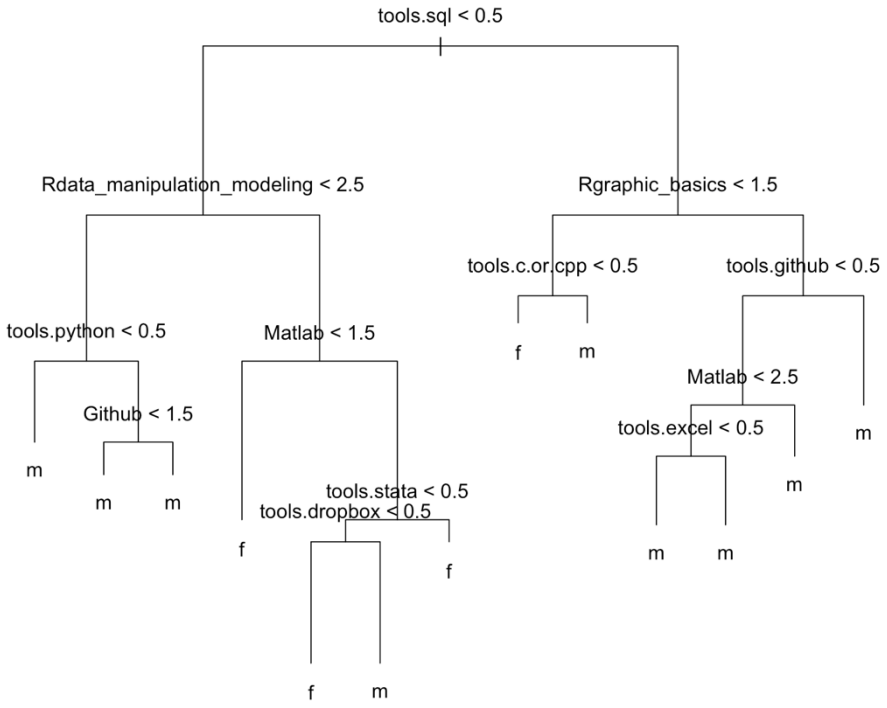
Decision Tree Explanation

We first fit the entire training set (the set with 112 cases of known gender) and derive the tree with 13 terminal nodes:



We notice the misclassification error rate of this single tree before any pruning or tuning is about 17.0%; this is already better than the error rate of 28.6%

```
> 1-sum(data_clean_training$pronoun=='m')/nrow(data_clean_training)
[1] 0.2857143
```

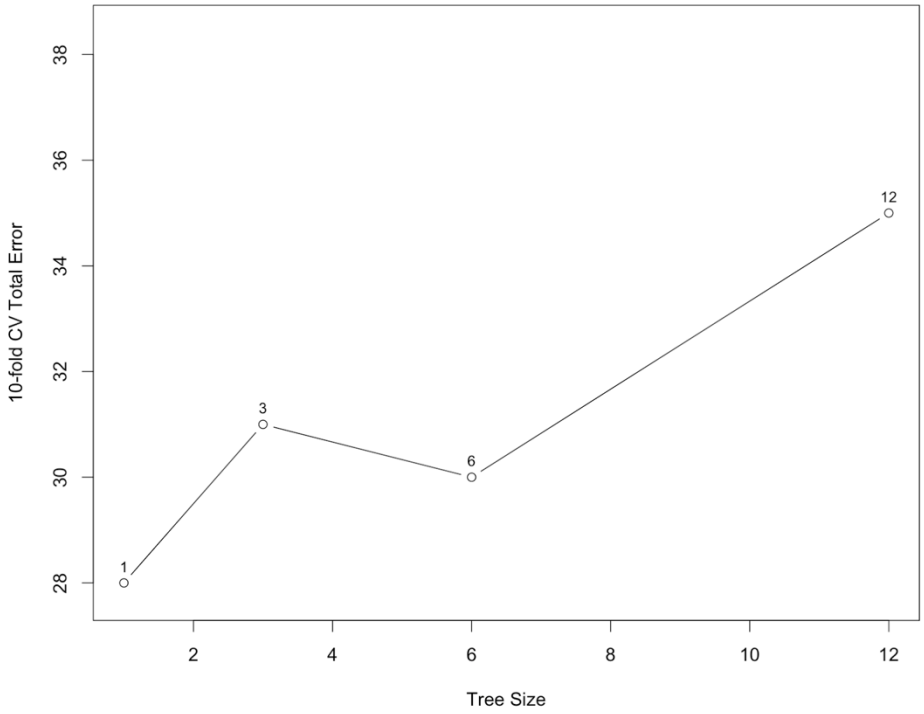
when we uniformly choose the gender to be male, which gives us reasons to continue to use tree classifier. But this is only the training error; we need to determine tree with the lowest test error. Thus we randomly select about 80% of observations (90 people) to be our training data and the rest 22 people to be test data.

```
> table(tree.test.pred,pronoun.test)
              pronoun.test
tree.test.pred  f  m
             f  2  5
             m  3 12
```

