DS1030 Final report

“Predicting the cannabis usage”

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

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 22 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

To verify the problem, the types of each class has been printed out as shown in Table. 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 2. Data type of each column

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

The EDA was performed for the dataset. First, the percentage of the target variable usage was checked as shown in Figure 1. 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 1. Cannabis usage

Next, the correlation between Education and Cscore was checked, using the boxplot as shown in Figure 2.

Chart, bar chart

Description automatically generated

Figure 2. 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 3.).

Chart, treemap chart

Description automatically generated

Figure 3. 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 and it turned out that the data such as education and age are not evenly distributed.

1. **Methods**

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 for which 80 percent is used for the training and 20 percent is left for the test set. 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, along with the 24 features.

The machine learning (ML) pipeline was developed using the random forest classifier and cross validated with the GridsearchCV with the parameter grid of maximum depth 1,3,10,30 and 100. The maximum features for the random forest classifier were chosen to be 0.5, 0.75, and 1.0.

For the ML algorithm, logistic regression, ridge classifier, random forest classifier and support vector machine were used. For the logistic regression, two penalties were tested, l2 and the none. For the ridge classifier, the cholesky and saga solvers were tested. For the random forest classifier, depth were tested starting from none to five. Lastly, for the support vector machine, gamma values of one and auto was tested, along with the C value of 0.5, 1, and 2.

1. **Results**

With the used ML algorithms, the calculated value of the model scores were obtained by using the accuracy score and the f\_beta score. Each model and the calculated score is shown below in the Table 3.

Table

Description automatically generated

Table 3. Accuracy and Fbeta score of each ML algorithm, along with the tuned parameters.

As we can see from the table, Random Forest Classifier with the default value performs the best, giving the accuracy score of almost 84 percent. This value is presented in the below figure using the confusion matrix.

Chart, treemap chart

Description automatically generated

Figure 4. Confusion matrix of the best value.

Next, finding the most important features were also performed using the mutual\_info\_classif function in Sklearn, and SHAP. First, the mutual\_info\_classif of top 5 most important value is shown in the below figure.

Chart, bar chart

Description automatically generated

Figure 5. Top 5 values of the mutual\_info\_classif

We can see that the UK has the highest contribution, followed by the Age and the Sensation Seeking (SS) value. However, although similar, it is a bit different in the SHAP values for the local feature importance. The SHAP values are shown below.

Chart

Description automatically generated

Figure 6. Top 5 values by SHAP.

Somewhat similar but different, SHAP shows us the highest importance is due to the Age, followed by the UK and the Oscore. The least important feature is the SS, which is very different to the confusion matrix and the Figure 6. It is interesting that Oscore is included in SHAP, whereas in the mutual\_info\_classif it is not. This could be a problem, since depending on the variable, the result of the prediction could change. Another unexpected thing is that SS importance being the lowest.

1. **Outlook**

Although the random forest classifier has already shown accuracy of 84 percent, this could be more improved in a few ways. First of all, it could yield better results if we change the criterion and max features in the model. Second of all, we could be able to get a better result if we include those features that were dropped, i.e. chocolate and alcohol. Last but not least, if we have a better dataset, it is likely that the accuracy of the prediction will be increased. It would also have been interesting to try out other techniques such as AdaBoostClassifier, StackingClassifier, and VotingClassifier.

One of the weak spots of this study is that the education and the age distribution was not ideal, which could have resulted in the biased results. By collecting some additional data to evenly distribute the education and age, more accurate prediction could be possible. Also, instead of the NEO-FFI-R personality scores in the dataset, it could have been better if the collection was based on the Myers-Briggs Type Indicator (MBTI).

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