

Predicting Consumption of Illegal Drugs Using Demographic and Personality Features

Riya Bhaskar
rubha02@gmail.com

ABSTRACT

Drug use among UK school-aged children is rising sharply, with more young people entering treatment, starting earlier, and using higher-risk substances like ketamine and ecstasy. The combination of early onset, polydrug use, and widespread mental-health difficulties shows this is a growing public-health crisis rather than simple experimentation. It is important to reduce this as drug usage can be harmful especially to those younger. Additionally drugs nowadays are commonly mixed with other harmful substances that can cause further harm to children. I aimed to find models that showed very clearly the correlation between drug usage and personality factors so those can be used to decrease the problem. To approach this, a dataset containing 1885 samples, 12 features, and drug usage levels for 18 drugs was used. The models used were decision trees, logistic regression and KNN. Each model predicted the drug usage level from CL0 (never used) to CL6 (used in the last day) within 4 categories: legal, mildly addictive, moderately addictive and highly addictive. Most models showed a similar accuracy for each target, legal being approximately 82%, mildly addictive being roughly 60%, moderately addictive being approximately 35% and highly addictive being roughly 48%. Heatmaps were used to analyse where the model was making mistakes and it was shown through the heatmaps, though not predicting the correct value, were largely within a close range to the actual value, with exceptions for the CL0 labels and the legal substances target due to data class imbalances as it is highly unlikely for people to take hard drugs during their lifetime and many do consume legal substances such as chocolate and caffeine. It can be seen that illegal drug usage cannot be perfectly predicted with our model, indicating that we would need to have more specific factors beyond just personality and demographic traits to get a clearer understanding and increase the predictive accuracy of the models.

INTRODUCTION

Many public health agencies now refer to the drug use among school-aged children in the UK as a youth drug-use pandemic, characterized by an increase in treatment rates, an earlier onset of use, and increased exposure to high-risk substances. According to the most recent national data, 16,212 minors under the age of 17 enrolled in drug or alcohol treatment in 2024–2025, a 13% rise from the year before. 86% of adolescents in treatment report cannabis difficulties, indicating that cannabis is still the most common drug. However, there is a concerning trend of ketamine usage, which rose from 5% of cases in 2021–2022

Sep 2026
Vol 11. No 6.

to 9% in 2024–2025. Over the same time span, inhalant usage, including nitrous oxide, has also almost tripled, from 329 to 919 cases [2].

The magnitude of the problem is also supported by school survey data, which shows that 17.4% of 15-year-olds reported using drugs more than once in 2023 and that drug experimenting rises sharply with age throughout secondary school years [1]. In addition to increasing prevalence, the pandemic is also increasing vulnerability. 53% of children in treatment report polydrug use, which is the frequent combining of various substances, and 82% of children started using substances before the age of 15. With 48% of children entering treatment indicating a need for mental health therapy and girls exhibiting notably high rates of self-harm and emotional distress, mental health needs are intricately linked [2]. When taken as a whole, these statistics point to a fast worsening public health crisis: youngsters are using drugs younger, more dangerously, and more frequently are seeking treatment. This is an actual pandemic, not merely an increase in experimenting, because of the early onset, polydrug usage, and mental health susceptibility.

Background

Due to the fact that younger people, men, those with less education, and members of specific ethnic groups are more prone to take illegal drugs, prior research has demonstrated that both demographic characteristics and personality features significantly predict drug usage. Cannabis users exhibit high Openness and low Agreeableness/Conscientiousness, while tobacco users exhibit a milder version of the same traits, and cocaine and heroin users score extremely high on Neuroticism and very low on Conscientiousness, according to the Five-Factor Model of personality traits. [3]

Previous literature has demonstrated the substantial correlation between some personality qualities and risky, impulsive behaviors, such as the use of illegal drugs. According to their review, substance use is more common among those with high neuroticism and low agreeableness or conscientiousness. Specifically in the paper absolute risk prediction for cannabis use disorder in adolescence and early adulthood using Bayesian Machine Learning by Tingfang Wang, Joseph M Boden, Swati Biswas and Pankaj K Choudhary, using their model predicted that lower conscientiousness, higher neuroticism and higher openness lead to substantially higher risk of cannabis use disorder. The model predicted a 0.34 hazard ratio for conscientiousness, relating towards higher levels of conscientiousness reduces chances of cannabis use disorder. The model predicted a hazard ratio of 5.64 for neuroticism, the high number effectively showing how neuroticism can be used to predict the chance of cannabis use disorder. The model also predicts a hazard ratio of 5.16 for openness, again showing a greater chance of cannabis use disorder. This model 5 year prediction had an AUC of 0.68 in cross-validation and 0.64 and 0.75 in two external validation datasets. Alongside this, the model is a prediction of cannabis use disorder rather than a causal link between the factors, making the estimation not accurate[5]. The essay provides a strong psychological explanation to go along with demographic trends, even if it doesn't specifically address demographics. It makes it evident that psychology plays a significant role in determining who is more susceptible to drug use. [4] Both of these studies shed light on my study issue because they highlight particular personality-related side effects and indicators of drug use. The first article provides me with a wealth of background material to address my study problem by displaying both demographic and psychological traits that are directly related to it. [3]

