Prediction on Drug Consumption Based on Demographic Information and Personality Traits

Junhao Lin, Minghua Zhang, and Mengting Tang

Abstract

The goal of the project is to assess the risk of drug use based on people's personality traits, impulsivity, sensory seeking, and their demographic information. By training the classification model, individuals will be labeled as "users" or "non-users" based on their learned characteristics. This project will use a data set from the UCI machine learning library, which contains the demographic information, psychological characteristics and drug history of the respondents.

1 Introduction

Drug use is a dangerous behavior that sometimes leads to health issues, drug abuse, and many other personal and social consequences, and are sometimes considered criminal. However, drug abuse cannot happen for no reason. Factors related to drug use must be investigated to prevent drug users from causing further damage. Specifically, in this project, factors such as personality characteristics will be analyzed to predict individual drug consumption behavior.

1.1 Problem Statement

The goal of our project was to create a program that can classify the influence of different factors to the drug consumption. For each individual respondent, there are 12 attributes that we have to determine the influence to the respondent including level of education, age, gender, country of residence, ethnicity, sensation seeking, impulsivity, neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. And each participant are questioned about 18 legal and illegal drugs including alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron). For the 18 drugs we group them together into three different categories to evaluate the risk to be drug consumer for each category of drugs.

1.2 Potential Ethical issue

In the dataset we are using, its demographic data contains ethnicity and country of origin, which could potentially cause discrimination. Also, after we investigated the data, these 2 features rarely give substantial impact on general result hence, in order to erase the potential discrimination issue, and to elevate model performance, we removed these 2 features.

2 Technical Approach

The objective of this project is to implement three Machine Learning methods to classify and evaluate the risk of each category of drugs. By applying three different algorithms, choosing the most effective model to return best performance of the data and predict the drug consumption.

2.1 Method 1: Logistic Regression

We considered using sklearn LogisticRegression model to perform the task. Logistic Regression was first used in the biological science and then transform to be used in many social science applications. It is

used when the dependent variable is ategorical, Mathematical representation of our model would be

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

Where t is the possibility of whether the instance had drug usage or not.

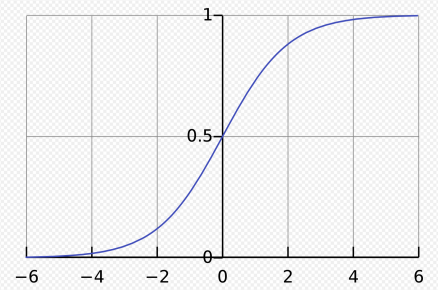

Figure 1: The standard logistic function.[2]

2.2 Method 2: Support Vector Machine(SVM)

We considered using SVM model using sklearn.svm SVC model to perform the task. The support vector machine (SVM) is a supervised machine learning technique that is commonly employed in classification tasks. However, it may also be utilized to solve regression problems. The goal of the support vector machine technique is to discover a hyperplane in an N-dimensional space (where n is the amount of attributes you have) that categorizes a data point clearly. There are several hyperplanes from which to choose to split the two kinds of data points. Our goal is to discover a plane with the greatest margin, or the greatest distance between data points from both classes. Maximizing the margin distance gives some reinforcement, making it easier to classify subsequent data points[5].

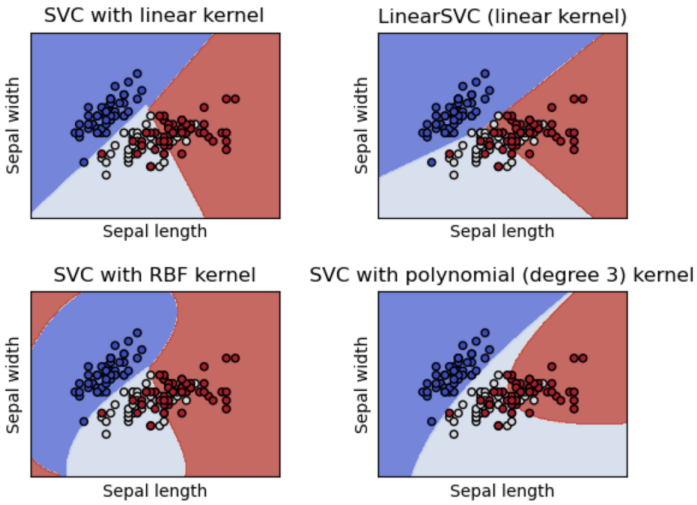

Figure 2: Example of SVM classifing multiple classes[4]

