Extra Credit Assignment

For MIS2502

**Overview and Purpose:**

The goal of this project is to provide students with additional hands-on experience in data analysis and reinforce the concepts and methods covered in class.

For this project, students should find a new suitable dataset on the internet (that was never used in the class before) and apply Decision Tree analysis to build the prediction of the outcome variable. The process should be very similar to the regular assignment on Decision Trees (i.e., start with the same Jupyter Notebook) but applied to a new dataset (e.g., <https://opendataphilly.org/>) .

**Requirements:**

To successfully complete this assignment, students must:

1. Describe the Data: Select a dataset and describe the outcome variable and features for prediction. Explain how the outcome variable relates to the features and what insights can be gained from analyzing the data.
2. Find the Best Value for Minimum Split: Use the decision tree algorithm to find the best value for the minimum split for the dataset. Explain your reasoning for selecting the optimal value.
3. Find the Node with the Highest and Lowest Probability: After building the decision tree model, identify the node with the highest and lowest probability. Explain what insights can be gained from analyzing these nodes and how they relate to the outcome variable and features.

**What to submit to Canvas:**

* 1. Your dataset.
  2. Your Jupyter Notebook.
  3. A Word document with answers to questions 1-3.

**For all students:** Everyone who successfully completes the assignment will receive **0, 25 (average), 50 (outstanding) points extra credit** added to the last assignment based on completeness and correctness.

**For MIS majors**, you will get 25-50 professional achievement points when you submit your work through community website:

**When you are ready to request your PRO points, do the following:**

1. Make sure that you have uploaded your work to the canvas class server.
2. Log in to the <https://community.mis.temple.edu/> site. Open your Dashboard. Open “Professional Achievement Points” and “Add New”. Select the option to submit a PRO point request relevant to this course, MIS2502. Your PRO points request should include a brief description of your work and your data.

**Evaluation:**

Students will receive 0, 25, or 50 PRO points based on the instructor’s assessment of their work.

**Describe the Data:** The data in the decision tree focuses on features such as Sweetness, HarvestTime, and Ripeness to determine banana quality. These attributes are key indicators of the fruit's maturity and desirability in the market. Sweetness is a primary factor, with further decision splits highlighting the significance of HarvestTime and Ripeness in classifying quality levels.

**Find the Best Value for Minimum Split:** The decision tree's structure suggests that a minimum split size around 2000 effectively balances model complexity and learning capacity, preventing overfitting while maintaining sufficient detail in predictions. Using 1000 would result in over complexity while 3000 over simplified the tree.

**Find the Node with the Highest and Lowest Probability**: Node #6, with a high classification for quality (83.6%), suggests optimal banana conditions with higher Sweetness and appropriate HarvestTime. Conversely, Node #3, where only 8.4% of samples are high quality, highlights the negative impact of lower Ripeness on quality.