Dataset

The data set relates personality and broad demographic features to legal and illegal drug consumption, all represented numerically. This data contains records from 1885 responders, allowing for a wide range of different values to illustrate major patterns clearly and effectively. For each respondent 12 attributes were known personality measures such as neuroticism, extraversion, openness to experience, agreeableness, conscientiousness, impulsiveness and sensation seeking, and then there were demographic features known such as level of education, age, gender, country of residence and ethnicity. The data set went over the drugs alcohol, amphetamines, amyl nitrate, benzodiazepine, caffeine, cannabis, chocolate, cocaine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, magic mushrooms, nicotine, semeron and VSA. Semeron and VSA are fictitious drugs and were used to ensure that the candidates filling out the form were not lying about their answers. Each candidate was to answer whether they had taken said drug and if so how recently they had taken it. The drugs tried were to show the consumption of such drugs, they were then categorised from CL0 - CL6. The CL0 - CL6 were used as simplified values for time periods of when the drug was last taken, CL0 being never used, CL1 shows that it was used over a decade ago, CL2 shows that it was used in the last decade, CL3 shows that it was used in the last year, CL4 shows that it was used in last month, CL5 shows that it was used in the last week, CL6 shows that it was used in the last day.

This data set is from the 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". All data was collected from a survey taken by people from the public of ages ranging from 18 - 65+. However this caused class imbalance within the data taken. To overcome this they did two sampling techniques Random Oversampling, a method which tackles class imbalance by randomly sampling variables within minority classes essentially copying the observations, Synthetic Minority Over-sampling Technique using Euclidean Norm (SMOTEN). SMOTEN allows minority classes to be represented better, reducing class imbalance within the training data. This is particularly useful as many of the substances such as chocolate have a severe class imbalance which can impact the accuracy of the model. These methods take the raw data and turn it into something that is usable for machine learning by reducing class imbalance.

Data Preprocessing

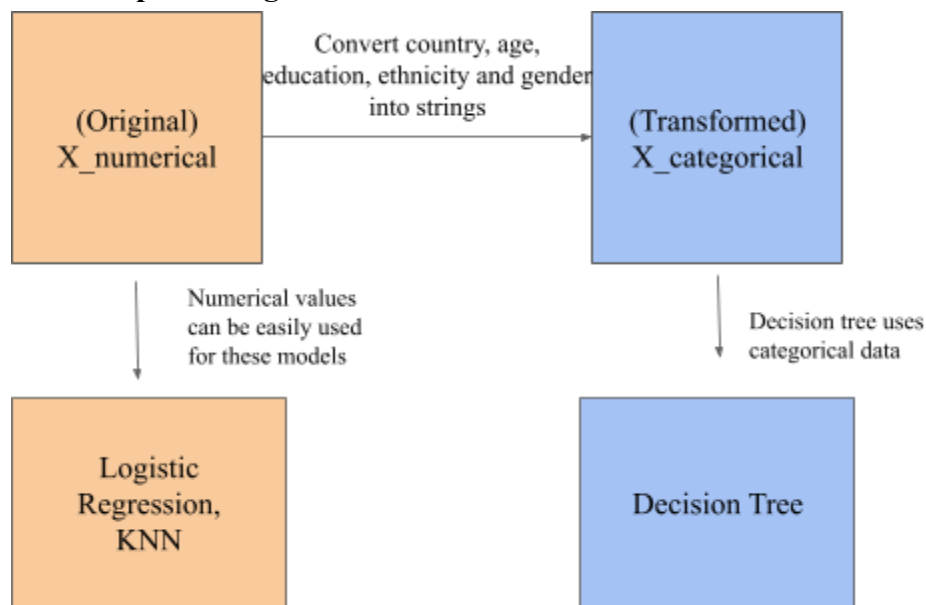


Figure 1. Data pre-processing overview.

