# Predicting Drug Consumption Using Data Mining Classifiers

Trung N. Nguyen

## ABSTRACT

Addiction derives from the Latin *addīcere* " to surrender" or "to enslave to". Addicts are people who surrender either or altogether their cognition and/or motor function. The reason for the desire is due to reinforcement and appetitive stimuli. The communal element in addictive drugs is the release of dopamine, which is a neurotransmitter that causes excitatory responses and aids in feelings of reward[3]. Although not all addictive drugs are illegal, addicts show similar tolerance and withdrawal symptoms. Tolerance causes the body to become desensitized to the effects of drugs resulting in larger doses to reach the same effects as the initial doses, while withdrawal symptoms are the opposite effects of drugs[3]. Addicts crave the positive reinforcement gained from drugs, which over time, their only purpose is to subside the withdrawal symptoms turning the positive into a negative reinforcement[3]. Although some addicts are able to stop the addiction, this does not guarantee the loss of cravings or the probability of relapse.This paper evaluates various classifiers to predict different types of drug consumption based on patient's age, gender, education, country, and personality traits based on NEO-FFI-R, BIS-11, and ImpSS. Experimental results will display the accuracy of the model applied on the Drug Consumption Dataset[1].

## 1. INTRODUCTION

Aristotle defines virtue as the disposition to act morally and righteously and as a mean between deficiency and excess. Similarly, drug consumption should be consumed in moderation, where the deficiency is the lack of drug consumption, when ill, while excess is drug addiction. Drug addiction is the result of the lack of stimulation and release of dopamine from small doses of the drug[3]. As a result, addicts overconsume the drug in order to reach the same state of euphoria or eliminate the withdrawal symptoms. However, even though drug addicts can abstain from the drug resulting in a lack of cupidity, they can still crave or relapse into addiction. The brain still remembers the memories of reward from the dopamine thus, resulting in the craving[3]. This paper attempts to predict different types of drug consumption based on subject's age, gender, education, country, and personality traits using different classifying algorithms – naïve Bayes, k-nearest neighbors, ZeroR, and J48 decision tree.

The data set contains 1885 subjects with 13 different class attributes – mostly consisting of participants from the UK[1]. The 13 attributes include age, gender, education, country, ethnicity, nscore, escore, oscore, ascore, cscore, impulsiveness, sensation (ss), and 19 different drugs[1]. The 19 drugs consist of alcohol, amphetamines (amphet), amyl nitrite (amyl), benzodiazepine (benzos), caffeine (caffe), cannabis, chocolate (choc), cocaine (coke), ecstasy, heroin, ketamine, legal highs (legalh), LSD, methadone (meth), magic mushrooms (mushrooms), nicotine, semeron (semer), and VSA[1]. WEKA was used for data analysis, and the calculation of accuracy, precision, recall, TP rate, and F-measurement were recorded on a spreadsheet[4]. The data from the original data set has been transformed using a shell script changing the categorical numeric values into categorical string values. For some classifiers, different pre-processing techniques were used in order to maximum accuracy.

## 2. RELATED WORK

Fehrman, Mirks, Muhammad, Egan, and Gorban[2] presented a comparison of different classification accuracies of decision tree, random forest, k-nearest neighbors, linear discriminant analysis, Gaussian mixture, probability density function estimation, logistic regression, and naïve Bayes. The accuracy of the results was measured from the specificity and sensitivity resulting in accuracies around 70 percent[2]. However, for each classification, many attributes were removed from the data set and prediction model. For data analysis, Ferhrman et al. compared T-scores and CI-scores of the different drugs[2].

## 3. BACKGROUND

From the data set, some variables were collected through a NEO-FFI-R, BIS-11, and ImpSS test. NEO-FFI-R measures a person's openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism[1]. These five personality traits were measured in the data set via. a personality test[1]. The initial letter of the personality pertains to the scores in the data set – nscore, cscore, ascore, escore, and oscore. BIS-11 measures impulsiveness thus, evident of the impulsive attribute, and ImpSS measures sensation, which is the ss attribute. Regarding the different drugs, participants were asked questions concerning their drug history, and they had to answer: never used, used it over a decade ago, in the last decade, year, month, week, or day[1].

