***For regression models please use Mean Squared Error (MSE) as your evaluation metric moving forward, and for classification models please use F1-Score.***

**Part A:**

1. Describe the importance of training and test data. Why do we separate data into these subsets?
2. What is k-fold cross validation and what do we use it for?
3. How is k-fold cross validation different from stratified k-fold cross validation?
4. Name the 4 types of supervised learning models that we have learned thus far that are used to predict *categorical* dependent variables like whether an email is labeled “spam” or “not spam.”
5. Name the 3 types of supervised learning models that we have learned thus far that are used to predict *continuous* dependent variables like test scores.

**Part B:**

1. Import the spam dataset and print the first six rows.
2. Read through the documentation of the original dataset here: <http://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.names>
3. [Links to an external site.](http://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.names)
4. . The dependent variable is "spam" where one indicates that an email is spam and zero otherwise. Which three variables in the dataset do you think will be important predictors in a model of spam? Why?
5. Visualize the univariate distribution of each of the variables in the previous question.
6. Choose one model from Part A Question 4. Split the data into training and test subsets. Build a model with the three variables in the dataset that you think will be good predictors of "spam". Run the model and evaluate prediction error using k-fold cross-validation. Describe why you chose any particular parameters for your model (e.g., if you used KNN how did you decide to choose a specific value for k).
7. Repeat the previous question but with a *different* model from Part A Question 4.
8. Repeat the previous question but with a *different* model from Part A Question 4.
9. Repeat the previous question but with a *different* model from Part A Question 4.
10. Now rerun all 4 models with 3 additional variables that you think will help the prediction accuracy. Did this cause the performance to improve over your previous models?
11. What is a variable that *isn't* available in this dataset but you think *could* increase your final model's predictive power if you had it? Why do you think it would improve your model?