Supervised Learning

Problem statement: We are given car data with different car attribute. The data set is giving us the acceptance rate of cars on the basis of these attributes. We are told to do different classifiers and test which one gives better result out of the classifiers.

Attribute Values:

Buying (Buying price)	:	v-high, high, med, low
Maint (Maintenance price)	:	v-high, high, med, low
Doors (Number of doors)	:	2, 3, 4, 5-more
Persons (accommodated person)	:	2, 4, more
Lug_boot (luggage boot)	:	small, med, big
safety	:	low, med, high

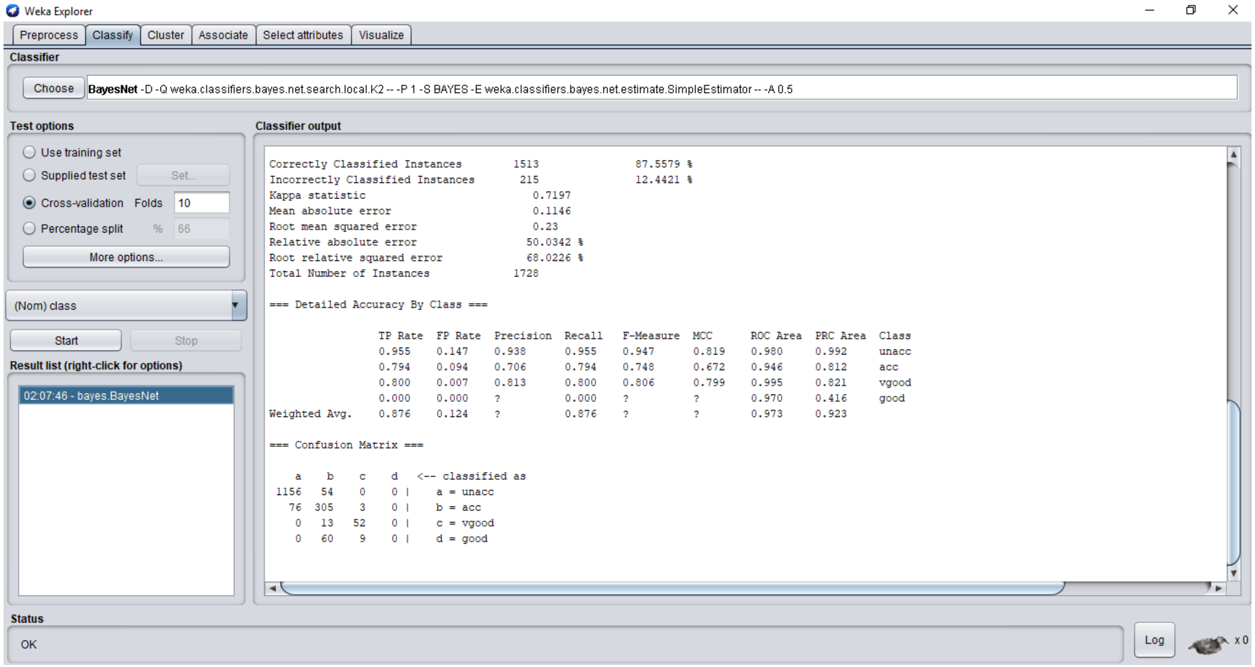
Missing Attribute Values: No missing value

Class Distribution (number of instances per class)

class	N	N [%]
unacc	1210	(70.023 %)
acc	384	(22.222 %)
good	69	(3.993 %)
v-good	65	(3.762 %)

order to evaluate the car data, set we had to convert the data file into csv format. After opening the data.csv file we notice that some of the instances in had string value or combination of numeric value in the instances. So in order to data set in Weka we had to assume the strings as presumable numeric values. We had to leave the class attribute as string because the testing in weka depends on the type of attribute we are trying to predict. After taking the class as the attribute we want to predict we choose five classifiers and got their confusion matrix with TP and FP rate.

