CSC 400, Exam #1, 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. Based on your correlation table, which two variables have the highest Pearson correlation? What is the correlation value?

2. Based on your correlation table (or graphs), which two variables have the highest Spearman correlation? What is the correlation value?

3. Describe how your correlation plots for the Pearson and Spearman correlations differ, qualitatively. What do you notice about how the plots differ?

4. In your pairplot, do any of the scatterplots appear to be particularly nonlinear? If so, which pair? Describe the shape and give the correlation value. If you were building a model of the data, is there a transformation you could make to try to improve the fit based on the scatterplot?

5. Does the pairplot suggest that there may be outliers in the data that may need to be analyzed and removed? If so, in which variable(s)?

6. For your decision tree, what rule makes the first split?

7. For your decision tree model, what is the accuracy of the model?

8. For your KNN model, which value of k produced the highest accuracy?

9. What was the confusion matrix for k=5 (for KNN)?

10. Did rescaling the variable improve the accuracy of your model? Why or why not?

11. After applying the LDA algorithm to the data, what was the accuracy of that classification model?

12. Given the available data, comparing the Decision Tree model, the KNN model and the LDA model, which model would you choose and why? You may want to consider more than accuracy, for example, interpretation may be important if the goal is to understand the classification scheme.

For the questions that follow, use the your analysis of the Pima Indians Diabetes data.

13. In your SVM model with a linear kernel (the default), what was the accuracy achieved with this model?

14. Were you able to improve the accuracy with a nonlinear kernel? If so, which one, and what was the improvement?

15. Compare the results of the SVM model with your simple neural network model. Was the accuracy better, worse or similar?

16. What is the confusion matrix you obtained from the XGBoost model?

17. Of the three models applied to the diabetes data, which model performed the best and why?

The questions that follow refer to the Income data and association rule analysis.

18. Based on support, what is the top rule? What are the lift and support values?

Part II:

19. Give three examples of applications for neural networks aside from regression and classification.

20. Convolutional neural networks are particularly useful for which types of applications? Give at least two examples.

21. Long Short-Term Memory networks are a type of Recurrent Neural Network (RNN) designed to address what type of issue? To what kinds of problems are they typically applied?

22. What kind of data or structures are graph neural networks designed to work best on? Give examples.

23. What are three factors to consider when choosing a neural network design?

24. Some classification methods are designed around binary classification, such as logistic regression. If you wanted to apply such a model to data with three or more classes, such as the wine dataset from the at-home portion of the exam, what would you need to do? Describe the process (in some detail).

25. What advantages and disadvantages are there to applying ensemble methods to a classification (or regression) problem? Give at least two of each.

26. How do boosting and bagging differ from each other? Explain both.

27. What simple assumption does Naïve Bayes make that earns it the name “naïve”?

28. K-Means is technically a clustering algorithm that is often used for classification. Explain how that works and why it is different from traditional classification techniques.

29. What is the main difference between data mining and machine learning? What do they have in common?

30. What is the purpose of the Gini index in a decision tree model?

31. In association rule mining, what is the purpose of a parallel coordinates plot?