MET CS699 SO1 DATA MINING

FINAL PROJECT REPORT

Data Use

project-2018-BRFSS-arthritis.arff

Tool Use

Weka: Using Weka to analynize the dataset

Excel: Using Excel to do the comparation to choose best model

Preprocess

Remove Missing Value: Before I proceeded with data processing, I checked the data and found many "?" in the dataset. These "?" had a significant impact on my output results. Therefore, I preprocessed the dataset by removing all data rows containing "?".

Data Reduction: For this dataset, there are about 108 attributes. But, not all attributes are useful. So I do the data reduction for this dataset, and make it remaining 22 attribute by different attribute selection methods.

Attribute Selection Methods

CfsSubsetEval:

CfsSubsetEval is a feature selection method used in machine learning. This method, standing for "Correlation-based Feature Selection", operates on the premise that a good feature subset contains features highly correlated with the classification, yet uncorrelated with each other. The CfsSubsetEval evaluation function measures the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy that exists among them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. This method has been implemented in popular machine learning tools like Weka, and it's often used in the preprocessing stage of a machine learning pipeline, helping to reduce dimensionality and improve model performance.

CorrelationAttributeEval:

CorrelationAttributeEval is a feature selection technique used in machine learning, specifically implemented in the Weka machine learning library. The method evaluates the worth of an attribute by measuring the correlation between it and the class. The correlation metric used can be either Pearson's correlation, which measures linear relationships, or Spearman's correlation, which measures monotonic relationships. The absolute value of the correlation signifies the strength of the relationship with the class, with 1 being a perfect correlation, -1 a perfect inverse correlation, and 0 no correlation. In the context of feature selection, this method can be used to rank attributes by their relevance to the class. It's a univariate filter method, meaning it considers each attribute independently of the others. Attributes with high absolute correlation values are considered more important for prediction tasks. This method is often used in the preprocessing stage of a machine learning pipeline to select features that have the most predictive power, thereby reducing the dimensionality of the dataset and potentially improving the performance of the model.

InfoGainAttributeEval

InfoGainAttributeEval is a feature selection method used in machine learning and specifically implemented in the Weka machine learning library. This method evaluates the worth of an attribute by measuring the information gain in relation to the class.Information gain is a metric that measures how much information is gained about the class by knowing the value of an attribute. It's based on the concept of entropy from information theory, which quantifies the uncertainty or impurity in a set of instances. The information gain of an attribute is calculated as the difference in entropy before and after the attribute is given. In the context of feature selection, this method can be used to rank attributes by their information gain values. Attributes with high information gain are considered more important for prediction tasks because they provide more useful information about the class.This method is often used in the preprocessing stage of a machine learning pipeline to select features that have the most predictive power. By reducing the dimensionality of the dataset and focusing on the most informative attributes, InfoGainAttributeEval can help improve the performance of the model.

OneRAttributeEval

OneRAttributeEval is a feature selection method used in machine learning, specifically implemented in the Weka machine learning library. This method evaluates the worth of an attribute by using the OneR (One Rule) algorithm, which generates a single rule for each attribute in the data and then selects the rule with the smallest total error rate. The OneR algorithm, or "One Rule", is a simple, yet often surprisingly accurate, classification algorithm that generates one rule for each predictor in the data, then selects the one rule with the smallest total error.In the context of feature selection, this method can be used to rank attributes based on the accuracy of their corresponding OneR rules. Attributes whose OneR rules have low error rates are considered more important for prediction tasks.This method is often used in the preprocessing stage of a machine learning pipeline to select features that have the most predictive power. By focusing on the most informative attributes,OneRAttributeEval can help improve the performance of the model.

SymmetricalUncertAttributeEval

SymmetricalUncertAttributeEval is a feature selection method used in machine learning, specifically implemented in the Weka machine learning library. This method evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class.Symmetrical Uncertainty is a measure derived from Information Theory. It is a symmetric measure of association between two random variables, providing a normalized value in the range of 0 (no association between the variables) to 1 (complete association). It is based on the concept of entropy and information gain.In the context of feature selection, SymmetricalUncertAttributeEval can be used to rank attributes based on their symmetrical uncertainty values in relation to the class. Attributes with high symmetrical uncertainty are considered more important for prediction tasks because they share a strong association with the class.

This method is often used in the preprocessing stage of a machine learning pipeline to select features that have the most predictive power. By reducing the dimensionality of the dataset and focusing on the most informative attributes, SymmetricalUncertAttributeEval can help improve the performance of the model.

Attributes After Attribute Selections(Order by Relevant)

CfsSubsetEval:

employ1	*pneuvac4*	*diffwalk*	*diffdres*
diabete3	*physhlth*	*chckdny1*	*iday*
persdoc2	*checkup1*	*chcocncr*	*addepev2*
chcscncr	*cvdstrk3*	*qstver*	*x.age80*
x.ageg5yr	*x.chldcnt*	*x.rfbing5*	*x.exteth3*
x.casthm1	*havarth3*		

CorrelationAttributeEval

x.ageg5yr	x.age80	checkup1	*diffwalk*
x.rfbing5	*pneuvac4*	*physhlth*	*employ1*

x.exteth3	x.phys14d	x.rfhlth	*diabete3*
x.age65yr	x.hcvu651	x.bmi5	addepev2
persdoc2	chcscncr	flushot6	x.age.g
x.totinda	exerany2		

InfoGainAttributeEval

iday	*diffwalk*	x.age80	*x.age.g*
x.phys14d	*qstver*	x.ageg5yr	*genhlth*
imonth	*pneuvac4*	*physhlth*	checkup1
x.rfbing5	persdoc2	*x.incomg*	*employ1*
rmvteth4	*x.chldcnt*	*diabete3*	*x.exteth3*
chcscncr	*x.rfhlth*		

OneRAttributeEval

x.state	wtkg3	*genhlth*	dispcode
iyear	hlthpln1	menthlth	*physhlth*
rmvteth4	lastden4	x.psu	fmonth
imonth	*iday*	qstlang	x.metstat
cvdinfr4	x.casthm1	sleptim1	persdoc2
exerany2	chcocncr		

SymmetricalUncertAttributeEval

x.ageg5yr	x.age80	*diffwalk*	x.rfbing5
checkup1	persdoc2	*chcscncr*	*physhlth*
x.age.g	x.phys14d	*x.chldcnt*	*pneuvac4*
employ1	diffdres	*diabete3*	chckdny1
cvdstrk3	*x.exteth3*	x.rfhlth	addepev2
genhlth	qstver		

Classification Algorithm

NaiveBayes: This is a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. It's particularly suited for high-dimensional datasets and is commonly used in natural language processing and spam filtering.

*Bagging:*Short for Bootstrap Aggregating, Bagging is a general-purpose procedure for reducing the variance of a machine learning method. It works by creating multiple subsets of the original dataset (with replacement), training a separate model on each subset, and then averaging the predictions (for regression) or voting for the most popular class (for classification).

*Logistic:*Logistic Regression is a statistical model that uses a logistic function to model a binary dependent variable. In machine learning, it's often used for binary classification problems - predicting one of two possible outcomes such as 'yes' or 'no'.

*Kstar:*K* (K Star) is an instance-based classifier, which means it doesn't explicitly build a model. Instead, it memorizes the training instances which are subsequently used as knowledge for the prediction phase. The Kstar algorithm uses entropy-based distance measure.

*MultipleClassClassifier:*This refers to a classification problem with more than two classes. Each sample can only be labeled as one class. For instance, predicting the type of fruit (apple, orange, banana, etc.) based on certain features is a multi-class classification problem. Various algorithms can be used to solve this problem, including the ones mentioned above, as well as others like Decision Trees, Random Forests, and SVMs.

What I learned

Through the process of building these 25 models, I learned a great deal about the performance of different feature selection methods and machine learning models on my dataset. Here's what stood out to me: Effectiveness of

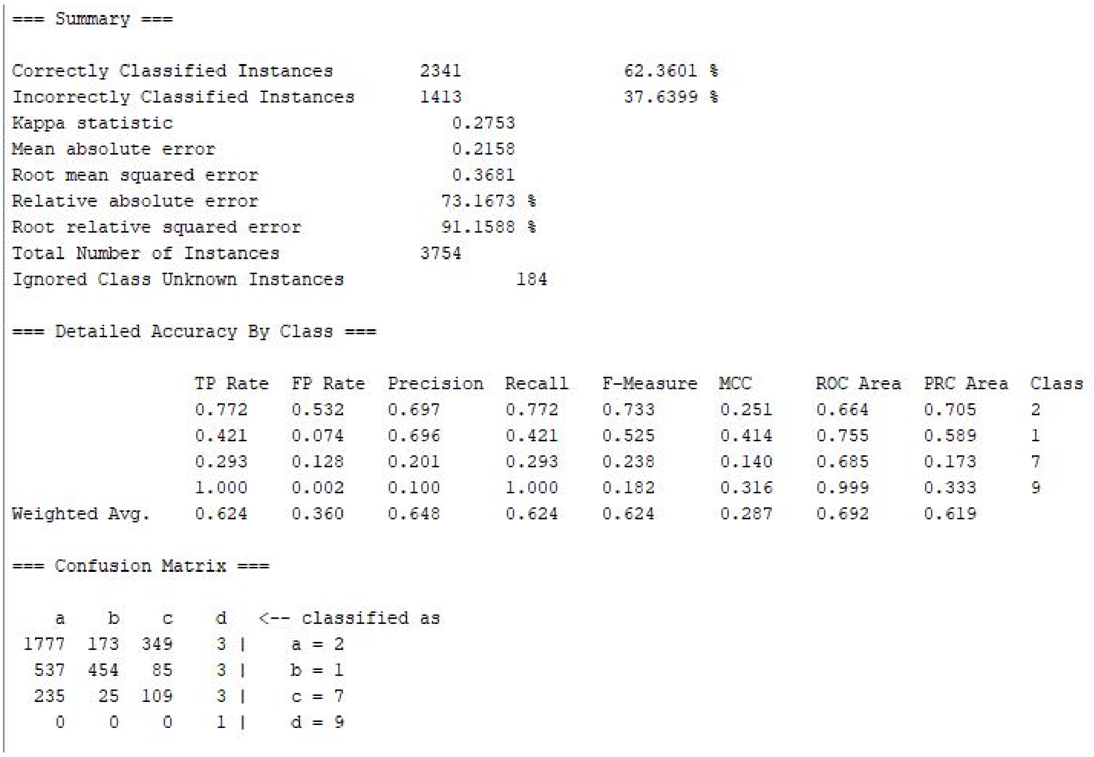
Feature Selection Methods: I observed that the feature selection methods had a significant impact on the performance of the models. InfoGainAttributeEval seemed to consistently result in high accuracy across different models, indicating its effectiveness for this dataset. On the contrary, OneRAttributeEval consistently resulted in lower accuracy, suggesting that the one-rule heuristic might not be the best approach for feature selection with this particular data. Consistency of Certain Models: Certain models, notably Logistic and Bagging, performed well regardless of the feature selection method used. This consistency suggests that these models might be more robust or better suited to the characteristics of my dataset. Model-Specific Synergies: There were instances where specific combinations of models and feature selection methods stood out. For instance, when I combined Bagging with InfoGainAttributeEval, the model achieved 100% accuracy. This suggests a potential synergy between this specific model and feature selection method. Risk of Overfitting: I noticed that some models achieved perfect or near-perfect accuracy, such as the Bagging and Logistic models when paired with InfoGainAttributeEval. While at first glance this might seem like an ideal outcome, it also raised concerns about potential overfitting. These models might be overly tailored to the training data and could perform poorly when introduced to new, unseen data. Sensitivity of Models to Feature Selection: The variation in model performance across different feature selection methods provided valuable insights. For example, NaiveBayes showed a substantial range in performance, from just 38.192% accuracy with OneRAttributeEval to 99.5937% with InfoGainAttributeEval. This suggests that the choice of feature selection method can heavily influence the performance of certain models. As for the most interesting models, I was particularly intrigued by those that yielded the highest accuracy, such as Bagging and Logistic models with InfoGainAttributeEval. Additionally, models that showed significant differences in performance depending on the feature selection method, like NaiveBayes, offered fascinating insights into the sensitivity of models to feature selection methods. In the appendix of my report, I will provide detailed information on all 25 models, including the feature selection method used, the accuracy achieved, and other relevant metrics and observations, such as precision, recall, F1-score, training time, complexity, and so forth.

