DS1030 Midterm report

“Predicting the cannabis usage”

1. **Introduction**

Due to the appearance of opioids and the dark side of them affecting our society nowadays, people are more and more getting used to the drugs. To see which feature makes people more addictive to it and predict the future usage, this dataset, “The drug consumption” has been chosen. The prediction of the drug usage is important since this could be one of the research projects, which foresee the features that prevent people from getting addicted.

The dataset is from University of California Irvine data archive, which contains 1,884 individual participants survey data, with total of 23 features.[1] The dataset has been previously used to check the new algorithms such as, Non-Rigid Structure-From-Motion(NRSfM), Extreme Learning Machine(ELM), and Extreme Gradient Boosting(XGBoost), of which predicts the vulnerability of the drug addiction, potential abusers, and usage classification.[2,3,4] The accuracies of the previous research were average of 80%. Within the features, there exists 5 personality features, which are obtained by Revised NEO Personality Inventory (NEO-FFI-R).[5] 5 personalities which represents the NEO-FFI-R is used in this dataset, and they are shown in Table 1. For this research, the target variable is the Cannabis, which is one the most widely used drug throughout the world.



Table . Big 5 personality traits of NEO-FFI-R

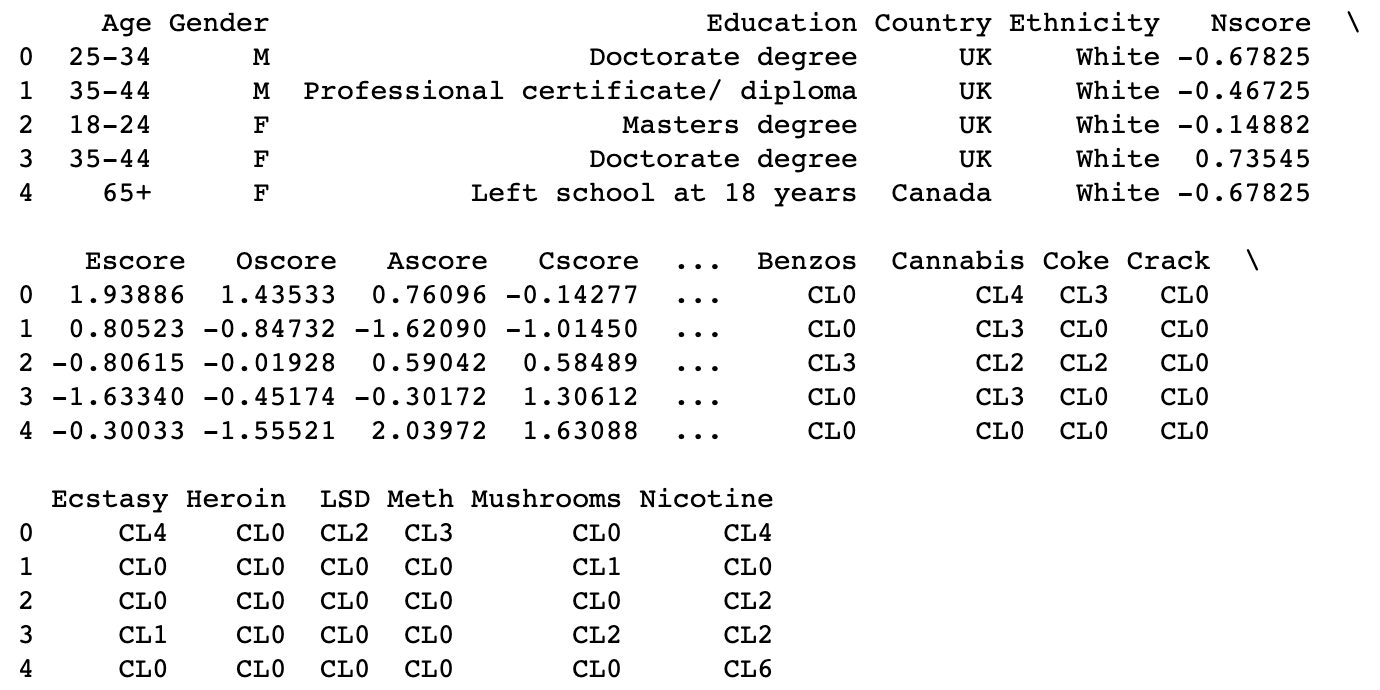


Figure . Printed head of the dataset

To verify the problem, the types of each class has been printed out as shown in Fig.2. All the data are either floats or object, so this will be a classification problem. We can also notice that in the drug section of the data, the values are written in string, from CL0 to CL6. These represents the usage frequency of the drugs, where CL0 = Never used, CL1 = Used 10 years ago, CL2 = Used in 10 years, CL3 = Used last year, CL4 = Used last month, CL5 = Used last week, CL6 = Used yesterday.

Table

Description automatically generated with medium confidence

Figure . Data type of each column

1. **Exploratory Data Analysis (EDA)**

The EDA was performed for the dataset. First, the fraction of the target variable usage was checked as shown in Figure 3. Surprisingly, most variety of people took cannabis the day before the survey, followed by those that have never tried.

Chart, bar chart, histogram

Description automatically generated

Figure . Cannabis usage

Next, the correlation between Education and Cscore was checked, using the boxplot as shown in Figure 4.

Chart, bar chart

Description automatically generated

Figure . Education vs Conscientiousness score

