**Assigning Students to the Proper Drug Awareness Training**

**HYPOTHESIS:** We can determine the likelihood of college student abuse of drugs based on personality characteristics.

**PILOT STUDY OBJECTIVE:** To be able to use the predictive model to assign incoming college students to one of two drug awareness programs:

1. One day generic drug awareness for those with a low likelihood of substance abuse
2. One week intensive drug awareness and counselling program for those our research indicates are more likely to abuse substances. This program will seek to help identify those vulnerable to drug usage when under stress, and teach them coping skills.

**STUDY FRAMEWORK:** Incoming freshmen will be randomly assigned to one of two groups. One group, the study group, will be tested and assigned to one of the above training programs. The control group will automatically be sent to the one-day course. So, every student gets at least the one-day course and some from the study group get the one-week intensive course. After the first semester, we’ll assess whether the control group and the study group differ in (a) their assessment of training utility, or (b) their drug abuse patterns. If the study group performance indicates that the assessment leading to assignment helped reduce drug abuse, the program will be expanded to include all incoming freshmen the following year.

**DATA SCIENCE PIPELINE PROCESS**

Data Source: Drug consumption (quantified) Data Set located at: <https://archive.ics.uci.edu/ml/machine-learning-databases/00373/>

Ingestion: We will collect the dataset from the repo at UCI, bring the dataset into a PostGRES RDB for cleaning and preliminary analysis. The data doesn’t need normalization

Munging and Wrangling: Very little wrangling is required for this dataset, but for future data we may need more. We’ll identify the features we need, the demographic and personality characteristics.

Computation and Analyses: From the historical data, we’ll compress all data regarding drug usage into one or more key variables to reflect drug usage (yes/no) and, if “yes”, whether a heavy or light user.

Modeling and Classification: We’ll train a classification algorithm to use the personality characteristics to classify whether someone is likely to use drugs, and if so, whether they’ll be a heavy or light user. (Light users and non-users will go into the one-day course.)

Reporting and Visualization: We shall show which demographic and personality factors are most strongly correlated with drug use classification – who is least likely to use and who is most likely to become a heavy user.

Feedback: We’ll reach out to each student after their first semester and determine whether or not they found the training effective and if they are using drugs. In addition, we’ll access student health records to look for objective indicators, such as health, absence record from class, or self-reporting of problems with drugs.