MSiA 421: Data Mining
Assignment -2

Individual Assignment (100 points)

Instructions:

- Submit the paper review as a word or pdf file.
- Submit code as a Python notebook (.ipynb) file along with the HTML version.
- Write elegant code with substantial comments. If you have referred to or reused code from a website add the links as reference.

1. Paper Review – Following the guidelines review any one of the technical papers from Group2 **(30)**

2. Generate random multidimensional (n=1000, D >= 15) data using *sklearn*. **(30)**

    - Build a K-means function from scratch (without using *sklearn*) and make assumptions to simplify the code as needed.
    - Use the elbow method to find an appropriate value for k
    - Use the silhouette plot to evaluate your clusters
    - Re-cluster the data to see if you can improve your results
    - Perform PCA on the original dataset and retain the most important PCs.
    - Run K-means on the PCA output, compare results with respect to cluster quality and time taken

3. Data mining and Cluster analysis of the following dataset **(60)**

    U.S. Chronic Disease Indicators (CDI)
    Data Dictionary

    This dataset relates to chronic disease indicator in the US and a set of behavioral and demographic factors such as race, gender, and location.

    As a data science consultant, your goal is to mine the dataset and extract meaningful insights for your clients in the health care industry. The course of action is as follows:

    - Review and understand the structure of the data (details in data dictionary)
        - Disease indicators, questions, demographic data, etc.

    - Data Transformation
        - Retain a limited subset of the data for the year 2020
        - Retain a limited subset of Topics (e.g., Alcohol, Diabetes, Cancer, etc.)
        - Group and summarize data as needed for subsequent analysis
        - Convert values in the StratificationCategory1 into new columns to help with analysis

- Exploratory Data Analysis (10)
    - Create statistical summaries
    - Create boxplots, correlation/pairwise plots
    - Perform basic outlier analysis

- Clustering (25)
    - In a few lines create a plan that describes the 3-4 questions that are suitable for cluster analysis
    - List the various clustering algorithm(s) you'd use and why:
        - E.g., K-means, K-medians, K-modes, Hierarchical methods, DBSCAN, etc.
    - Apply the above algorithms to the filtered dataset based on your plan
    - Report on the quality of the clusters, pros/cons, and summarize your findings

- Bias/Fairness Questions (25)

    Data
    - In the dataset under study, from a bias/fairness (b/f) perspective, there are 2 sensitive features: race and gender.
    - Analyze the data by a combination (2) of features (sensitive and other). Example features to include in the analysis: location (county, state), and other features you consider relevant. Though these features may not be considered sensitive they can be a proxy for sensitive features.
    - Determine feature groupings that are relevant for your analysis and explain your choices.
    - Do you detect bias in the data?
    - Present the results visually to show salient insights with respect to to bias.
    - Based on the EDA and your project objective, develop a hypothesis about where b/f issues could arise in the modeling (cluster analysis).

    Modeling
    - Based on your hypothesis, assess the fairness of your model/analysis by applying the fairness-related metrics that are available in any of the following tools: Python *Fairlearn* package, R *Fairness*/*Fairmodels* package, or other similar tools.
    - Explain the reasoning for the groups that you selected for the fairness metrics.
    - Compare the fairness metrics for the different groups.
    - If you developed multiple models compare the fairness metrics for the models.
    - Comment on the results.
    - Suggest how the bias/fairness issues could be mitigated.
    - Present the results visually to show salient insights.

Note: In the Fall Quarter you attended lectures on Bias/Fairness. Additionally, the following is a useful resource for analyzing b/f in data and modeling: Fairness & Bias Metrics