Project 1 Questions

Question 1

Perform exploratory data analysis. Create visualisations that illustrate the relationship between substance use and the predictor variables.

Question 2

2.1

Build a classifier that predicts whether an individual's substance use level will be "high" or "low" based only on the person's background (age, gender, education, etc) and on the personality measurements.Use only the first 1500 rows to train the model.

2.2

Make predictions on the remaining rows, 1501:1885 and create a table (or confusion matrix) to compare the predictions to the truth. What is the accuracy?

Question 3

3.1

Use another method or combination of methods to solve the same problem (namely: predict the UseLevel based on some or all of the predictors in the first 16 columns). If you like, you can make a copy of the data frame in which you convert the non-numerical columns to numerical columns. (Some of the methods we used in the course don't work for factors or character predictors, like knn).

3.2

Estimate how well you would expect your method to perform on new test data.

Question 4

4.1

Create a variable that is "yes" or "no", representing whether the patient reports that they ever used heroin. Use a random forest to predict whether someone has ever used heroin. For predictors, use the first 16 columns (as before), and now also the other illicit drugs. Think about whether you should use the any, UseLevel or severity columns as predictors.

4.2

Based on your interests and your EDA, find another classification, regression or feature selection question that you can ask with this dataset. For example, you could see whether you can predict reported use other drugs. You could create new outcomes like whether someone uses both crack and cocaine, or either crack or cocaine. You could try to predict the level of alcohol consumption based on the personality measurements. There are many interesting options.

Build an appropriate machine learning model to perform this task.

Why did you choose the method you did? How does it perform? Show the results with a table or plot and an estimate of the loss.