Coding Challenge III

Learning Python Modules, AI & Data Processing Relevant to Biomedical Questions

Due Tuesday, 11.17 at 11:59 pm

Please submit your answers in PDF form, and programs in .py format only. Screenshots can be saved and converted into a PDF.

NOTES: Justify each step in your code. Define the mathematical steps you took. Comment each line of code. Define input and output of your code. State all assumptions and simplifications.

1. ***Using a Python program you create:*** **(a) Split your team project dataset into a test and training set. (b) Identify the target attribute you want to predict. (c) Build a linear regression model to predict the target attribute.**
    a. What did you pick as the target attribute (output) and predictors (input)? Justify why.
    b. Did you normalize or weight the data beyond what was provided? Why or why not?
    c. Define mathematically how you determined the accuracy of your predictive model, and score your model predictions on two versions of your test dataset (picking 75% of the data each time).

2. ***Using Python code:*** **Develop a random forest model to predict your target attribute from (1).**
    a. What were the predictors in your model, and were they different from your linear regression model?
    b. Score your model predictions on two versions of your test dataset (picking 75% of the data each time). Which model was more accurate, the linear regression or random forest model?

3. **Using Python code, import the 3 cell images provided.**
    a. Identify 5 image-based metrics that you want to compare the images by (e.g., color, shape, size, number of nuclei)
    b. Apply hierarchical clustering to the three images' metrics, and identify which two cell images (Cell A, B or C) are most similar mathematically. How do the results compare to your expectations by looking at the three images?