Assignment 2

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

I. Introduction

A. Objectives

Alcohol consumption patterns vary significantly across individuals and understanding the factors that influence drinking frequency can be used for addiction prevention, and personalized intervention strategies. This study looks at 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 drinking behaviors and create targeted prevention strategies.

B. 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 features.

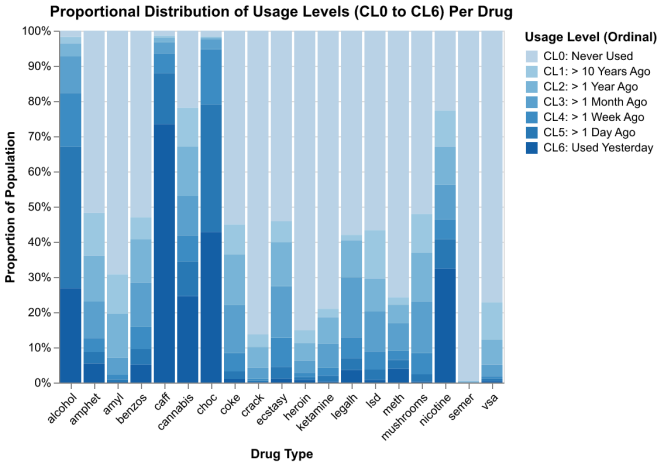
- nscore: Neuroticism
  - Tendency to experience negative emotions such as anxiety and depression
- escore: Extraversion
  - Outgoing, cheerful and in search of stimulation characteristics
- oscore: Openness to experience
  - General appreciation for unusual ideas and creativity
- ascore: Agreeableness
  - Cooperative, polite, kind and friendly
- cscore: Conscientiousness
  - Tendency to be organized and dependable
- imp: Impulsivity
  - Tendency to act with little to no forethought or consideration of the consequences
- ss: Sensation seeking
  - Search for new experiences and feelings that are intense or exciting

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. This experiment will mainly focus on alcohol.

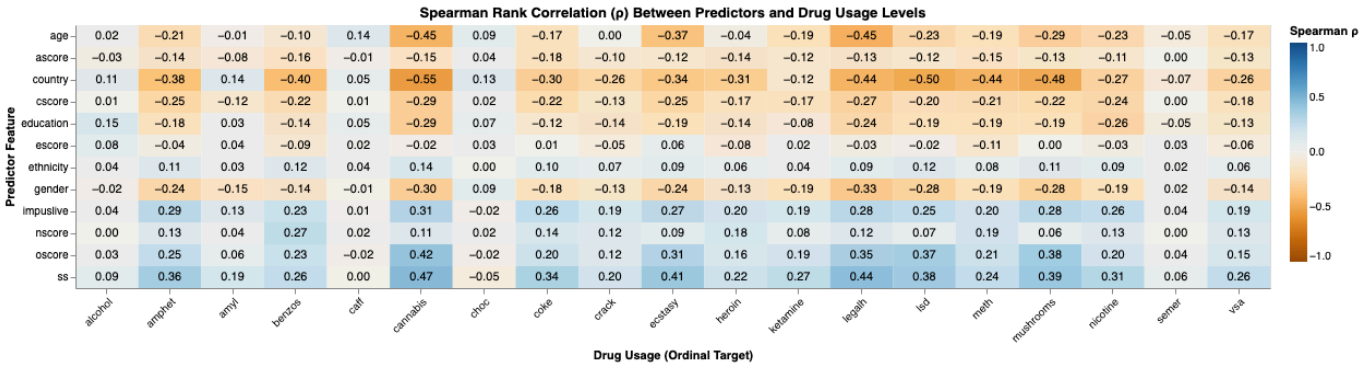
II. Methodology

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

As pointed out in the dataset section, the data was already preprocessed when we retrieved it. Although the authors employed a complex and intensive preprocessing methods that we had a hard time understanding, given the time constraint, the resulting values are standardized and suitable for training classification models since it was the author's goal to experiment with various classification methods.

