Predicting Illegal Substance Use

Cynthia Zhang

December 9, 2022

GitHub repository: https://github.com/CynthiaCZ/DATA1030_project.git

1 Introduction

This study aims to produce a classification model that can predict the number of illegal substances used within a year from demographics, personality traits, as well as the frequency of legal substance use.

The data is accessed via https://www.kaggle.com/datasets/obeykhadija/drug-consumptions-uci. It was collected in 2015 through an anonymous online survey [2]. The Revised NEO five-factor inventory (NEO-FFI-R) was employed to assess traits of neuroticism (N), extraversion (E), openness to experience (O), agreeableness (A), and conscientiousness (C). In addition to that, two additional scales were included to gauge impulsiveness (impulsive) and sensation-seeking (SS). Respondents were asked to provide demographic information including age, gender, level of education, country, and ethnicity, and then to complete the personality evaluation questionnaire. The test subjects were also questioned about their drug consumption history within a list of 19 legal and illegal substances. For each one of the said substances, there are 7 ordinal categories indicating the frequency of substance use. In total, there are 1885 unique study subjects and 32 attributes with no missing values.

Most existing studies using this dataset focus on only one or a few drug categories at a time. Usually, the goal is to find predictive trends between personality traits and those specific substances. For example, a study by Qiao et al. concludes that Light Gradient Boosting Machine (LightGBM) model performs well in predicting potential users and usage time of two illegal drugs, Amyl and Meth [3]. Another study by Adinugroho et al. shows that Extreme Learning Machine (ELM), a type of artificial neural network, has varying accuracy, ranging from 37% to 86% in the prediction of the usage duration of each specific substance, legal or illegal [1].

However, this study proposes a more generalized model, focusing on the number of illegal drugs used within a timeframe, making the model potentially useful for psychologists to identify patients prone to drug abuse of any kind. Due to the sensitive nature of illegal drug use, many patients are less willing to disclose such information. Legal substances such as

nicotine and alcohol, on the contrary, are less stigmatized and more openly discussed. That is why in this study, five of the legal substances are included as features and are used to predict illegal substance abuse.

2 Exploratory Data Analysis (EDA)

The target variable, i.e. the number of illegal substances used within a year is right-skewed (Fig. 1). It ranges from 0 to 7, with a mean of 0.79 and a median of 0.

