**Introduction to Data Analysis (DATA 1200)**

**Assignment #3 – Decision Tree (15% of Final Grade)**

John Hughes wants to leverage the **ricedataset.csv**. He is looking at using a Decision Tree model to help predict the type of rice (Cammeo or Osmancik).

The dataset has 3,810 observations and 8 variables:

**Independent Variables:**

Area: Returns the number of pixels within the boundaries of the rice grain.   
Perimeter: The circumference by calculating the distance between pixels around the boundaries of the rice grain.   
Major Axis Length: The longest line that can be drawn on the rice grain, i.e. the main axis distance, gives.   
Minor Axis Length: The shortest line that can be drawn on the rice grain, i.e. the small axis distance, gives.   
Eccentricity: It measures how round the ellipse, which has the same moments as the rice grain, is.   
Convex Area: Returns the pixel count of the smallest convex shell of the region formed by the rice grain.   
Extent: Returns the ratio of the region formed by the rice grain to the bounding box pixels.

**Dependent Variable:**

Class: Cammeo and Osmancik

**The Ask:**

1. **Create a Python Script using Jupyter Notebook (then convert to .html)** ***– 2%***
   1. Using Python develop a **Decision Tree algorithm** script to predict Class. Attach the HTML copy of your Python Code with your submission

**Note: All steps need to be annotated (i.e. Wk6a-DTExample)**

1. **Create a PowerPoint (PPT or PPTX) presentation that includes the following:**
   1. Cover Page (Title, Name (1st and last) and Student Number)
   2. Rational Statement (summary of the problem or problems to be addressed by the PPT) ***– 2%***
   3. Present the Correlation Heatmap and Explain **two (2) key insights** with associated explanations ***– 2%***
   4. Present the Confusion Matrix/Classification Report and Explain **three (3) key insights** from the Model Metrics (i.e., Precision, Recall, F1, Support for both summary and detailed metrics) ***- 6%***
   5. Explain **three (3) ways** to help improve the performance of the Decision Tree model. Please justify each of your answers. ***– 3%***

**Hint: Leverage the Wk6a-DTExample**

**Please post your PowerPoint Document (.ppt or .pptx) and Jupyter Notebook in HTML (.html) format via assignments under Assignment #3 by**

**Friday, March 17th, 2023 @ 11:59 p.m.**