1. Describe the importance of training and test data. Why do we separate data into these subsets?

Training data is used to train the model. The model uses the training data to learn patterns and relationship between the data. The model updates its variables according to the data in order to predict unseen data.

Test data is used to test how model preforms when predicting unseen data.

This helps the model to avoid overfitting and progressively review performance of the model.

1. What is k-fold cross validation and what do we use it for?

K-fold cross-validation is a technique used to access model performance while training it to avoid overfitting. Initially data is divided in k subsets and model is trained using k-1 folds k times. Each time the model is trained using k-1 fold and the remaining fold used for testing and validation

1. How is k-fold cross validation different from stratified k-fold cross validation?

K-Fold Cross-Validation:

* Splits the dataset into *k* folds randomly.
* Each fold is used once as the test set while the others are used for training.

Stratified K-Fold Cross-Validation:

* Maintains the distribution of classes in each fold.
* This is especially important for imbalanced datasets, where certain classes may be underrepresented

1. 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.”

* Logistic Regression
* Decision Trees
* Naive Bayes
* Support Vector Machines (SVM)
* K-Nearest Neighbors (KNN)

1. 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.

* Linear Regression
* Decision Trees for Regression
* Support Vector Regression (SVR)

1. 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?

A variable not in the dataset that could potentially improve the model's predictive power might be "sender\_domain\_reputation". This could be a numerical score based on the reputation of the sender's email domain. It could be valuable because spammers often use domains with poor reputations or newly created domains, while legitimate senders tend to use established, reputable domains.