Models Detail

1. CfsSubsetEval

1. NaiveBayes

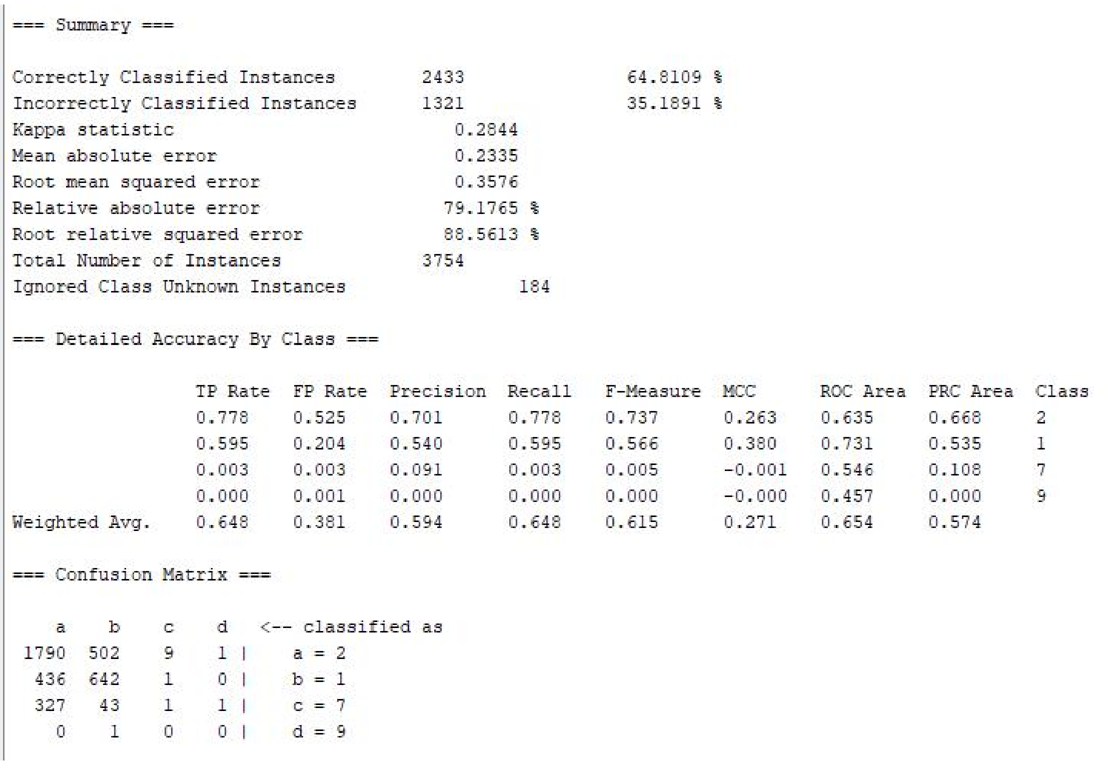
Accuracy: 62.3601%



```
=== Summary ===

Correctly Classified Instances        2341               62.3601 %
Incorrectly Classified Instances      1413               37.6399 %
Kappa statistic                          0.2753
Mean absolute error                      0.2158
Root mean squared error                  0.3681
Relative absolute error                 73.1673 %
Root relative squared error             91.1588 %
Total Number of Instances             3754
Ignored Class Unknown Instances                184

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.772    0.532    0.697      0.772    0.733      0.251    0.664     0.705     2
                 0.421    0.074    0.696      0.421    0.525      0.414    0.755     0.589     1
                 0.293    0.128    0.201      0.293    0.238      0.140    0.685     0.173     7
                 1.000    0.002    0.100      1.000    0.182      0.316    0.999     0.333     9
Weighted Avg.    0.624    0.360    0.648      0.624    0.624      0.287    0.692     0.619

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1777  173  349    3 |    a = 2
  537  454   85    3 |    b = 1
  235   25  109    3 |    c = 7
    0    0    0    1 |    d = 9
```

2. Bagging

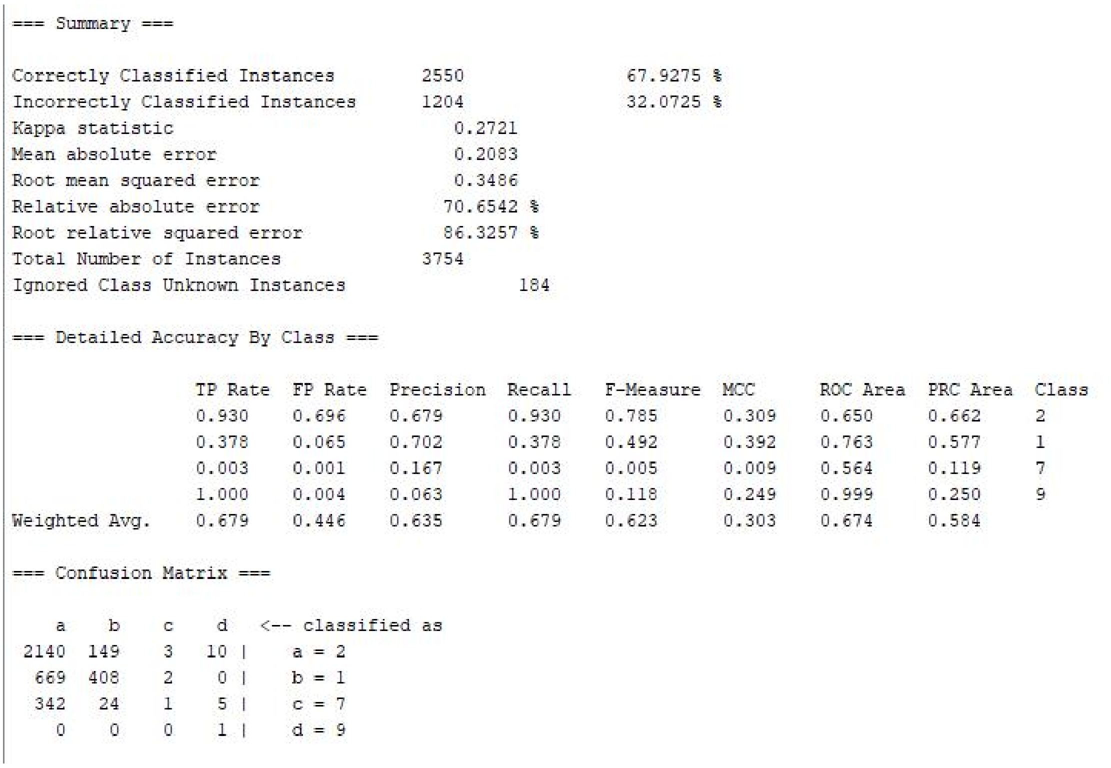
Accuracy: 64.8109%



```
=== Summary ===

Correctly Classified Instances        2433               64.8109 %
Incorrectly Classified Instances      1321               35.1891 %
Kappa statistic                          0.2844
Mean absolute error                      0.2335
Root mean squared error                  0.3576
Relative absolute error                 79.1765 %
Root relative squared error             88.5613 %
Total Number of Instances             3754
Ignored Class Unknown Instances                184

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.778    0.525    0.701      0.778    0.737      0.263    0.635     0.668     2
                 0.595    0.204    0.540      0.595    0.566      0.380    0.731     0.535     1
                 0.003    0.003    0.091      0.003    0.005      -0.001   0.546     0.108     7
                 0.000    0.001    0.000      0.000    0.000      -0.000   0.457     0.000     9
Weighted Avg.    0.648    0.381    0.594      0.648    0.615      0.271    0.654     0.574

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1790  502    9    1 |    a = 2
  436  642    1    0 |    b = 1
  327   43    1    1 |    c = 7
    0    1    0    0 |    d = 9
```