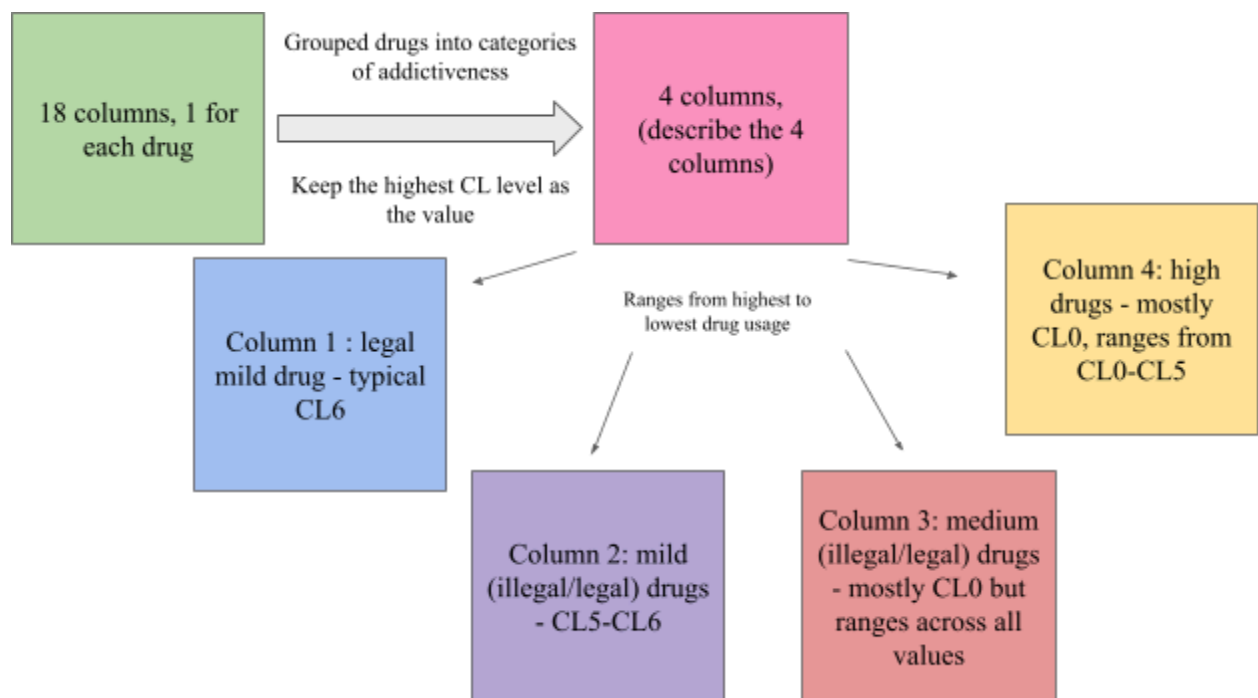


Figure 2. Label data pre-processing overview.

The data provided to me stored some values as numerical values representing the true string values. This made the data set more confusing as I would have to refer back to the key provided to fully understand what each value was. As part of my preprocessing I changed this by reidentifying each numerical value with an appropriate string value within a dictionary. However I only did this for indices that required this change which were country, age, education, ethnicity and gender. I also did this to implement a decision tree model which uses categorical values, whereas Logistic Regression and KNN require numerical values so I used the original version of the indices for this. Some features did not require this as there wouldn't be an appropriate string value for this. This separated my features from originally fully being numerical, to being numerical and categorical (Fig. 1).

The targets I tested were the hardness of the drug and organized into 4 categories: legal mild drugs, illegal/legal mild drugs, illegal medium drugs and illegal high drugs with a total of 18 drugs. Into the legal category I moved chocolate and caffeine as these are substances that are typically used on a daily basis by most people and aren't especially harmful and they are both fully legal with little implications. In the illegal/legal mild drug category I moved the drugs alcohol, cannabis, legal highs and nicotine, as all of these drugs are relatively addictive and can cause harm, but not as severe as some of the other drugs present in the dataset. I moved amphetamines, amyl nitrate, benzodiazepine, ecstasy, ketamine, methadone and mushrooms into the illegal medium category due to their moderate addictiveness and moderate harmfulness both mentally and physically. In the illegal high category I moved cocaine, crack, heroine and LSD as these were the most harmful of the drugs tested. Seremon was provided in the dataset although it was a fictitious drug, this was to ensure truthfulness of the statements being provided by the respondents. I decided to drop this as I don't need to test the fictitious drug to find the correlation between personality and drug usage (Fig. 2). Alongside these rankings were backed up by Professor David J Nutt in the paper drug harms in the UK : a multicriteria decision analysis. This paper doesn't include all the drugs discussed within this paper however does give an accurate attestation towards how the drugs have been ranked[6]. Also in the UK government classification under the misuse of drugs acts we can see drugs such amphetamines are ranked as class B directly corresponding to the rankings shown in this paper [7].

METHODOLOGY AND MODELS

For my models I used Logistic Regression, decision trees and KNN from the sklearn library to model the probabilities and accuracy of my targets for my research question. I also used these models to optimize the hyperparameters for each model at each drug level , and visualised these results in a heatmap within each model for each drug hardness level. To evaluate class imbalance I used confusion matrices as they show how many observations from each actual were predicted in each class. It also immediately the model is simply predicting the majority class.

Logistic Regression

Logistic Regression is a machine learning algorithm that is used for classification tasks where the goal is to predict the probability that an observation belongs to a given class or not. Although commonly used for binary classification it can also be utilised for multiclass classification making it appropriate for my categorical variable from CL0 - CL6. I used this model because the values such as CL0 and CL6 are categorical and logistic regression can estimate the probability of an observation belonging to a given class based on a set of predictors. The dataset includes numerical personality scores and encoded demographic features, which can be incorporated well into the model after adequate and appropriate pre-processing. Another advantage of logistic regression is that the model is interpretable, you can easily examine how traits such as neuroticism, openness, age, or education influence the likelihood of belonging to each drug-use category. I choose to use a confusion matrix and heatmap to examine the distribution of correct and incorrect predictions and precision of each target. This is particularly important as my dataset contained substantial class imbalance. For example my model showed that in the target category 'legal' the accuracy was artificially high as the model predicted CL6 only and most of the actual values were CL6 as it was the dominant class, despite the model performing poorly at identifying the other classes. The confusion matrix therefore provides more detailed information than accuracy alone by showing which classes are being correctly or incorrectly predicted, while the heatmap makes these patterns easier to interpret visually. As a hyperparameter I used `max_iter()` set at 1000. `Max_iter()` ascertains the maximum number of repeats required for the optimisation of the algorithm used to estimate the model's coefficient. The model's predictive performance does not directly improve by increasing `max_iter()`, instead it allows the optimisation algorithm more iterations to approach a solution. This helps prevent convergence warnings and ensures that the model has sufficient opportunity to find the optimal coefficients for the chosen settings.