## 4. DATA PRE-PROCESSING

The original data set contained categorical numeric values for every attribute, which was incomprehensible without comments in the data set. Therefore, a shell scripted was implemented in order to change the numeric values into string values. Additionally, the .txt file was converted into an .arff file to process into WEKA. An .arff requires three lines of code: @relation, @attribute, and @data. The @relation line describes the relation name; the @attribute line defines all of the attributes and data types; and @data defines the start of the data set[4].

In order to obtain a higher accuracy and allow the machine to predict user and non-user, the data set was divided into two different sets. One set defined user as those who used the drug within the year; therefore, those participants who answered "used in the last decade", "used it over a decade ago", and "never used" were considered non-users. In the second data set, participants who answered "used it in the past year" were also considered non-users. Since there is no definitive temporal distinction that separates users from non-users, the selection was arbitrary. The first set considers those who consume drugs within the past year to be users, while the second set considers those who consume drugs within the past month to be users.

The experiment used different classifiers to determine the accuracy of each drug. When applying the classifier, the data set contained all attributes except for the 19 different drugs. The only drug attribute was the drug that was predicted. Additionally, some drug attributes had a class imbalance problem. WEKA provides a pre-processing technique called SMOTE, which resamples the data set applying Synthetic Minority Oversampling Technique[4]. This filter was applied to amyl, choc, crack, heroin, ketamine, semer, and VSA. Regarding the different classifiers, KNN (Ibk) used $k$ = 45, which was derived from the square root of the total number of participants, and J48 had true assigned to reducedErrorPuning. The other classifiers used default parameters.