C. Classification Algorithms

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

- Logistic Regression
  - Ordinal type specifically for this experiment, uses a sigmoid function to convert feature inputs into a probability between 0 and 1.
- Random Forest
  - Random Forest creates multiple decision trees, each one different from the rests by using random features.
  - Each random tree makes a prediction and a vote is made based on which prediction that most trees agree on.
- Gradient Boost
  - An ensemble technique that builds models iteratively, where a new model or tree tries to correct the errors of the previous ones. At each iteration, the algorithm fits a new model to the errors of the combined previous models, minimizing the loss function using gradient descent.
  - By default, sklearn's GradientBoostingClassifier uses multinomial deviance as its loss function since we're training a multiclass classifier.

D. Accuracy Metrics

To evaluate the performance of each classification model, we use sklearn's confusion_matrix and f1_score functions. Accuracy is calculated from the confusion matrix by taking the sum of the diagonal elements using numpy's trace function, which counts the correctly classified instances and divides it by the total number of samples, obtained with numpy's sum function.

III. Results

Algorithm	F_1
Logistic Regression	0.310
Random Forest	0.309
Gradient Boost	0.307

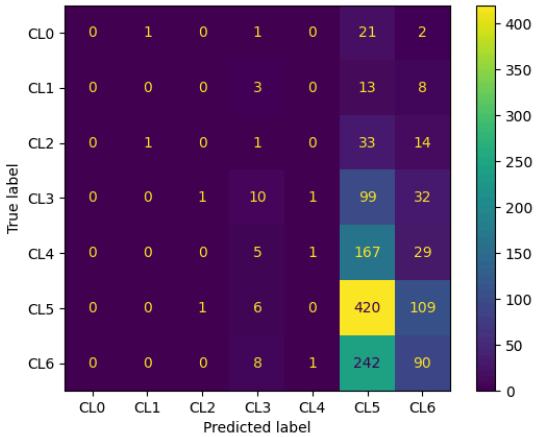


Fig. 2. Confusion matrix for Logistic Regression model. Accuracy of 39%

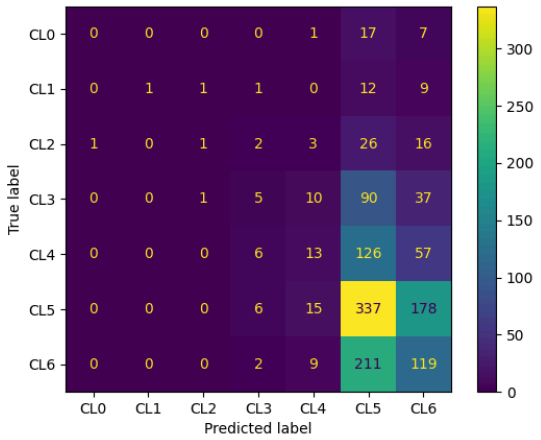


Fig. 3. Confusion matrix for Random Forest model. Accuracy of 36%

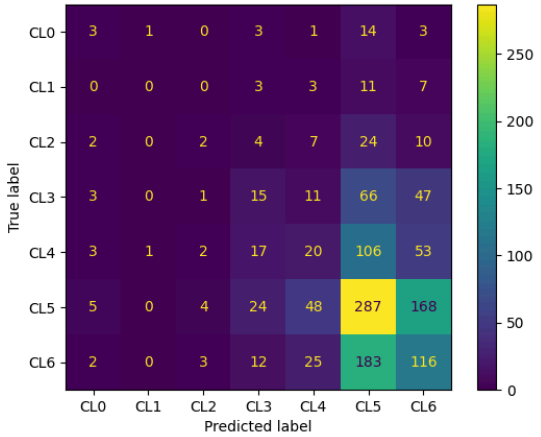
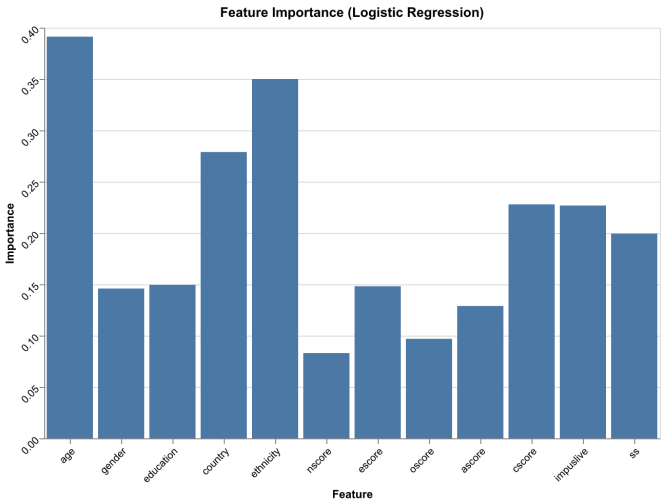
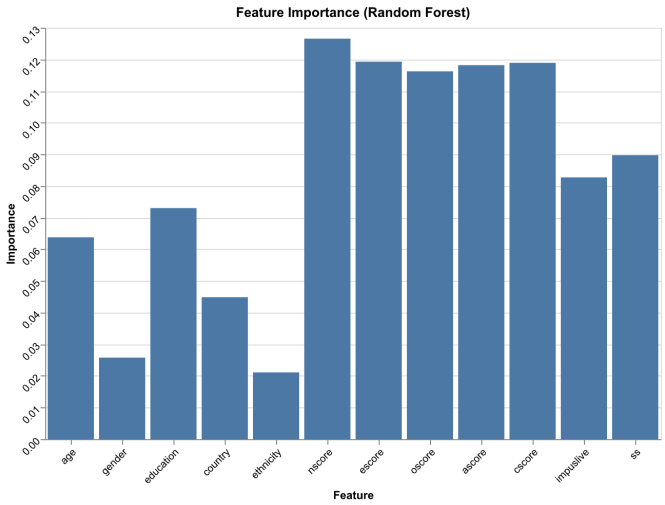
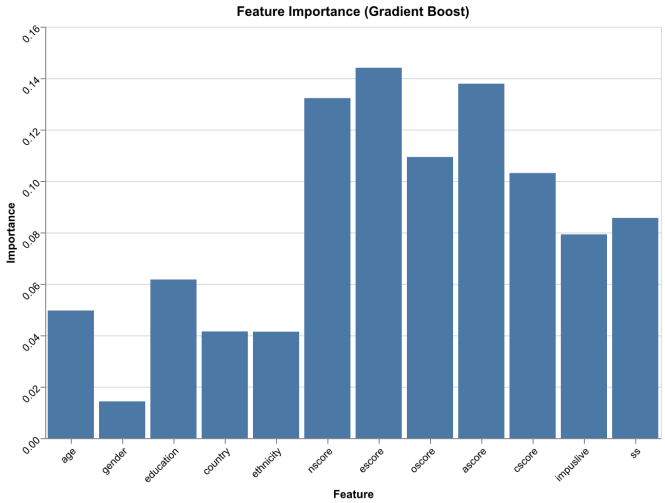


Fig. 4. Confusion matrix for Gradient Boost model. Accuracy of 33%

We wanted to see which features affected the prediction the most for each classification algorithm.







In the logistic regression model, age stands out as the most influential predictor of alcohol consumption frequency, as it has the highest coefficient among all features. Personality traits, by contrast, do not exhibit strong predictive power in this model. However, in both the Gradient Boost and Random Forest classification models, personality related features seem to be the most important factors.

IV. Conclusion

This study explored the use of logistic regression, random forest and gradient boosting classification algorithms to predict the frequency of alcohol consumption based on personality traits and demographic features. Among the models evaluated, Logistic Regression achieved the highest accuracy and F_1 score, although not by much. Interestingly, age emerged as the most influential predictor in the logistic regression model, while personality-related features played a more significant role in the tree-based models.

The results highlight the complexity of modeling ordinal outcomes in behavioral data and suggest that different algorithms may capture distinct aspects of the underlying relationships. While the overall predictive performance is fairly low, these findings provide valuable insights into the factors associated with alcohol use frequency and demon-

strate the potential of machine learning approaches in this domain. Future work could explore the creation of a multi-output classifier to predict ordinal classes across the rest of the drugs to get a better understanding of what features of a person causes them to partake in drug use.

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