Decision Trees

A Decision Tree is a supervised machine learning model that predicts an outcome asking a series of threshold-based questions about the input variable. Each question divides the data into smaller groups, and this process continues with further splits until the model makes a final prediction at a leaf node. I used this model because the dataset includes numerical personality scores, encoded demographic variables, and categorical drug-use classes, and decision trees can handle these variables without needing feature scaling. They are also able to represent non-linear relationships and interactions between predictors, which logistic regression might not capture unless interaction terms or other transformations are specifically added. Decision trees are quite interpretable, since the sequence of splits can be visualized, making it clear how features such as neuroticism, openness, age, or education influence each classification outcome and showing their importance. I chose to use grid search, since decision trees are sensitive to their chosen structure and may overfit if they become too complex. The grid search tested different combinations of the parameters `max-depth` and `criterion` (entropy versus gini), using cross-validation to identify the combination that gave the best validation performance. I also used two accuracy measures: cross-validation accuracy (CVA) and the overall accuracy of the model. The CVA is the model's average validation performance across multiple validation folds during hyperparameter selection, while test-set accuracy measures how the final model performs on previously unseen data. The test set accuracy gives a more realistic estimate of how the final model will generalize to new

observations, assuming the test set was kept separate from the model-selection process. The relatively shallow trees, especially the one with a maximum depth of 1, could suggest a simpler decision boundary and possibly higher bias with lower variance. Still, whether this points to underfitting should be evaluated using the cross-validation and test-set results instead of relying on depth alone.

KNN

KNN is a non-parametric, instance-based learning method. Instead of learning a formula or coefficients, most similar observations in the training dataset, rather than learning a predefined equation or set of coefficients. It does this through cross validation by going through a range of potential k-values and seeing which one is optimal for the model. N_neighbours parameters determine the number of neighbouring observations considered when making each prediction which impacts the balance between variance, bias and complexity. A low value of k can make the model more responsive towards local patterns and noise, increasing variance and possibly generalizing the data (overfitting). A higher value of k could increase bias but does produce a more stable prediction. Choosing a good value for n_neighbours makes the difference between a model that overreacts to tiny fluctuations in your drug-consumption dataset and one that captures the broader behavioural patterns you need. I used cross-validation to assess a large number of values for k and identify which one would be best for each category. There was a high amount of class imbalance within the data set, the accuracy of the model was clarified cautiously, as a high accuracy can be misleading when some classes are much more common than others. So I used cross-validation to determine the best k value rather than starting at a low value. Through trial I found different ranges of values for each n_neighbours to find the max optimum instead of just staying at a low starting number, especially if the optimum number output was high. I found that the optimal n-neighbours of 19, 23, 100, 22 were best for legal, mildly addictive, moderately addictive, and highly addictive substances, respectively. These results demonstrate that the optimal number of neighbours varied between categories reflecting differences in the data. Overall, n_neighbours is an important hyperparameter because it controls the trade-off between sensitivity to individual observations and generalisation to unseen data

Each model produces an accuracy for each category legal, low, medium and high addictiveness. For logistic regression the model produced a test accuracy and a confusion matrix which is visualised as a heatmap to present the instances in each class. In the decision tree models an accuracy_score is shown after using the hyperparameter max_depth and cross validation to produce a score which is equal to the accuracy of the model. Finally for KNN a best k is produced which refers to the number of nearest neighbours. The mean accuracy for each k is calculated and plotted as best k. The number of neighbours that yields the highest cross validation accuracy. A heatmap is also produced for decision trees and KNN.

