DM LAB 4

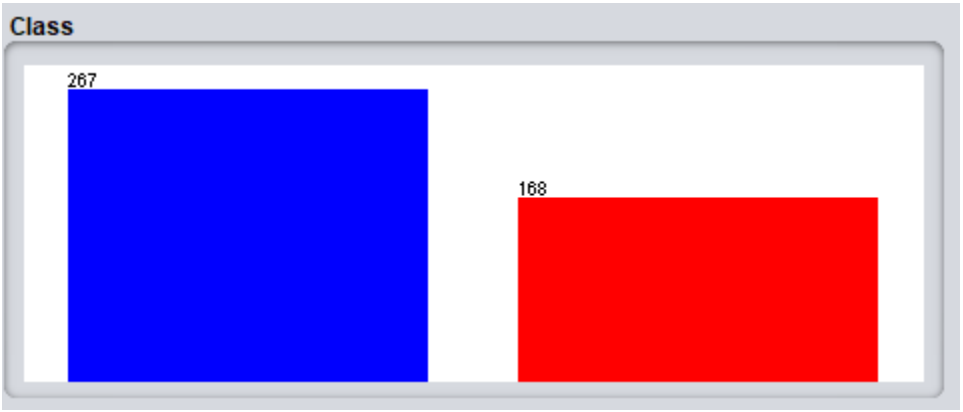
Name : Zain Al Abidin
Roll No : 21L-6260

Basic Introduction of Dataset
My data consists of a vote of the US that the democrats and the republicans voted on. They had different issues and they voted either yes or no in each case.

a) Total instances : 435

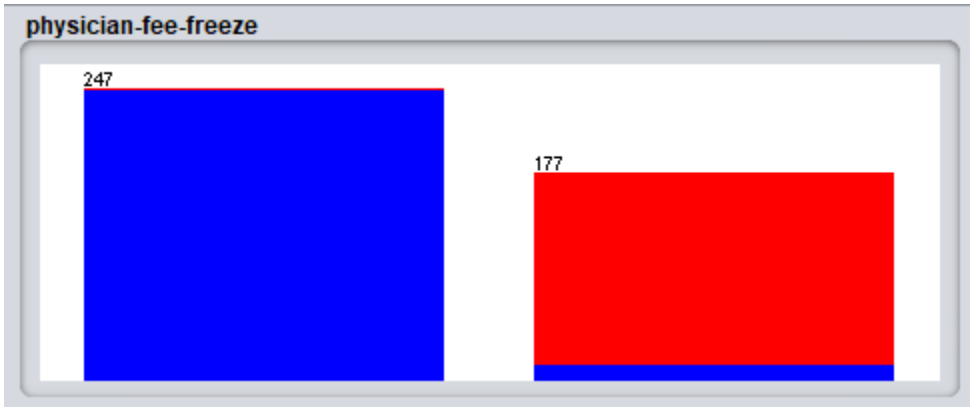
b) Total Attributes : 16 + 1 (for class name) = 17 total attributes

c)



This is the class attribute where we can consider the democrats(blue) as yes label and the republicans(red) and the no label.

d)



This attribute better separates the data and tells us that the majority of people are in the favor of physician fee freeze. The blue color represents the democrats and the red color represents the republicans. The left column shows the people that voted against the physician fee freeze while the right column represents the people that voted for the physician fee freeze.
Therefore we can conclude that most democrats voted against the physician fee freeze while most republicans voted for the physician fee freeze.

II)