Bayes Net classifier:



Classifier: **BayesNet** -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

● Cross-validation Folds 10

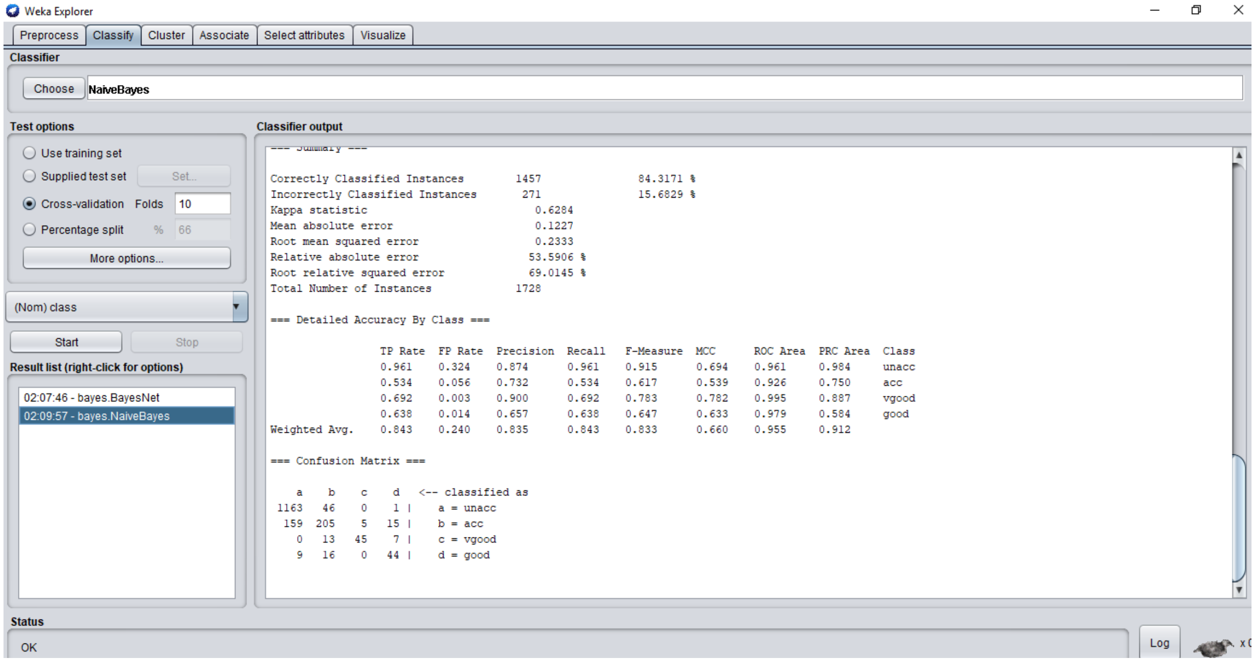
(Nom) class

02:07:46 - bayes.BayesNet

```
Correctly Classified Instances        1513               87.5579 %
Incorrectly Classified Instances       215               12.4421 %
Kappa statistic                          0.7197
Mean absolute error                      0.1146
Root mean squared error                  0.23
Relative absolute error                 50.0342 %
Root relative squared error             68.0226 %
Total Number of Instances             1728

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.955    0.147    0.938      0.955    0.947      0.819    0.980     0.992     unacc
                 0.794    0.094    0.706      0.794    0.748      0.672    0.946     0.812     acc
                 0.800    0.007    0.813      0.800    0.806      0.799    0.995     0.821     vgood
                 0.000    0.000    ?          0.000    ?          ?        0.970     0.416     good
Weighted Avg.    0.876    0.124    ?          0.876    ?          ?        0.973     0.923

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1156   54    0    0 |    a = unacc
   76  305    3    0 |    b = acc
    0   13   52    0 |    c = vgood
    0   60    9    0 |    d = good
```

Status: OK

Naïve Bayes:



