**Jesse, We Need To Train: A Machine Learning Approach to Predict Personality Impact on Methamphetamine Abuse**

**Karen Li, Andres Rivera, Christian Barbosa**

Department of Computer Science, Northeastern University

{li.kar, rivera.and, barbosa.c}@husky.neu.edu

**Abstract**

Much ink has been expended on the personality traits, sociological factors, and family risk factors that predispose people to substance abuse. While most of these studies have been qualitative, focusing on case studies and self-reported values, we take a machine learning approach to drug risk analysis. In this project we propose three methods to build classifiers that identify the most significant predictors of substance abuse. Input features are age, gender, country, education, ethnicity, impulsivity, neuroticism, extraversion, openness to new experiences, agreeableness, and conscientiousness. However, we focus on the Big Five (OCEAN) traits. Our models—a logistic regression classifier, SVM, and a neural network—predict the likelihood of a person using methamphetamine. Additionally, we briefly analyze this field’s extant literature, explain impact, and give recommendations for future study.

1. **Introduction**

In order to properly allocate resources to harm reduction in affected communities, academics and policymakers alike have studied the personality and socioeconomic characteristics of drug use. Lloyd writes that the main focus of risk research has been the “prediction of the onset of illicit drugs use,” though what should be the priority is the “factors associated with the development of *problem* drug use.” We agree. While people begin misusing drugs for a variety of reasons—teenage angst, a desire to fit in with peers, a lack of economic prospects, etc.—preventive measures are best focused on preventing what keeps them going back to substance abuse. After all, sixty-one percent of meth abusers will relapse within the first year after treatment, with 25% more relapsing in years 2-5; this returns them to re-navigate a patchwork, underfunded, and cynical network of methadone clinics, therapy sessions, and social workers. These underlying factors are what keep meth users returning to the product, not the initial factors that drove them to start using at a younger age.

Additionally, the impact of this field of research is immense. The human and economic impact meth has on families, neighborhoods, and communities are untold. Meth tears apart families, separates parents from children, sends breadwinners to children and hurting a family’s economic prospects. Children whose parents are methamphetamine users are often flagged by CPS and likely to be separated from their parents, a profoundly traumatizing experience. Low-income children, who already are disproportionately likely to have user parents, are more likely to be placed in foster care. This again is traumatizing and disruptive to their schooling patterns; the latter, ironically, is a risk factor for methamphetamine abuse, continuing the cycle once more. Addiction is a disease. In the US alone, methamphetamine abuse spiked from 1.9 million to 5.1 million users between 2017 and 2018. The World Drug Report estimates that the intangible cost of addiction—disability-years lost, child endangerment, hospital bills, and crime— is between 16 and 48 billion dollars. We should examine meth addiction’s causes and risk factors with the same intensity and funding we would allocate to cancer or Alzheimer’s.

In this project, we propose three different classifier models to predict the correlation between a personality trait in the Big Five and methamphetamine use in order to determine if we can classify non-users and users by their personality traits, and if so, which trait is the most effective in doing so. We use a logistic regression classifier, a SVM classifier, and a multi-layered FNN on a dataset of about two thousand user/nonusers from the UCI Machine Learning Repository. Additionally, while we focus on the Big Five (OCEAN) traits, we do evaluate the effect of age, gender, nationality, ethnicity, and education on methamphetamine use.

* 1. **Related Work**

Social sciences have experienced difficulties in conducting studies on problem drug use’s risk factors. Lloyd identifies two main bodies of research: retrospective studies of users’ life histories and backgrounds, and longitutdinal studies. However, the first form runs the risk of conflating causal events for events occurring at the same time of drugs use. For instance, poor school attendance is associated with drug abuse, but it’s unsure whether truancy *led to* drug use or drug users, over time, become truant. Longituinal studies, commonly conducted in the United States, are time-intensive as they follow users for decades in their lives. It’s difficult to put forth legislation in an election cycle when the math backing you up won’t be done for another five years.

Therefore, we turn to machine learning techniques instead. Qiao et. al, who used random forest (RF), extreme gradient boosting (XGBoost), and other models to predict our drug of choice, compared their models’ performance to KNN. They found that LightGBM was the most efficient, with all models having roughly the same accuracy (~70%), but that feature selection is crucial and neuroticism was the most important factor. Similarly, Shahriar et. al atempst to find social factors that lead to substance use, but expand their reach to cover nicotine and alcohol. Though Sharhriar’s feature count is extremely high, which might lead to overfitting, their neural network yields an extremely high accuracy of 93.33%, with ‘age of first drug use’ as the most significant feature. This makes sense; drug users who start early risk adversely affecting neurological development during the critical aldohesenct years, tend to skip classes, and engage in riskier behaviors. Finally, we look at Nath et. al’s work, which utilizes the same dataset to predict binary drug consumption with a 77.2% accuracy.

* 1. **Big Five**

The Big Five personality traits are a proposed taxonomy for personality traits developed by J.M. Digman and Lewis Goldberg in the early 1990s. It lists five personality traits (styled OCEAN or CANOE), each calculated on a sliding scale continuum. The five are listed as follows:

* **Openness**— characterized by a receptiveness to new experiences, adventurousness when it comes to decision making, and imagination and insight. On the low end of the spectrum, people are more traditional and conservative.
* **Conscientiousness**— characterized by high levels of thoughtfulness, organizational and planning skills, and high impulse control. On the low end of this spectrum, people are likely to be procrastinators and dislike routine.
* **Extraversion**— characterized by assertiveness and high amounts of emotional expression, these people are energized by spending time in others’ company.
* **Agreeableness**— characterized by high levels of trust, altruism, empathy, and connection. On the low end of the spectrum, people can be socially unpleasant.
* **Neuroticism—** characterized by high levels of anxiety, irritability, and mood swings. People at the high end of this spectrum may find it difficult to deal with high levels of stress, while those on the lower end may experience a more emotionally resilience attitude to life.

We choose to focus on these traits because of the storied history of literature indicating that they are relevant to becoming a drug user. Fehrman et. al recounts how there is a positive correlation between neuroticism and openness and drugs user; possibly because depressant drugs such as cannabis and opioids reduce anxiety, and well—taking illicit drugs is the ultimate new experience. As stated above, most literature on this issue attempts to evaluate an individual’s risk of illicit drug consumption, then recommend the most efficient changes in the individual’s social environment to reduce that risk. Though no personality test is perfect, OCEAN is sufficient to evaluate a user’s restraint and emotionality, which is why this project works with it.

1. **Technical Approach**

In this section, we give a high-level technical overview of our three chosen classifiers and describe why we selected them.

* 1. **Classifier 1: Logistic Regression**
  2. **Classifier 2: Vanilla Support Vector Machines (SVM)**
  3. **Classifier 3: Feedforward Neural Network (FNN)**

1. **Experimental Results**
   1. **Dataset**
   2. **Ss**
2. **Conclusions and Future Work**
3. **Participant Contribution**

References