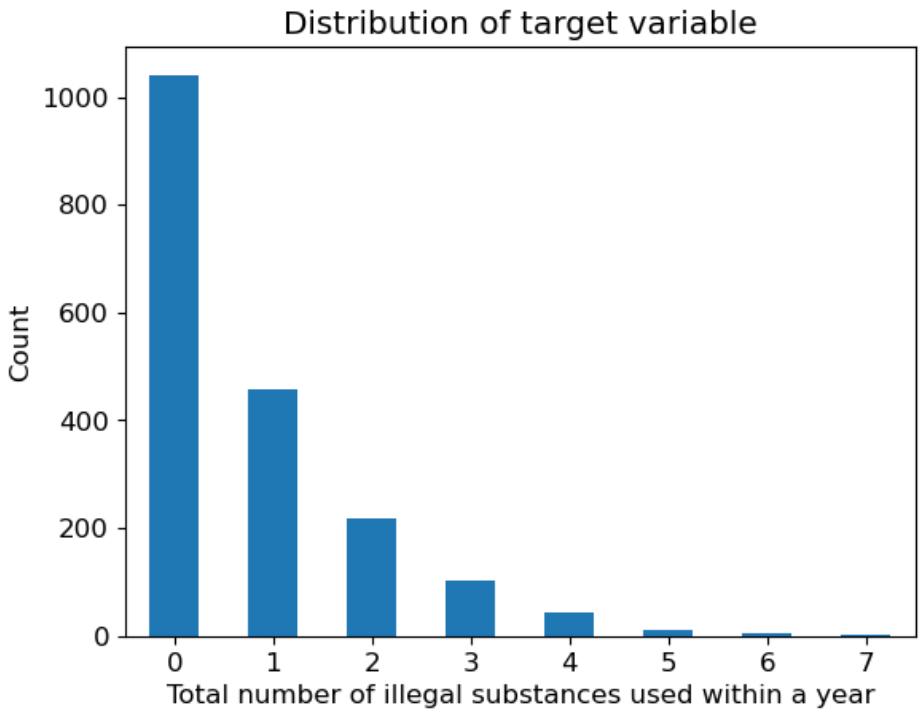

Figure 1: Bar plot showing the target variable is right-skewed

The majority (more than 50%) of the respondents are from the UK, and more than 25% are from the US (Fig. 2). The remaining study subjects are from Canada, Australia, the Republic of Ireland, New Zealand, as well as a diversity of other countries. Due to this imbalance, the country feature was excluded from the original study [2], and this study did the same correspondingly.

According to the linear correlation indicated by the f-statistic, one feature that is strongly correlated to the number of illegal drugs used within a year is the weekly use of nicotine. The left column of the violin plot (Fig. 3) has a wider base, indicating that most non-smokers use zero illegal drugs within the year, whereas the right column shows that smokers tend to use one or more illegal drugs.

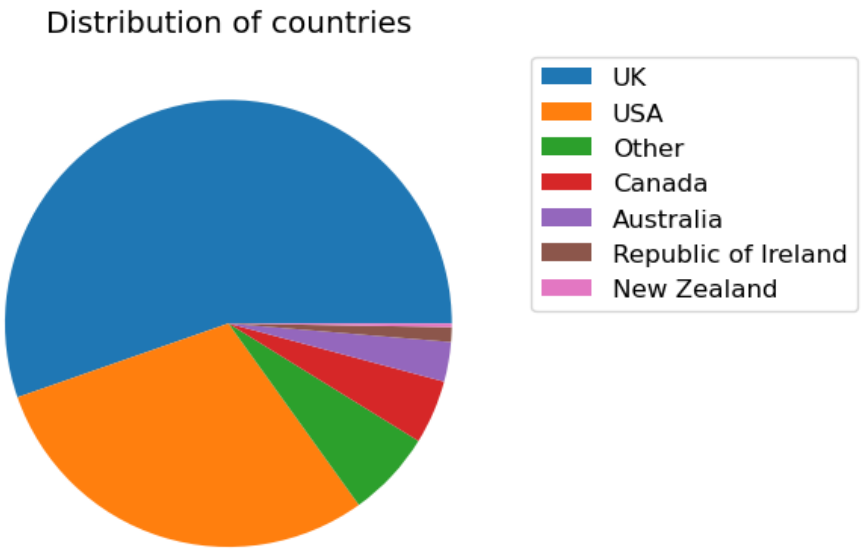

Figure 2: Pie plot showing most of the respondents are from either the UK or the US

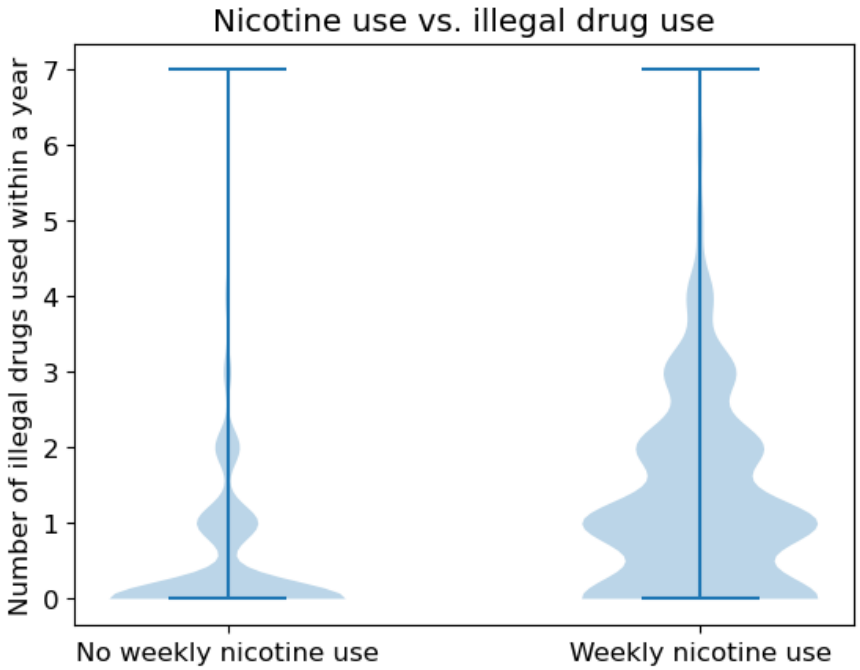

Figure 3: Violin plot showing smokers are more likely to use illegal drugs than non-smokers

