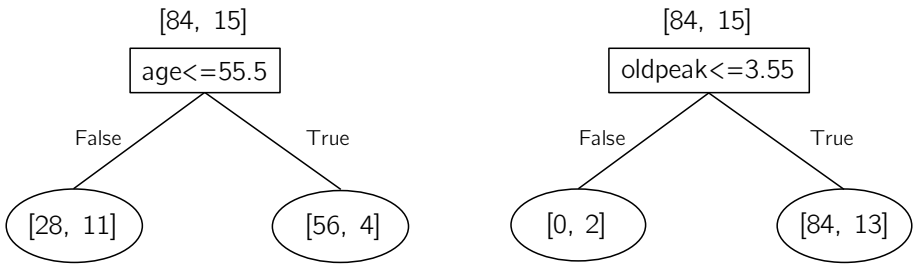
Midterm 1 Practice Problems *(find and work with a partner)*

1. *Entropy.* Consider the two feature choices below (for the heart disease dataset), and their associated splits. Counts of label -1 vs. 1 are shown in brackets.



   (a) After splitting the data based on each feature, what is the *classification error* for each tree?

   (b) Before considering the feature, what is $H(Y)$, the entropy of the initial partition? (don't need to find a value, just set up the equation)

   (c) Which tree do you think produces more information gain?

2. *Confusion Matrices.* Consider the confusion matrix below:

True class \ Predicted class	1	2
1	18	2
2	10	30

   (a) What is the classification error? What is the classification accuracy?

   (b) Normalize the confusion matrix (each row should sum to 1).

3. *Linear Regression.* Say we have $p = 1$ and two training examples: $(x_1, y_1) = (3, 0)$ and $(x_2, y_2) = (7, 1)$, and we would like to fit a linear model to this dataset.

   (a) Draw these two examples on a coordinate system and sketch the linear function that would fit them. What are the optimal weights? ($\hat{w}_0$ and $\hat{w}_1$)

   (b) Say in our SGD method, we choose to analyze (x_2, y_2) first. Before starting SGD, we set $w_0 = 0$ and $w_1 = 0$. After analyzing (x_2, y_2), what are w_0 and w_1? Choose $\alpha = 0.1$. Plot this updated line on your graph above.

   (c) Next we consider (x_1, y_1). What are w_0 and w_1 be after this second data point? Plot this line on your graph above. At this point we have finished *one* iteration of SGD.

   (d) Say we wanted to perform polynomial regression on this dataset with $d = 2$ (i.e. fit a quadratic function). What does the matrix $\mathbf{\Phi}$ look like after creating the polynomial features?

4. How did we define *model complexity* for:

   (a) Decision Trees:

   (b) Polynomial Regression: