CS/OREM 5/7331 Data Mining
Project 3: Classification

Due Date: See Canvas

UG Points: out of 100

GR Points: out of 110

Please submit your report in **PDF format.** If you want to submit code, then please submit it in an additional file containing sufficient comments to make it understandable.

Experts are talking about the possibility of a fourth Corona virus wave hitting the world and the US (see https://www.aarp.org/health/conditions-treatments/info-2021/covid-4th-wave.html). Use the data that you have identified for projects 1 and 2 and, if available, new data. We would like to classify counties or states in high/low or low/medium/high risk in terms of how affected they would be by a fifth waive.These results can be used to prepare the infrastructure and plan possible interventions (e.g., mask mandates, temporarily closing businesses and schools, etc.). Early interventions based on data might dampen a severe outbreak and therefore save lives and shorten the length of necessary closings.

Follow the CRISP-DM framework

1. Data Preparation [40 points]

- Define your classes (e.g., more than x corona-related cases or fatalities per a population of 10000 per week). Explain why you defined the classes this way (maybe you want to look at the data first).
- Combine files as needed to prepare the data set for classification. You will need a single table with a class attribute to learn a model.
- Identify predictive features, create additional features, and deal with missing data (for classification models that cannot handle missing data).

2. Modeling [50 points]

- Prepare the data for training, testing and hyper parameter tuning.
- Create at least 3 different classification models (different techniques or using different class variables) using the training data. Discuss each model and the advantages of each used classification method for your classification task.
- Assess how well each model performs (use training/test data, cross validation, etc. asappropriate).

3. Evaluation [5 points]

- How useful is your model for your stakeholder? How would you assess the model's value if it was used?

4. Deployment [5]

- How would your model be used in practice? What actions would be taken based on your model? how often would the model be updated? Etc.

Gradates / Exceptional Work [10 points]

Examples: Use more classification algorithms, insightful visualization of results, in-depth explanation why one method works better than another, and using and explaining a method not covered in class.