Another feature with a strong correlation to drug use is age. As shown in the normalized stacked bar plot (Fig. 4), most people in the youngest age group (18-24) have tried one or more illegal drugs within the year, while people in older age groups tend to use zero or very few illegal drugs.

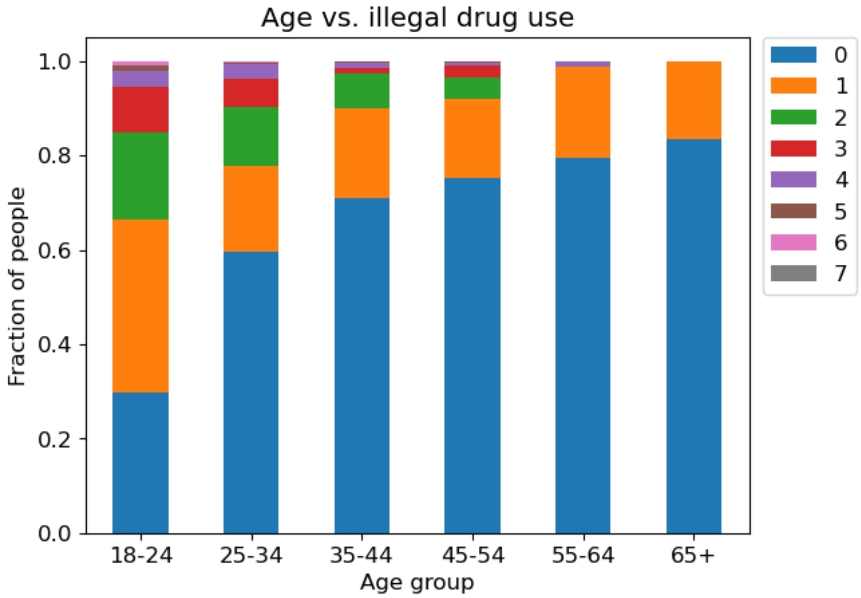

Figure 4: Stacked bar plot showing younger people are more likely to use illegal drugs than older people

Among the seven personality traits, the one with the highest f-score is Sensation Seeking (SS). This is illustrated by the heatmap (Fig. 5). People with negative SS scores tend to fewer illegal drugs. As the SS score increases, the number of drugs used within the year increases accordingly, forming the triangular shape at the lower left corner of the heatmap.