3. *Logistic*

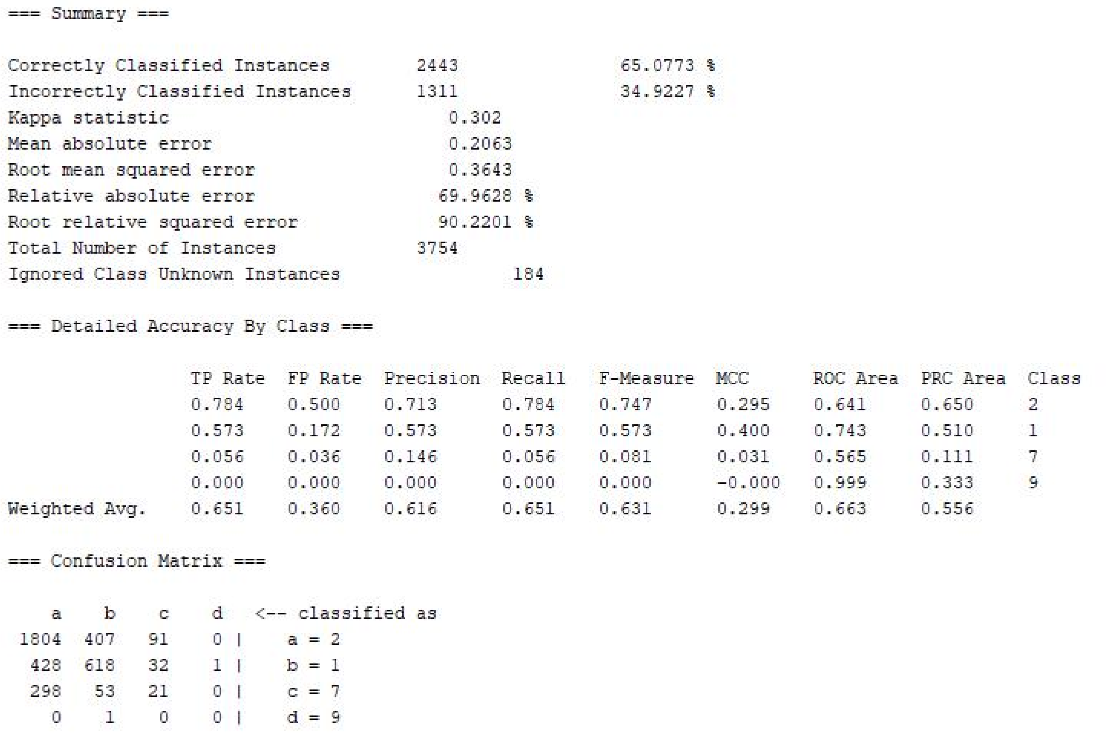
Accuracy: 67.9275%



```
=== Summary ===

Correctly Classified Instances        2550               67.9275 %
Incorrectly Classified Instances      1204               32.0725 %
Kappa statistic                          0.2721
Mean absolute error                      0.2083
Root mean squared error                  0.3486
Relative absolute error                 70.6542 %
Root relative squared error             86.3257 %
Total Number of Instances             3754
Ignored Class Unknown Instances                184

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.930    0.696    0.679      0.930    0.785      0.309    0.650     0.662     2
                 0.378    0.065    0.702      0.378    0.492      0.392    0.763     0.577     1
                 0.003    0.001    0.167      0.003    0.005      0.009    0.564     0.119     7
                 1.000    0.004    0.063      1.000    0.118      0.249    0.999     0.250     9
Weighted Avg.    0.679    0.446    0.635      0.679    0.623      0.303    0.674     0.584

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2140  149    3   10 |    a = 2
  669  408    2    0 |    b = 1
  342   24    1    5 |    c = 7
    0    0    0    1 |    d = 9
```

4. *Kstar*

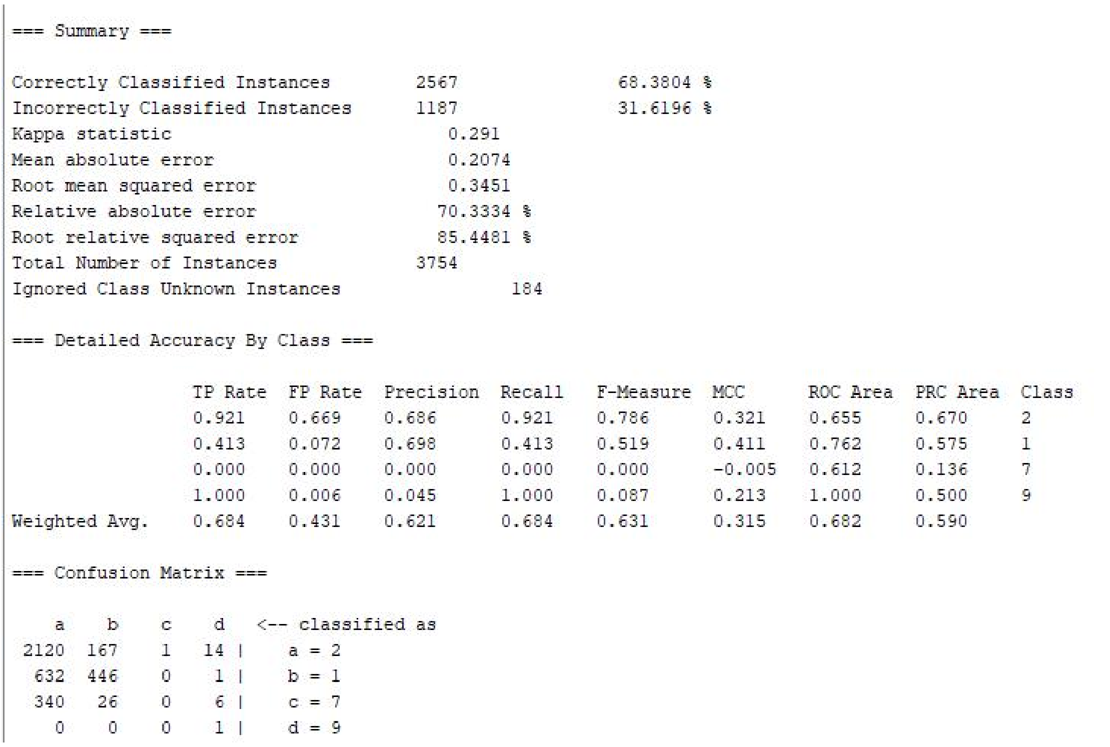
Accuracy: 65.0773%



```
=== Summary ===

Correctly Classified Instances        2443               65.0773 %
Incorrectly Classified Instances      1311               34.9227 %
Kappa statistic                          0.302
Mean absolute error                      0.2063
Root mean squared error                  0.3643
Relative absolute error                 69.9628 %
Root relative squared error             90.2201 %
Total Number of Instances             3754
Ignored Class Unknown Instances                184

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.784    0.500    0.713      0.784    0.747      0.295    0.641     0.650     2
                 0.573    0.172    0.573      0.573    0.573      0.400    0.743     0.510     1
                 0.056    0.036    0.146      0.056    0.081      0.031    0.565     0.111     7
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.999     0.333     9
Weighted Avg.    0.651    0.360    0.616      0.651    0.631      0.299    0.663     0.556

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1804  407   91    0 |    a = 2
  428  618   32    1 |    b = 1
  298   53   21    0 |    c = 7
    0    1    0    0 |    d = 9
```

5. *MultipleClassClassifier*

Accuracy: 68.3804%

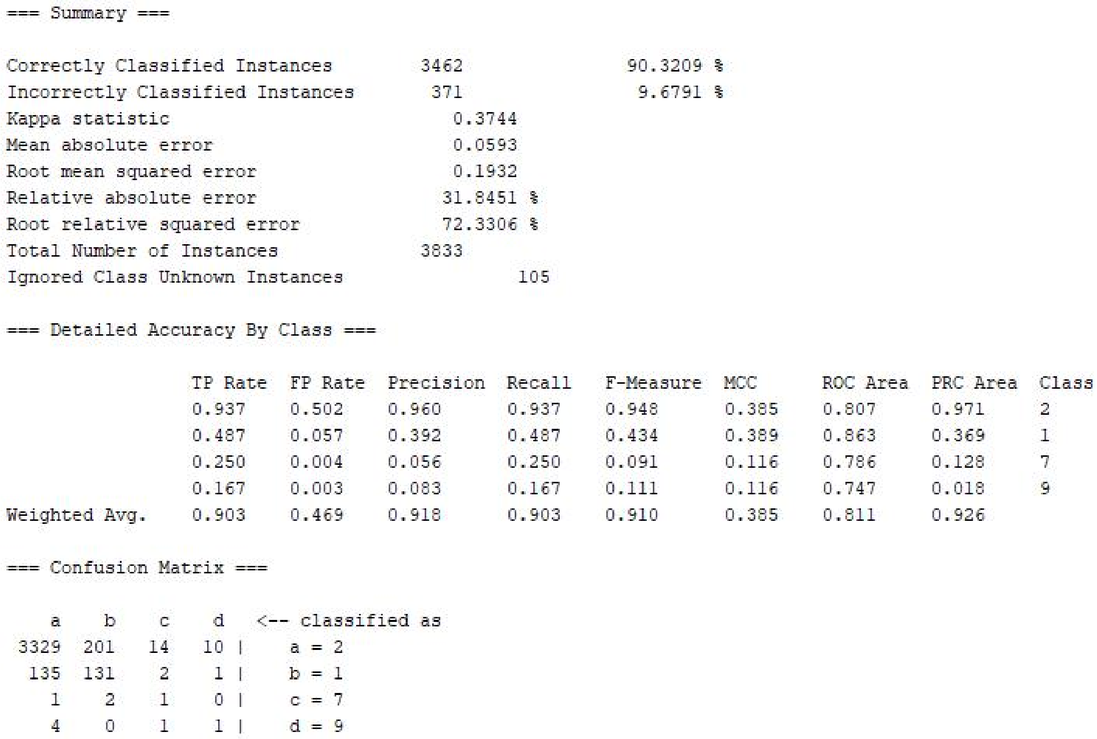


```
=== Summary ===

Correctly Classified Instances        2567               68.3804 %
Incorrectly Classified Instances      1187               31.6196 %
Kappa statistic                          0.291
Mean absolute error                      0.2074
Root mean squared error                  0.3451
Relative absolute error                 70.3334 %
Root relative squared error             85.4481 %
Total Number of Instances             3754
Ignored Class Unknown Instances                184

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.921    0.669    0.686      0.921    0.786      0.321    0.655     0.670     2
                 0.413    0.072    0.698      0.413    0.519      0.411    0.762     0.575     1
                 0.000    0.000    0.000      0.000    0.000      -0.005   0.612     0.136     7
                 1.000    0.006    0.045      1.000    0.087      0.213    1.000     0.500     9
Weighted Avg.    0.684    0.431    0.621      0.684    0.631      0.315    0.682     0.590

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2120  167    1   14 |    a = 2
  632  446    0    1 |    b = 1
  340   26    0    6 |    c = 7
    0    0    0    1 |    d = 9
```