# 5. EXPERIMENTAL RESULTS

## Table 1. Monthly Drug Consumption Using Naïve Bayes, KNN, J48, and ZeroR

| DRUG | DM Technique | TP Rate (NU) | TP Rate(U) | Precision(NU) | Precision(U) | Recall(NU) | Recall(U) | F-Measure(NU) | F-Measure(U) | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| | Naive-Bayes | 0.066 | 0.95 | 0.22 | 0.825 | 0.066 | 0.95 | 0.101 | 0.883 | **79.31** |
| | KNN | 0 | 1 | ? | 0.823 | 0 | 1 | ? | 0.903 | **82.28** |
| | J48 | 0.024 | 0.988 | 0.296 | 0.825 | 0.024 | 0.988 | 0.044 | 0.899 | **81.7** |
| Alcohol | ZeroR | | | | | | | | | **82.28** |
| | Naive-Bayes | 0.798 | 0.661 | 0.887 | 0.496 | 0.798 | 0.661 | 0.84 | 0.566 | **76.6** |
| | KNN | 0.911 | 0.378 | 0.83 | 0.561 | 0.911 | 0.378 | 0.868 | 0.452 | **78.78** |
| | J48 | 0.881 | 0.388 | 0.827 | 0.494 | 0.881 | 0.388 | 0.853 | 0.434 | **76.66** |
| Amphet | ZeroR | | | | | | | | | **76.87** |
| | Naive-Bayes | 0.858 | 0.383 | 0.902 | 0.291 | 0.858 | 0.383 | 0.879 | 0.331 | **79.53** |
| | KNN | 0.973 | 0.214 | 0.891 | 0.543 | 0.973 | 0.214 | 0.93 | 0.307 | **87.26** |
| | J48 | 0.969 | 0.226 | 0.892 | 0.526 | 0.969 | 0.226 | 0.929 | 0.316 | **87.12** |
| Amyl | ZeroR | | | | | | | | | **86.82** |
| | Naive-Bayes | 0.772 | 0.591 | 0.826 | 0.506 | 0.772 | 0.591 | 0.798 | 0.545 | **72.04** |
| | KNN | 0.91 | 0.342 | 0.777 | 0.6 | 0.91 | 0.342 | 0.838 | 0.436 | **74.85** |
| | J48 | 0.896 | 0.325 | 0.77 | 0.552 | 0.896 | 0.325 | 0.828 | 0.409 | **73.37** |
| Benzos | ZeroR | | | | | | | | | **71.62** |
| | Naive-Bayes | 0.107 | 0.976 | 0.232 | 0.942 | 0.107 | 0.976 | 0.146 | 0.959 | **92.19** |
| | KNN | 0 | 1 | ? | 0.937 | 0 | 1 | ? | 0.968 | **93.73** |
| | J48 | 0.016 | 0.996 | 0.2 | 0.938 | 0.016 | 0.996 | 0.03 | 0.966 | **93.42** |
| Caff | ZeroR | | | | | | | | | **93.73** |
| | Naive-Bayes | 0.836 | 0.79 | 0.779 | 0.845 | 0.836 | 0.79 | 0.807 | 0.816 | **81.17** |
| | KNN | 0.85 | 0.755 | 0.755 | 0.85 | 0.85 | 0.755 | 0.799 | 0.8 | **79.95** |
| | J48 | 0.796 | 0.803 | 0.782 | 0.816 | 0.796 | 0.803 | 0.789 | 0.809 | **79.95** |
| Cannabis | ZeroR | | | | | | | | | **52.3** |
| | Naive-Bayes | 0 | 1 | 0 | 0.953 | 0 | 1 | 0 | 0.976 | **95.23** |
| | KNN | 0 | 1 | ? | 0.953 | 0 | 1 | ? | 0.976 | **95.34** |
| | J48 | 0 | 1 | 0.231 | 0.955 | 0.033 | 0.995 | 0.058 | 0.974 | **95** |
| Choc | ZeroR | | | | | | | | | **95.34** |
| | Naive-Bayes | 0.792 | 0.566 | 0.865 | 0.435 | 0.792 | 0.566 | 0.827 | 0.492 | **94.16** |
| | KNN | 0.989 | 0.053 | 0.786 | 0.579 | 0.989 | 0.053 | 0.876 | 0.097 | **78.2** |
| | J48 | 0.936 | 0.156 | 0.796 | 0.409 | 0.936 | 0.156 | 0.86 | 0.226 | **76.34** |
| Coke | ZeroR | | | | | | | | | **77.88** |
| | Naive-Bayes | 0.881 | 0.62 | 0.965 | 0.313 | 0.881 | 0.62 | 0.92 | 0.416 | **86** |
| | KNN | 0.998 | 0.013 | 0.92 | 0.4 | 0.98 | 0.013 | 0.958 | 0.025 | **91.9** |
| | J48 | 0.985 | 0.146 | 0.929 | 0.46 | 0.985 | 0.146 | 0.956 | 0.221 | **91.75** |
| Crack | ZeroR | | | | | | | | | **91.96** |
