Homework # 4 (200 points)

Due on May 07, 11:59pm EST

For this assignment, we will be working with a very interesting mental health dataset from a real-life research project. All identifying information, of course, has been removed. The attached spreadsheet has the data (the tab name “Data”). The data dictionary is given in the second tab. You can get as creative as you want. The assignment is designed to really get you to think about how you could use different methods.

1. Please use a clustering method to find clusters of patients here. Whether you choose to use k-means clustering or hierarchical clustering is up to you as long as you reason through your work. Can you come up with creative names for the profiles you found? (60)

2. Let’s explore using Principal Component Analysis on this dataset. You will note that there are different types of questions in the dataset: column: E-W: ADHD self-report; column X – AM: mood disorders questionnaire, column AN-AS: Individual Substance Misuse; etc. Please reason through your work as you decide on which sets of variables you want to use to conduct Principal Component Analysis. (60)

3. Assume you are modeling whether a patient attempted suicide (column AX). Please use support vector machine to model this. You might want to consider reducing the number of variables or somehow use extracted information from the variables. This can be a really fun modeling task! (80)