CSC 400, Exam #2, Spring 2024 Name _____

Instructions: Answer each question thoroughly. For questions in Part 1, use the work you did at home to answer the questions. Be sure to answer each part of each question. In Part 2, report exact answers unless directed to round.

Part I:
Use the work you did at home to answer these questions on the wine dataset.

1. How did rescaling your variables impact the accuracy of your predictions in the k-means model (or any of the clustering models)?

2. What was the accuracy of your (best) k-means (semi-supervised) model with k=3?

3. What was the accuracy of your (best) spectral clustering (semi-supervised) model with k=3?

4. What was the accuracy of your (best) DBSCAN (semi-supervised) model with three clusters (this might be 2 clusters plus noise)?

5. Describe the hyperparameters you had to set to obtain improvements in your models for spectral clustering and DBSAN.

6. What was the accuracy of your (best) mean-shift clustering (semi-supervised) model with k=3?

7. Given the four methods you tried here, which algorithm was the most successful (and efficient) and which was the least successful (or least efficient)?

8. For your outlier analysis, which variables contained outliers based on looking at boxplots?

9. For the Ash variable, which observations were potential outliers (based on the boxplots)?

10. Which two variables appeared to be the most non-normal?

11. Which statistical test did you perform on the data to test for possible outliers? Why this test?

12. Based on the statistical test you chose, how many of the outliers flagged by the boxplots (for all variables), still considered outliers after the test? Give the observation numbers.

13. Based on the outlier analysis you did, how would you proceed? Would you remove the outliers? If so, based on which part of the analysis? If not, why not?

For the questions that follow, use your analysis of the beersales data.

14. Is the original beersales data stationary? Or does it show a trend or seasonal pattern?

15. After decomposing the beersales time series, describe the trend? Is it increasing or decreasing? Is it linear or non-linear?

16. Describe the differences between the standard decomposition output and the LOESS decomposition?

17. How many differences did you have to take to obtain a stationary series? Explain how you knew when to stop.

18. How many lags should be included in your ARIMA model based on the ACF graph? Based on the PACF graph?

19. What ARIMA model did you settle on? What were the parameters? How did you determine the best fit?

20. For your exponential smoothing model, what was the (approximate) value of the best smoothing parameter?

Part II:

21. What are some potential drawbacks of using (non-linear) regression methods for irregular time series analysis (for interpolation).

22. Why is cross validation essential for developing robust models?

23. What is image segmentation? How can we use it to process images?

24. Give three advantages of being able to incorporate spatial information into your data mining analysis.

25. How can clustering be used to identify potential outliers or anomalies? How do these outliers differ from more traditional statistical methods?

26. In image processing, pixels in the image are typically turned into an array of pixel values (these may be in just one color or multiple colors). Describe some examples of feature engineering in this context. You may use a specific example, such as character recognition, as the basis for your answer.

27. How does the hierarchical clustering algorithm work?

28. In the at-home portion of the exam, we looked at how to use clustering methods in a semi-supervised way to create a classification model. How do clustering and classification differ?

29. In the at-home portion of the exam, we applied DBSCAN. Explain the general algorithm steps?

30. Describe the algorithm steps for Fuzzy-C Means clustering.

31. Why does rescaling improve performance in algorithms that use a Euclidean distance metric?

32. Give two examples of other distance metrics that could be employed instead of the standard Euclidean distance metric.