Classifier: **NaiveBayes**

● Cross-validation Folds 10

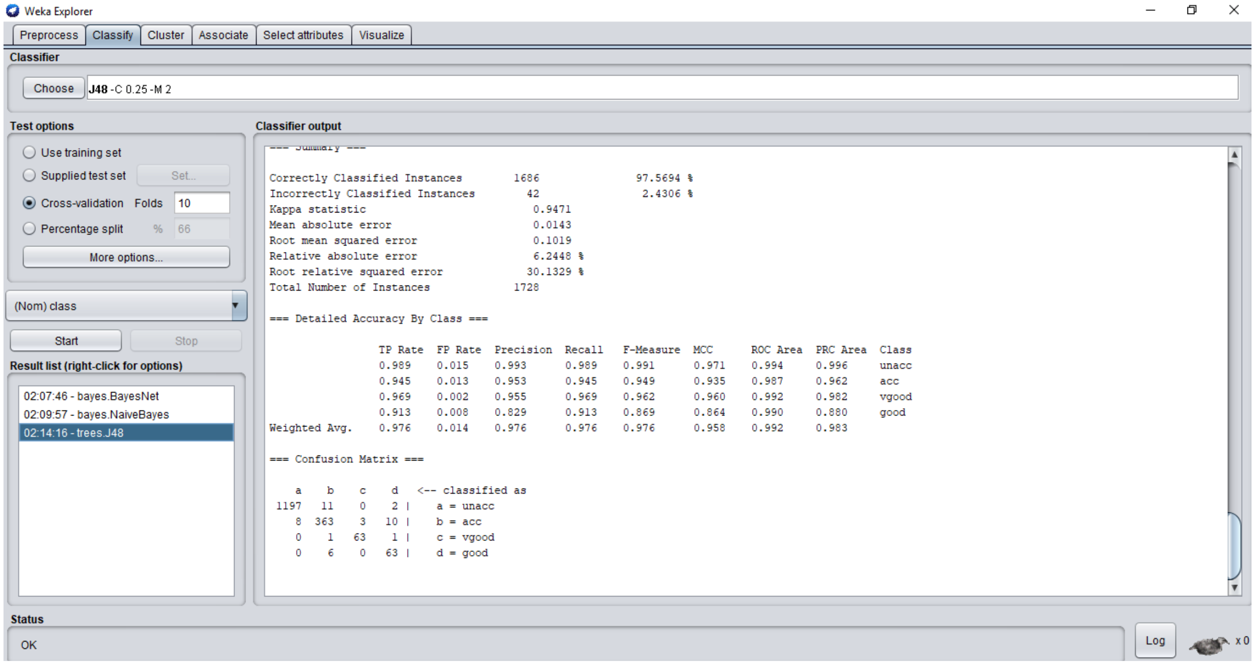
(Nom) class

02:07:46 - bayes.BayesNet
02:09:57 - bayes.NaiveBayes

```
=== Summary ===

Correctly Classified Instances        1457               84.3171 %
Incorrectly Classified Instances       271               15.6829 %
Kappa statistic                          0.6284
Mean absolute error                      0.1227
Root mean squared error                  0.2333
Relative absolute error                 53.5906 %
Root relative squared error             69.0145 %
Total Number of Instances             1728

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.961    0.324    0.874      0.961    0.915      0.694    0.961     0.984     unacc
                 0.534    0.056    0.732      0.534    0.617      0.539    0.926     0.750     acc
                 0.692    0.003    0.900      0.692    0.783      0.782    0.995     0.887     vgood
                 0.638    0.014    0.657      0.638    0.647      0.633    0.979     0.584     good
Weighted Avg.    0.843    0.240    0.835      0.843    0.833      0.660    0.955     0.912

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1163   46    0    1 |    a = unacc
  159  205    5   15 |    b = acc
    0   13   45    7 |    c = vgood
    9   16    0   44 |    d = good
```

Status: OK

J48:



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

Result list (right-click for options)

