuals are users or non-users of illegal drugs. The model has achieved an overall accuracy of approximately 76.51%, which indicates the proportion of total predictions (both user and non-user classifications) that the model made correctly.

The precision for predicting non-users is 71%, meaning that when the model predicts someone as a non-user, there's a 71% chance the prediction is correct. Conversely, the precision for users is higher at 79%, suggesting that the model is more reliable in identifying users of illegal drugs compared to non-users.

The recall, or sensitivity, measures the model's ability to identify actual positives accurately. For non-users, the recall is 58%, indicating that the model correctly identifies 58% of all actual non-users. For users, the recall is much higher at 87%, showing that the model is particularly effective at identifying actual users, but it may miss identifying a fair portion of non-users.

The F1 score stands at 64% for non-users and 83% for users. This metric shows that the model is more balanced and effective in classifying users compared to non-users, as evidenced by the higher F1 score for users. This suggests that while the model is generally good at identifying who uses drugs, it struggles somewhat more with accurately identifying individuals who do not use drugs.

Overall, the model's performance across these metrics—reflected in the macro and weighted averages—shows a decent level of effectiveness, with a balance between recognizing drug users and non-users, although it performs better at the former. The macro average F1 score across both categories is 73%, indicating an overall solid performance, while the weighted average considers the support, or the number of instances for each class, resulting in a slightly higher average precision and recall of 76%.

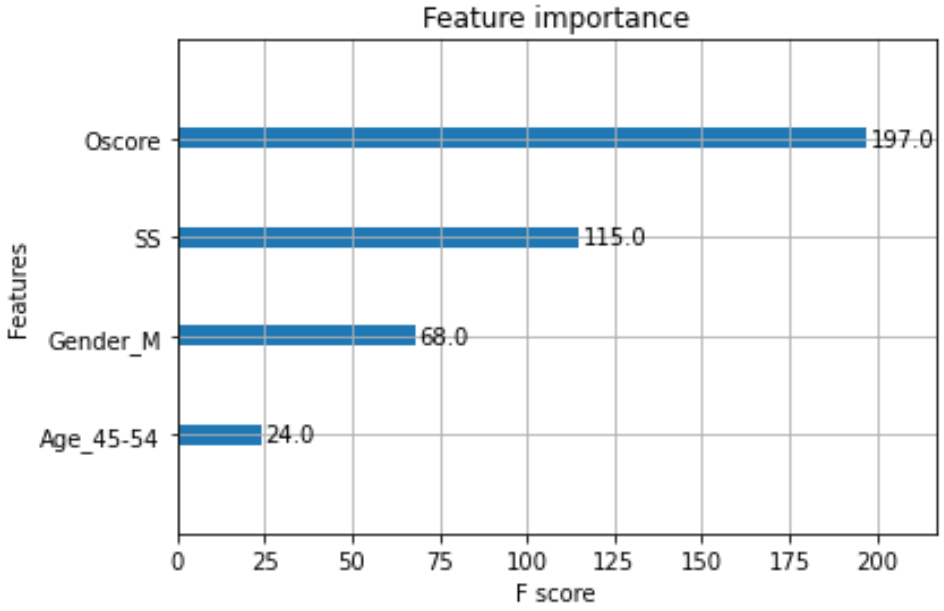

Figure 6 Feature Importance

The feature corresponding to the highest value (197.0) is the most influential in the model's decisions which is the Oscore or the Openness to Experience. This suggests that changes in this feature have the largest impact on the model's prediction outcome. The next significant feature, marked by 115.0, also plays a crucial role, though its impact is noticeably less than the top feature. Features corresponding to 68.0 and 24.0 have progressively less influence on the model's output. While still important, their changes result in less significant shifts in predictions compared to the top features.

References

[1] H. Who, “THE PUBLIC HEALTH DIMENSION OF THE WORLD DRUG PROBLEM.”

[2] E. Fehrman, V. Egan, A. N. Gorban, J. Levesley, E. M. Mirkes, and A. K. Muhammad, *Personality Traits and Drug Consumption*. .

[3] W. Kang, “Big Five personality traits predict illegal drug use in young people,” *Acta Psychol. (Amst).*, vol. 231, no. October, p. 103794, 2022, doi: 10.1016/j.actpsy.2022.103794.