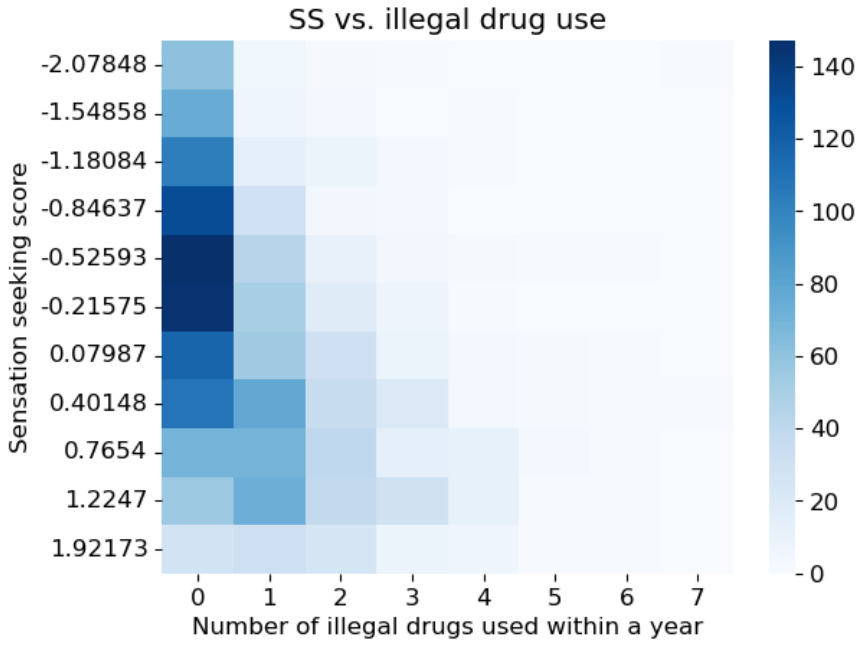

Figure 5: Heatmap showing Sensation Seeking score is positively correlated to drug use

3 Methods

3.1 Featuring Engineering

The original dataset is manipulated using sklearn in the following ways to suit the purpose of this study. First, the ID column and the Semeron column are removed. Semeron is a fictitious drug introduced to identify over claimers in the survey [1]. Then, as mentioned above, the country column is removed to reduce the bias in sampling.

In the original dataset, each substance has seven frequency categories: never used, used over a decade ago, or in the last decade, year, month, week, or day. This study uses a week-based classification for the five legal substances, i.e. alcohol, caffeine, chocolate, legal high, and nicotine. Respondents who use these legal substances on a weekly basis or more frequently receive a score of 1, and otherwise they receive a score of 0. Those five features were then appended to the feature matrix, totaling 17 features. For the thirteen illegal substances, on the other hand, the classification is year-based. Using the substance in the last year or more frequently correspond to a score of 1, and using it less frequently than that corresponds to a score of 0. The illegal substances columns are then summed, resulting in the target variable: the number of illegal substances used within the year.

3.2 Splitting and Data Preprocessing

Since the data was collected anonymously from unique study subjects, for this study, the data is assumed to be IID. And since it also does not have an apparent group structure or time stamp, the standard train_test_split and K-fold validation from sklearn is used to split the dataset into train, validation, and test sets with a 60-20-20 proportion for each validation fold. However, the size of the dataset is relatively small and there are very few points in the higher label classes, so it was not possible to apply stratification to the splitting process.

The ordinal features, age group and education, are encoded using OrdinalEncoder. The categorical features, gender and ethnicity, are encoded using OneHotEncoder. Further, for the numeric personality scores, StandardScaler is employed.

3.3 Training Pipeline

Five different classification models are trained on the dataset, including one linear model, logistic regression with l2 penalty, and four non-linear models, K neighbors classifier, random forest classifier, support vector classifier, and XGBoost classifier. The hyperparameters tuned for each of those models are listed in Table 1 below. GridSearchCV is used to loop through all combinations of the hyperparameters for each model. To measure the uncertainties due to splitting and non-deterministic ML methods, for each model, model fitting and testing are replicated for ten different random states.

F1 score with macro averaging is the chosen evaluation metric for the models. Because the classes are imbalanced, the F1 score takes the true negative values into consideration,

Table 1: models and hyperparameters tuned

Model	Hyperparameters
Logistic regression with l2 penalty	C: [10, 2, 1.5, 1.25]
K neighbors classifier	n_neighbors: [3, 5, 10, 50]
	weights: ['uniform', 'distance']
Random forest classifier	max_depth: [5, 10, 15, 20]
	max_features: [0.1, 0.3, 0.5, 0.7]
Support vector classifier	gamma: [0.0001, 0.001, 0.01]
	C: [10, 50, 100, 150]
XGBoost classifier	learning_rate: [0.01, 0.03, 0.05]
	max_depth: [5, 10, 15]

preventing the inflation of the true positive score due to falsely predicting the majority class labels to points from minority classes. Macro averaging calculates the unweighted mean F1 score across different classes. Thus, each class contributes equally to the evaluation, making this approach better at capturing the classifier's performance on smaller classes.

4 Results

4.1 Model Comparison

As shown in Figure 6, all five models perform similarly well in terms of their average test F1 score across different random states. All five F1 scores are at least four standard deviations above the baseline F1 score. The baseline is calculated by predicting a class label of 0 (the majority class) to all of the data points without considering any feature.