The table tells us that out of 22 test cases, 3 females are misclassified as males, and 5 males are misclassified as females. The test error rate is about 36.4%:

```
> 1-sum(tree.test.pred==pronoun.test)/length(tree.test.pred)
[1] 0.3636364
```

We now apply cross-validation to determine the optimal level of tree complexity. We observe from the plot that the 10-fold cross-validation error rates with 6 terminal nodes are lowest, and lower than that of the original tree (the one built with training data subsetted from data_clean_training) with 12 terminal nodes. The y-axis denotes the number of total misclassification cases among all 10 folds.



We thus pick 6 to be the number of terminal nodes and prune the tree, see if the result is better:

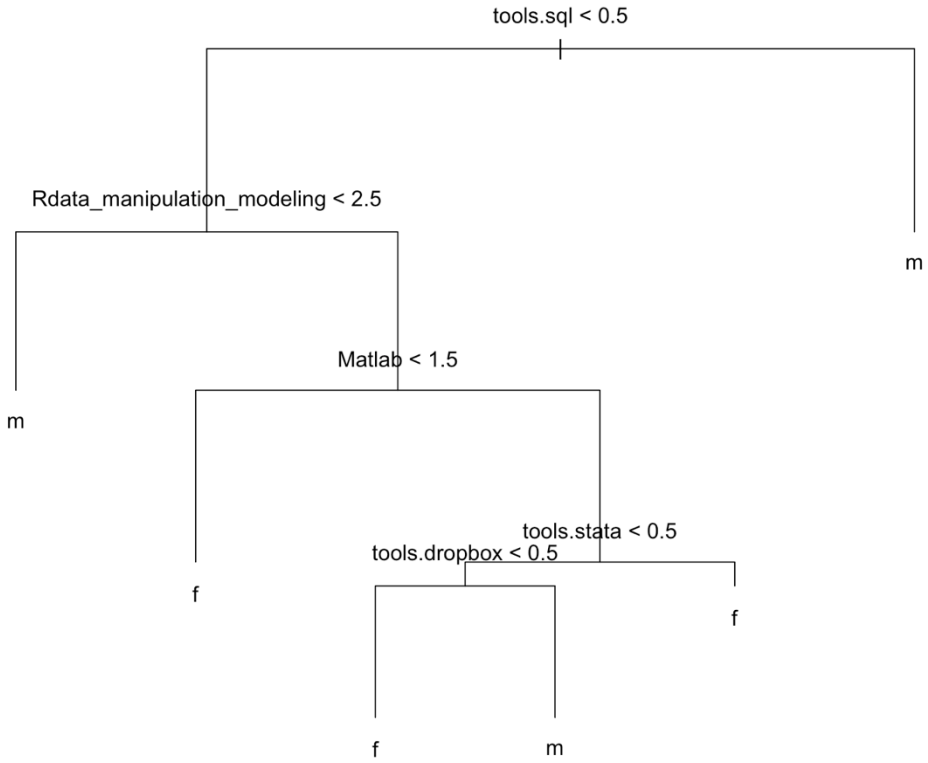
```
> table(tree.test.pred,pronoun.test)
              pronoun.test
tree.test.pred  f  m
             f  2  3
             m  3 14
```

The table tells us that out of 22 test cases, 3 females are misclassified as males, and 3 males are misclassified as females. The test error rate is about 27.3%.

```
> 1-sum(tree.test.pred==pronoun.test)/length(tree.test.pred)
[1] 0.2727273
```

Therefore we decide the number of terminal nodes to be 6, and fit the entire dataset with 112 cases to this single tree model, and use the resulting tree to predict the gender of two people.

Here's the tree we use:



And this is our results:

```
> test.pred
[1] m f
```

which means that the one who wrote "doesn't matter" is a male student, and the one who didn't write anything is a female student. We then updated the dataset data_clean with the corresponding gender.