| | Naive-Bayes | 0.789 | 0.667 | 0.863 | 0.545 | 0.789 | 0.667 | 0.824 | 0.6 | **75.6** |
| | KNN | 0.865 | 0.524 | 0.828 | 0.594 | 0.865 | 0.524 | 0.846 | 0.557 | **77.14** |
| | J48 | 0.845 | 0.499 | 0.817 | 0.549 | 0.845 | 0.499 | 0.831 | 0.523 | **75** |
| Ecstasy | ZeroR | | | | | | | | | **72.57** |
| | Naive-Bayes | 0.848 | 0.682 | 0.952 | 0.374 | 0.848 | 0.682 | 0.897 | 0.483 | **82.83** |
| | KNN | 0.972 | 0.216 | 0.903 | 0.505 | 0.972 | 0.216 | 0.936 | 0.303 | **88.27** |
| | J48 | 0.967 | 0.237 | 0.905 | 0.487 | 0.967 | 0.237 | 0.935 | 0.319 | **88.07** |
| Heroine | ZeroR | | | | | | | | | **88.22** |
| | Naive-Bayes | 0.79 | 0.644 | 0.9 | 0.432 | 0.79 | 0.644 | 0.841 | 0.517 | **76.11** |
| | KNN | 0.92 | 0.41 | 0.863 | 0.561 | 0.92 | 0.41 | 0.891 | 0.474 | **81.89** |
| | J48 | 0.943 | 0.337 | 0.851 | 0.593 | 0.943 | 0.337 | 0.895 | 0.429 | **82.23** |
| Ketamine | ZeroR | | | | | | | | | **80.12** |
| | Naive-Bayes | 0.806 | 0.725 | 0.873 | 0.615 | 0.806 | 0.725 | 0.838 | 0.666 | **78.2** |
| | KNN | 0.847 | 0.64 | 0.846 | 0.641 | 0.847 | 0.64 | 0.847 | 0.641 | **78.51** |
| | J48 | 0.863 | 0.564 | 0.823 | 0.637 | 0.863 | 0.564 | 0.842 | 0.598 | **77.35** |
| Legalh | ZeroR | | | | | | | | | **70.08** |
| | Naive-Bayes | 0.831 | 0.758 | 0.931 | 0.531 | 0.831 | 0.758 | 0.879 | 0.625 | **81.64** |
| | KNN | 0.912 | 0.505 | 0.879 | 0.591 | 0.912 | 0.505 | 0.895 | 0.545 | **82.97** |
| | J48 | 0.906 | 0.542 | 0.887 | 0.594 | 0.906 | 0.542 | 0.896 | 0.567 | **83.29** |
| LSD | ZeroR | | | | | | | | | **79.84** |
| | Naive-Bayes | 0.814 | 0.609 | 0.911 | 0.401 | 0.814 | 0.69 | 0.86 | 0.484 | **77.93** |
| | KNN | 0.976 | 0.153 | 0.849 | 0.563 | 0.976 | 0.153 | 0.908 | 0.241 | **83.61** |
| | J48 | 0.953 | 0.172 | 0.849 | 0.426 | 0.953 | 0.172 | 0.898 | 0.245 | **82.02** |
| Meth | ZeroR | | | | | | | | | **83.02** |
| | Naive-Bayes | 0.819 | 0.733 | 0.911 | 0.547 | 0.819 | 0.733 | 0.862 | 0.627 | **79.89** |
| | KNN | 0.866 | 0.615 | 0.883 | 0.579 | 0.866 | 0.615 | 0.874 | 0.597 | **80.85** |
| | J48 | 0.886 | 0.482 | 0.851 | 0.559 | 0.886 | 0.482 | 0.868 | 0.517 | **79.31** |
| Mushrooms | ZeroR | | | | | | | | | **76.98** |
| | Naive-Bayes | 0.72 | 0.689 | 0.643 | 0.76 | 0.72 | 0.689 | 0.679 | 0.722 | **70.24** |
| | KNN | 0.695 | 0.701 | 0.644 | 0.747 | 0.695 | 0.701 | 0.668 | 0.723 | **69.81** |
| | J48 | 0.615 | 0.743 | 0.651 | 0.712 | 0.615 | 0.743 | 0.732 | 0.728 | **68.7** |
| Nicotine | ZeroR | | | | | | | | | **56.23** |
| | Naive-Bayes | 0.999 | 0.5 | 0.998 | 0.6 | 0.999 | 0.5 | 0.999 | 0.545 | **99.74** |
| | KNN | 1 | 0 | 0.997 | ? | 1 | ? | 0.998 | ? | **99.68** |
| | J48 | 1 | 0 | 0.997 | ? | 1 | 0 | 0.998 | ? | **99.68** |
| Semer | ZeroR | | | | | | | | | **99.68** |
| | Naive-Bayes | 0.851 | 0.647 | 0.958 | 0.316 | 0.851 | 0.647 | 0.902 | 0.425 | **83.18** |
| | KNN | 0.987 | 0.095 | 0.911 | 0.095 | 0.987 | 0.095 | 0.948 | 0.156 | **90.15** |
| | J48 | 0.984 | 0.2 | 0.921 | 0.567 | 0.984 | 0.2 | 0.951 | 0.296 | **90.86** |
| VSA | ZeroR | | | | | | | | | **90.4** |