02:07:46 - bayes.BayesNet
02:09:57 - bayes.NaiveBayes
02:14:16 - trees.J48

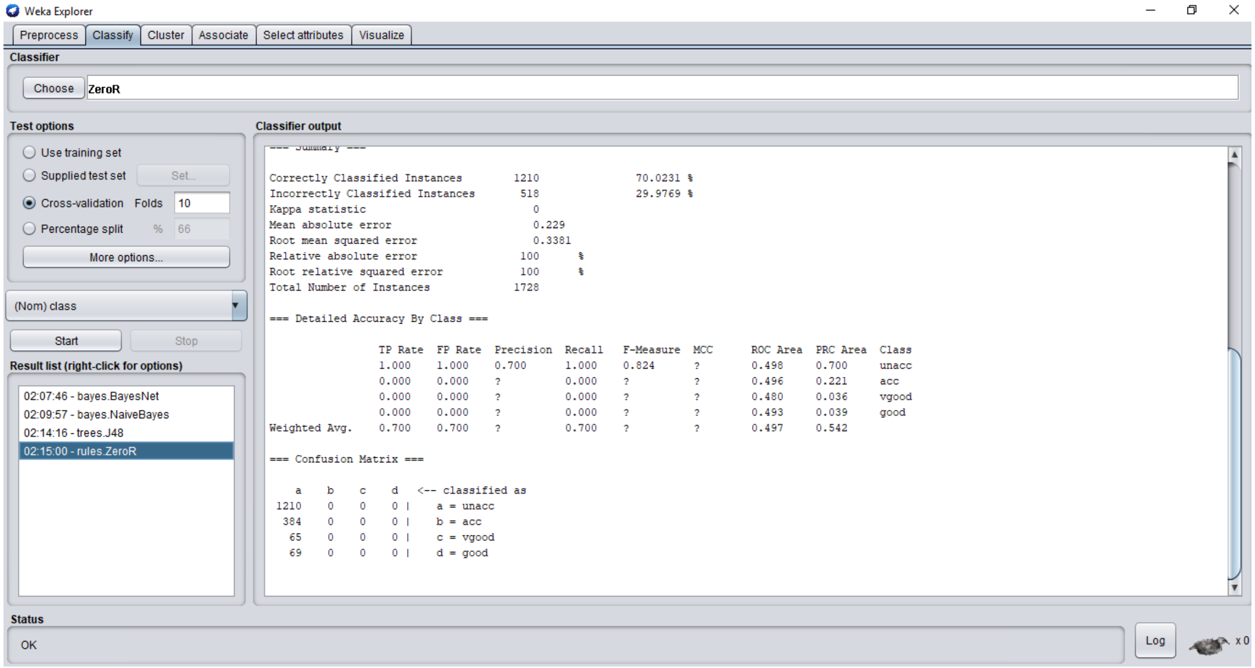
Classifier output

```
=== Summary ===

Correctly Classified Instances        1686               97.5694 %
Incorrectly Classified Instances        42                2.4306 %
Kappa statistic                          0.9471
Mean absolute error                      0.0143
Root mean squared error                  0.1019
Relative absolute error                  6.2448 %
Root relative squared error             30.1329 %
Total Number of Instances             1728

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.989    0.015    0.993      0.989    0.991      0.971    0.994     0.996     unacc
                 0.945    0.013    0.953      0.945    0.949      0.935    0.987     0.962     acc
                 0.969    0.002    0.955      0.969    0.962      0.960    0.992     0.982     vgood
                 0.913    0.008    0.829      0.913    0.869      0.864    0.990     0.880     good
Weighted Avg.    0.976    0.014    0.976      0.976    0.976      0.958    0.992     0.983

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1197   11    0    2 |    a = unacc
    8  363    3   10 |    b = acc
    0    1   63    1 |    c = vgood
    0    6    0   63 |    d = good
```

Status

OK

ZeroR:



Classifier

ZeroR

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

Result list (right-click for options)

02:07:46 - bayes.BayesNet
02:09:57 - bayes.NaiveBayes
02:14:16 - trees.J48
02:15:00 - rules.ZeroR

