CS 699
Assignment 3

Note: R is recommended since that is what is taught in class, but you may use any software you choose. As explained in submission guidelines on the next page, please provide your “final answers” in a document and provide a separate file containing your code or other detail that explains how you obtained your answers.

Problem 1 (25 points) Consider the following dataset:

ID	A1	A2	A3	Class
1	Medium	Mild	East	Y
2	Low	Mild	East	Y
3	High	Mild	East	N
4	Low	Mild	West	N
5	Low	Cool	East	Y
6	Medium	Hot	West	N
7	High	Hot	East	Y
8	Low	Cool	West	N
9	Medium	Hot	East	Y
10	High	Cool	East	Y
11	Medium	Mild	East	Y
12	Low	Cool	West	N

Suppose we have a new observation *X* = (*A1* = Medium, *A2* = Cool, *A3* = East). Use the formulas presented in the Blackboard Module 2 reading (and reinforced during live class) to predict the class label of *X* using Naïve Bayes classification. You may use code to help you count how many you have of each level of each feature and do the math you need to do, but **you may not** use a Naïve Bayes function in any library, package, or other software. Reminder: your calculation must include probability information obtained from each of the three attributes. Be sure to submit your code with the assignment.

Problem 2 (25 points). Use the same dataset as in Problem 1. Calculate the information gain (based on change of entropy) for A2 and A3 and determine which of these two is better as the test attribute at the root. You may use code to help you with this task, but **you may not** use a decision tree algorithm in any library, package, or other software. Be sure to submit your code with the assignment.

(Assignment continues on the next page)

Problem 3 (20 points). Use the dataset *autism-adult.csv*. This data set has 704 observations and 19 features.

(1). Generate training and holdout partitions on the data set. Use 1/3 of the data in the holdout.
(2). Fit a Naïve Bayes model. Use the model to make predictions on the holdout data. Generate a confusion matrix and measure the model's performance using one or more appropriate metrics of your choice.
(3). Does this seem to be a good model? Discuss why or why not.

Problem 4 (30 points). Use the dataset *restaurantdata-small.csv*. This data set has 8368 observations and 6 features. The numeric value to be predicted is named Revenue.

(1). Generate training and holdout partitions on the data set. Use 1/3 of the data in the holdout.
(2). Fit a regression tree model with Complexity Parameter of 0. Use the tree to make predictions on the holdout data. Measure the tree's performance using one or more appropriate metrics of your choice.
(3). Prune the tree by implementing the minimum cross-validation error method. Use the pruned tree to make predictions on the holdout data. Measure the tree's performance using one or more appropriate metrics of your choice.
(4). Compare the pruned tree to the full tree. Which, if any, seems to perform better? Discuss.
(5.) Calculate variable importance for the pruned tree. What are the most important considerations if you want to start a restaurant that generates a large amount of revenue?

Submission Guidelines:

- **Submit the solutions in a single Word or PDF document and upload it to Blackboard. You should separately upload the code or another document that explains how you arrived at the answers. Make sure each filename includes your name.**
    - **If you need more than 3-4 separate files, you might ZIP/RAR these files. Please still upload the Word/PDF separately. The separate Word/PDF greatly simplifies the facilitator's ability to comment on your assignment.**
- **Your submission should be organized well and easy to follow. Clearly list which question you are answering and provide the answer requested (one or more numbers, a screenshot, a plot, an explanation, whatever the question asks you to do). If a question contains multiple parts, you should clearly provide your answer for each part.**
- **If your facilitator tells you to submit the files differently than the above guidelines, you are expected to respect your facilitator's wishes starting on the next assignment.**

- **Facilitators can deduct up to 20% if you fail to follow these requirements (more if the questions are not actually answered).**
- **Facilitators can deduct 5% for each day the assignment is late.**
- **Unless your facilitator or the professor agrees, your assignment will not be graded if it is more than 3 days late (e.g., no credit will be given after Friday at 6 AM Boston time). The professor will usually ask the facilitator to make the decision but in rare cases (<1% of the time) has overridden a facilitator. Do not expect the professor to override in most cases.**