CS 211 PA #4

For this assignment, you will write a program that builds a decision tree based on a supplied CSV file. This document does not explain how to construct decision trees. For that, it is suggested that you consult the other files included in this folder as well as the lecture notes on the day in which decision trees were covered.

Program Flow

Your program should provide the user with the following options:

1. Build decision tree from file

When the user selects this option, your program will prompt the user for a CSV file. The first row of the CSV file is expected to be a header describing each column. All subsequent rows will be data. Upon parsing the file, your program will prompt the user for the outcome variable. All other data rows will be considered potential factors (predictor variables) for your decision tree. Having established both factors and outcome, your program will then construct a decision tree.

2. Write decision tree to file

Your program will write the decision tree constructed in option #1 to a text file. Each line in the text file will represent a single node. The line will be comprised of the node's edge value (e.g. Strong), name (e.g. Wind), and the number of the node's children N (e.g. 2). This information will be separated by a single pipe "|". Each child will follow its parent as is done when representing a tree structure using a vector format. For example, the file on the left corresponds to the decision tree on the right.

|  |  |
| --- | --- |
| NULL|Outlook|3  Sunny|Humidity|2  Overcast|Yes|0  Rain|Wind|2  High|No|0  Normal|Yes|0  Strong|No|0  Weak|Yes|0 |  |

3. Predict outcome

Your program will attempt to predict outcomes using the decision tree constructed in option #1. Your program will prompt the user for a CSV file to open. The first row of the CSV file is expected to be a header describing each column. All subsequent rows will be data. Upon parsing the file, your program will prompt the user for the outcome variable. All other data rows will be considered potential factors (predictor variables) for your decision tree. Note that unlike option #1, the outcome variable for this task is used to judge the accuracy of your decision tree. You will output the results of your prediction to a separate CSV file specified by the user. This CSV file will look exactly the same as the input CSV file obtained earlier except it will have an additional "prediction" column added to the end. This column will contain the predicted outcome determined by your decision tree.

If, during your prediction, you encounter a previously unseen predictor value (e.g. Wind -> "Very Strong"), you may halt prediction for that given row of data.

4. Read tree from file

Given that menu option #1 will generate the same tree on the same dataset, it is wasteful to continually reconstruct the tree. This menu option will build an in-memory representation from a user-specified file that matches the format of the file you generate in menu option #2.

Header Comment, and Formatting

* Be sure to modify the file header comment at the top of your program to indicate your name, student ID, completion time, and the names of any individuals that you collaborated with on the assignment.
* Remember to follow the basic coding style guide. For a list of basic rules, [see my website](http://adamcarter.com/teaching/cpts121/style) or examine my example files from previous assignments and labs.

Reflection Essay

In addition to the programming tasks listed above, your submission must include an essay that reflects on your experiences with this homework. This essay must be at least 350 words long. Note that the focus of this paper should be on your reflection, ***not*** on structure (e.g. introductory paragraph, conclusion, etc.). The essay is graded on content (i.e. it shows deep though) rather than syntax (e.g. spelling) and structure. Below are some prompts that can be used to get you thinking. Feel free to use these or to make up your own.

* Describe a particular struggle that you overcame when working on this programming assignment.
* Conversely, describe an issue with your assignment that you were unable to resolve.
* Provide advice to a future student on how he or she might succeed on this assignment.
* Describe the most fun aspect of the assignment.
* Describe the most challenging aspect of the assignment.
* Describe the most difficult aspect of the assignment to understand.
* Provide any suggestions for improving the assignment in the future.

Deliverables

You must upload your assignment through Canvas no later than midnight on Monday, April 8, 2019.

Grading Criteria

Your assignment will be judged by the following criteria:

Reflection essay (10pts)

* Your reflection meets the minimum requirements as specified earlier in this document.

Tier 1 (60 points)

* You complete menu option #1 (generate decision tree) and #3 (predict outcomes)

Tier 2 (30 points)

* You complete menu options #2 and #4

Tier 3 (20 BONUS)

* You apply your decision tree to other real-world data that you find. Be sure to include the source CSV file so that I can check your output.