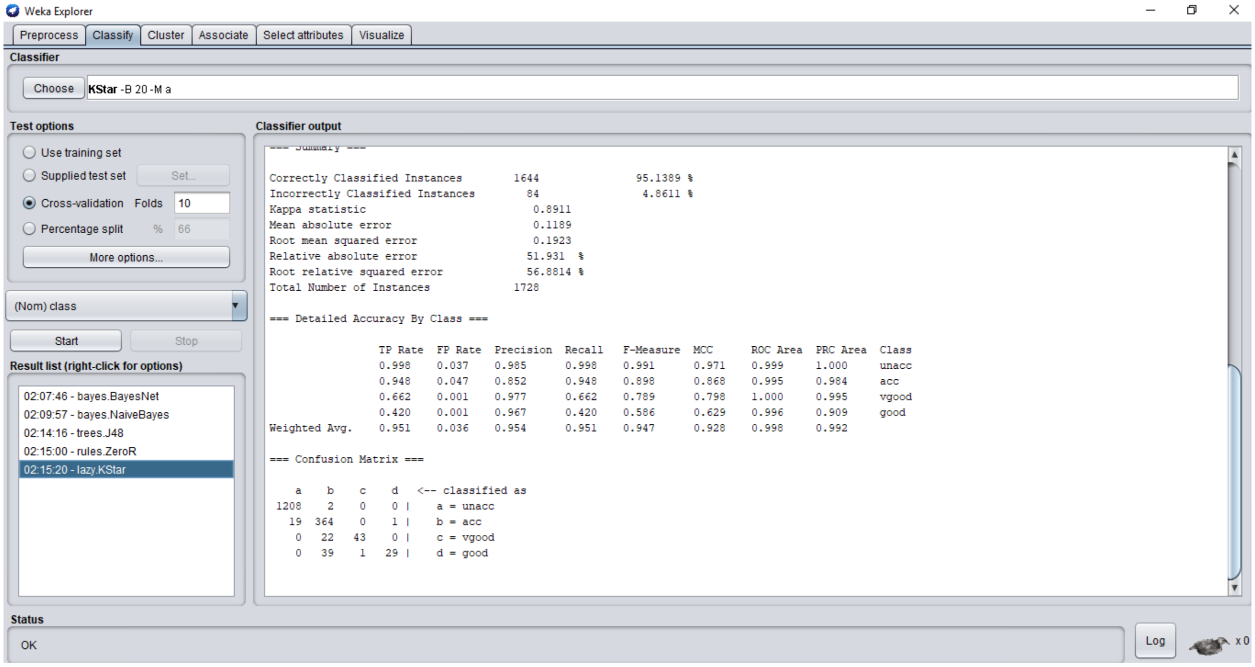
Classifier output

```
=== Summary ===

Correctly Classified Instances        1210               70.0231 %
Incorrectly Classified Instances       518               29.9769 %
Kappa statistic                          0
Mean absolute error                      0.229
Root mean squared error                  0.3381
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             1728

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.700      1.000    0.824      ?        0.498     0.700     unacc
                 0.000    0.000    ?          0.000    ?          ?        0.496     0.221     acc
                 0.000    0.000    ?          0.000    ?          ?        0.480     0.036     vgood
                 0.000    0.000    ?          0.000    ?          ?        0.493     0.039     good
Weighted Avg.    0.700    0.700    ?          0.700    ?          ?        0.497     0.542

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1210    0    0    0 |    a = unacc
  384    0    0    0 |    b = acc
   65    0    0    0 |    c = vgood
   69    0    0    0 |    d = good
```

Status

OK

Kstar:



Classifier

KStar -B 20 -M a

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

Result list (right-click for options)

02:07:46 - bayes.BayesNet
02:09:57 - bayes.NaiveBayes
02:14:16 - trees.J48
02:15:00 - rules.ZeroR
02:15:20 - lazy.KStar

