Lab #1: An Example Data Science Process (You whale really love this!)

The goal of this lab is to gain familiarity with the process of doing data science by studying an example. There are many such examples to be found in published scientific research. Accordingly, we will read "Predicting seasonal movements and distribution of the sperm whale using machine learning algorithms" during the lab, discuss it, and compare it to the data science process.

1. What were the authors trying to accomplish?
2. Describe their data. How many whales did they track? On average, how long was each whale tracked? During that time, how many locations did they visit on average and how far did they travel on average?
3. In section 2.5, the authors describe their approach to applying machine learning models to perform species distribution modeling. In this problem, the area of interest is divided into a grid, and the model is applied to each grid cell to make a prediction. What were the eight input values for each grid cell?
4. Is this framed as a classification or regression problem? Describe what the dependent (output) variable represents and how it is interpreted.
5. In section 3.4, the authors describe the results of testing 14 different ML models plus a combined "stacking" model. Which models performed best?
6. To generate a probability map (see Figure 6), the authors generated 10 versions of each grid (they call them simulations), applied the model to each grid, and averaged the model predictions. What differences do you observe between the maps in Figure 6 for the wet and dry seasons? What do these differences mean in terms of the movement of the whales?
7. Do you think that it was worthwhile to evaluate 14 different types of ML models?
8. In section 4.3, the authors discuss the advantages of their approach versus more traditional approaches. Summarize their argument.
9. What might you do differently, if anything, if you were given this data set?
10. Explain each of the six phases of the CRoss Industry Standard Process for Data Mining (CRISP-DM) model. For each phase, describe its purpose and what activities or decisions are typically involved. Additionally, illustrate how these phases interconnect to guide the entire data science process.

Submission Requirements:

- Please type complete answers to these questions into a file entitled Lab_1_INITITIALS.doc. That is, if your name is Albert Brian Carter save your file as Lab_1_ABC.doc. Convert this file to pdf and upload it to canvas.
- Submissions will not be accepted in a different format.
- Late Submission Policy: One day late: maximum 95 points; two days late: maximum 90 points; three days late: maximum 85 points. Submissions will not be accepted more than three days late.