2. CorrelationAttributeEval

1. *NaiveBayes*

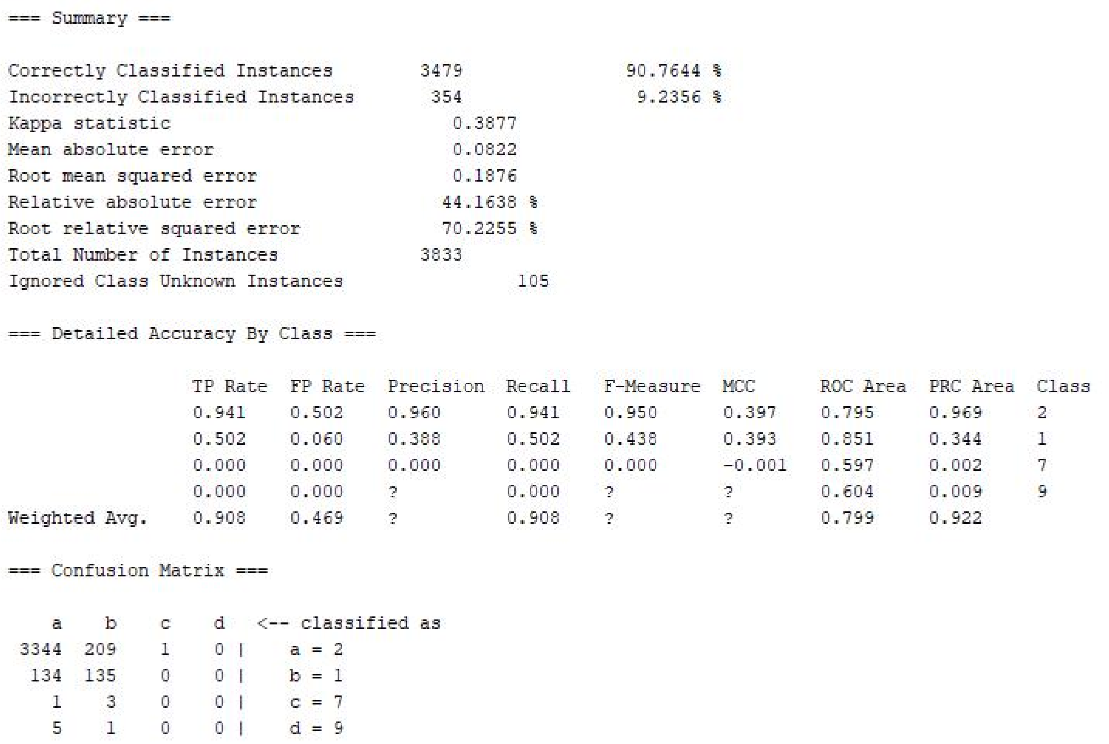
Accuracy: 90.3209%



```
=== Summary ===

Correctly Classified Instances        3462               90.3209 %
Incorrectly Classified Instances       371                9.6791 %
Kappa statistic                          0.3744
Mean absolute error                      0.0593
Root mean squared error                  0.1932
Relative absolute error                 31.8451 %
Root relative squared error             72.3306 %
Total Number of Instances             3833
Ignored Class Unknown Instances                105

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.937    0.502    0.960      0.937    0.948      0.385    0.807     0.971     2
                 0.487    0.057    0.392      0.487    0.434      0.389    0.863     0.369     1
                 0.250    0.004    0.056      0.250    0.091      0.116    0.786     0.128     7
                 0.167    0.003    0.083      0.167    0.111      0.116    0.747     0.018     9
Weighted Avg.    0.903    0.469    0.918      0.903    0.910      0.385    0.811     0.926

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 3329  201   14   10 |    a = 2
  135  131    2    1 |    b = 1
    1    2    1    0 |    c = 7
    4    0    1    1 |    d = 9
```

2. *Bagging*

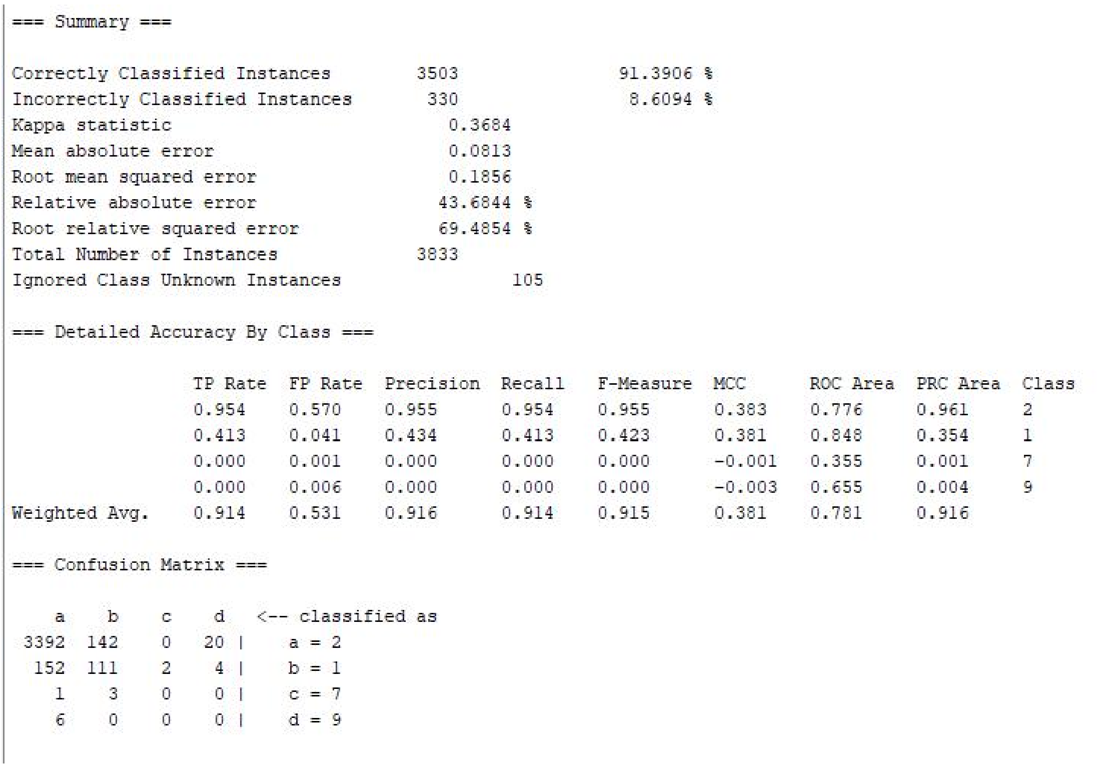
Accuracy: 90.7644%



```
=== Summary ===

Correctly Classified Instances        3479               90.7644 %
Incorrectly Classified Instances       354                9.2356 %
Kappa statistic                          0.3877
Mean absolute error                      0.0822
Root mean squared error                  0.1876
Relative absolute error                 44.1638 %
Root relative squared error             70.2255 %
Total Number of Instances             3833
Ignored Class Unknown Instances                105

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.941    0.502    0.960      0.941    0.950      0.397    0.795     0.969     2
                 0.502    0.060    0.388      0.502    0.438      0.393    0.851     0.344     1
                 0.000    0.000    0.000      0.000    0.000      -0.001   0.597     0.002     7
                 0.000    0.000    ?          0.000    ?          ?        0.604     0.009     9
Weighted Avg.    0.908    0.469    ?          0.908    ?          ?        0.799     0.922

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 3344  209    1    0 |    a = 2
  134  135    0    0 |    b = 1
    1    3    0    0 |    c = 7
    5    1    0    0 |    d = 9
```

3. Logistic

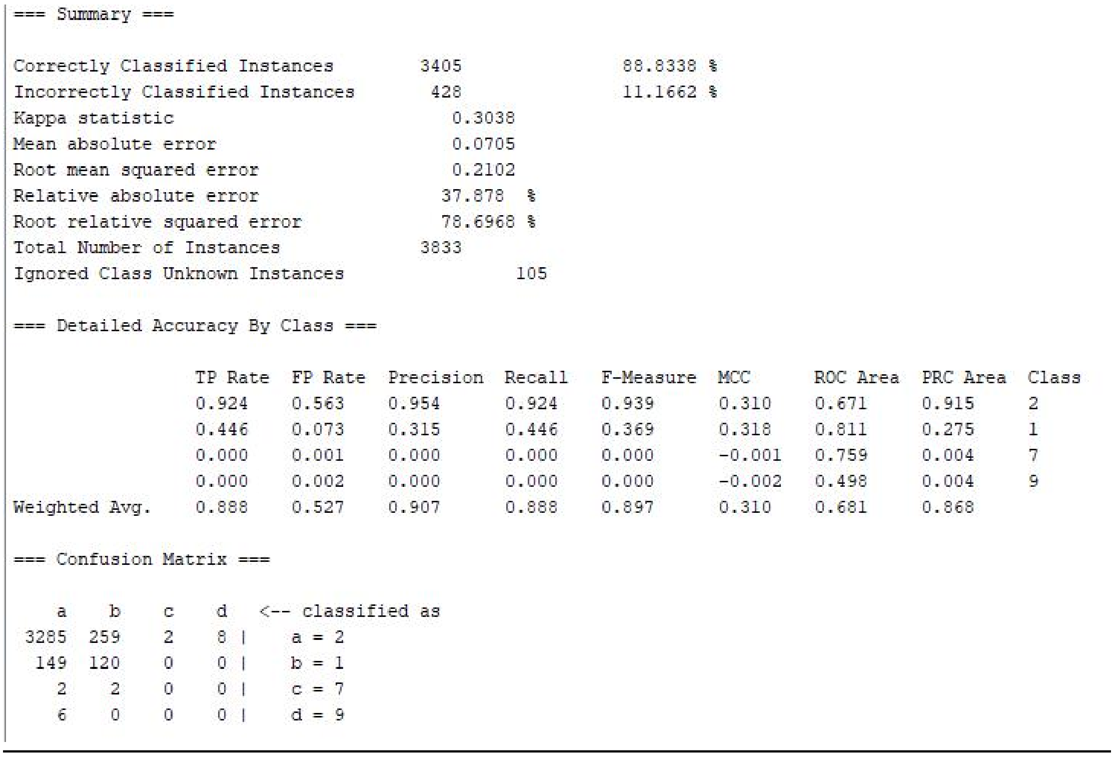
Accuracy: 91.3906%



```
=== Summary ===

Correctly Classified Instances        3503               91.3906 %
Incorrectly Classified Instances       330                8.6094 %
Kappa statistic                          0.3684
Mean absolute error                      0.0813
Root mean squared error                  0.1856
Relative absolute error                 43.6844 %
Root relative squared error             69.4854 %
Total Number of Instances             3833
Ignored Class Unknown Instances                105

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.954    0.570    0.955      0.954    0.955      0.383    0.776     0.961     2
                 0.413    0.041    0.434      0.413    0.423      0.381    0.848     0.354     1
                 0.000    0.001    0.000      0.000    0.000      -0.001   0.355     0.001     7
                 0.000    0.006    0.000      0.000    0.000      -0.003   0.655     0.004     9
Weighted Avg.    0.914    0.531    0.916      0.914    0.915      0.381    0.781     0.916

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 3392  142    0   20 |    a = 2
  152  111    2    4 |    b = 1
    1    3    0    0 |    c = 7
    6    0    0    0 |    d = 9
```

