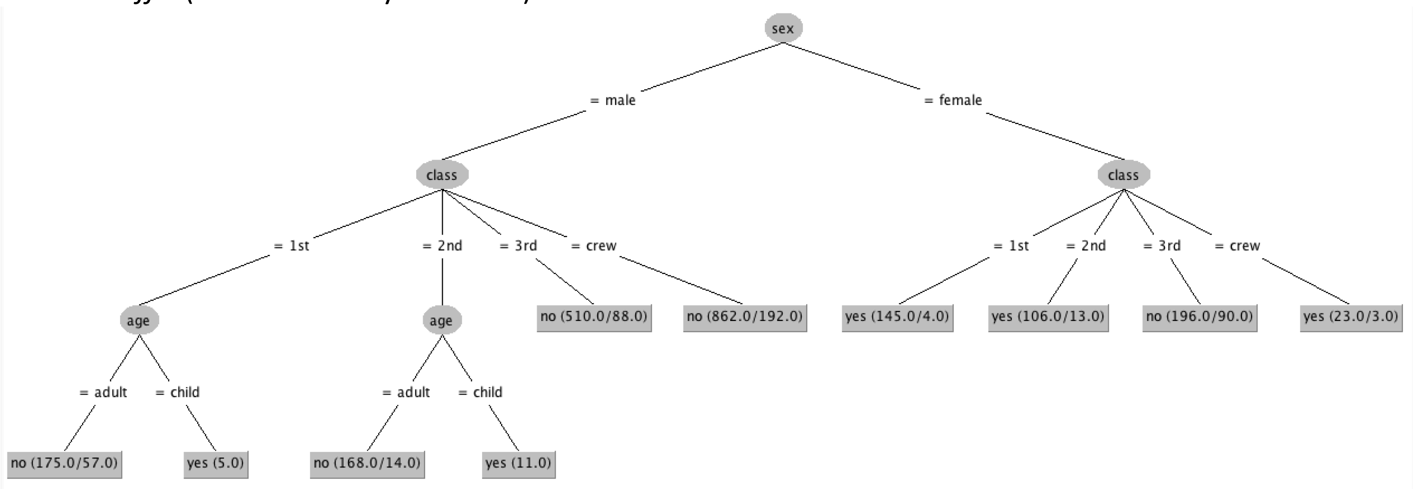
11-344 PS1 Connor W. Colombo
cwcolomb

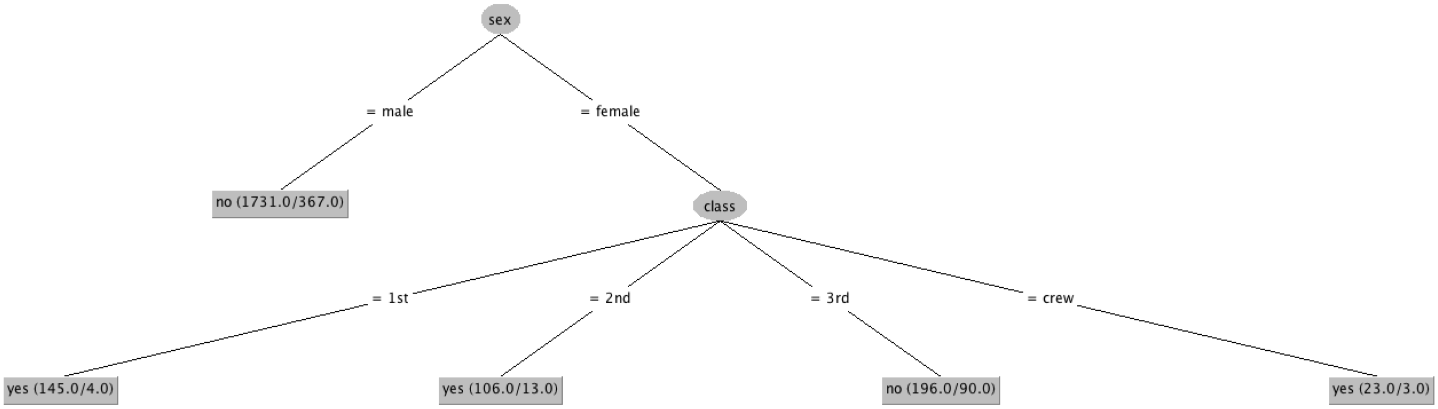
A.

1.

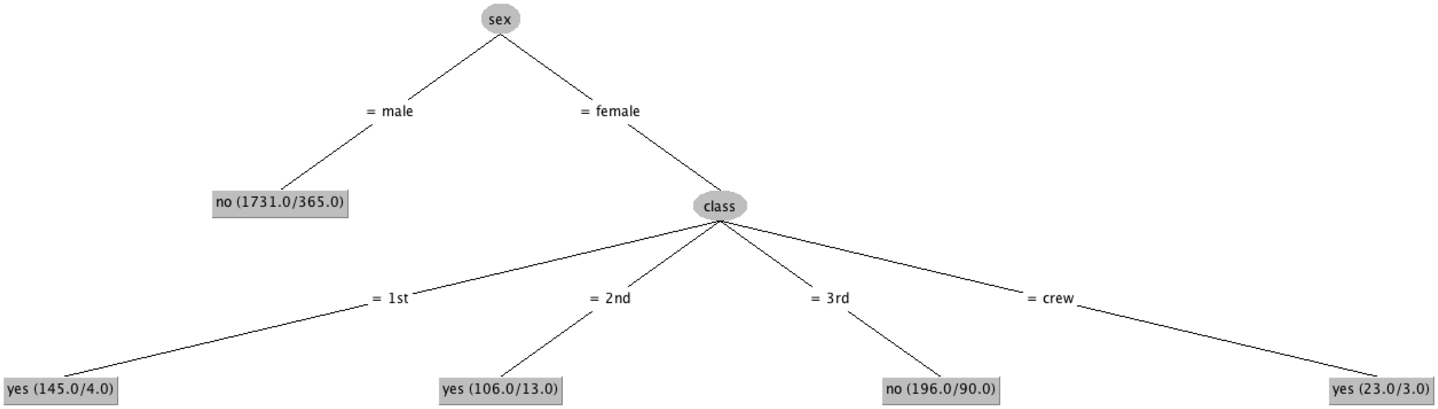
a. titanic.arff (78.9% correctly classified)



b. titanic-age-numeric.arff (78.3% correctly classified)



c. titanic_noise.arff (78.4% correctly classified)



2.

With respect to that generated for the fully nominal dataset, the tree generated for the dataset where age is numeric lacks branching on age. Since the lack of this distinction this seems to have minimally impacted the accuracy of the model, this suggests that information gain from branching on age is minimal. So, when the number of possible states for this factor was expanded from 2 to probably somewhere close to 70, the number of branches needed to maintain accuracy increased since the information gain of just branching on 18 would like be pretty low since there are likely to be many members whose ages are above and below 18 who didn't survive. Thus, branching on age was not considered worth the risk of overfitting, reducing the confidence factor, causing the entire subtree which branches on age to be removed (which J48 does as its default method of pruning).

3.
With respect to that generated for the original dataset, the tree generated for the noisy dataset lacks branching on age. This suggests that noise added to the set of ages caused the information gain of branching on age to decrease, making it likely that branching either didn't occur or the subtrees which branched on age where removed to prevent overfitting.

Also of note is that J48 learned the same model here as it did for the numeric age dataset, which seemed to have a similar issue with low information gain for branching on age.

B.
1. Python already installed.
2. Program executed. *output.csv* created and appears to match *outputReference.csv*.
3. Relevant files will be uploaded to Canvas.