The best-performing model is the XGBoost classifier. Its average F1 score across different random states is 0.178, which is 5.49 standard deviations above the baseline. Out of the ten different random states, there are some variations among the best-performing hyperparameter combinations found through K-fold validation. However, the most commonly recurring combination is learning_rate = 0.03, max_depth = 10, and n_estimators = 100.

4.2 Feature Importance

4.2.1 Global Feature Importance

To illustrate the differences and similarities in global feature importance of different models, sklearn's permutation_importance is applied to the first and second best-performing models, i.e. XGBoost classifier and logistic regression. For the XGBoost model, the top five most important features are sensation-seeking (SS), age, openness to experience (Oscore), education, and conscientiousness (Cscore) (Fig. 7). On the other hand, the logistic regression

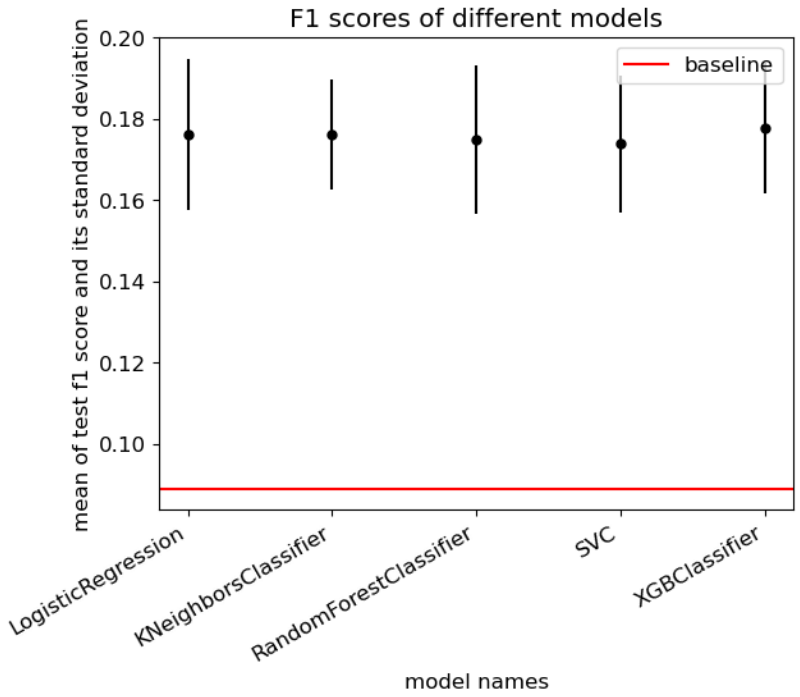

Figure 6: F1 scores and the associated standard deviation for all models

model included openness to experience (Oscore), age, education, conscientiousness (Cscore), and impulsiveness as the five most globally important features (Fig. 8). From this it is safe to conclude that there is significant overlap between what features these two models consider to be important globally. And overall for these models, the psychological traits, age, and education have more predictive power over other features.

