Capstone Project 1 Proposal

Problem/Usefulness of Solution: We still don't know how exactly certain drugs are related to development of mental health-related issues. Drugs such as alcohol, marijuana, cocaine, and hallucinogens are frequently used by college students. Use of these substances is sometimes carried into their young adult lives after they graduate. At the same time, mental health disorders (such as anxiety and depression) begin to develop for certain individuals. If we could establish a strong correlation between psychological distress and substance use, we could alert young adults that use of these substances could have deleterious effects on their mental health. This could have applications in the treatment of mental health disorders as well; people may be delighted to hear that they could alleviate their depression simply by quitting their use of drugs. On the other hand, if we discover that use of substances has little correlation with development of these disorders, we could choose to focus our treatment efforts elsewhere.

Data: The data comes from the Substance Abuse and Mental Health Services Administration. Specifically, it comes from the 2015 National Survey on Drug Use and Health. Because I am mostly interested in the effect that drugs have on young adults, I have chosen to focus on individuals between the ages of 18 and 25. There are several factors that the survey uses to delineate substance use, such as the number of days that the individual has used the substance in the past year. There are two dependent variables that I want to focus on in particular: the survey has devised its own sophisticated metrics to indicate whether an individual is psychologically distressed, in addition to whether the individual has had a major depressive episode in the past year. The data comes with a codebook that describes how the data was collected and how certain

metrics were calculated; it also clarifies the meaning of all ambiguous numbers within the dataset. Hopefully, this will make data cleaning a relatively simple process.

Problem Approach: This is a supervised, classification problem. All the inputs we are using have labels. Moreover, the dependent variables that we are trying to predict are both discrete 'yes' or 'no' answers. I will use predictors involving the frequency of use of alcohol, marijuana, cocaine, and hallucinogens (LSD and mushrooms). I will also use predictors that the survey has included to indicate clear abuse of these drugs. We are trying to predict whether or not the individuals have significant psychological distress (indicated by a score greater than or equal to 13) and/or a major depressive episode in the last year.

Deliverables:

- Code for:
    - Data wrangling (acquisition and cleaning)
    - Initial exploratory data analysis
    - Statistical analysis
    - Machine Learning modeling
- Written report
- Presentation and Slide Deck