CSC 401, Modeling Projects Directions, Fall 2024 Name_____

Instructions: The goal of these projects is to give students an opportunity to apply both custom-built functions, and built-in functions to an analysis of real-world datasets. The goal of machine learning is to automate predictions with the highest degree of accuracy possible. A thorough analysis of the models including metrics and visualizations will be expected whether explicitly stated in the directions or not.

Think about your analysis and plan before you code. Not every dataset is amenable to every method of analysis. It is important to do some initial exploration of the data before building your models. These explorations and testing of the data should be included in your written analysis, at least a short summary of them, and any interesting results. Do not wait until the last minute to start these projects. Initial data exploration does not need to wait.

The expectations for these projects is a minimum of three pages of written analysis (this can include graphs, but should not be only graphs). If you are doing a custom implementation, you may include small chunks of code to illustrate, but do not include snapshots of all your code. If a small chunk of output from functions is important, then you may include that, but do not just include R outputs without explanation. The explanations matter more than the specific outputs or code. Graphs should be accompanied with an explanation of what they show, and what they mean for your analysis. You should aim for professional communication of your analysis.

Add comments to your code to explain what you are doing to me, and to yourself later when you haven't looked at it in a while. All your code should execute. The only thing I should have to change to make it run is the location of the data file.

If you include mathematical function expressions, use an equation editor (there is one built into Word, or there are others online to make your equations look nice).

Cite sources. If you use code you obtained from online sources, that is fine, but cite the source appropriate (while I don't have a specific format requirement, I suggest using APA or another style you are familiar with). Include references for any information you use that is not common knowledge. For some datasets, original sources are provided in a Dataset Sources file posted in the course. You should reference the original source. Those sites may also provide additional information about the variables that may be of some use to your analysis.

Submission: Write up a work document describing your code and the results obtained. Include with your submission your R code file. (You may use Quarto to construct your reports, but I don't need to qmd file.)