2.3 Method 3: Neural Network

We considered using NN model provided by pytorch library. In the implementation, we used nn.Sequential to represent our NN model, for it is flexible enough and intuitive:

```
model = nn.Sequential(
    input_layer
    activation function
    hidden layers
    activation function
    output layer
    activation function
    )
```

We choose to use **ReLu** activation function for all hidden layers. As Figure 2 demonstrated, ReLu only gives positive output when given positive input, otherwise it would produce 0. Advantage of this activation function is that, model that uses it is easier to train and often achieves better performance.[3]

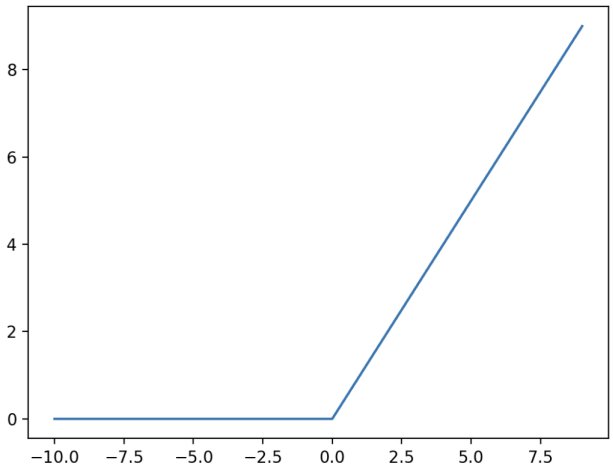

Figure 3: Line Plot of Rectified Linear Activation for Negative and Positive Inputs[3]

Whereas for the output layer, we used sigmoid function as activation function. Because eventually we are doing binary classification. The behavior of sigmoid function is demonstrated in Figure 1.

3 Experimental Results

The general procedure of this project consists of pre-processing the dataset, training the models by different approaches while finding optimal parameters, generating accuracy score for test set and evaluation. All the processes were done by using Python based on libraries such as Sklearn and PyTroch.

3.1 Raw Dataset

The data we used is Drug consumption (quantified) Data Set[6] from UCI Machine Learning Repository. The dataset contains records for 1885 respondents. Each record has 12 attributes. For each drug, the respondents have been selected to ask about when they have drug and then separated into seven classes which is "Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day". As shown in figure 5, classes "age", "gender", "education", "country", "ethnicity" are represented by floats which are encoded, and as we mentioned in **1.2** and shown in figure 4, country and ethnicity may raise ethical issue, and did not contribute much in result, and thus during processing we removed class "country" and "ethnicity".

3.2 Processing Data

Inside the raw data, all instances of different classes are represented by float, which completely makes no sense. We implemented several method to transform them into integer so it would makes more sense. At the same time, we dropped country and ethnicity as shown in figure 6.

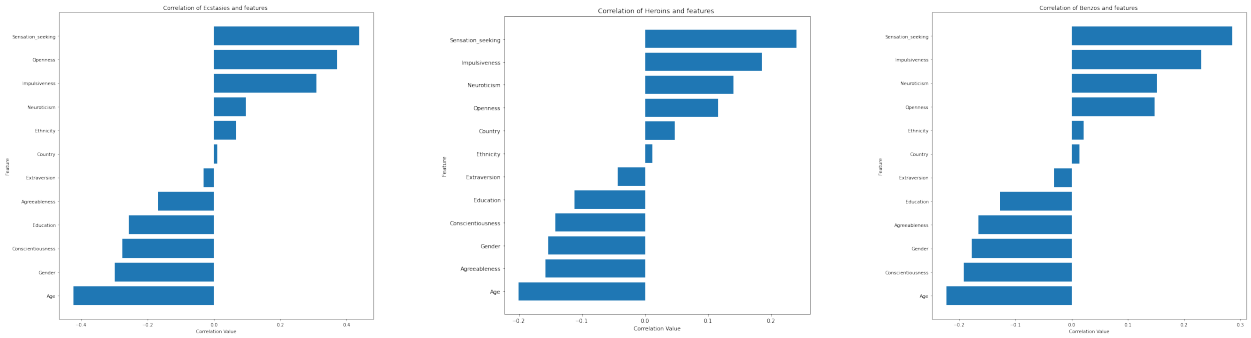

Figure 4: Classes correlation to 3 kinds of drug consumption.

