Special Topics: Data Analytics and Visualization in Healthcare
CSCI-GA.3033-096 (19635)
Lab assignment 5

Instructions:

Part I

- Investigate a supervised learning classifier (not reviewed in lectures). It can be an algorithm your group will use in the final project.
- Identify the advantages and disadvantages of the selected classifier.
- Construct a model in either R or Python to investigate the relationship between US acute-care hospital characteristics and Clostridium difficile infections (CDI). Use the dataset hospitals_infections.csv. The dataset description is as follows:
    - provider_id: The code number assigned to an individual inpatient hospital.
    - hospital_ownership: The category that defines the status of each hospital's ownership type.
    - emergency_services: Whether a hospital provides emergency services or not.
    - cpcd: Clinical process of care domain score.
    - pecd: Patient experience of care domain score.
    - hsbp: Spending per hospital patient with Medicare (Medicare spending per beneficiary [MSPB]). Specifically, a hospital's MSPB measure is calculated as the hospital's average MSPB amount divided by the median MSPB amount across all hospitals.
    - readmission: Hospital-wide 30-day readmission rate.
    - c_diff_compared: Healthcare-associated clostridium difficile infections measures hospital-level results.
- Evaluate your model.
- Explain in a MS Word/PDF document the analysis of your results.

Part II

- Select a current (no more than 5 years), journal/conference article pertaining to machine learning in healthcare.
- The paper should have at least one supervised machine learning method.
- Summarize the key ideas discussed in the paper.
- Critique the overall content of the paper.
- Include a reference to the paper in your write-up using a citation format.

Notes:
- This assignment is individual.
- You need to submit 2 files (the .R or .ipynb file and the MS Word/PDF document).
- Please do not copy/paste information. Plagiarism will be penalized.