ECON860 Data Analysis For Economics(Fall 2023)
Final Exam 1
Due: 17th December 2023 (1:00pm)

Name: _____
UID: _____

---

1. You are given a dataset with 21644 individuals. The dataset contains the answers to a questionnaire with 40 questions to evaluate their personality traits and a measure of the math ability of the individuals. Your task is to cluster these individuals into groups and relate the personality traits to their math ability.
   (a) The questionnaire is similar to the "Big Five Inventory" in the lecture (But not the same, so they do NOT necessarily correspond to the Five personality traits we mentioned in the lecture). However, you do not have the cookbook, so you do not know which questions correspond to which personality traits.
   (b) The answers are on a scale of 1 to 5. If the individual refuses to answer that particular question, the value would be 0.
   (c) Use factor analysis to get a measure of several personality traits from the questionnaire. Notice that this questionnaire is not the same as the "Big Five Inventory", so you may not have exactly five traits. You need to follow the procedure introduced in the lecture to find out what is the suitable number of traits (factors).
   (d) Use the personality traits to cluster the individuals. You may use KMean clustering, Gaussian mixture model, or any other unsupervised learning techniques.
   (e) Which algorithm gives you a better result? Explain your answer or explain why it is not possible to evaluate which algorithm is better.
   (f) Use the personality traits to predict the math ability of the individuals. You may use linear regression, logistic regression, or any other supervised learning techniques.
   (g) Which model gives you a better result? Explain your answer or explain why it is not possible to evaluate which algorithm is better.
   (h) Now you are assembling a team of 30 individuals to work on a math project. You want to choose the individuals with the best math ability. However, you cannot choose those people who are in the original dataset. You can only choose 30 individuals from the population. Also, you do not have the resources to do a math test nor to collect 40 answers from those new recruits. You can only collect 20 from them. Which 20 questions should you choose among the 40 questions in the original questionnaire? And how will you use the information you collect from this new questionnaire to assemble your team? Explain your answer.
   (i) Suppose instead of a math project, you are assembling a team of 30 individuals to work on a project that requires a variety of different personality traits. Which 20 questions should you choose among the 40 questions in the original questionnaire? Is your answer different from the previous question? Explain your answer.
   (j) You must hand in your homework via Github. Create a repository named "ECON860_final". In your repository, you should have the code and a .gitignore file. You should also include a file named README, which includes step-by-step instructions on how to run your Python code to collect the data you collected. This is especially important if you have multiple Python files. You should have a written answer committed in your repository. It can be included in the README file or it can be a separate file.
   (k) You can commit and push to Github as many times as you like. Only your last commit before the deadline is graded. I can read your previous commits, but they will not be graded.

(l) It is more important to hand in partial work than to not hand in anything. For example, if you are not able to get a nice set of personality traits in earlier parts, you can still hand in the code you used to cluster the individuals. Also, you can use whatever imperfect traits you have to predict the math ability of the individuals. You will get partial credit for the parts you have done.

(m) Bonus question: It is also possible to use the questionnaire answers themselves to predict the math ability of the individuals. Explain whether this is a good idea or not and why.