[4] U. of Leicester, “Personality traits of drug users,” 2019. https://neurosciencenews.com/personality-substance-abuse-14209/.

[5] R. R. McCrae and P. T. Costa, “A contemplated revision of the NEO Five-Factor Inventory,” *Pers. Individ. Dif.*, vol. 36, no. 3, pp. 587–596, 2004, doi: 10.1016/S0191-8869(03)00118-1.

[6] M. E. Patrick, J. E. Schulenberg, P. M. O’Malley, L. D. Johnston, and J. G. Bachman, “Adolescents’ reported reasons for alcohol and marijuana use as predictors of substance use and problems in adulthood,” *J. Stud. Alcohol Drugs*, vol. 72, no. 1, pp. 106–116, 2011, doi: 10.15288/jsad.2011.72.106.

[7] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban, “The Five Factor Model of personality and evaluation of drug consumption risk,” pp. 1–38, 2014.

[8] J. Patton, M. Stanford, and E. Barratt, “Factor Structure of the Barratt Impulsiveness Scale,” 1995.

[9] M. Zuckerman, “Behavioral expressions and biosocial bases of sensation seeking,” vol. 16, no. 1, pp. 1–23, 2022, doi: https://doi.org/10.1017/CBO9780511527937.

[10] W. Y. Ahn, D. Ramesh, F. G. Moeller, and J. Vassileva, “Utility of machine-learning approaches to identify behavioral markers for substance use disorders: Impulsivity dimensions as predictors of current cocaine dependence,” *Front. Psychiatry*, vol. 7, no. MAR, pp. 1–11, 2016, doi: 10.3389/fpsyt.2016.00034.

[11] S. Center for Behavioral Health Statistics, “2014 National Survey on Drug Use and Health: Methodological Summary and Definitions,” no. November 2023, 2014.

[12] A. M. Jeffers, S. Glantz, A. Byers, and S. Keyhani, “Sociodemographic Characteristics Associated With and Prevalence and Frequency of Cannabis Use Among Adults in the US,” *JAMA Netw. Open*, vol. 4, no. 11, p. E2136571, 2021, doi:

10.1001/jamanetworkopen.2021.36571.

[13] J. A. Swartz *et al.*, “Associations among drug acquisition and use behaviors, psychosocial attributes, and opioid-involved overdoses,” *BMC Public Health*, vol. 24, no. 1, pp. 1–11, 2024, doi: 10.1186/s12889-024-19217-y.

[14] U. I. Islam, E. Haque, D. Alsalman, M. N. Islam, M. A. Moni, and I. H. Sarker, “A Machine Learning Model for Predicting Individual Substance Abuse with Associated Risk-Factors,” *Ann. Data Sci.*, vol. 10, no. 6, pp. 1607–1634, 2023, doi: 10.1007/s40745-022-00381-0.

[15] M. A. S. Catharine E Fairbairn, Kasey G Creswell, Jeffrey F Cohn, John M Levine, Aidan G C Wright, “Extraversion and the Rewarding Effects of Alcohol in a Social Context,” *Journal of Abnormal Psychology*, 2015. https://discovery.researcher.life/article/extraversion-and-the-rewarding-effects-of-alcohol-in-a-social-context/d12b1331f7813d93b25e323fa920e36d (accessed Sep. 30, 2024).

[16] D. Kotov, R., Gamez, W., Schmidt, F., & Watson, “Linking ‘big’ personality traits to anxiety, depressive, and substance use disorders: A meta-analysis,” *Psychol. Bull.*, vol. 136, no. 5, pp. 768–821, 2010.

[17] S. Seedat *et al.*, “Disorders in the WHO World Mental Health Surveys,” *Arch. Gen. Psychiatry*, vol. 66, no. 7, pp. 785–795, 2009, doi: 10.1001/archgenpsychiatry.2009.36.Cross-national.

[18] L. Degenhardt *et al.*, “Toward a global view of alcohol, tobacco, cannabis, and cocaine use: Findings from the WHO world mental health surveys,” *PLoS Med.*, vol. 5, no. 7, pp. 1053–1067, 2008, doi: 10.1371/journal.pmed.0050141.