Homework 2: Evaluation and Model Selection

Complete the following exercises. Remember to explain your answers.

1. We'd like to identify models for some real-world scenarios.

   (a) The following questions ask you to choose three models for your workflow.

      i. Before we select any models, we'd like to get a baseline model for comparison purposes. What is this baseline model called? Explain how it works (non-technical).

      ii. We'd like to pick a simple model for our workflow. Will this model have higher bias or higher variance? Explain why.

      iii. We'd like to pick a more complex model for our workflow. Will this model have higher bias or higher variance? Explain why.

   (b) Scenario1: A music streaming service wants to group songs from its massive catalog into distinct genres based on their acoustic features, such as tempo, key, loudness, dancea-bility, energy, and mood. The goal is to automatically identify patterns in the music and suggest personalized playlists or genres to users based on these patterns.

      i. What type of model approach should we take?

      ii. What is our target variable?

      iii. What is our feature set?

      iv. Give an example of an algorithm we could use for our simple model:

      v. Give an example of an algorithm we could use for our more complicated model:

   (c) Scenario2: A biologist is conducting an experiment to study the effect of various environmental factors on the growth of a specific type of plant. The goal is to predict the plant's growth (measured in height) based on factors such as temperature, humidity, soil pH, light intensity, and water availability.

      i. What type of model approach should we take?

      ii. What is our target variable?

      iii. What is our feature set?

      iv. Give an example of an algorithm we could use for our simple model:

      v. Give an example of an algorithm we could use for our more complicated model:

2. Suppose we have trained a classifier that attempts to predict whether or not a person has COVID based off of their symptoms. When we compare the results predicted by our classifier to each person's actual COVID status, we get the following results on the testing set. Here, 1 represents "has COVID", while 0 represents "does not have COVID".

	$y_{true} = 1$	$y_{true} = 0$
$y_{pred} = 1$	128	56
$y_{pred} = 0$	72	285

As an example for how to read this table, this means that there were 56 people who did not have COVID, that the classifier predicted had COVID.

(a) What is the accuracy of this classifier? Write a sentence explaining what this means to someone who is not in this course.

(b) What is the precision of this classifier? Write a sentence explaining what this means to someone who is not in this course.

(c) What is the recall of this classifier? Write a sentence explaining what this means to someone who is not in this course.

(d) What is the F1-score of this classifier?

(e) What do you think of the performance of this classifier? Consider the problem - how easy or difficult is it to predict COVID status based on symptoms?

3. Suppose we have trained a regression model that attempts to predict house prices based on their location and their characteristics. When we compare the houses' actual selling price with their predicted price, we get the following results on the (very small) testing set.

actual price ($)	399,000	157,000	223,000	347,000
predicted price($)	410,000	150,000	225,000	350,000

(a) What is the mean squared error for the model?

(b) What is the root mean squared error for the model?

(c) What is the mean absolute error for the model?