4. Kstar

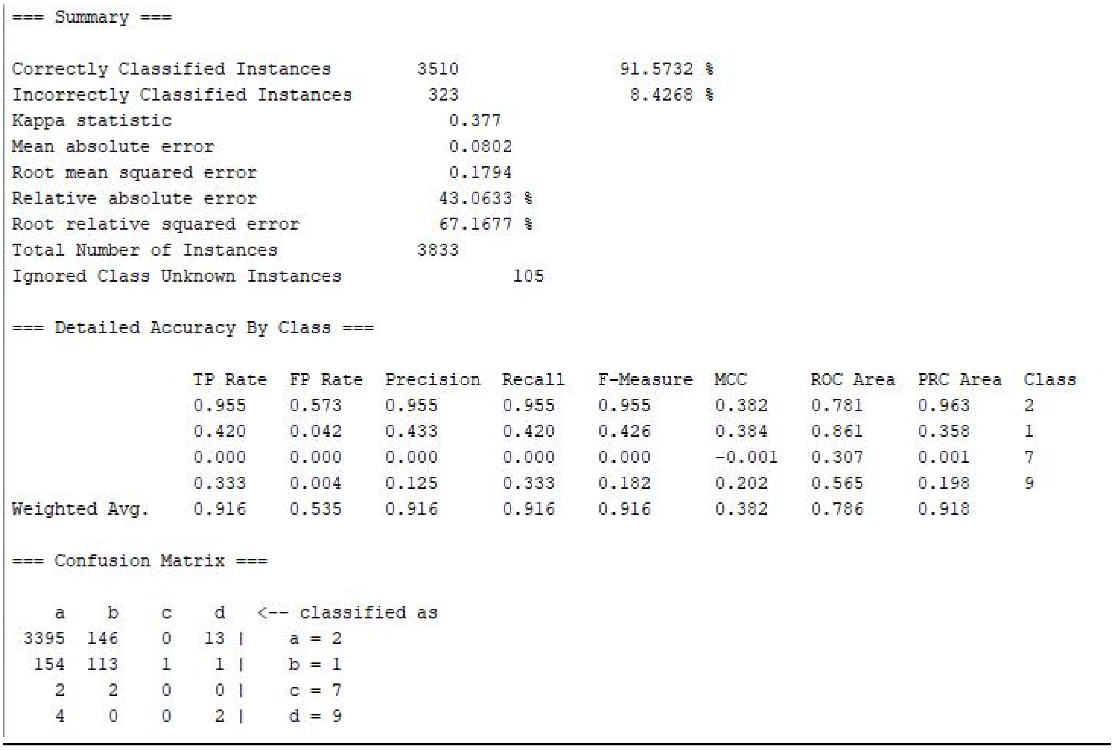
Accuracy: 88.8338%



```
=== Summary ===

Correctly Classified Instances        3405               88.8338 %
Incorrectly Classified Instances       428               11.1662 %
Kappa statistic                          0.3038
Mean absolute error                      0.0705
Root mean squared error                  0.2102
Relative absolute error                 37.878  %
Root relative squared error             78.6968 %
Total Number of Instances             3833
Ignored Class Unknown Instances                105

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.924    0.563    0.954      0.924    0.939      0.310    0.671     0.915     2
                0.446    0.073    0.315      0.446    0.369      0.318    0.811     0.275     1
                0.000    0.001    0.000      0.000    0.000      -0.001   0.759     0.004     7
                0.000    0.002    0.000      0.000    0.000      -0.002   0.498     0.004     9
Weighted Avg.   0.888    0.527    0.907      0.888    0.897      0.310    0.681     0.868

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 3285  259    2    8 |    a = 2
  149  120    0    0 |    b = 1
    2    2    0    0 |    c = 7
    6    0    0    0 |    d = 9
```

5. *MultipleClassClassifier*

Accuracy: 91.5732%

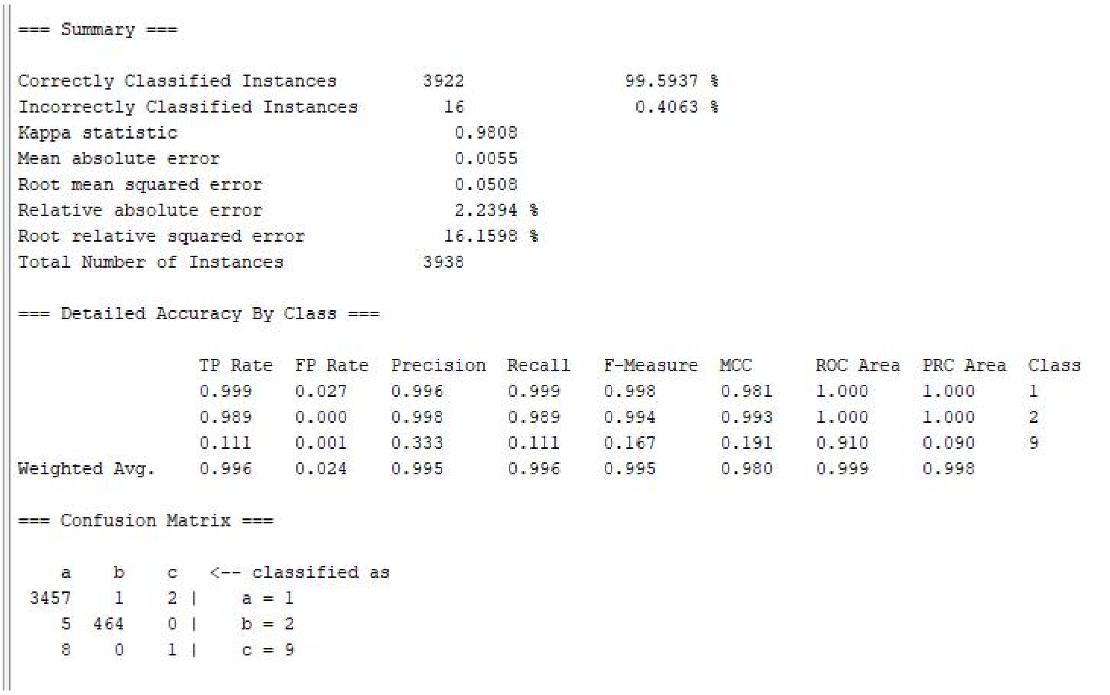


```
=== Summary ===

Correctly Classified Instances        3510               91.5732 %
Incorrectly Classified Instances       323                8.4268 %
Kappa statistic                          0.377
Mean absolute error                      0.0802
Root mean squared error                  0.1794
Relative absolute error                 43.0633 %
Root relative squared error             67.1677 %
Total Number of Instances             3833
Ignored Class Unknown Instances                105

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.955    0.573    0.955      0.955    0.955      0.382    0.781     0.963     2
                0.420    0.042    0.433      0.420    0.426      0.384    0.861     0.358     1
                0.000    0.000    0.000      0.000    0.000      -0.001   0.307     0.001     7
                0.333    0.004    0.125      0.333    0.182      0.202    0.565     0.198     9
Weighted Avg.   0.916    0.535    0.916      0.916    0.916      0.382    0.786     0.918

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 3395  146    0   13 |    a = 2
  154  113    1    1 |    b = 1
    2    2    0    0 |    c = 7
    4    0    0    2 |    d = 9
```

3. InfoGainAttributeEval

1. *NaiveBayes*

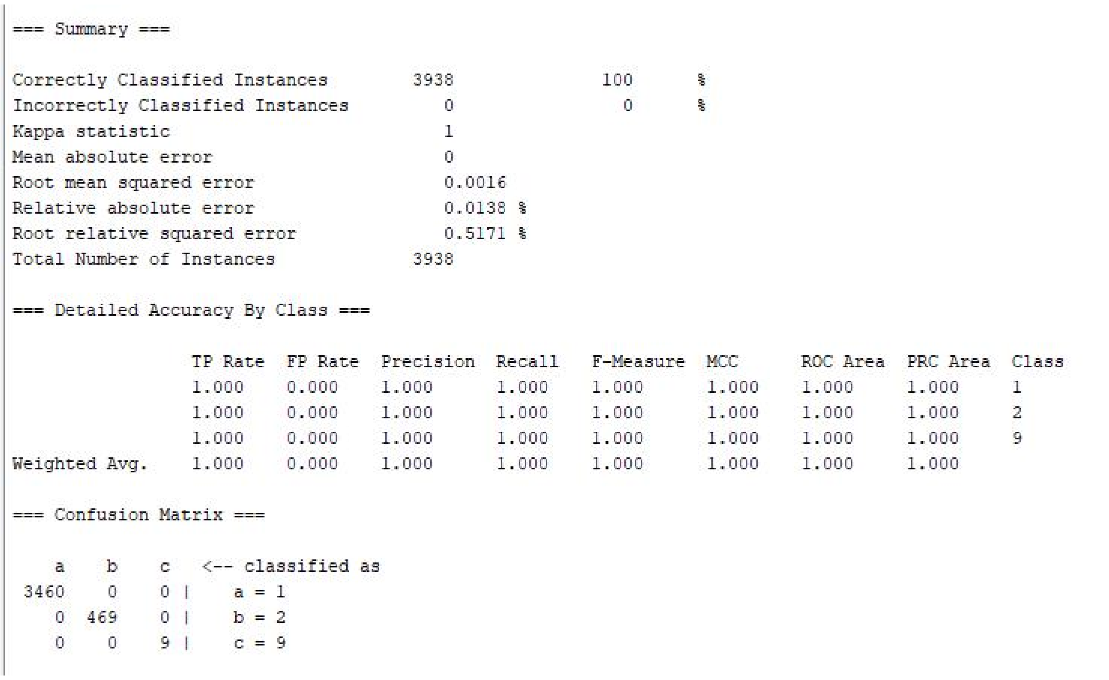
Accuracy: 99.5937%



```
=== Summary ===

Correctly Classified Instances        3922               99.5937 %
Incorrectly Classified Instances        16                0.4063 %
Kappa statistic                          0.9808
Mean absolute error                      0.0055
Root mean squared error                  0.0508
Relative absolute error                  2.2394 %
Root relative squared error             16.1598 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.999    0.027    0.996      0.999    0.998      0.981    1.000     1.000     1
                 0.989    0.000    0.998      0.989    0.994      0.993    1.000     1.000     2
                 0.111    0.001    0.333      0.111    0.167      0.191    0.910     0.090     9
Weighted Avg.    0.996    0.024    0.995      0.996    0.995      0.980    0.999     0.998

=== Confusion Matrix ===

    a    b    c   <-- classified as
 3457    1    2 |    a = 1
    5  464    0 |    b = 2
    8    0    1 |    c = 9
```

2. *Bagging*

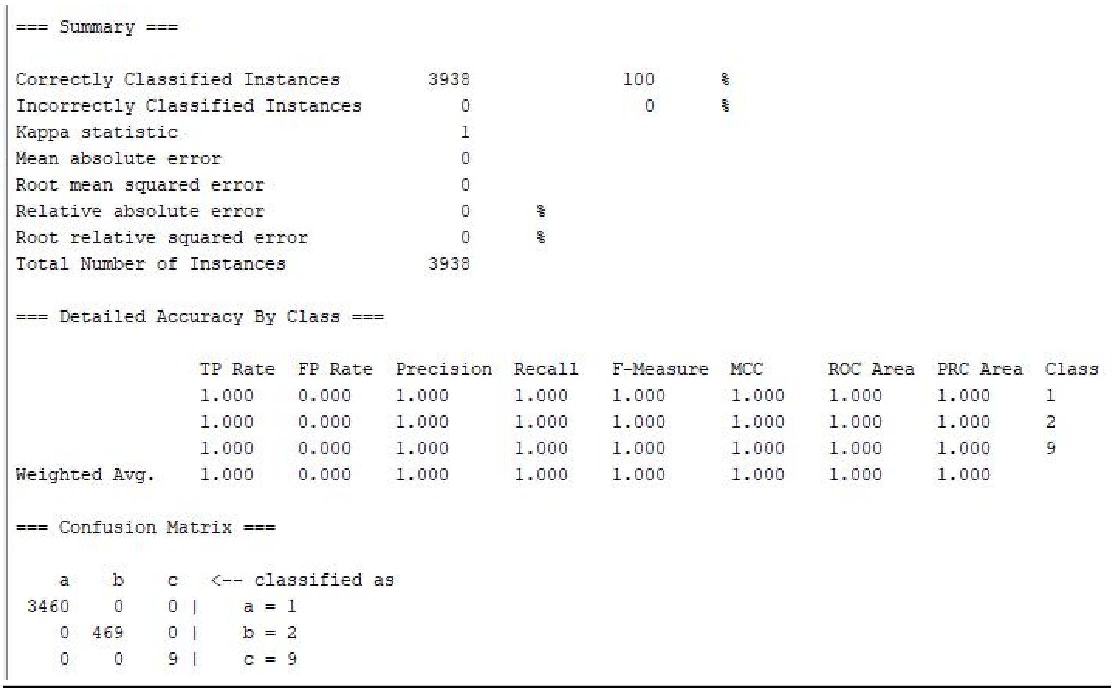
Accuracy: 100%



```
=== Summary ===

Correctly Classified Instances        3938              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0.0016
Relative absolute error                  0.0138 %
Root relative squared error              0.5171 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     1
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     2
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     9
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b    c   <-- classified as
 3460    0    0 |    a = 1
    0  469    0 |    b = 2
    0    0    9 |    c = 9
```

3. *Logistic*

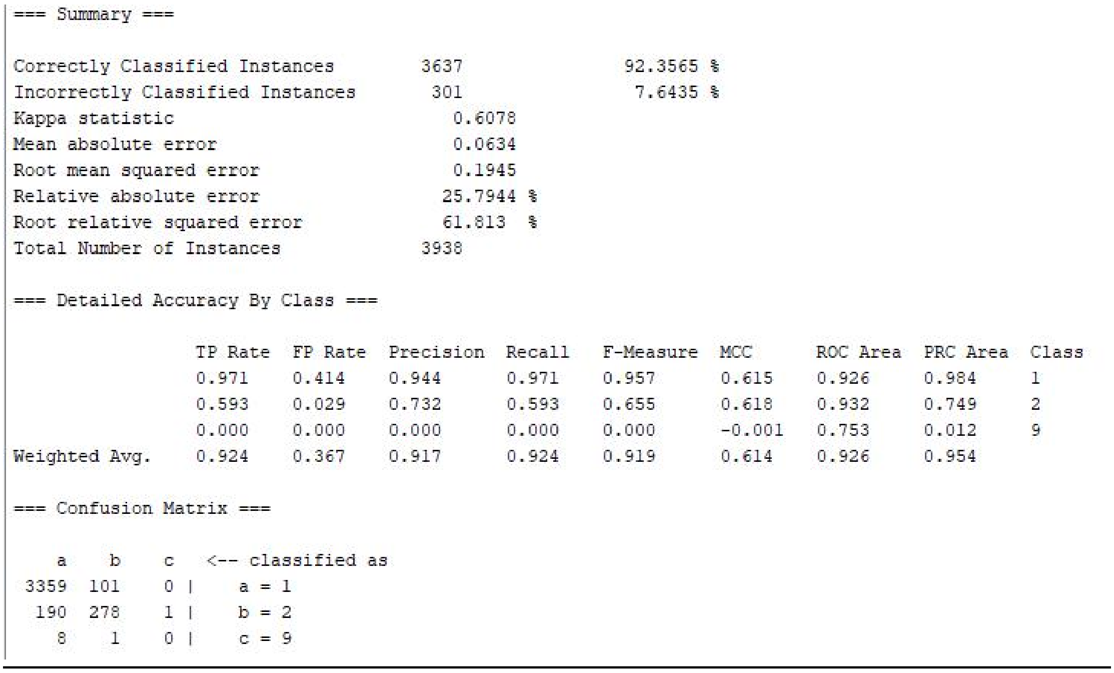
Accuracy: 100%



```
=== Summary ===

Correctly Classified Instances        3938              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     1
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     2
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     9
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b    c   <-- classified as
 3460    0    0 |    a = 1
    0  469    0 |    b = 2
    0    0    9 |    c = 9
```

4. *Kstar*

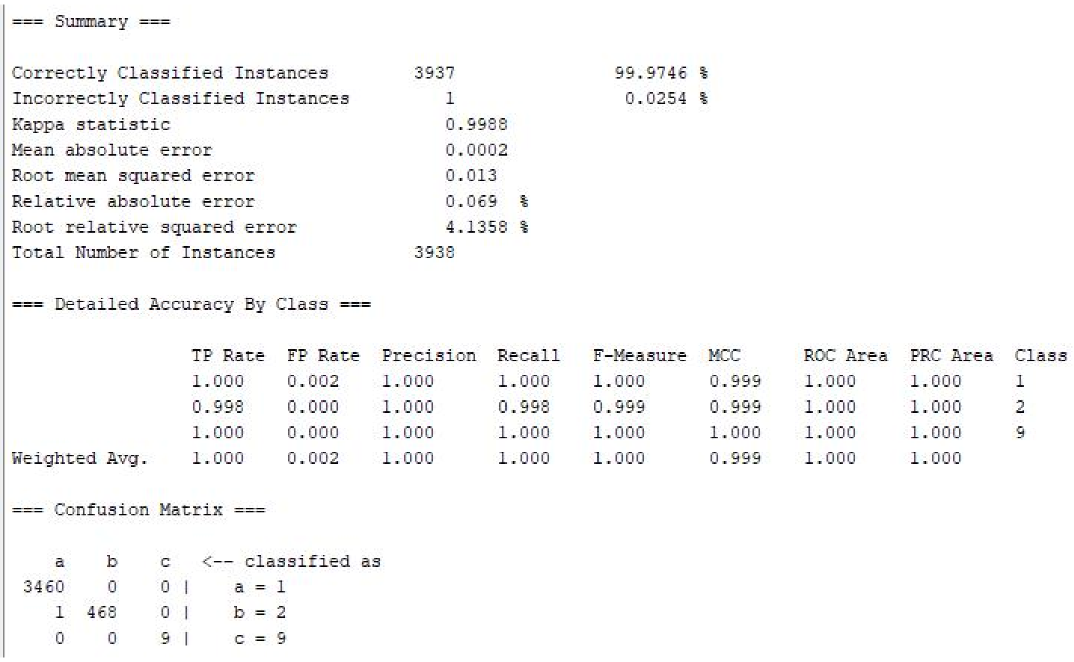
Accuracy: 92.3565%



```
=== Summary ===

Correctly Classified Instances        3637               92.3565 %
Incorrectly Classified Instances       301                7.6435 %
Kappa statistic                          0.6078
Mean absolute error                      0.0634
Root mean squared error                  0.1945
Relative absolute error                 25.7944 %
Root relative squared error             61.813  %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.971    0.414    0.944      0.971    0.957      0.615    0.926     0.984     1
                 0.593    0.029    0.732      0.593    0.655      0.618    0.932     0.749     2
                 0.000    0.000    0.000      0.000    0.000      -0.001   0.753     0.012     9
Weighted Avg.    0.924    0.367    0.917      0.924    0.919      0.614    0.926     0.954

=== Confusion Matrix ===

    a    b    c   <-- classified as
 3359  101    0 |    a = 1
  190  278    1 |    b = 2
    8    1    0 |    c = 9
```

5. *MultipleClassClassifier*

Accuracy: 99.9746%

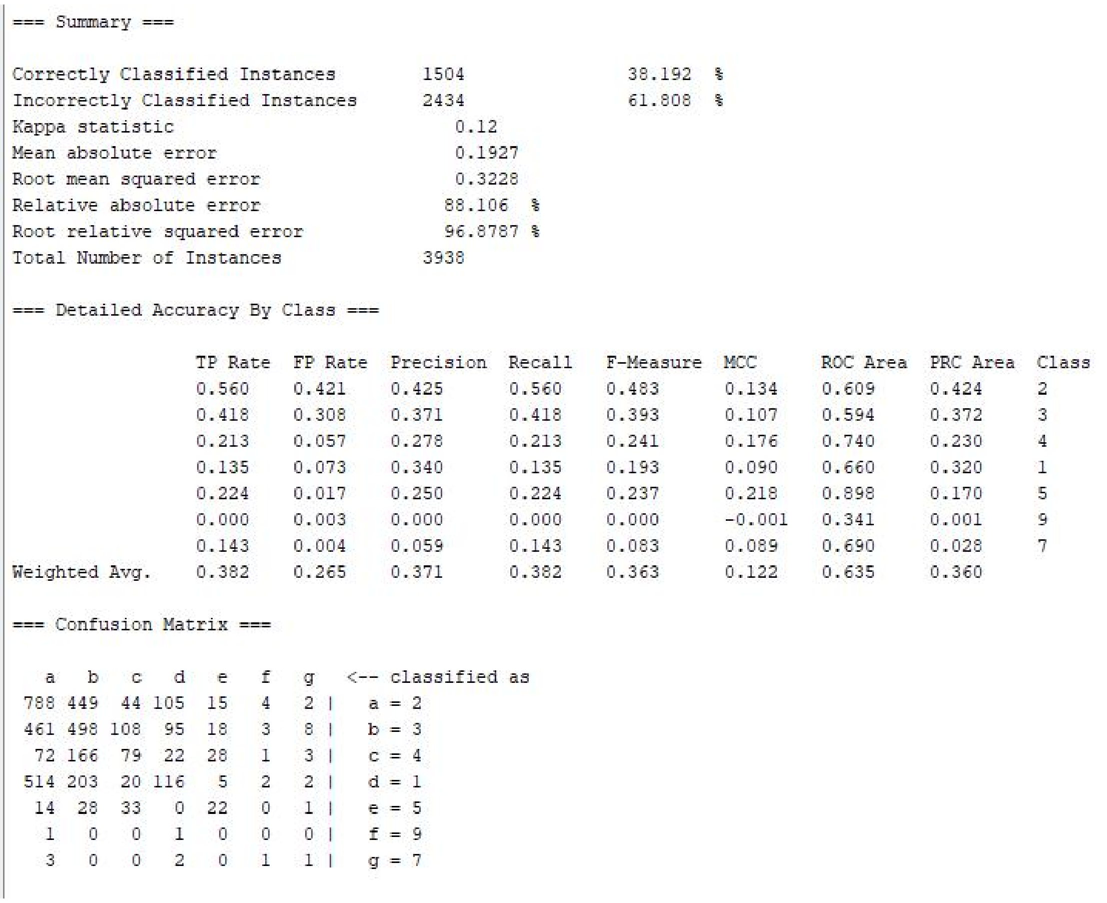


```
=== Summary ===

Correctly Classified Instances        3937               99.9746 %
Incorrectly Classified Instances         1                0.0254 %
Kappa statistic                          0.9988
Mean absolute error                      0.0002
Root mean squared error                  0.013 
Relative absolute error                  0.069  %
Root relative squared error              4.1358 %
Total Number of Instances             3938     

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.002    1.000      1.000    1.000      0.999    1.000     1.000     1
                 0.998    0.000    1.000      0.998    0.999      0.999    1.000     1.000     2
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     9
Weighted Avg.    1.000    0.002    1.000      1.000    1.000      0.999    1.000     1.000     

=== Confusion Matrix ===

    a    b    c   <-- classified as
 3460    0    0 |    a = 1
    1  468    0 |    b = 2
    0    0    9 |    c = 9
```

4. OneRAttributeEval

1. <u>NaiveBayes</u>

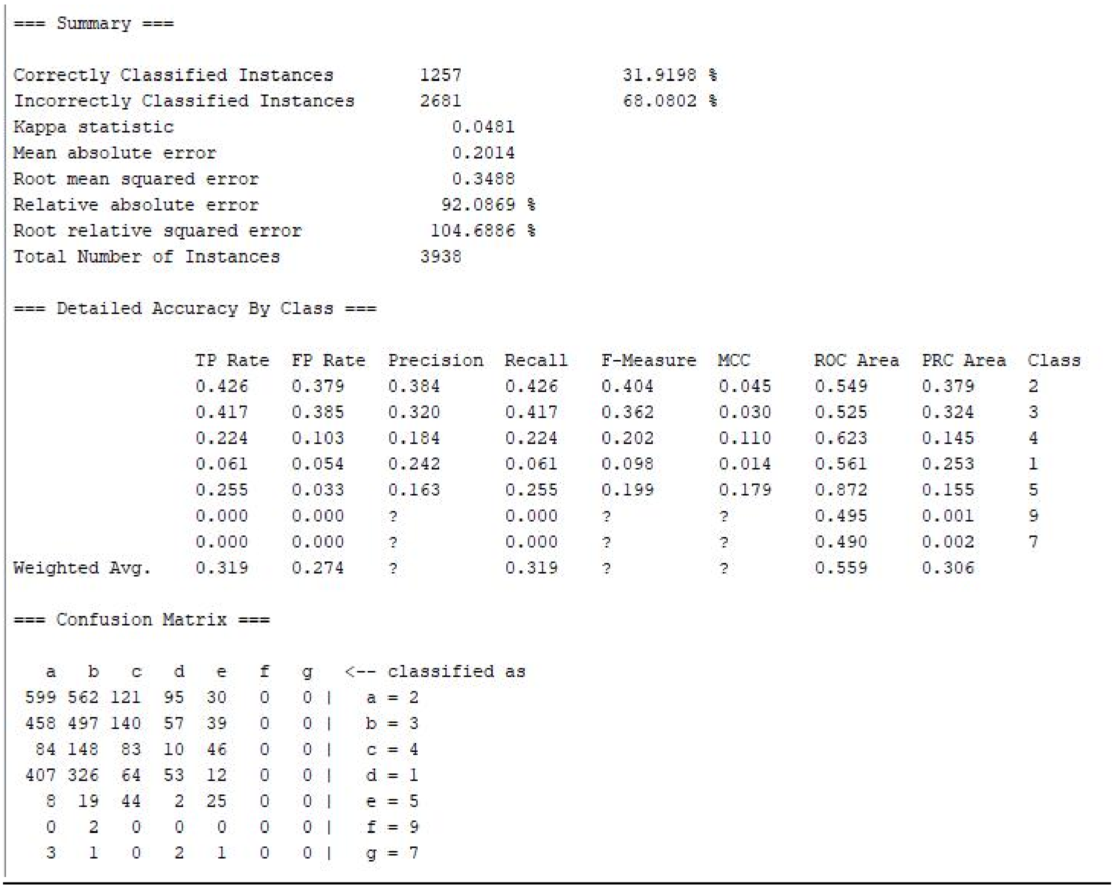
Accuracy: 38.192%



```
=== Summary ===

Correctly Classified Instances        1504               38.192  %
Incorrectly Classified Instances      2434               61.808  %
Kappa statistic                          0.12
Mean absolute error                      0.1927
Root mean squared error                  0.3228
Relative absolute error                 88.106  %
Root relative squared error             96.8787 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.560    0.421    0.425      0.560    0.483      0.134    0.609     0.424     2
                 0.418    0.308    0.371      0.418    0.393      0.107    0.594     0.372     3
                 0.213    0.057    0.278      0.213    0.241      0.176    0.740     0.230     4
                 0.135    0.073    0.340      0.135    0.193      0.090    0.660     0.320     1
                 0.224    0.017    0.250      0.224    0.237      0.218    0.898     0.170     5
                 0.000    0.003    0.000      0.000    0.000      -0.001   0.341     0.001     9
                 0.143    0.004    0.059      0.143    0.083      0.089    0.690     0.028     7
Weighted Avg.    0.382    0.265    0.371      0.382    0.363      0.122    0.635     0.360

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 788 449  44 105  15   4   2 |   a = 2
 461 498 108  95  18   3   8 |   b = 3
  72 166  79  22  28   1   3 |   c = 4
 514 203  20 116   5   2   2 |   d = 1
  14  28  33   0  22   0   1 |   e = 5
   1   0   0   1   0   0   0 |   f = 9
   3   0   0   2   0   1   1 |   g = 7
```

2. Bagging

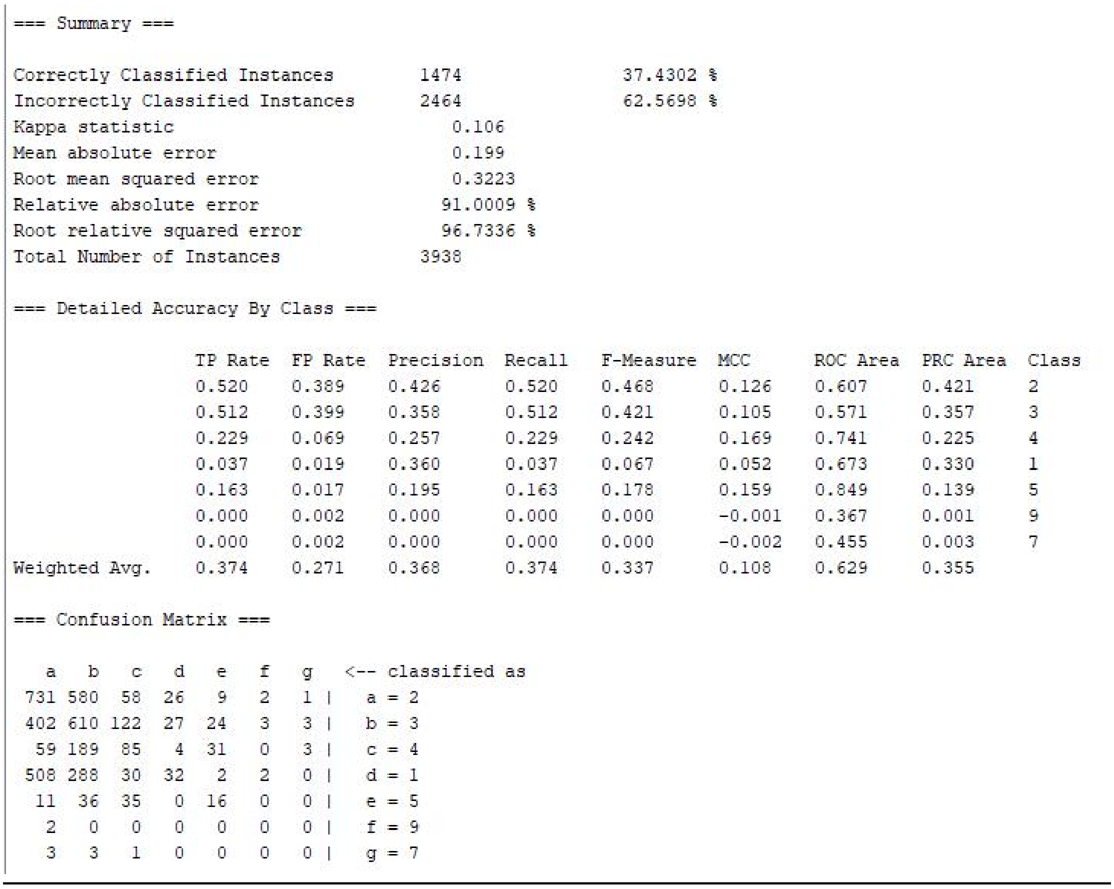
Accuracy: 31.9198%



```
=== Summary ===

Correctly Classified Instances        1257               31.9198 %
Incorrectly Classified Instances      2681               68.0802 %
Kappa statistic                          0.0481
Mean absolute error                      0.2014
Root mean squared error                  0.3488
Relative absolute error                 92.0869 %
Root relative squared error            104.6886 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.426    0.379    0.384      0.426    0.404      0.045    0.549     0.379     2
                 0.417    0.385    0.320      0.417    0.362      0.030    0.525     0.324     3
                 0.224    0.103    0.184      0.224    0.202      0.110    0.623     0.145     4
                 0.061    0.054    0.242      0.061    0.098      0.014    0.561     0.253     1
                 0.255    0.033    0.163      0.255    0.199      0.179    0.872     0.155     5
                 0.000    0.000    ?          0.000    ?          ?        0.495     0.001     9
                 0.000    0.000    ?          0.000    ?          ?        0.490     0.002     7
Weighted Avg.    0.319    0.274    ?          0.319    ?          ?        0.559     0.306

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 599 562 121  95  30   0   0 |   a = 2
 458 497 140  57  39   0   0 |   b = 3
  84 148  83  10  46   0   0 |   c = 4
 407 326  64  53  12   0   0 |   d = 1
   8  19  44   2  25   0   0 |   e = 5
   0   2   0   0   0   0   0 |   f = 9
   3   1   0   2   1   0   0 |   g = 7
```