Age	Gender	Education	Country	Ethnicity	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
-0.95197	-0.48246	-0.61113	-0.57009	0.12600	-0.05188	-1.76250	0.58331	-0.76096	-0.14277
-0.07854	-0.48246	-0.61113	0.24923	0.11440	-0.14882	-0.57545	1.43533	-0.91699	-0.78155
-0.95197	-0.48246	-1.43719	-0.57009	-0.31685	1.49158	-1.92173	-0.58331	-1.77200	0.58489
-0.95197	-0.48246	0.45468	0.24923	-0.31685	-0.05188	-1.76250	0.88309	-0.76096	2.33337
-0.95197	-0.48246	-0.61113	-0.28519	-0.31685	-0.79151	0.32197	0.29338	-0.30172	-0.27607
-0.95197	0.48246	-0.61113	-0.57009	-0.31685	-1.19430	1.74091	1.88511	0.76096	-1.13788
-0.95197	-0.48246	-0.61113	-0.57009	-0.31685	-0.24649	1.74091	0.58331	0.76096	-1.51840
-0.07854	0.48246	0.45468	-0.57009	-0.31685	1.13281	-1.37639	-1.27553	-1.77200	-1.38502
-0.95197	0.48246	-0.61113	-0.57009	-0.31685	0.91093	-1.92173	0.29338	-1.62090	-2.57309
-0.95197	-0.48246	-0.61113	0.21128	-0.31685	-0.46725	2.12700	1.65653	1.11406	0.41594

Figure 5: Raw Data pre-processed

3.3 Hyper-parameter Tuning and Cross-Validation

We considered using GridSearchCV library from sklearn. This model provides a more visually-understandable way for hyper-parameters tuning. The GridSearchCV instance implements the usual estimator API: when "fitting" it on a dataset all the possible combinations of parameter values in the given grid are evaluated and the one giving the best accuracy score is retained.Moreover, the search also involves cross validation and stratified 5-folds was used in our models. For example, in the following parameter grid:

```
param_grid = [
  {'C': [1, 10, 100, 1000], 'kernel': ['linear']},
  {'C': [1, 10, 100, 1000], 'gamma': [0.001, 0.0001], 'kernel': ['rbf']},
 ]
```

it specifies that two grids should be explored: one with a linear kernel and C values in [1, 10, 100, 1000], and the second one with an RBF kernel, and the cross-product of C values ranging in [1, 10, 100, 1000] and gamma values in [0.001, 0.0001][1].

3.4 Results from all ML methods

Logistic Regression: In building LR model, four parameters are controlled to evaluate and find the one gives the best accuracy score: penalty, C(Inverse of regularization strength), class weight, and the solver used in optimization problem.Using GridSearch, the following optimal parameters in predicting **Heroins** consumption are given:

```
{'C': 1e-20,
 'class_weight': 'balanced',
 'penalty': 'l1',
 'solver': 'liblinear'}
```

Then, these parameters were used to build the best logistic regression model, and then predicted the given test set. Finally, testing accuracy and training accuracy scores were calculated to evaluate model's performance. Here are the best tuned parameters for other two predicting models: **Ecstasies:**

```
{'C': 0.1, 'class_weight': 'balanced', 'penalty': 'l1', 'solver': 'liblinear'}
```

Age	Gender	Education	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness	Impulsiveness	Sensation_seeking
2	1	5	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1.18084
1	0	8	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575
2	0	5	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0.40148
0	1	7	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084
2	1	8	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	-0.21575
5	1	3	-0.67825	-0.30033	-1.55521	2.03972	1.63088	-1.37983	-1.54858
3	0	7	-0.46725	-1.09207	-0.45174	-0.30172	0.93949	-0.21712	0.07987
2	0	1	-1.32828	1.93886	-0.84732	-0.30172	1.63088	0.19268	-0.52593
2	1	5	0.62967	2.57309	-0.97631	0.76096	1.13407	-1.37983	-1.54858
4	0	7	-0.24649	0.00332	-1.42424	0.59042	0.12331	-1.37983	-0.84637

Figure 6: Processed Data

Benzos:

```
{'C': 1e-17, 'class_weight': 'balanced', 'penalty': 'l2','solver': 'liblinear'}
```

Support Vector Machine(SVM): Similarly to LR, two paremeters were controlled in SVM: C and the kernel type used in this model. Then, using GridSearch, the following optimal parameters were generated and their corresponding testing and training accuracy score.