# Table 2. Annual Drug Consumption Using Naïve Bayes, KNN, J48, and ZeroR

| DRUG | DM Technique | TP Rate (NU) | TP Rate(U) | Precision(NU) | Precision(U) | Recall(NU) | Recall(U) | F-Measure(NU) | F-Measure(U) | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| | Naive-Bayes | 0.037 | 0.987 | 0.313 | 0.868 | 0.037 | 0.987 | 0.066 | 0.924 | **85.95** |
| | KNN | 0 | 1 | ? | 0.865 | 0 | 1 | ? | 0.928 | **86.54** |
| | J48 | 0.143 | 0.974 | 0.459 | 0.88 | 0.143 | 0.974 | 0.218 | 0.924 | **86.2** |
| Alcohol | ZeroR | | | | | | | | | **86.54** |
| | Naive-Bayes | 0.798 | 0.661 | 0.887 | 0.496 | 0.798 | 0.661 | 0.84 | 0.566 | **76.6** |
| | KNN | 0.911 | 0.378 | 0.83 | 0.561 | 0.911 | 0.378 | 0.868 | 0.452 | **78.78** |
| | J48 | 0.881 | 0.388 | 0.827 | 0.494 | 0.881 | 0.388 | 0.853 | 0.434 | **76.66** |
| Amphet | ZeroR | | | | | | | | | **76.87** |
| | Naive-Bayes | 0.858 | 0.38 | 0.902 | 0.291 | 0.858 | 0.383 | 0.879 | 0.331 | **79.53** |
| | KNN | 0.973 | 0.214 | 0.891 | 0.543 | 0.973 | 0.214 | 0.93 | 0.307 | **87.26** |
| | J48 | 0.969 | 0.226 | 0.892 | 0.526 | 0.969 | 0.226 | 0.929 | 0.316 | **87.12** |
| Amyl | ZeroR | | | | | | | | | **86.82** |
| | Naive-Bayes | 0.772 | 0.591 | 0.826 | 0.506 | 0.772 | 0.591 | 0.798 | 0.545 | **72.04** |
| | KNN | 0.91 | 0.342 | 0.777 | 0.6 | 0.91 | 0.342 | 0.838 | 0.436 | **74.85** |
| | J48 | 0.896 | 0.25 | 0.77 | 0.552 | 0.896 | 0.325 | 0.828 | 0.409 | **73.37** |
| Benzos | ZeroR | | | | | | | | | **71.62** |
| | Naive-Bayes | 0.107 | 0.976 | 0.232 | 0.942 | 0.107 | 0.976 | 0.146 | 0.959 | **92.19** |
| | KNN | 0 | 1 | ? | 0.937 | 0 | 1 | ? | 0.968 | **93.73** |
| | J48 | 0.016 | 0.996 | 0.2 | 0.938 | 0.016 | 0.996 | 0.03 | 0.996 | **93.42** |
| Caff | ZeroR | | | | | | | | | **93.73** |
| | Naive-Bayes | 0.836 | 0.79 | 0.779 | 0.845 | 0.836 | 0.79 | 0.807 | 0.816 | **81.17** |
| | KNN | 0.85 | 0.756 | 0.755 | 0.85 | 0.85 | 0.756 | 0.8 | 0.8 | **80** |
| | J48 | 0.796 | 0.803 | 0.782 | 0.816 | 0.796 | 0.803 | 0.789 | 0.809 | **79.95** |
| Cannabis | ZeroR | | | | | | | | | **53** |
| | Naive-Bayes | 0 | 1 | 0 | 0.953 | 0 | 1 | 0 | 0.98 | **95.23** |
| | KNN | 0 | 1 | ? | 0.953 | 0 | 1 | ? | 0.976 | **95.34** |
| | J48 | 0.033 | 0.995 | 0.231 | 0.955 | 0.033 | 0.995 | 0.058 | 0.974 | **94.97** |
| Choc | ZeroR | | | | | | | | | **95.34** |
| | Naive-Bayes | 0.792 | 0.566 | 0.865 | 0.435 | 0.792 | 0.566 | 0.827 | 0.492 | **74.16** |
| | KNN | 0.989 | 0.053 | 0.786 | 0.579 | 0.989 | 0.053 | 0.876 | 0.097 | **78.2** |
| | J48 | 0.936 | 0.156 | 0.796 | 0.409 | 0.936 | 0.156 | 0.86 | 0.226 | **76.34** |
| Coke | ZeroR | | | | | | | | | **77.88** |
| | Naive-Bayes | 0.881 | 0.62 | 0.964 | 0.313 | 0.881 | 0.62 | 0.92 | 0.416 | **86** |
| | KNN | 0.998 | 0.013 | 0.92 | 0.4 | 0.998 | 0.013 | 0.958 | 0.025 | **91.9** |
| | J48 | 0.985 | 0.146 | 0.929 | 0.46 | 0.985 | 0.146 | 0.956 | 0.221 | **91.75** |
| Crack | ZeroR | | | | | | | | | **91.96** |
| | Naive-Bayes | 0.789 | 0.667 | 0.863 | 0.545 | 0.789 | 0.667 | 0.824 | 0.6 | **75.6** |