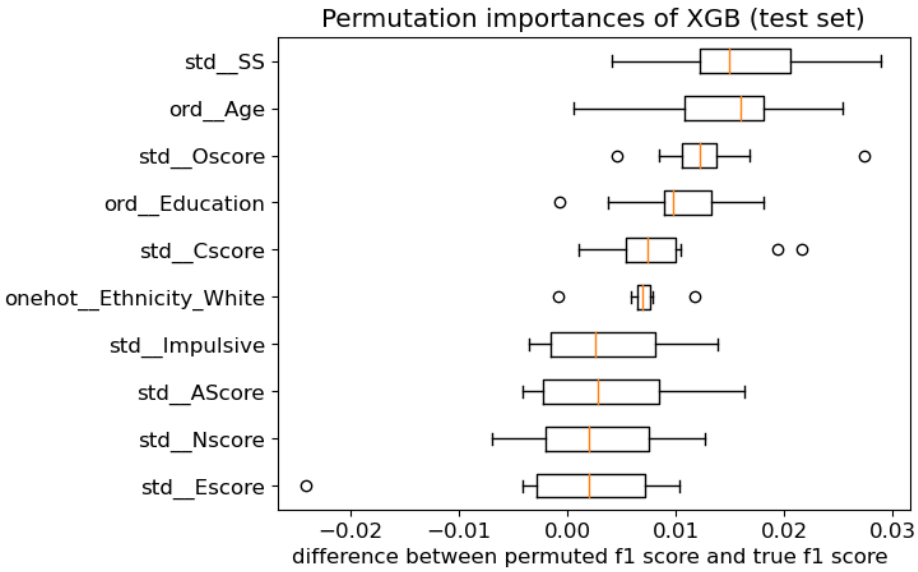

Figure 7: Important features of XGBoost classifier from permutation importance

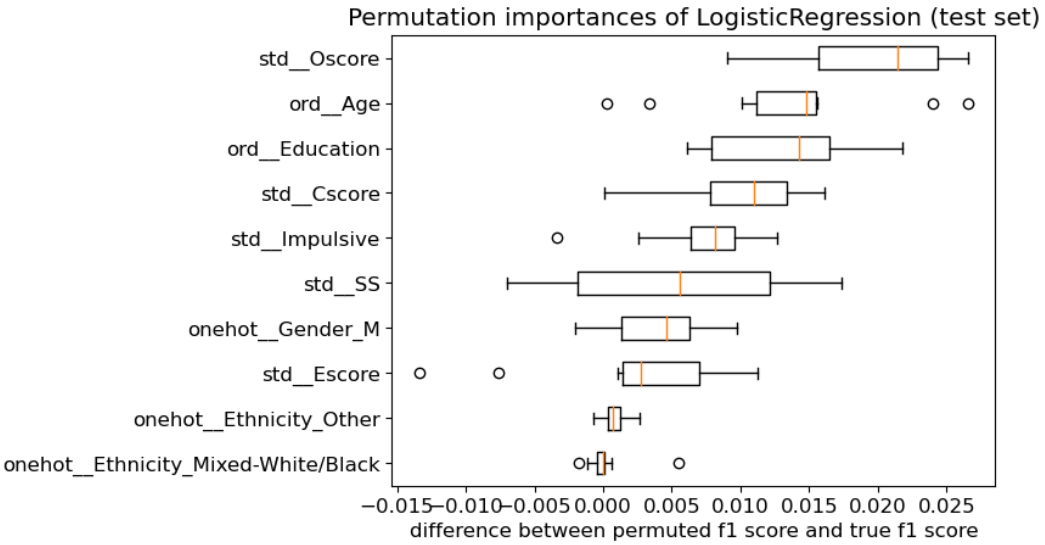

Figure 8: Important features of logistic regression from permutation importance

To further illustrate the global feature importance of the model with the highest predictive power, the shap.TreeExplainer package is used to calculate the SHAP values of the XGBoost classifier, and a summary plot is created (Fig. 9). With this measure of global importance, the top five features of interest are sensation-seeking (SS), age, openness to experience (Oscore), conscientiousness (Cscore), and agreeableness (A). These coincide with many of the important features measured with feature perturbation, further confirming their contribution to the model.

