MA678 Midterm Project Proposal

Personal Statement:

My career plan after graduation is to work as a data scientist in public health related filed. I think it is very important to practice how to analyze medical health related data because when I look at some medical related data, I think I understand what each column is about in medical terms, but I don’t understand the correlation or interaction between each variable and how they can affect outcome. Therefore, by choosing this heart failure prediction dataset can prepare me with fundamental workflow of how to analyze public health related data in the future.

Question:

* Out of all 12 variables, which variable(s) can affect the prediction of heart failure (significantly)?
* This dataset contains many categorical variables, how can I define them in a better way?

Data source: https://www.kaggle.com/fedesoriano/heart-failure-prediction

Proposed timeline of work:

* Data Processing: 11/04/2021 -- 11/09/2021
* EDA: 11/09/2021 -- 11/16/2021
* Modeling and Validation: 11/16/2021 -- 11/23/2021
* Write up: 11/23/2021 -- 11/28/2021
* Review and edit: 11/28/2021 -- 12/02/2021