| | KNN | 0.865 | 0.524 | 0.828 | 0.594 | 0.865 | 0.524 | 0.846 | 0.557 | **77.14** |
| | J48 | 0.845 | 0.499 | 0.817 | 0.549 | 0.845 | 0.499 | 0.831 | 0.523 | **75.01** |
| Ecstasy | ZeroR | | | | | | | | | **72.57** |
| | Naive-Bayes | 0.848 | 0.682 | 0.952 | 0.374 | 0.848 | 0.682 | 0.897 | 0.843 | **82.83** |
| | KNN | 0.915 | 0.492 | 0.931 | 0.437 | 0.915 | 0.492 | 0.923 | 0.462 | **86.52** |
| | J48 | 0.967 | 0.237 | 0.905 | 0.487 | 0.967 | 0.237 | 0.935 | 0.319 | **88.07** |
| Heroine | ZeroR | | | | | | | | | **88.22** |
| | Naive-Bayes | 0.79 | 0.644 | 0.9 | 0.432 | 0.79 | 0.644 | 0.841 | 0.517 | **76.11** |
| | KNN | 0.92 | 0.411 | 0.863 | 0.561 | 0.92 | 0.411 | 0.891 | 0.474 | **81.89** |
| | J48 | 0.943 | 0.337 | 0.663 | 0.057 | 0.943 | 0.337 | 0.895 | 0.429 | **82.23** |
| Ketamine | ZeroR | | | | | | | | | **80.12** |
| | Naive-Bayes | 0.806 | 0.725 | 0.873 | 0.615 | 0.806 | 0.725 | 0.838 | 0.666 | **78.2** |
| | KNN | 0.847 | 0.64 | 0.846 | 0.641 | 0.847 | 0.64 | 0.847 | 0.841 | **78.51** |
| | J48 | 0.863 | 0.564 | 0.823 | 0.637 | 0.863 | 0.564 | 0.842 | 0.598 | **77.35** |
| Legalh | ZeroR | | | | | | | | | **70.08** |
| | Naive-Bayes | 0.831 | 0.758 | 0.931 | 0.531 | 0.831 | 0.758 | 0.879 | 0.625 | **81.64** |
| | KNN | 0.912 | 0.505 | 0.879 | 0.591 | 0.912 | 0.505 | 0.895 | 0.545 | **82.97** |
| | J48 | 0.906 | 0.542 | 0.887 | 0.594 | 0.906 | 0.542 | 0.896 | 0.567 | **83.29** |
| LSD | ZeroR | | | | | | | | | **79.84** |
| | Naive-Bayes | 0.813 | 0.609 | 0.911 | 0.4 | 0.813 | 0.69 | 0.859 | 0.483 | **77.88** |
| | KNN | 0.971 | 0.159 | 0.85 | 0.531 | 0.971 | 0.159 | 0.906 | 0.245 | **83.34** |
| | J48 | 0.94 | 0.197 | 0.852 | 0.429 | 0.946 | 0.197 | 0.897 | 0.27 | **81.91** |
| Meth | ZeroR | | | | | | | | | **83.02** |
| | Naive-Bayes | 0.819 | 0.733 | 0.911 | 0.547 | 0.819 | 0.733 | 0.862 | 0.627 | **79.89** |
| | KNN | 0.866 | 0.615 | 0.883 | 0.579 | 0.866 | 0.615 | 0.874 | 0.597 | **80.45** |
| | J48 | 0.886 | 0.482 | 0.851 | 0.559 | 0.886 | 0.482 | 0.868 | 0.517 | **79.31** |
| Mushrooms | ZeroR | | | | | | | | | **76.98** |
| | Naive-Bayes | 0.72 | 0.689 | 0.643 | 0.76 | 0.72 | 0.689 | 0.679 | 0.722 | **70.24** |
| | KNN | 0.695 | 0.701 | 0.644 | 0.747 | 0.695 | 0.701 | 0.668 | 0.623 | **69.81** |
| | J48 | 0.588 | 0.723 | 0.623 | 0.693 | 0.588 | 0.723 | 0.605 | 0.707 | **66.37** |
| Nicotine | ZeroR | | | | | | | | | **56.23** |
| | Naive-Bayes | 0.999 | 0.5 | 0.998 | 0.6 | 0.999 | 0.5 | 0.999 | 0.545 | **99.74** |
| | KNN | 1 | 0 | 0.997 | ? | 1 | 0 | 0.998 | ? | **99.68** |
| | J48 | 1 | 0 | 0.997 | ? | 1 | 0 | 0.998 | ? | **99.68** |
| Semer | ZeroR | | | | | | | | | **99.68** |
| | Naive-Bayes | 0.851 | 0.647 | 0.958 | 0.316 | 0.851 | 0.647 | 0.902 | 0.425 | **83.18** |
| | KNN | 0.987 | 0.095 | 0.911 | 0.439 | 0.987 | 0.095 | 0.948 | 0.156 | **90.15** |
| | J48 | 0.984 | 0.2 | 0.921 | 0.567 | 0.984 | 0.2 | 0.951 | 0.296 | **90.86** |
| VSA | ZeroR | | | | | | | | | **90.4** |

