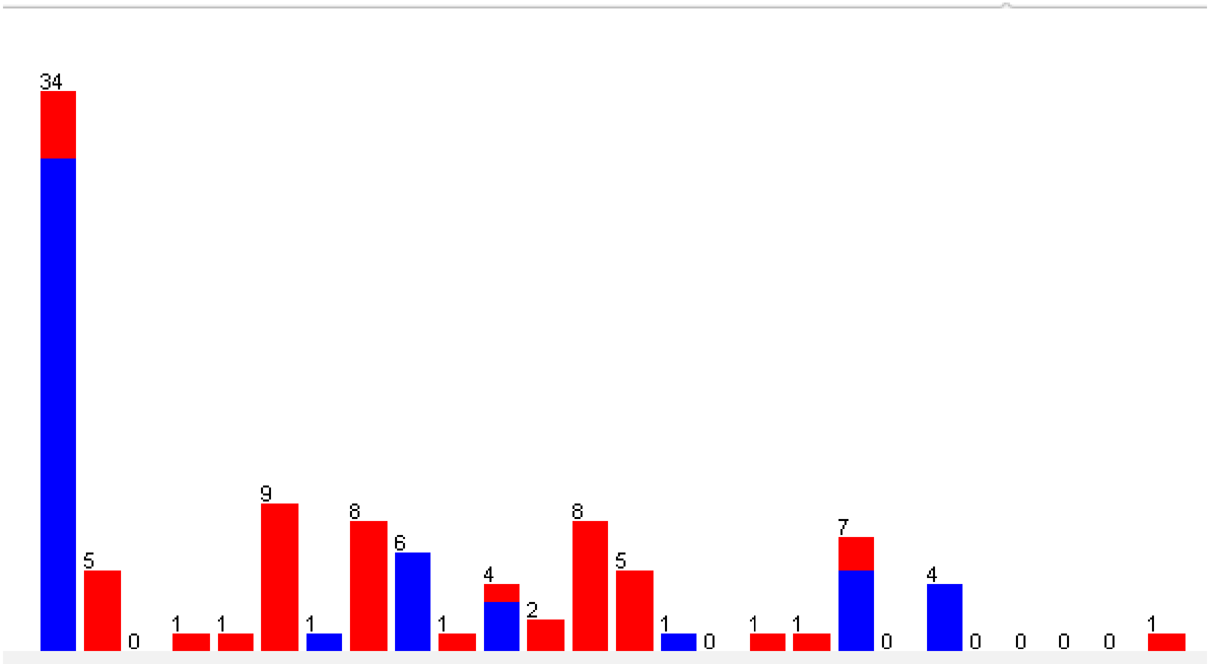
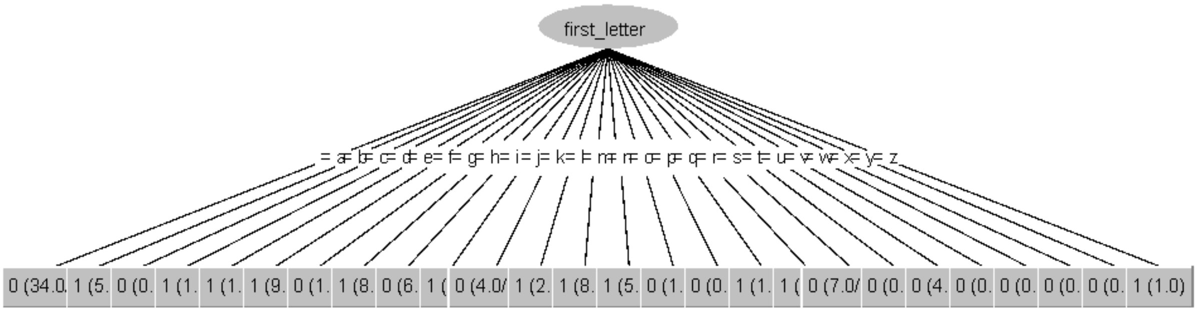
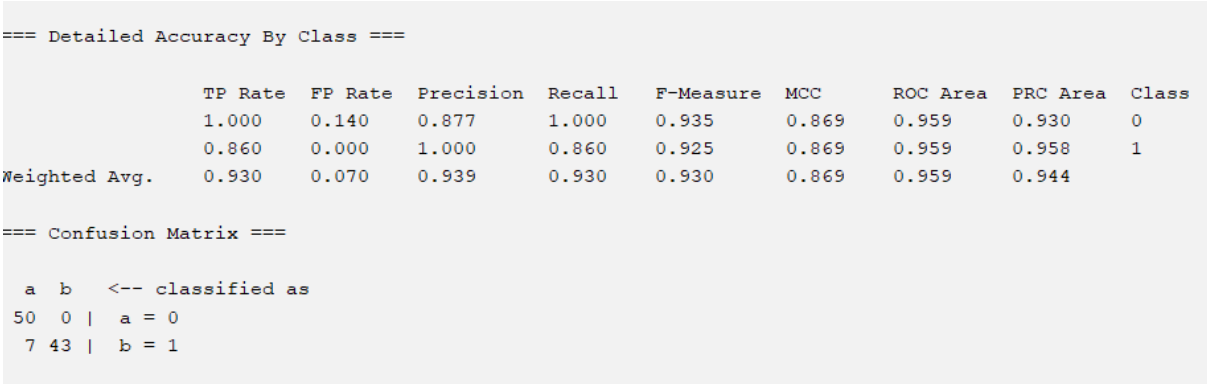
Its noticeable that a large majority of positves fall on the letter a