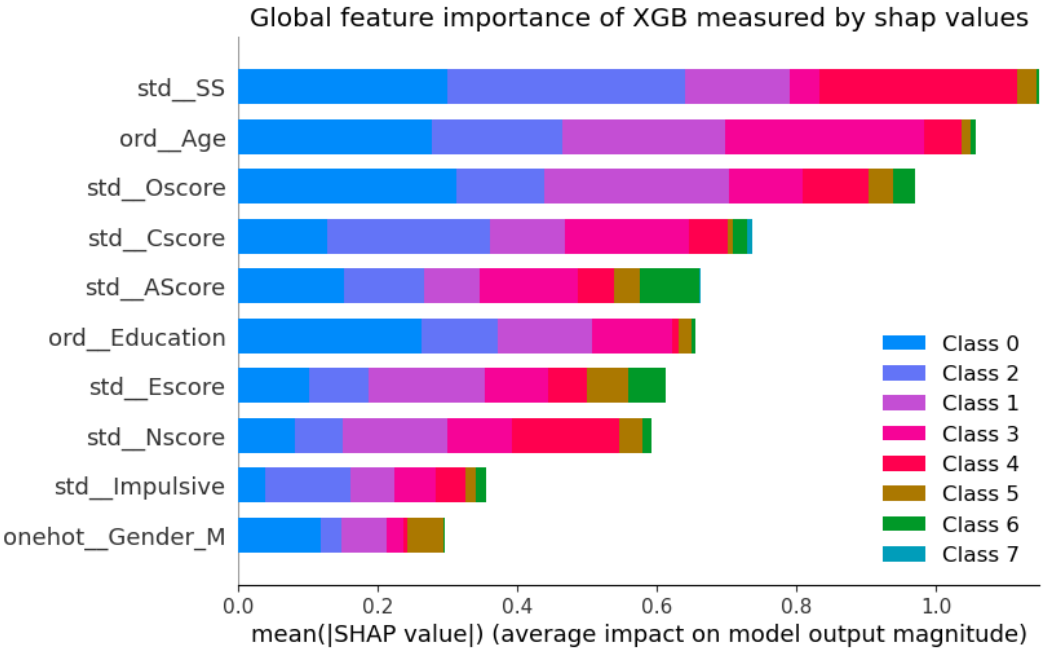

Figure 9: Important features of XGBoost classifier from SHAP

4.2.2 Local Feature Importance

The above-mentioned SHAP values are also used to interpret the XGBoost model's prediction on single data points. Two examples of local feature importances plotted using SHAP force_plot are shown in Figures 10 and 11. For both of these plots, the class of interest is set to 0. The first data point has a true class label of 3 and a SHAP value smaller than the base value for class 0. Attributes that are influencing this prediction are younger age, higher sensation-seeking (SS) score, less education, and higher openness to experience (Oscore). Whereas the second data point has a true class label of 0 and a SHAP value greater than the base value for class 0. Factors positively contributing to its higher-than-average SHAP value include lower sensation-seeking (SS) score, older age, greater amount of education, and higher conscientiousness (Cscore) score.

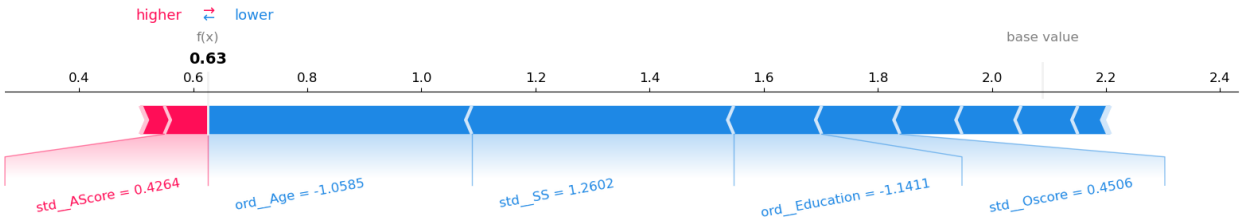

Figure 10: Important features of XGBoost classifier's prediction on one data point

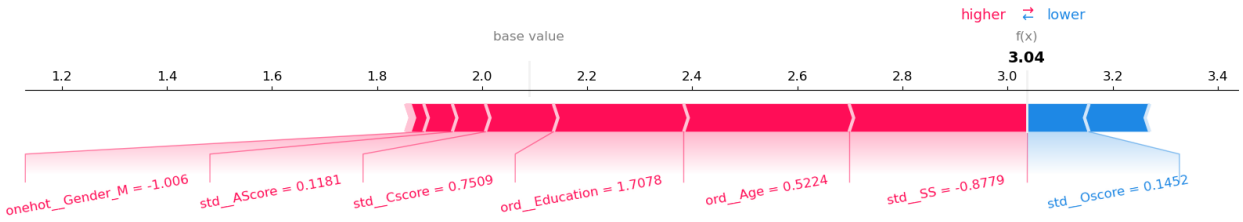

Figure 11: Important features of XGBoost classifier's prediction on another data point