Classifier output

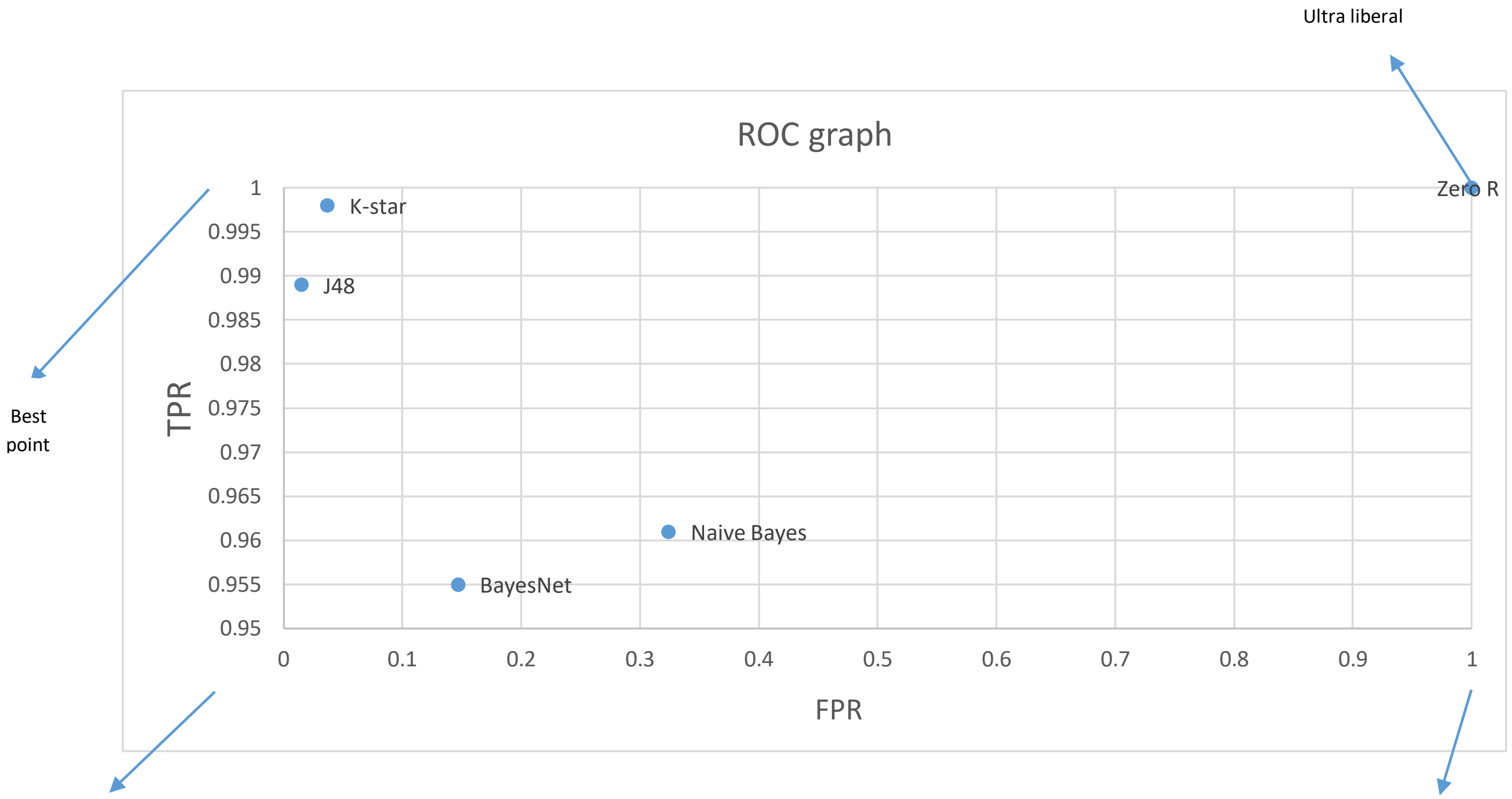
```
=== Summary ===

Correctly Classified Instances        1644               95.1389 %
Incorrectly Classified Instances        84                4.8611 %
Kappa statistic                          0.8911
Mean absolute error                      0.1189
Root mean squared error                  0.1923
Relative absolute error                 51.931  %
Root relative squared error             56.8814 %
Total Number of Instances             1728

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.037    0.985      0.998    0.991      0.971    0.999     1.000     unacc
                 0.948    0.047    0.852      0.948    0.898      0.868    0.995     0.984     acc
                 0.662    0.001    0.977      0.662    0.789      0.798    1.000     0.995     vgood
                 0.420    0.001    0.967      0.420    0.586      0.629    0.996     0.909     good
Weighted Avg.    0.951    0.036    0.954      0.951    0.947      0.928    0.998     0.992

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1208    2    0    0 |    a = unacc
   19  364    0    1 |    b = acc
    0   22   43    0 |    c = vgood
    0   39    1   29 |    d = good
```

Status

OK

We are considering the 'unacc' class positive interest.

Classifiers	True positive Rate (TPR)	False Positive Rate (FPR)
Bayes Net	0.955	0.147
Naïve Bayes	0.961	0.324
ZeroR classifier	1.0	1.0
J48	0.989	0.015
KStar Classifier	0.998	0.037



Ultra conservative

Fig: ROC plot

Worst Point

Evaluation: From the ROC graph I should choose the classifier that gives best result out of these five classifiers. Each classifier's point is labeled in the ROC. The selected point shows the best point, worst, UL, UC points of the ROC plot. This point shows us which classifier is giving us the best relative best classification. Amongst all the classifier, Kstar and J48 classifier is the closes to the best point and farthest from the worst point possible. This worst point is very unreliable. Zero R is situated in the ultra-liberal point which is biased towards giving positive result more often. Naive Bayes and Bayes Net is close to worst point which tends to make a lot of mistake. So amongst all these classifiers the j48 classifier gives better classification of the car dataset. Amongst all these Kstaris closest to being the best point.

Decision: Kstar is the best classifier amongst these selected classifiers for car dataset.