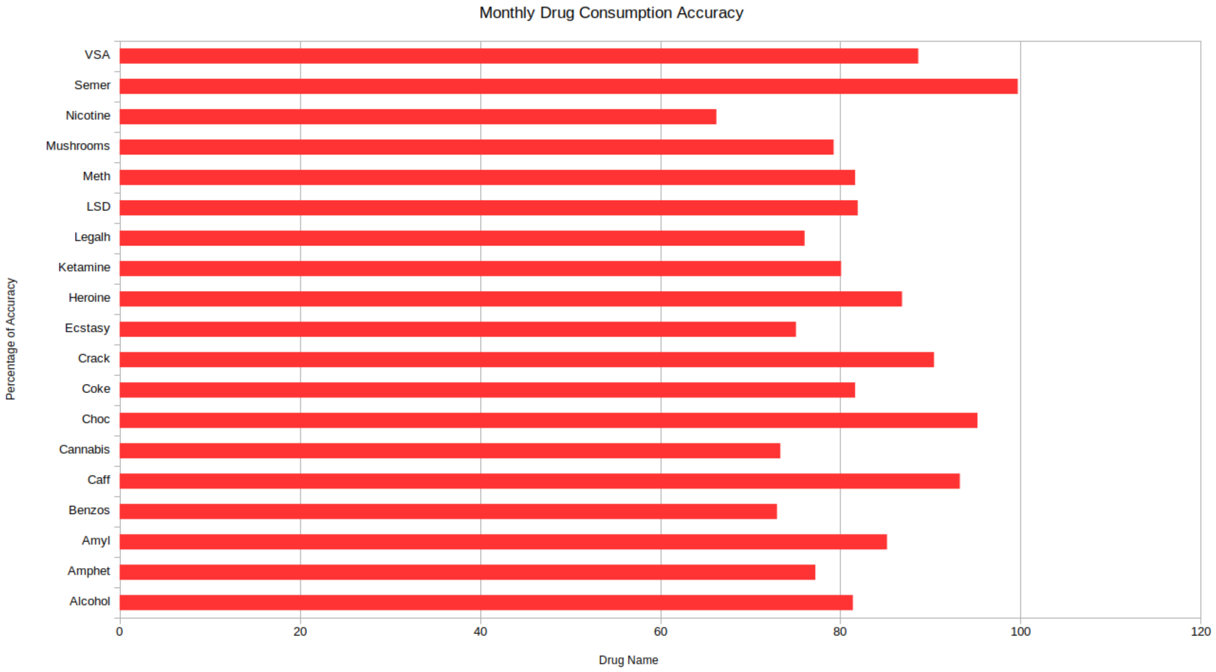

**Figure 1. Average Accuracies of Monthly Drug Consumption**

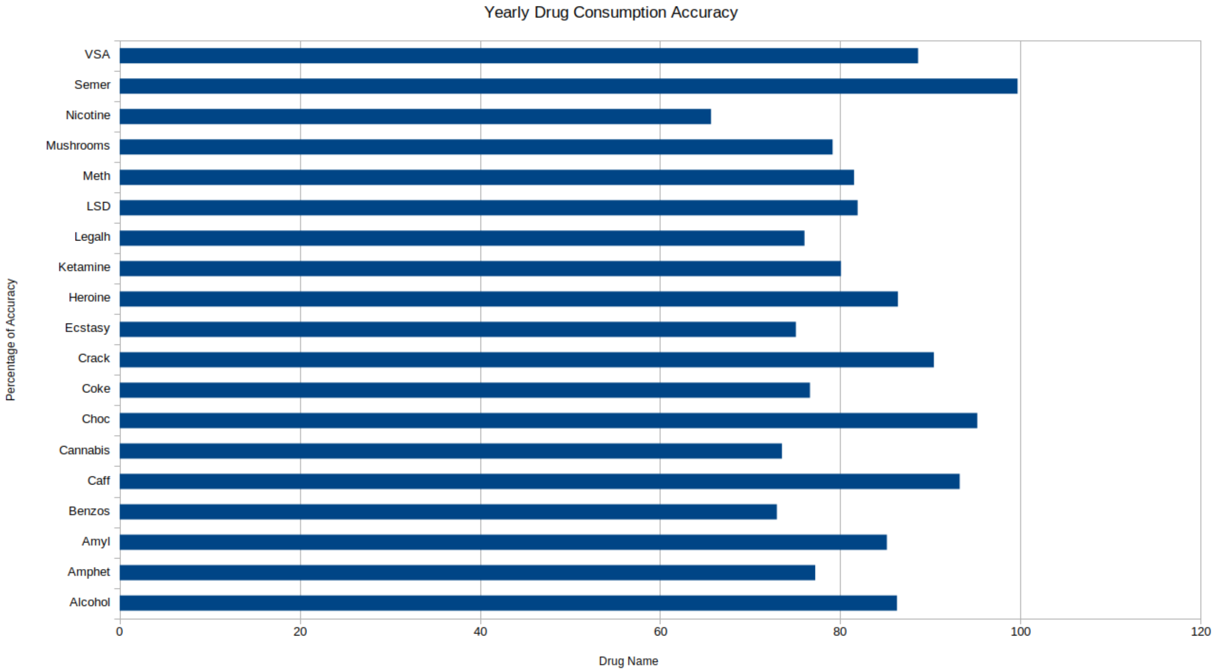

**Figure 2. Average Accuracies of Yearly Drug Consumption**

Table 1 shows the accuracy of drug consumption of those who used the drug within the past month, week, and day. Table 2 shows the accuracy of drug consumption of those who used the drug within the past year, month, week, and day. The accuracy ranges from 56.23-95.34. Although the model did predict semer correctly, the accuracies do not represent the true accuracy of the model due to biases in the data set. In both data sets, out of the 1885 participant, only 6 participants were considered 'user', while 1879 participants were considered 'non-user'. Furthermore, even after the application of SMOTE, the instances only increased to 12 users. Therefore, class imbalance will bias the results unless a new training set is created.

Although there are small discrepancies between annual and monthly drug consumption, the spreadsheets and graphs show similar results with the same drug. Drugs such as alcohol, nicotine, ecstasy, and others showed larger deviation due to their popular usage, while unfamiliar drugs such as semer had similar accuracies because the time differences was not larger enough to affect it. Additionally, some attributes have missing values (?) in KNN and J48 showing how the model was bad which was most evidently from oversampling. Additionally, none of the attributes for ZeroR were recorded because the classifier ignores all of the other attributes except for the predicted attribute. Since ZeroR is used to determine the baseline performance of the model, its sole purpose would be to compare the different accuracies and classifiers.

Other important measures are TP rate, precision, recall, and F-measure. Two different values were displayed to show the measures regarding the classification of 'user'(U) and 'non-user'(NU). Precision describes the number of correct classifications from the original set, while recall describes the correct classifications out of true positives and false negatives. For the majority of classifiers, both the precision and recall measures were higher than 60 percent. F-measure is combination of precision and recall:

