**MSIS 5193**

**Assignment 2 – Due Sept20th (11:59pm – nighttime)**

**Submit on Canvas**

**Assignments are to be your own work. If you have questions, you can ask your TA.**

Please finish the following tasks using the 2017CHR\_CSV\_Analytic\_Data-new.csv file in the “datasets” folder on canvas. Please submit your code and screenshots of the output of each question in the word file to submit to Canvas.

**Task 1.** Use 2017CHR\_CSV\_Analytic\_Data-new.csv file to complete the following tasks

1. Check if any columns have missing value (show the result). If there are columns having missing value, impute the columns (with missing values). If there is no missing value, leave the data frame as is. (5 points)
2. Drop (remove) the identifier columns: 5-Digit FIPS Code, statecode, countycode, county (5 points)
3. Use z-score normalization to normalize these columns: Poor physical health days Value, Poor mental health days Value, Food environment index Value (5 points)
4. Create a new column “Diabetes-level” by coding the “Diabetes Value” into four groups, and label them as low, median low, median high, and high (Hint: refer to data coding – categorizing coding) (20 points)
5. Apply feature selection to find the top 5 features relevant to “Diabetes-level”. (Hint: refer to feature selection, use new column “Diabetes-level” as target column y, the rest columns (excluding “Diabetes Value” and “Diabetes-level”) as input X)) (20 points) (optional)