Maximum conscientiousness value tended to be higher for those with higher education as expected but having similar mean values was something not expected. Pearson correlation matrix was calculated to see the dependency of each feature (Figure 5.).

Chart, treemap chart

Description automatically generated

Figure . Pearson correlation matrix

We can see here, with the target variable cannabis, Sensation Seeking (SS) shows the highest value.

Since this project is a classification problem, balance of each class was checked as shown in Figure 6. From this, it can be noticed that data such as education and age are not evenly distributed.

Chart, pie chart

Description automatically generated

Figure . Balance of each class

1. **Data Splitting/Preprocessing**

As we have noticed in the previous chapter, the data is imbalanced. To overcome this problem, first, stratified train\_test\_split was used to split the data into other and test sets. Next, StratifiedKFold split was used on the other section to split them into train and validation set. The dataset used here is IID, since it is based on individual participants of the survey. There is no group structure nor time series in the data set.

For the preprocessing, OrdinalEncoder was used for the categorical values that can be ranked, such as the education and the age. For those categorical values that cannot be ranked, such as cannabis, ethnicity, country, and more, OneHotEncoder was used. For the personality scores, which contains values in numbers, StandardScaler was used. After the preprocessing, the data points were obtained as shown in Figure 7., along with the 24 features.

Text, letter

Description automatically generated

Figure . Shape of X\_train, X\_train\_prep, X\_val, X\_val\_prep, X\_test, X\_test\_prep

1. **References**

1)<https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29>

2) Z. T. Qiao, Q. Chai, X. Zhang, et al. "Predicting potential drug abusers using machine learning techniques," 2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Shanghai, China, 2019, pp. 283-286. IEEE. [[Web Link]](https://doi.org/10.1109/ICIIBMS46890.2019.8991550)

3) 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), Lombok, Indonesia, 2019, pp. 33-37. IEEE. [[Web Link]](https://doi.org/10.1109/SIET48054.2019.8986131)

4) A. Shahriar, F. Faisal, S. U. Mahmud, et al. "A Machine Learning Approach to Predict Vulnerability to Drug Addiction." In 2019 22nd International Conference on Computer and Information Technology (ICCIT) (pp. 1-7) 2019. IEEE. [[Web Link]](https://doi.org/10.1109/ICCIT48885.2019.9038605)

5) A. Alujaa, O. Garcı́ab, J.Rossierc, L.F.Garcı́aa, “Comparison of the NEO-FFI, the NEO-FFI-R and an alternative short version of the NEO-PI-R (NEO-60) in Swiss and Spanish samples” 2005, Vol 38, pp. 591-604

6) https://github.com/Youbin-K/DS1030\_project