It can also be noticed that the decision tree completely relied on the first letter to make decisions



the true positive rate was 1 and false positive was 0.140. just by completely relying on first letter it achieved and precision of 0.877 and recall of 1



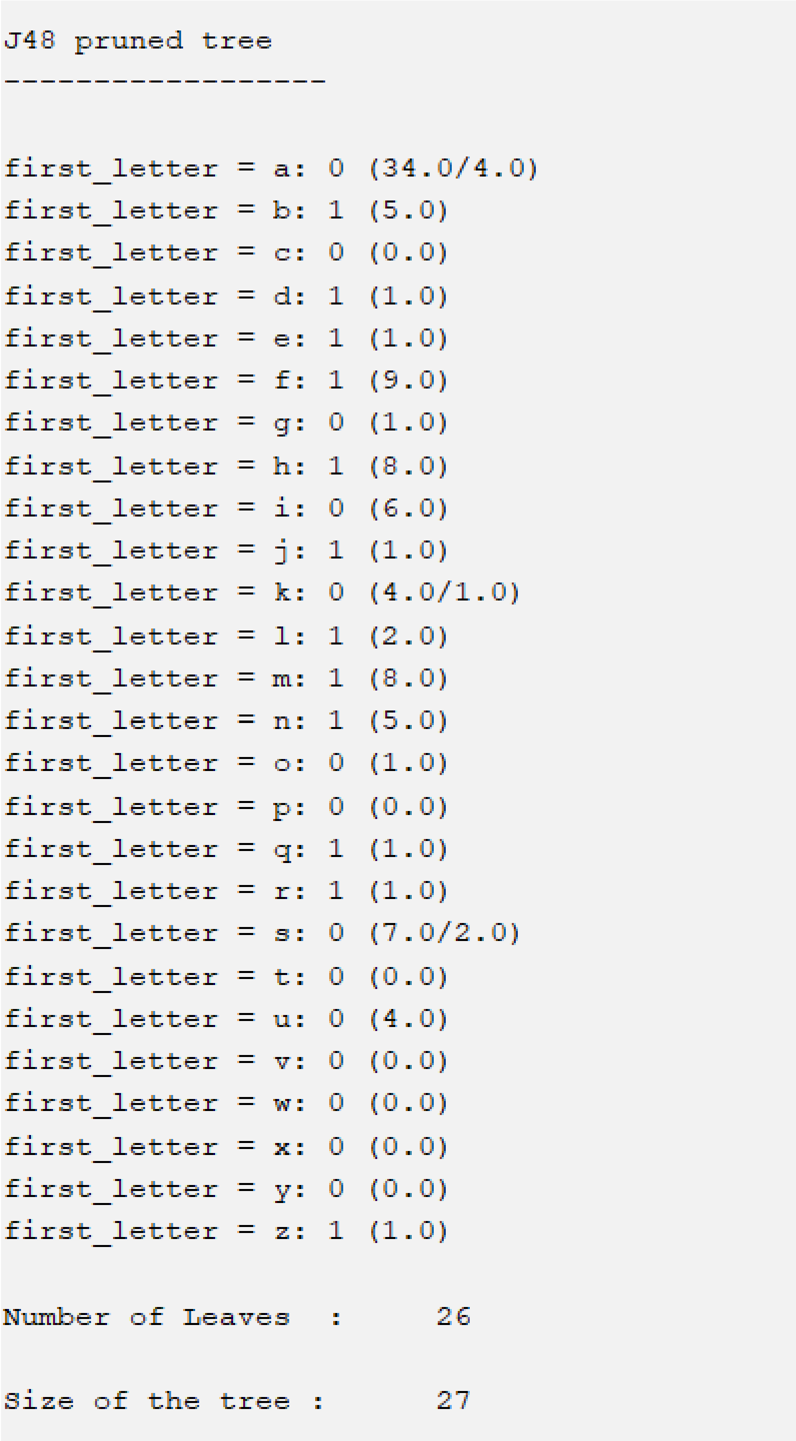
```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.140    0.877      1.000    0.935      0.869    0.959     0.930     0
                 0.860    0.000    1.000      0.860    0.925      0.869    0.959     0.958     1
Weighted Avg.    0.930    0.070    0.939      0.930    0.930      0.869    0.959     0.944

=== Confusion Matrix ===

  a  b   <-- classified as
 50  0 |  a = 0
  7 43 |  b = 1
```

This could suggest that there might have been some mislabels within the data.



```
J48 pruned tree
------------------

first_letter = a: 0 (34.0/4.0)
first_letter = b: 1 (5.0)
first_letter = c: 0 (0.0)
first_letter = d: 1 (1.0)
first_letter = e: 1 (1.0)
first_letter = f: 1 (9.0)
first_letter = g: 0 (1.0)
first_letter = h: 1 (8.0)
first_letter = i: 0 (6.0)
first_letter = j: 1 (1.0)
first_letter = k: 0 (4.0/1.0)
first_letter = l: 1 (2.0)
first_letter = m: 1 (8.0)
first_letter = n: 1 (5.0)
first_letter = o: 0 (1.0)
first_letter = p: 0 (0.0)
first_letter = q: 1 (1.0)
first_letter = r: 1 (1.0)
first_letter = s: 0 (7.0/2.0)
first_letter = t: 0 (0.0)
first_letter = u: 0 (4.0)
first_letter = v: 0 (0.0)
first_letter = w: 0 (0.0)
first_letter = x: 0 (0.0)
first_letter = y: 0 (0.0)
first_letter = z: 1 (1.0)

Number of Leaves  :     26

Size of the tree :      27
```

The tree overall had 27 node

Experience

Training the approximation function was interesting. Weka was an easy to use tool and visualized things quickly though python might have been quicker and allowed more flexible analysis and visualizations.

The resultant function did not completely grasp the actual concept but was able to approximate it. Reason for the inaccuracies could be mislabels and unreliability within the inherent dataset or the limitations of the features extracted. Full 100% accuracy was not achieved but a close enough approximation function was achieved.