4.2.3 Interpretations

These globally and locally important features agree for the most part, and remain consistent across different measurement strategies, e.g. feature perturbation and SHAP. The importance of different psychological traits and demographics as well as the direction in which each attribute influences the prediction aligns with the expectation of this study. For example, as shown in Figure 5 of the EDA section, the Sensation Seeking (SS) score is positively correlated to drug use. This correlation reoccurred as the most important feature for XGBoost in both permutation importance and global SHAP, as well as in the local SHAP force plots. However, during EDA, the use of legal substances, especially tobacco, was strongly linearly correlated with the target variable, but these features are unexpectedly among the least

important features in the final models. Broadly speaking, in the context of the problem, the model interpretations are also in line with the general stigma around people who tend to become drug abusers.

5 Outlook

Given the time limit of this study, there remains great scope for improvement. Preferably, different models could be applied to the dataset, and for the already trained models, further hyperparameter fine-tuning could be done.

The greatest limitation of the models is that, even though the F1 score with macro averaging is used to minimize false negative rates on smaller label classes, the models do not perform ideally on those classes as shown in the confusion matrix of the XGBoost model (Fig. 12). This is probably due to the severely unbalanced distribution of the dataset. An alternative, binary classification model that distinguishes non-drug users (class 0) and drug users (class 1 through 7) might be more suitable.

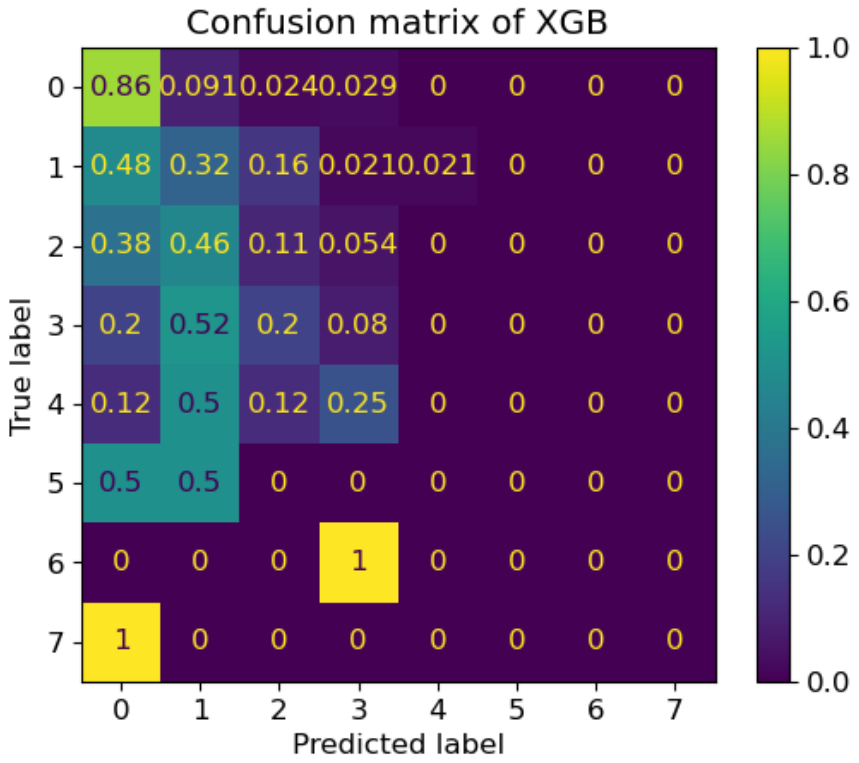

Figure 12: Confusion matrix showing the XGB model have lower predictive power for higher label classes

References

[1] S. Adinugroho, Y. A. Sari, and N. Hidayat, *Drug usage duration classification using extreme learning machine based on personality features*, 2019 International Conference on Sustainable Information Engineering and Technology (SIET), (2019).

[2] 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*, Data Science, (2017), p. 231–242.

[3] Z. Qiao, T. Chai, Q. Zhang, X. Zhou, and Z. Chu, *Predicting potential drug abusers using machine learning techniques*, 2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), (2019).