$$\text{F-Measure} = 2* \frac{precision * recall}{precision + recall}$$

The majority of F-measures range around 80 to 90 percent. Therefore, altogether, the accuracy, precision, recall, and F-measure evaluate the performance and model to be good.

# 6. CONCLUSION

The study demonstrates good models for testing the accuracy of various drugs using different classifiers based on age, gender, education, country, ethnicity, and personality traits[1]. Additionally, pre-processing techniques on the data set aided in maximizing the accuracy, precision, recall, and F-measure. Fig. 1 and 2 show the averages of the different classifiers on each drug based on the respective data sets.

Although this model was good, it is not the perfect model for classifying different drug consumptions evident from the absurd accuracy from semer and the missing values from KNN and J48 classifiers. Future studies should attempt to develop better models that would retain or improve the different measures, while classifying the various drugs without inaccurate results. Another possible study would be to do a correlational analysis to see which drugs are correlated in their usage, where if one person uses *x* drug, that person is also prone to develop an addiction to *y* drug.

# 7. REFERENCES

[1] Drug Consumption Data Set. <https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29> Last visited November 2018.

[2] Fehrman, Elaine, Mirkes, Evgeny M., Muhmmad, Awaz K., Egan, Vincent, Gorban, Alexander N., “The Five Factor Model of personality and evaluation of drug consumption risk”, pp. 1 – 67, arXiv:1506.069297v2 [stat.AP], Jan. 2017.

[3] Carlson, Neil R., *Foundations of Behavioral Neuroscience* (9th internal ed.), pp. 1 – 519, London, 2014.

[4] Weka 3 : Data Mining Software in Java. <https://www.cs.waikato.ac.nz/ml/weka/index.html> Last visited November 2018.