RESULTS AND DISCUSSION

Logistic Regression

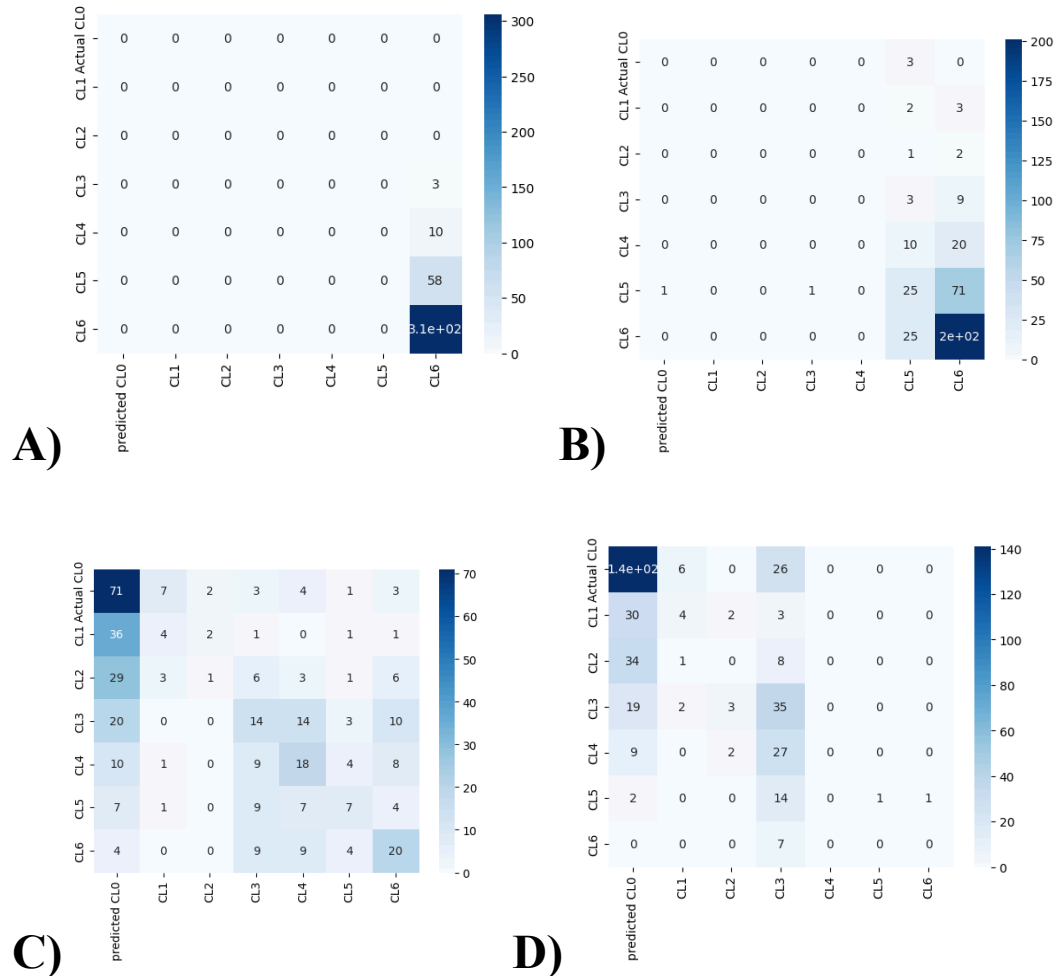


Figure 3: Heatmap showing predicted vs. actual drug usage levels for logistic regression

A) legal substances

B) low addictiveness

C) medium addictiveness

D) high addictiveness drugs

Figure 3 shows the heatmaps for the categories legal, low, medium and high addictiveness for the logistic regression model. There was a strong diagonal for CL6 with sparse amounts of the lower classes due to CL6's dominance. This caused the model to predict CL6 as the majority option when predicting. For low category CL6 is less dominant and the misclassification is more spread out. This signifies greater confusion between different usage levels (CL0 to CL6). For the medium drug usage heatmap there is no consistent pattern or clear dominance across the classes. This shows a high level of confusion, indicating widespread misclassification between various usage levels. Finally in the high category there is relative strength of CL0, causing the model to mostly predict CL0, creating misclassification towards CL0 and various other classes. The heatmap is more concentrated across the middle compared to medium.

The accuracy shown from the Logistic Regression isn't fully accurate as most accuracy scores are quite low excluding legal which showed a high accuracy value of 0.812(3sf), this is most likely due to the high class imbalance in legal as there is a high proportion of CL6 in the target legal causing the model to have a high bias causing a high accuracy. The model is not very good at predicting the drugs as most accuracies are low, this could be due to not adding in extra terms in my code that would allow the model to determine ideas such as drug usage only increases if both neuroticism and impulsivity are high. I also cannot get a perfect prediction from this as my factors are limited demographically and would require more specifics into this area to get a more accurate testament of the drug usage per hardness.

For predicting legal in logistic regression the most influential factors are sensation seeking, openness and impulsiveness. This can be explained as many of the variables in the legal category such as chocolates and caffeine, though they don't have major health implications in small amounts, they can be unhealthy in large amounts and are commonly consumed in smaller quantities mostly for the taste and enjoyment rather than as a meal or for specific health benefits. In the low category the factors impacting are age and impulsiveness. Drugs in this category such as methadone are typically taken by younger people due to peer pressure which can then lead to addiction from an early age causing major problems further in life [8]. In the medium category the most impactful factor is age, impulsiveness and sensation seeking. All the same attributes from the low category apply to the medium category as well, however these drugs have a greater impact on cognitive function and can have a greater impact on mental health issues such as anxiety which is why sensation seeking becomes a highly impactful factor for this category of drugs. For the high addictiveness category of drugs the factors that impact the predictedness the most are sensation seeking and impulsive. This is because drugs that are this dangerous and highly addictive can be taken impulsively after one time use due their addictiveness and highly impact cognitive function by increasing social outgoingness and causing hallucination, these sensations can be specifically wanted and a clear reason why these drugs can be taken.

