Assignment 2

Luis Dale Gascon
Computer Science
Towson University
lgascon1@students.towson.edu

I. Introduction

A. Dataset

The dataset is a collection of responses from 1885 people with 12 features. As the dataset has already been preprocessed and quantified, we needed more information to understand what the preprocessed values mean, and fortunately a research paper by [1] provided a detailed description of the process.

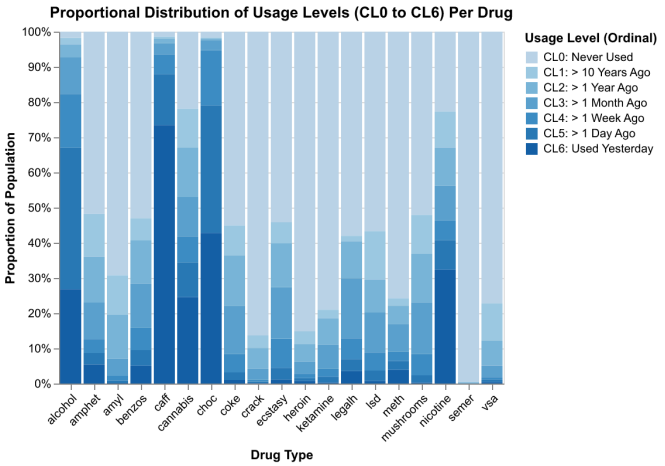
- nscore: Neuroticism
- escore: Extraversion
- oscore: Openness to experience

a) *Different classes by frequency of use:*

The dataset presents 19 different drugs with the following classes as the target features:

Class Label	Description
CL0	Never Used
CL1	Used over a Decade
CL2	Used in Last Decade
CL3	Used in Last Year
CL4	Used in Last Month
CL5	Used in Last Week
CL6	Used in Last Day

b) *Exploring the Dataset:*

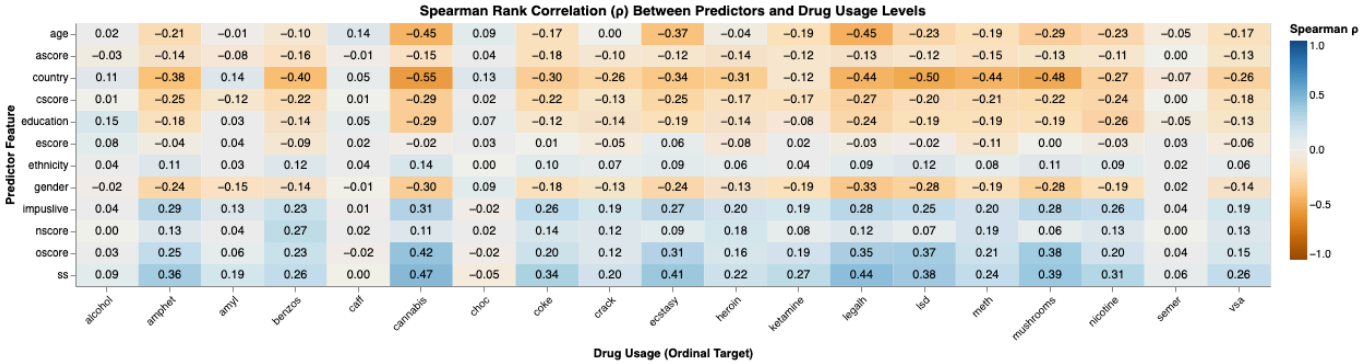


Looking at the stacked barchart, the majority of the recently used population falls under mainstream drugs such as alcohol, caffeine, chocolate, and nicotine.

B. Objectives

Alcohol consumption patterns vary significantly across individuals and understanding the factors that influence drinking frequency has important implications for public health, addiction prevention, and personalized intervention strategies. While previous research has explored relationships between personality traits and substance use, there remains a need for predictive models that can accurately classify individuals into specific alcohol usage frequency categories.

This study addresses the challenge of predicting alcohol consumption frequency based on an individual's personality and demographic characteristics. We aim to develop classification models that can predict alcohol usage patterns across seven ordinal categories: never used (CL0), used over a decade ago (CL1), used in the last decade (CL2), used in the



last year (CL3), used in the last month (CL4), used in the last week (CL5), and used in the last day (CL6).

Successfully solving this problem could enable early identification of individuals at risk for problematic drinking behaviors and inform targeted prevention strategies.

II. Methodology

A. Data Partition

To paritition the test and training datasets, we will be using the holdout method via train_test_split from the sklearn library with the following parameters:

- test_size = 0.70
- shuffle = True
- random_state = 42

For our feature selection, we plan to perform the 2 sample t-test on the feature that is highly correlated with my classification target. Based on the heatmap of the Spearman Rank Correlation, we will be using the education feature as that has the highest correlation with our target column of alcohol.

We will also perform the chi-squared test on the target column of alcohol.

Using ttest_ind from scipy, the function returned a p-value of 0.73, which fails to reject the null hypothesis that the mean of the partition aren't statistically similar

For the categorical target values, we used the chi2 function from scipy and received a p-value of 0.88, which fails to reject the null hypothesis that the mean of the partition aren't statistically similar.

B. Preprocessing

The dataset is already preprocessed when it was provided as stated in the dataset section. The following preprocessing steps are as follows:

a) *Ordinal Features:*

When the data was collected for the personality traits, the NEO-FFI-R questionnaire was used. The authors

b) *Nominal Features:*

The author implemented nonlinear categorical principal component analysis (CatPCA), which starts with a factor analysis to select the most informative components to keep.

C. Classification Algorithms

The classification algorithms that we will be using are the following:

- Logistic Regression
- Random Forest
- Gradient Boost

Logistic Regression, the ordinal type for this experiment, uses a sigmoid function to convert feature inputs into a probability between 0 and 1.

Random Forest creates multiple decision trees, each one different from the rests by using random features.

Gradient Boost tries to minimize the loss function using gradient descent. This iteration stops

By default, sklearn's GradientBoostingClassifier uses multinomial deviance as its loss function since we're training a multiclass classifier.

D. Experimental Setup

E. Accuracy Metrics

III. Results

IV. Conclusion

Logistic regression has age having the highest coefficient and thus the predictor that affects

Both Gradient Boost and Random Forest classification models has the features relating to personality as the important features that has made the most impact to the model's prediction

It's a surprise that Logistic Regression scored the highest in both the accuracy metrics.

References

[1] 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." [Online]. Available: https://arxiv.org/abs/1506.06297