Introduction to Classification: single feature models *(find and work with a partner)*

Consider the following dataset, modified from the Cleveland Heart Disease dataset here: `https://archive.ics.uci.edu/ml/datasets/heart+Disease`. This dataset has $p = 2$ features and $n = 16$ examples. Our goal is to predict the class label as negative (-1, no heart disease) vs. positive ($+1$, heart disease). We will consider this dataset as the *training data*.

```
@relation cleveland-14-heart-disease
@attribute 'slope' { up, flat, down}
@attribute 'thal' { fixed_defect, normal, reversable_defect}
@attribute 'class' { negative, positive}
@data
down,  fixed_defect,       negative
down,  normal,             negative
up,    normal,             negative
up,    normal,             negative
up,    normal,             negative
flat,  fixed_defect,       negative
flat,  normal,             negative
up,    reversable_defect,  negative
flat,  normal,             positive
flat,  reversable_defect,  positive
down,  normal,             positive
flat,  reversable_defect,  positive
down,  reversable_defect,  positive
flat,  fixed_defect,       positive
down,  reversable_defect,  positive
flat,  normal,             positive
```

1. Draw a decision tree for the feature `slope` (which relates to heart rate during exercise). Each edge corresponds to one feature value. At each leaf, compute the *probability* of a positive classification.

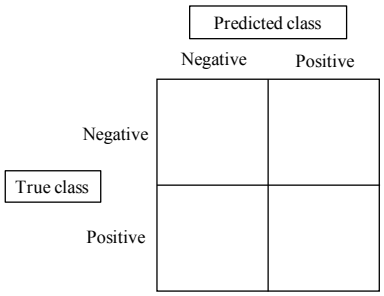
2. Do the same thing for the feature `thal` (which relates to the blood disorder *Thalassemia*).

3. If our threshold is 1/2, which feature values would classify a new test example as positive? What if the threshold is 1/12?

Evaluation Metrics *(find and work with a partner)*

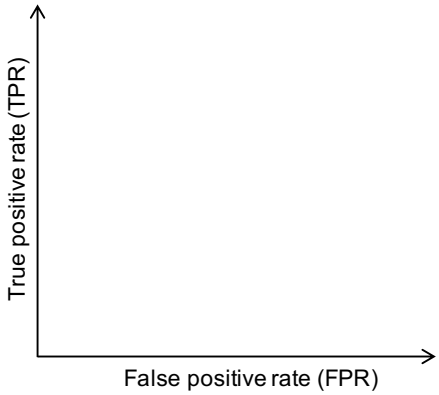
Say you google "ROC", wanting documents about ROC curves. Say there are 100 total documents to search through, and 10 are returned to you. Of these, 7 are about ROC curves. In reality, there are 20 documents about ROC curves.

1. Complete the confusion matrix below, adding the sum of each row to the right and the sum of each column to the bottom.



2. What is the precision? The recall (true positive rate)?

3. Compute the false positive rate and then plot this result on the ROC curve axes below.



4. With the same query, another search engine gives you 30 documents, 18 of which are about ROC curves. Add this point to your ROC curve above.

5. Add the other two points that are always on a ROC curve. In words, how would you describe the classifications strategies of these two points?