Decision tree

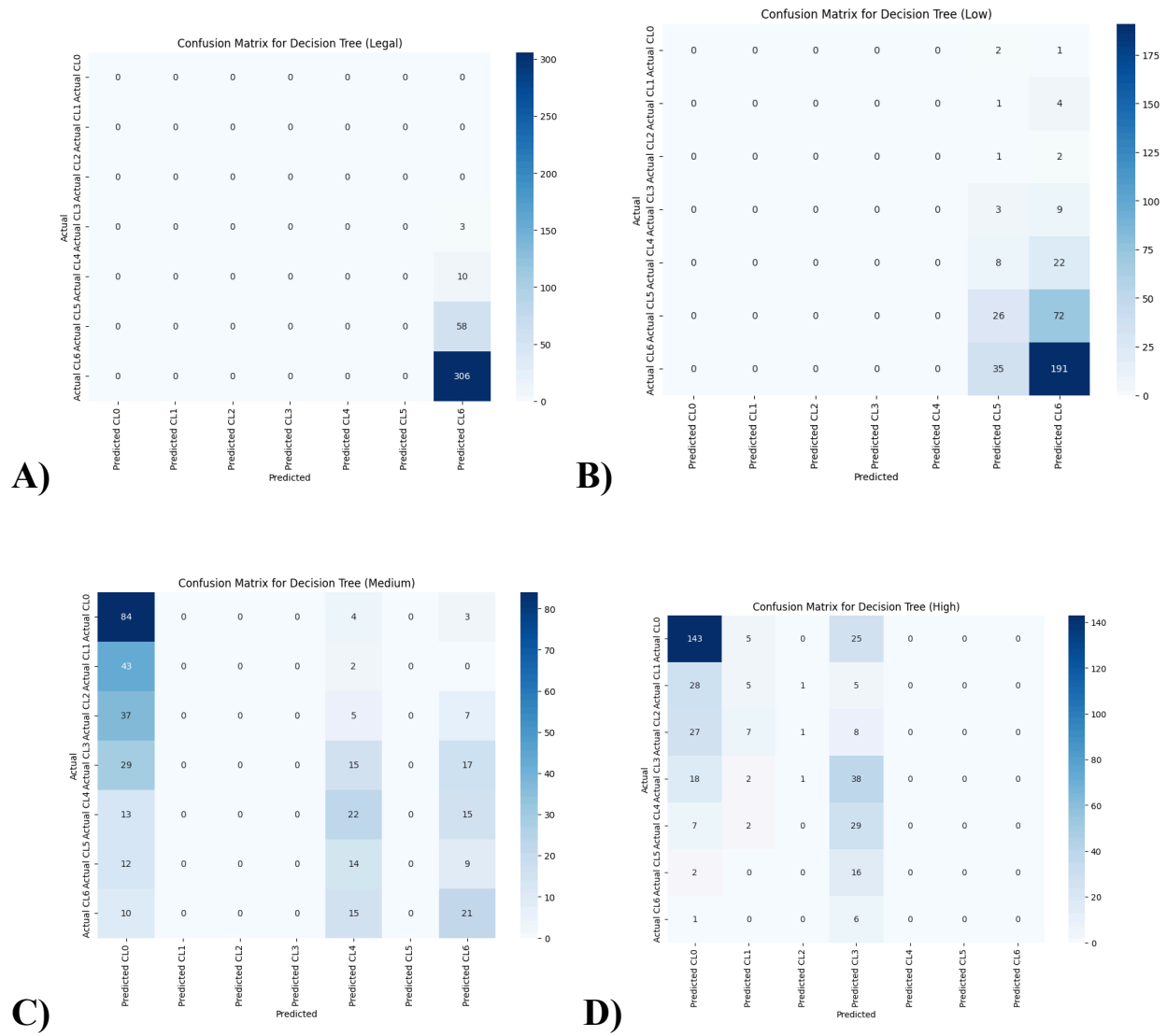


Figure 4: Heatmap showing predicted vs. actual drug usage levels for using Decision trees

- A) legal substances
- B) low addictiveness
- C) medium addictiveness
- D) high addictiveness

In figure 4 we see a heatmap for each category legal, low, medium and high for the decision tree model. Again, there was a clear dominance of CL6 in the legal decision tree with 306 CL6 instances being predicted accurately causing the model to correlate the legal class with CL6. Again there is a small amount of the lower classes such as CL0 and CL3. In the low decision tree, there is still a slight CL6 dominance however it is lower than the dominance seen within the legal decision tree. The model also still struggles with lower usage of classes such as CL0 and CL3 with very few correct predictions for these classes in the low category. In the medium category there was a dominance for CL6 and CL0, predictions were typically scattered across CL1 to CL5 and there was no clear diagonal pattern. In the high category there was a clear dominance to CL0 and very sparse amounts of higher classes from CL4-CL6.

The accuracy shown from decision trees show that the model is not fully accurate as most of the accuracy scores are again low excluding the target legal that has an accuracy of 0.812(3sf), this could be due to the class imbalance in legal again or due to legal having very specific personality traits allowing the tree to split it up easily. The model is again not very good at predicting the drugs due to mainly low accuracies, this could be due to the classes overlapping heavily in feature space, so the tree can't find clear thresholds. It could also be due to the tree missing key ideas as it can only split on one feature at a time, some real patterns are only shown through variables such as 'high neuroticism' compared to 'neuroticism' which would indicate a low accuracy as it was seen that most illegal drugs showed signs of high neuroticism however this would not be taken into account by the decision tree. I again cannot get a perfect prediction using a decision tree as I would require having to use more specific demographic features to get a true insight to an accurate prediction.

Seen within the legal decision tree the factor impacting the predictions most is the age_string as it is the root node of the decision tree so it was the most impacting feature in the initial split of the data for predicting legal data usage. This shows that people in different age categories tend to have different patterns of legal drug use with a clear distinction based on whether their value is above or below a certain threshold. Within the low decision tree the most major factors are the country_string and agreeableness, these are the first factors to be identified from a major split. Other factors such as impulsiveness and openness have small impacts in the low decision tree. The presence of agreeableness (ascore) high up in the tree for the false branch of the country string indicates its high predictive power. In the medium decision tree the factors impacting the predictions most is country string and sensation seeking. Similar to 'low' the medium decision tree first has major impact from the country string but instead of having a strong impact from agreeableness there is an influence from sensation seeking, as these drugs have a greater impact on cognitive focus and can help with mental health issues such as anxiety explaining why sensation seeking has a greater impact here. Finally at the high decision tree the most influencing factor is sensation seeking. As previously stated this is explained as drugs in this category have a greater impact on emotion and cognitive function and can be impactful for mental health issues.

