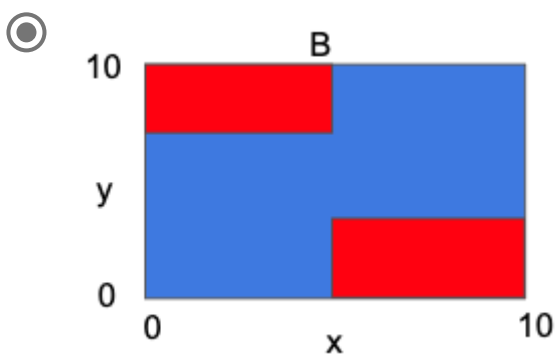# Week 2 Quiz

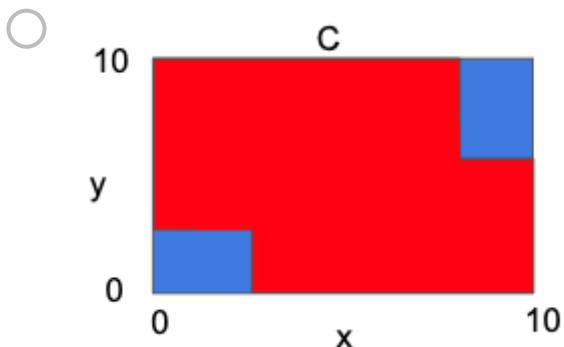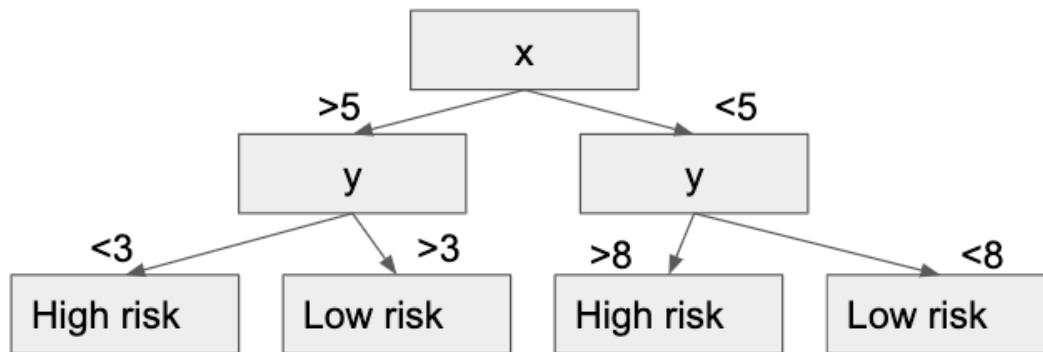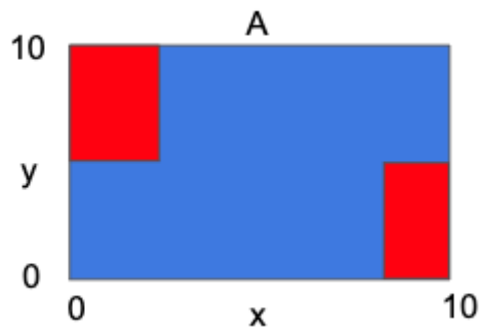**TOTAL POINTS 10**

---

1. Which decision boundary corresponds to the following decision tree?   **1 / 1 point**

    In the options, red indicates high risk, blue indicates low risk.

○

◉

✓ **Correct**

One way to approach this is to work your way back up from the leaves to see the regions that are classified as high risk. Doing this, we see that one high risk area is where x > 5 and y > 3, and where x < 5 and y > 8. This corresponds to the two rectangles shown here.

2. True or False: A tree of depth 1 is more expressive than a classical linear model.   **0 / 1 point**

    ○ False

    ● True

    ! **Incorrect**

    Imagine a tree that takes in 3 features and has a depth of 1. How many features can it use to make a prediction? Compare that to a linear regression model that takes in the same 3 features. How many features can the regression model use when making a prediction?

3. One way to aggregate predictions from multiple trees is by a majority vote. Using this aggregation rule, select the prediction of the following trees on the data point (x=4, y=7, z=2):   **1 / 1 point**
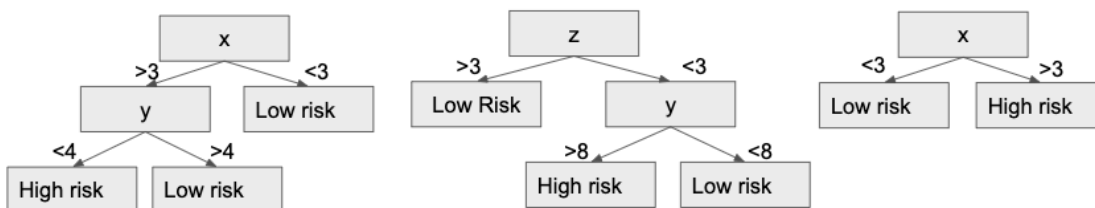
◯ High risk

⦿ Low risk

✓ **Correct**

The hardest part of this is just keeping track of the coordinates. Following the tree from the root down, we get low risk from the first tree, low risk from the second tree, and high risk from the third tree. Therefore the majority vote is low risk.

4. You've fit a random forest of 10 trees with max depth 20. Your training ROC is 0.99 and test ROC is 0.54. Which of the following is NOT a reasonable thing to try?   **0 / 1 point**

◯ Increasing the number of trees

⦿ Trying a linear model

◯ Increasing the max depth

◯ Decreasing max depth

! **Incorrect**

The model is overfitting, since it's performing poorly on the test set compared to the training set.

Using a linear model would reduce overfitting by constraining the decision boundary.

Which is NOT a reasonable thing to try?

5. True or False: When your data is missing at random, then whether or not you are missing a covariate is completely independent of your outcome.   **1 / 1 point**

HINT: the data is "missing at random" and not "missing completely at random".

⦿ False

◯ True

✓ **Correct**

This is False.

For example, you may be predicting death within 5 years, but some values are more likely to be missing if you are older.

Since missing data is correlated with old age, which is itself correlated with likelihood of death within 5 year, whether or not your data is missing is not independent of the outcome.

It is true, however, that missing data would be independent of the outcome when conditioned on age.

6.   When is complete case analysis least likely to bias your model?   **1 / 1 point**

  ○   Dropping data will bias your model in all of the above cases

  ○   Data is missing at random

  ◉   Data is missing completely at random

  ○   Data is missing not at random

  ✓   **Correct**
       If data is missing completely at random, then there is no difference in the distribution of your data with and without the rows with missing data, so as long as you have enough complete data, dropping those rows should not bias your model.

7.   You have created a model using mean imputation. At test time, you should fill in missing   **1 / 1 point**
     values with:

  ○   Mean from the test data

  ○   None of the above

  ◉   Mean from the train data

  ○   0.0

  ✓   **Correct**
       Explanation: When testing your model, you should use the mean from your training data. Otherwise, your test statistics will not be reflective of your training procedure and how they will perform in the real world, since in the real world you will not have access to the entire collection of test examples beforehand.

8. Let's say blood pressure (BP) measurements are more likely to be missing among young people, who generally have lower blood pressure. You use mean imputation to train your model. Which option correctly characterizes the mean BP (after imputation) in your training dataset?

1 / 1 point

● It is higher than the true mean

○ It is the same as the true mean

○ It is lower than the true mean

○ None of the above

✓ **Correct**

The BP measurements which will be present in your dataset will be higher than average, since many of the lower BP values from the younger people in the dataset will be missing.

Therefore the mean which you impute will be higher than average, so the overall mean will be higher. This is an example of bias introduced in your dataset through imputation.

9. You have trained the following tree and want to make a prediction on someone, but all you know is they are 40 years old, and do not have their blood pressure.

0 / 1 point

From you dataset, you learn the linear relationship between blood pressure and age is:

$BP = 1.7 \times Age + 60.$

Using regression imputation, and the decision tree shown here, what is your prediction for this person's risk of heart attack?
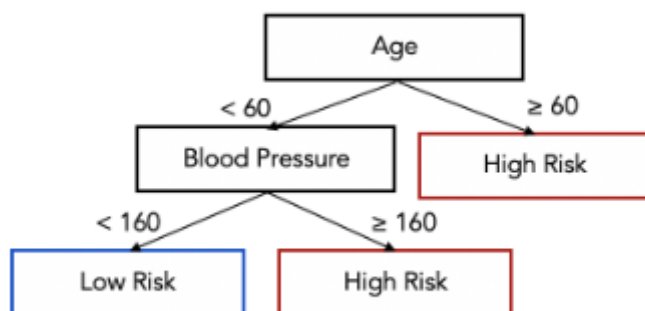
● High Risk

○ Low Risk

❗ **Incorrect**

First let's get our imputed BP value. Then use the BP and Age to follow the decision tree down to one of the leaves of the tree.

10. Assume you have missing data on one of your features, and are considering two options:

Option 1: Drop the feature that has missing values and fit a linear regression on the remaining features.

Option 2: Use imputation on the feature that has missing values, and fit a linear regression on all features.

True or False: "Both options have the same performance".

◉ False

○ True

✓ **Correct**

It seems like this might be true because you could plug in your imputation equation into the linear regression to get a linear regression based on all features but the imputed one.

However, since the imputed feature is not exactly a linear combination of the other features, the model that is learned will not be the same, since the model will still be able to take into account the variation in the feature when it is not missing. Recall that for some patients, the data of the feature is still measured, and not missing for all patients.

To convince yourself, think of a dataset with 1000 points where only one example is missing a measurement. Fitting a model on the whole dataset will not be the same as fitting it on a dataset that drops that entire column.