Heroins:

```
{'C': 0.5, 'kernel': 'linear'}
```

Ecstasies:

```
{'C': 1, 'kernel': 'rbf'}
```

Benzos:

```
{'C': 0.5, 'kernel': 'linear'}
```

Pytorch NN: When constructing a neural network, we first count how many features we're measuring, and that is the number of input neuron. Eventually, we decided to use this model:

```
model = nn.Sequential(nn.Linear(10, 20),
                      nn.ReLU(),
                      nn.Linear(20, 20),
                      nn.ReLU(),
                      nn.Linear(20, 20),
                      nn.ReLU(),
                      nn.Linear(20, 1),
                      nn.Sigmoid())
```

Since the data we're training on isn't that large, 2 layers of hidden layers is enough to process the data, where each of the layer has 20 neurons, and adding more neurons would not add to more performance. When training, we use a mini-batch of 60 to train the model in a total of 40 epochs, and use testing data to obtain the test accuracy in each epoch . After all epoch has been ran, we calculate the average of accuracy to present as conclusion.

After gaining the accuracy scores of each modeling approach in three predicted models, figure 7 below is made to comparing the performance of each machine learning model. Generally, all modeling approaches gave the similar performance, and comparatively SVM generates a higher accuracy score.

4 Participants Contribution

Junhao Lin was primarily focusing on the formulation of project report, and implemented the neural network methods.

Minghua Zhang and Mengting Tang was primarily focusing on pipe-lining raw data, and transform them into process-able form, and implemented the Logistic regression and SVM model, and the creation of plots.

All team members shared equal responsibility in idea and project data-set of this project.

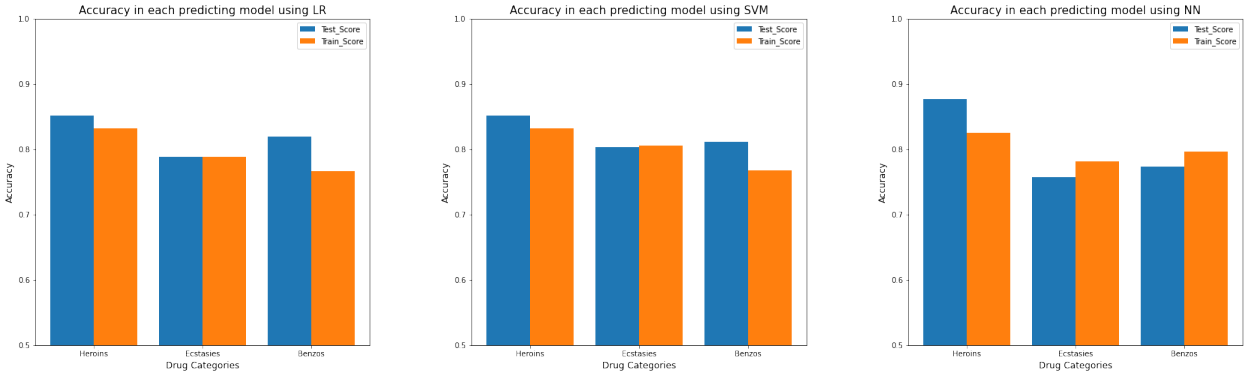

Figure 7: Accuracy of each ML model on different classes

5 References

[1] 3.2.1 Exhaustive Grid Search, https://scikit-learn.org/stable/modules/grid_search.html
[2] Logistic Regression, https://en.wikipedia.org/wiki/Logistic_regression#Formal_mathematical_specification
[3] A Gentle Introduction to the Rectified Linear Unit (ReLU), by Jason Brownlee, January 9, 2019. https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/
[4] 1.4.1 Support Vector Machine, https://scikit-learn.org/stable/modules/svm.html
[5] Support Vector Machine - Introduction to Machine Learning, June 7, 2018. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
[6] Drug consumption (quantified) Data Set, by Elaine Fehrman, Vincent Egan, and Evgeny M. Mirkes https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29