Contact Info

Deloitte Prompt

Our client at the NIH wants to build a program to address drug use among teenagers/young adults in the US. They are asking you to use existing data to understand factors that lead to drug use and make recommendations for the program, as well as help them understand how and where they should start to roll out these programs.

Overarching Question

What are the top <u>(fill number)</u> external factors that contribute to drug addiction (whether the use of drugs affects other aspects of their life + frequency of use).

Important Features

- Age Category: Filter to 12-25 years old
- Affecting emotional health
- Affecting school performance
-

Week 1 Tasks

- Each person selects 20+ features that they think are relevant and do EDA to determine/confirm this
- Drop any columns within your partition that are majority missing values
- Define how we want to deal with types of missing values we find (ex: 90, 99, etc)
- If we never heard of a drug, remove it from the columns

First 660 - Nate
Second - Calvin
Third - Mohit
Fourth - Zhen
Fifth - Gino

Rough Outline for Reference

- Explore the columns, filter out the ones that immediately are irrelevant (age too old, etc)
- Select about 100 columns out of the 800+ to work with
    - Drop columns that too many missingness within (>50%)
- Visualize Variables and correlation using plotly or matplotlib
- Then we can start to think about how we want to design a model later on
- Create a boolean column (0 = not likely to abuse, 1 = likely to abuse)
- Normalize the data between 0 and 1
- Find k most significant factors that lead to drug abuse
- Use model to output likelihood to abuse drugs based on these factors

What is Drug Use and Addiction?

Use Missingness Testing

- Some people are not going to answer truthfully to this