**Programming Assignment-5: Decision Tree Part I**

**First Name1: \_\_Richu\_\_\_\_\_ Last Name1: \_\_\_\_Mathew\_\_\_\_\_\_\_**

**First Name2:\_\_Andrew\_\_ Last Name 2: \_\_\_\_\_\_\_KING\_\_\_\_\_\_\_\_**

**Introduction**

In this programming assignment, you will implement two fundamental functions, entropy and information gain, required to build a decision tree. A dataset is provided.

**1: Graded Exercises**

Different from previous programming assignments, some examples are provided in the Jupyter notebook, which are built-in pandas and numpy functions needed to implement entropy and information gain.

After you have learned from these examples, the graded exercises will prepare you for the implementation of the two fundamental functions.

**2: Entropy**

You will implement the entropy function that returns the entropy for a given column vector.

**3: Information Gain**

You will implement the information gain function that returns the information gain after splitting a dataset by using an attribute *a*.

**4: Experiment**

1. What’s the entropy before splitting the dataset

**Ans**: 0.9965700521510767

1. What is information gain for the following attributes, 'wesley', 'romulan', 'poetry', 'honor', 'tea', 'barclay'

**Ans**:

|  |  |
| --- | --- |
| **Attributes** | **Information Gain** |
| 'wesley' | 0.1661272240561006 |
| 'romulan' | 0.14462841188220366 |
| 'poetry' | 0.1596491024146185 |
| 'honor' | 0.16666421278319188 |
| 'tea' | 0.1915245673772541 |
| 'barclay' | 0.16029078947930264 |

1. Which attribute you should choose to split the dataset?

**Ans**: I would choose tea to split the data because it has the highest information gain.

**Submission:**

* **Rule1:**
  + If you work with a partner, please name your zipped file as follows:

PA5\_P1\_LNAME1\_LNAME2.Zip for folder and PA5\_P1\_LNAME1\_LNAME2.docx for a word document, i.e., the file names should include both LAST NAMEs.

* + If you work on your own, the format should be

PA5\_P1\_LNAME.Zip for folder and PA5\_P1\_LNAME.docx for a word document.

* **Rule2:**
  + Put your FULL names whether working in a group or individual in the word document that answers all the questions in **4: Experiment**. Answers to Questions 1 – 3 can be found in the Jupyter Notebook.
* **Rule3:**
  + **EVERYONE** in the class should submit this Assignment, which should provide all files (like test excel files etc.. ) that are necessary for the execution of code in the submission folder.
* **Rule4:**
  + Please submit your Jupyter notebook as well as the Word document.