KNN

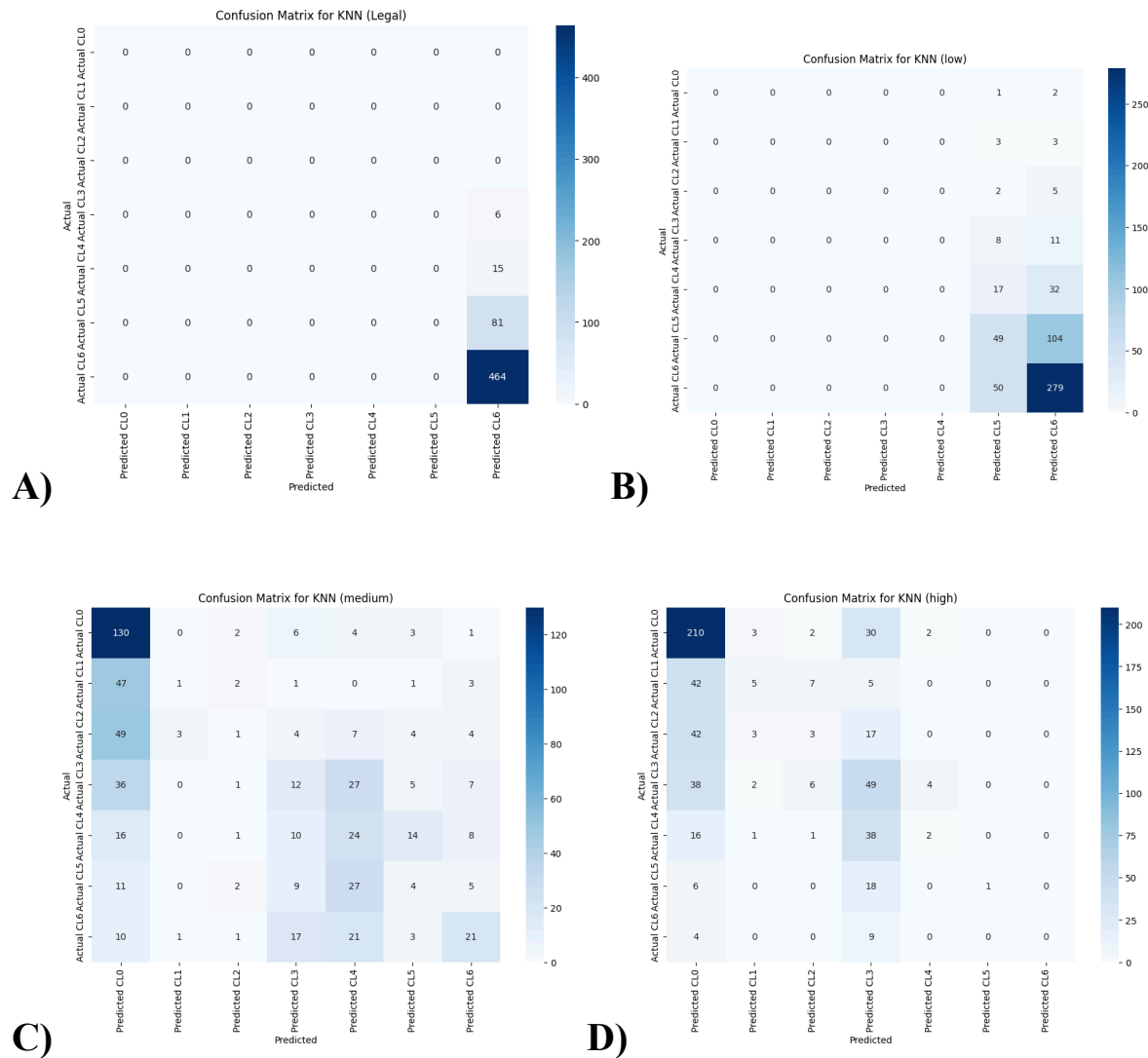


Figure 5: Heatmap showing predicted vs. actual drug usage levels for drugs using KNN

- A) legal substances
- B) low addictiveness
- C) medium addictiveness
- D) high addictiveness

In figure 5 we see heatmaps for each category of drug use presenting a visual for the confusion matrix produced. In the legal heatmap there was also clear dominance in CL6, however this caused some misclassification as there were 15 instances where the actual class was CL5 but the model predicted CL6

as there was such a dominance for CL6, we can also see that there is sparse with amounts of the lower classes making the model worse at predicting these classes. In the low heatmap the highest predictions, although not as close to legal, is CL6, however the sparseness between classes is smaller leading to fewer incorrect predictions. Although this didn't fully stop misclassification with 50 actual CL6 instances being predicted as CL5 and 32 actual CL4 instances being predicted as CL6. In the medium heatmap there is a dominance of both CL0 and CL6, besides this there is no clear pattern and misclassification has decreased with 130 CL0 has been correctly identified as CL0. In the high heatmap there is a dominance towards CL0, with a small amount of misclassification. Out of 42 actual CL1 instances, 3 are predicted as CL0, 5 as CL1, 7 as CL2 and 30 as CL3

The accuracy shown from KNN show that the model is not fully accurate as most of the accuracy scores are again low excluding the target legal that has an accuracy of 0.822, this could be due to the class imbalance in legal, as the k value has been optimised meaning there is little bias and less noise. The model is again not very good at predicting the drugs due to mainly low accuracies, this is can be due to my KNN model seeing the categorical values of CL0-CL6 as unordered, so instead of understanding that CL1 and CL2 are ever closer than CL1 and CL4 it sees them as 6 unrelated cases, decreasing accuracy. It could also be due to the severe class imbalance as some drugs have 80-90% of samples in CL0 and very little in CL3 and CL4, and as KNN is a class-majority classifier minority classes are overwhelmed and their accuracies collapse resulting in a low overall accuracy that is dependent on the majority class. I also cannot get a perfect prediction from this as my factors are limited demographically and would require more specifics into this area to get a more accurate testament of the drug usage per hardness.

In KNN, StandardScaler() is used causing each factor to have a mean of 0 and standard deviation of 1 causing each factor to contribute equally to the calculation of distance between data points. Therefore there is no specific factor impacting the predictiveness of the model majorly.