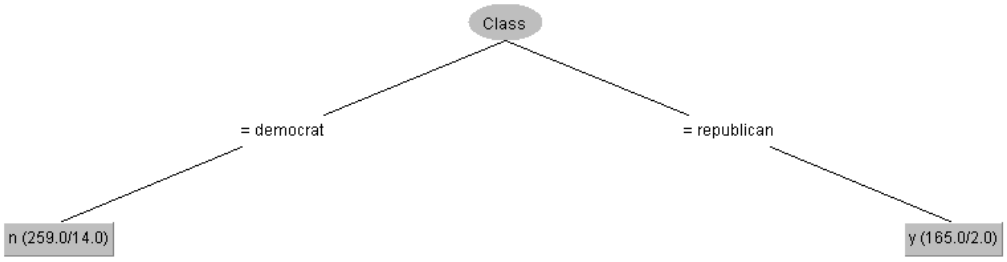
```
=== Summary ===

Correctly Classified Instances         423               97.2414 %
Incorrectly Classified Instances        12                2.7586 %
Kappa statistic                          0.9418
Mean absolute error                      0.0519
Root mean squared error                  0.1506
Relative absolute error                 10.9481 %
Root relative squared error             30.9353 %
Total Number of Instances              435
```

We can clearly see that our model has an accuracy of 97 percent.

Confusion Matrix

```
   a   b   <-- classified as
 261   6 |   a = democrat
   6 162 |   b = republican
```

This is the decision tree for the attribute physician-fee-freeze while indicates a clear class seperation



III)

```
=== Summary ===

Correctly Classified Instances         247               72.6471 %
Incorrectly Classified Instances        93               27.3529 %
Kappa statistic                          0.2687
Mean absolute error                      0.3351
Root mean squared error                  0.4836
Relative absolute error                 80.8066 %
Root relative squared error            108.7935 %
Total Number of Instances              340
```

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.836    0.578    0.801      0.836    0.818      0.270    0.605     0.782     good
               0.422    0.164    0.481      0.422    0.450      0.270    0.605     0.370     bad
Weighted Avg.  0.726    0.468    0.716      0.726    0.721      0.270    0.605     0.673
```

```
=== Confusion Matrix ===

   a   b   <-- classified as
 209  41 |   a = good
  52  38 |   b = bad
```

This is the summary of the dataset that shows that we have an accuracy of 72% which is not good but not that bad either.

ZeroR and the accuracy is somewhat the same however that is wrong because we can clearly see that the ZeroR classifier cannot predict the bad classes (clearly visible in the confusion matrix)

```
=== Summary ===

Correctly Classified Instances         250               73.5294 %
Incorrectly Classified Instances        90               26.4706 %
Kappa statistic                          0
Mean absolute error                      0.4147
Root mean squared error                  0.4445
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    1.000    0.735      1.000    0.847      ?        0.500     0.735     good
               0.000    0.000    ?          0.000    ?          ?        0.500     0.265     bad
Weighted Avg.  0.735    0.735    ?          0.735    ?          ?        0.500     0.611

=== Confusion Matrix ===

   a   b   <-- classified as
 250   0 |   a = good
  90   0 |   b = bad
```

One R also achieved the same accuracy because the dataset is not that good

```
=== Summary ===

Correctly Classified Instances         246               72.3529 %
Incorrectly Classified Instances        94               27.6471 %
Kappa statistic                          0.1581
Mean absolute error                      0.2765
Root mean squared error                  0.5258
Relative absolute error                 66.6681 %
Root relative squared error            118.2987 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.900    0.767    0.765      0.900    0.827      0.172    0.567     0.762     good
                 0.233    0.100    0.457      0.233    0.309      0.172    0.567     0.309     bad
Weighted Avg.    0.724    0.590    0.684      0.724    0.690      0.172    0.567     0.642

=== Confusion Matrix ===

   a   b   <-- classified as
 225  25 |   a = good
  69  21 |   b = bad
```

Used Naive Bayes and achieved a better accuracy

```
=== Summary ===

Correctly Classified Instances         260               76.4706 %
Incorrectly Classified Instances        80               23.5294 %
Kappa statistic                          0.3824
Mean absolute error                      0.2819
Root mean squared error                  0.4005
Relative absolute error                 67.9798 %
Root relative squared error             90.114  %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.852    0.478    0.832      0.852    0.842      0.383    0.804     0.921     good
                 0.522    0.148    0.560      0.522    0.540      0.383    0.804     0.592     bad
Weighted Avg.    0.765    0.390    0.760      0.765    0.762      0.383    0.804     0.834

=== Confusion Matrix ===

   a   b   <-- classified as
 213  37 |   a = good
  43  47 |   b = bad
```