Comparison

Model	Drug Hardness	Accuracy
Logistic Regression	Legal	0.81 (2dp)
	Low	0.60 (2dp)
	Med	0.36 (2dp)
	High	0.48 (2dp)
Decision Tree	Legal	0.81
	Low	0.58
	Med	0.34
	High	0.50

KNN	Legal	0.82
	Low	0.58
	Med	0.34
	High	0.48

Table 1: Summary of model performance across drug hardness levels for logistic regression (max_iter=1000), decision tree, and KNN models.

As analysed between each category's heatmaps for each model, there was little change between the models. Between the models we see that the accuracies for each target are very similar, with legal being around 0.8, mild around 0.6, medium around 0.3 and high around 0.5. This could be due to the class imbalance of the data set meaning that the data being used is causing similar biases within the model. Most people don't take drugs during their lifetime, which results in the model being biased towards predicted CL0 (never used). Using hard drugs is even less likely than soft drugs, which exacerbates the data imbalance and model bias. Legal substances like coffee and chocolate are very likely to be used, resulting in CL6 (used in the last day) being overrepresented in the dataset and predictions, as these legal substances are typically consumed on a daily basis.

CONCLUSION

The study's conclusions demonstrate that although personality and demographic traits do provide valuable information on trends in illicit drug use, predictive models based on these factors—especially KNN—have trouble achieving high accuracy. The low accuracy across models, however, indicates that the intricacy of drug-use behavior cannot be well captured by these variables alone. The model's predictive value is limited since drug usage is impacted by a variety of social, environmental, and psychological factors that are not included in the dataset.

The model is not yet complete as a forecasting tool. Additional characteristics that are known to significantly impact drug-use vulnerability, such as peer pressure, socioeconomic status, school involvement, mental health diagnoses, family history of substance use, and trauma exposure, could increase accuracy. The model might also be strengthened with more specific demographic factors, such as household income, parental education, or neighborhood deprivation. Other modeling techniques, such as ordinal regression or ensemble methods, might more successfully manage class imbalance and better represent the structure of the CL0–CL4 labels. All things considered, the study shows that while personality and demographics are important, more accurate predictions of illegal drug usage require richer data and more advanced modeling. Alongside this incorporating the use of macro F1-scoring could have improved the class imbalance prominent within the models as it would have shown a value from 0-1 of how well the precision and recall of each class which would give the equal importance of each class despite their observation. Also within the logistic regression and decision trees adding in class weighting which would give greater weight to underrepresented classes and ensures that the model doesn't automatically assume the majority class when predicting.

ACKNOWLEDGEMENTS

This paper was written with the help of the inspirit team specifically Khanh Ngyuen who helped me to analyze and interpret the data, create models and write this paper

REFERENCES

1. NHS England. "Smoking, Drinking and Drug Use among Young People in England, 2023." *Smoking, Drinking and Drug Use among Young People in England, 2023*, 2023, <https://digital.nhs.uk/data-and-information/publications/statistical/smoking-drinking-and-drug-use-among-young-people-in-england/2023>. Accessed 18 01 2026.
2. office for health, improvement and disparities. "Children's substance misuse treatment statistics 2024 to 2025: report." *Children's substance misuse treatment statistics 2024 to 2025: report*, 2025, <https://www.gov.uk/government/statistics/substance-misuse-treatment-for-children-2024-to-2025/childrens-substance-misuse-treatment-statistics-2024-to-2025-report>. Accessed 18 01 2026.
3. Kotov, Roman. "Linking "Big" Personality Traits to Anxiety, Depressive, and Substance Use Disorders: A Meta-Analysis." *Linking "Big" Personality Traits to Anxiety, Depressive, and Substance Use Disorders: A Meta-Analysis*, 2010, <https://static1.squarespace.com/static/5692fe4a4bf1182ee6045530/t/5820d216e58c62bd0992476a/1478545945761/KotovGamezSchmidtWatson%282010%29.pdf.pdf>. Accessed 22 11 2025.
4. Löckenhoff, Terracciano A. "Five-Factor Model personality profiles of drug users." *Five-Factor Model personality profiles of drug users*, 2008, <https://link.springer.com/article/10.1186/1471-244x-8-22>. Accessed 22 11 2025.

5. Wang, T., Boden, J. M., Biswas, S., & Choudhary, P. K. (2025). Absolute Risk Prediction for Cannabis Use Disorder in Adolescence and Early Adulthood Using Bayesian Machine Learning. *Drug and Alcohol Review*, 44(6), 1680–1690. <https://doi.org/10.1111/dar.14098>
6. Nutt, D. J., King, L. A., & Phillips, L. D. (2010). Drug harms in the UK: a multicriteria decision analysis. *The Lancet*, 376(9752), 1558–1565. [https://doi.org/10.1016/s0140-6736\(10\)61462-6](https://doi.org/10.1016/s0140-6736(10)61462-6)
7. Home Office. (2022, February 3). Seizures of drugs in England and Wales: user guide. Retrieved August 13, 2026, from GOV.UK website:
https://www.gov.uk/government/publications/seizures-of-drugs-in-england-and-wales-user-guide/seizures-of-drugs-in-england-and-wales-user-guide--2?preview=7289506&utm_source=chatgpt.com
8. Villines, Z. (2022, February 23). What to know about peer pressure and drugs. Retrieved August 13, 2026, from Medicalnewstoday.com website:
<https://www.medicalnewstoday.com/articles/peer-pressure-drugs>