3. *Logistic*

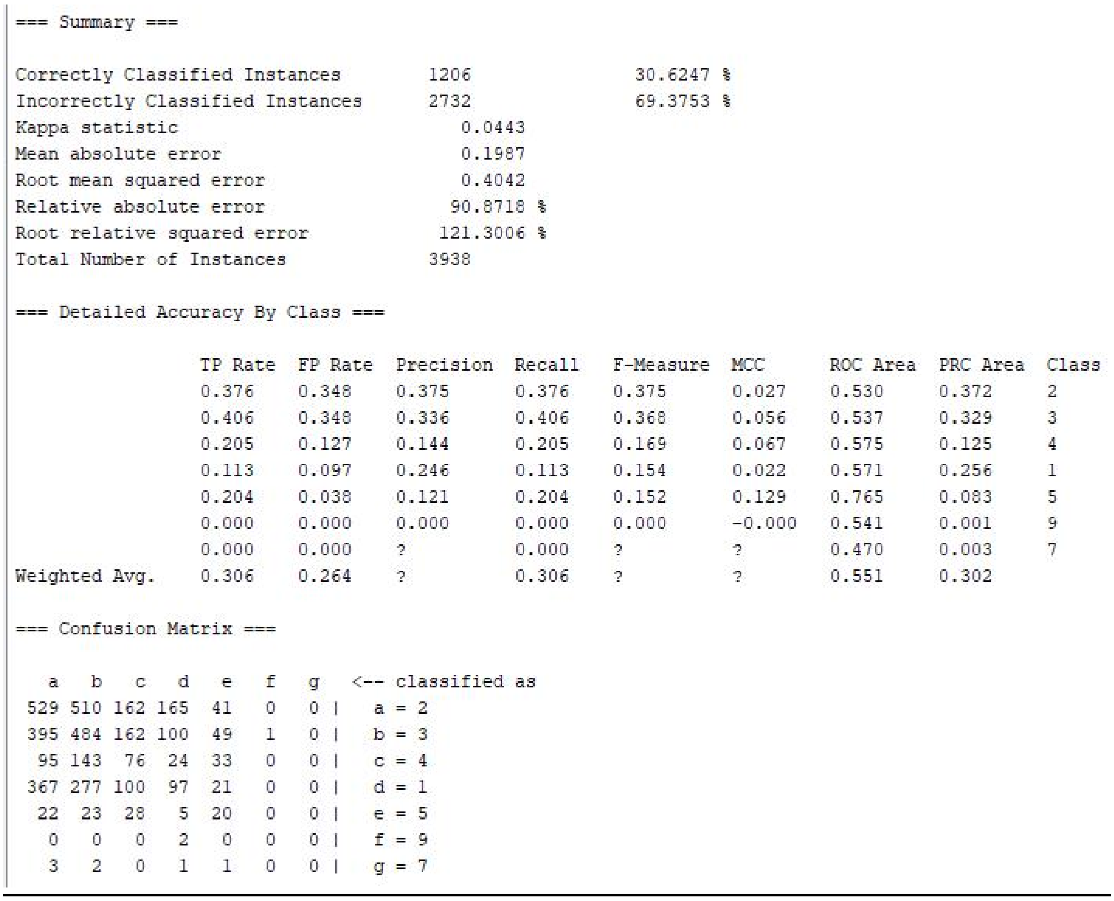
Accuracy: 37.4302%



```
=== Summary ===

Correctly Classified Instances        1474               37.4302 %
Incorrectly Classified Instances      2464               62.5698 %
Kappa statistic                          0.106
Mean absolute error                      0.199
Root mean squared error                  0.3223
Relative absolute error                 91.0009 %
Root relative squared error             96.7336 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.520    0.389    0.426      0.520    0.468      0.126    0.607     0.421     2
                 0.512    0.399    0.358      0.512    0.421      0.105    0.571     0.357     3
                 0.229    0.069    0.257      0.229    0.242      0.169    0.741     0.225     4
                 0.037    0.019    0.360      0.037    0.067      0.052    0.673     0.330     1
                 0.163    0.017    0.195      0.163    0.178      0.159    0.849     0.139     5
                 0.000    0.002    0.000      0.000    0.000      -0.001   0.367     0.001     9
                 0.000    0.002    0.000      0.000    0.000      -0.002   0.455     0.003     7
Weighted Avg.    0.374    0.271    0.368      0.374    0.337      0.108    0.629     0.355

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 731 580  58  26   9   2   1 |   a = 2
 402 610 122  27  24   3   3 |   b = 3
  59 189  85   4  31   0   3 |   c = 4
 508 288  30  32   2   2   0 |   d = 1
  11  36  35   0  16   0   0 |   e = 5
   2   0   0   0   0   0   0 |   f = 9
   3   3   1   0   0   0   0 |   g = 7
```

4. *Kstar*

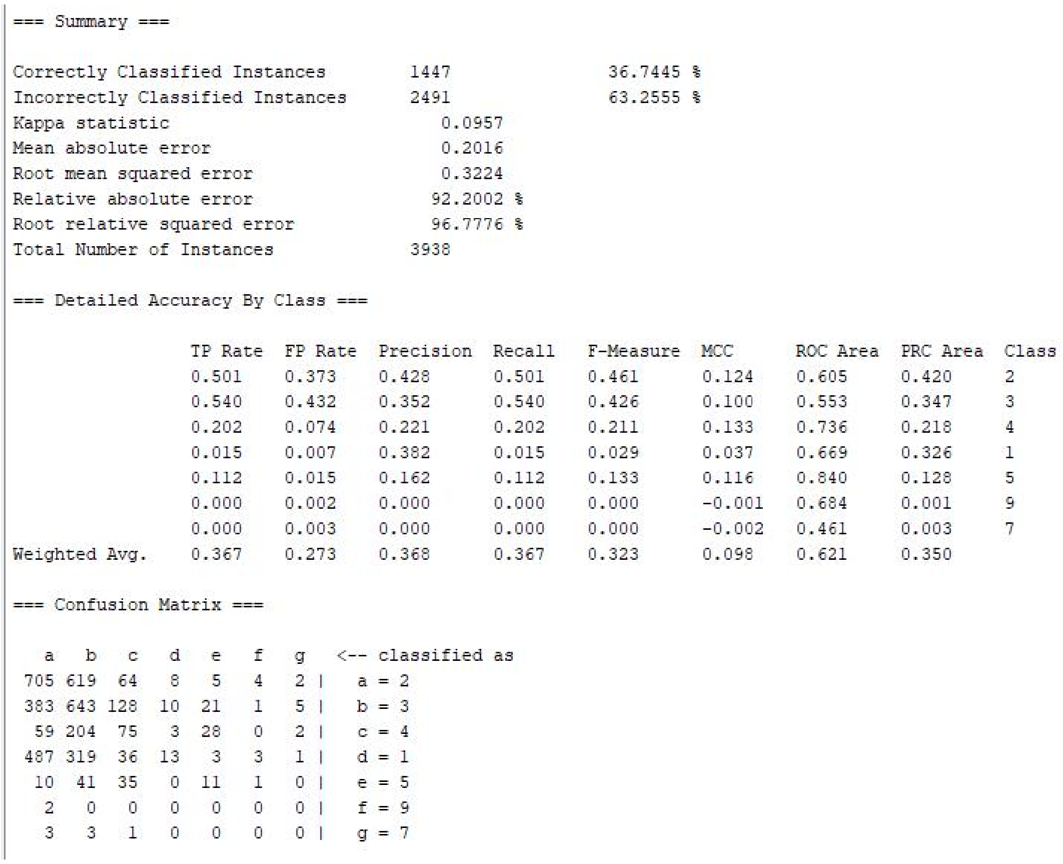
Accuracy: 30.6247%



```
=== Summary ===

Correctly Classified Instances        1206               30.6247 %
Incorrectly Classified Instances      2732               69.3753 %
Kappa statistic                          0.0443
Mean absolute error                      0.1987
Root mean squared error                  0.4042
Relative absolute error                 90.8718 %
Root relative squared error            121.3006 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.376    0.348    0.375      0.376    0.375      0.027    0.530     0.372     2
                 0.406    0.348    0.336      0.406    0.368      0.056    0.537     0.329     3
                 0.205    0.127    0.144      0.205    0.169      0.067    0.575     0.125     4
                 0.113    0.097    0.246      0.113    0.154      0.022    0.571     0.256     1
                 0.204    0.038    0.121      0.204    0.152      0.129    0.765     0.083     5
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.541     0.001     9
                 0.000    0.000    ?          0.000    ?          ?        0.470     0.003     7
Weighted Avg.    0.306    0.264    ?          0.306    ?          ?        0.551     0.302

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 529 510 162 165  41   0   0 |   a = 2
 395 484 162 100  49   1   0 |   b = 3
  95 143  76  24  33   0   0 |   c = 4
 367 277 100  97  21   0   0 |   d = 1
  22  23  28   5  20   0   0 |   e = 5
   0   0   0   2   0   0   0 |   f = 9
   3   2   0   1   1   0   0 |   g = 7
```

5. *MultipleClassClassifier*

Accuracy: 36.7445%

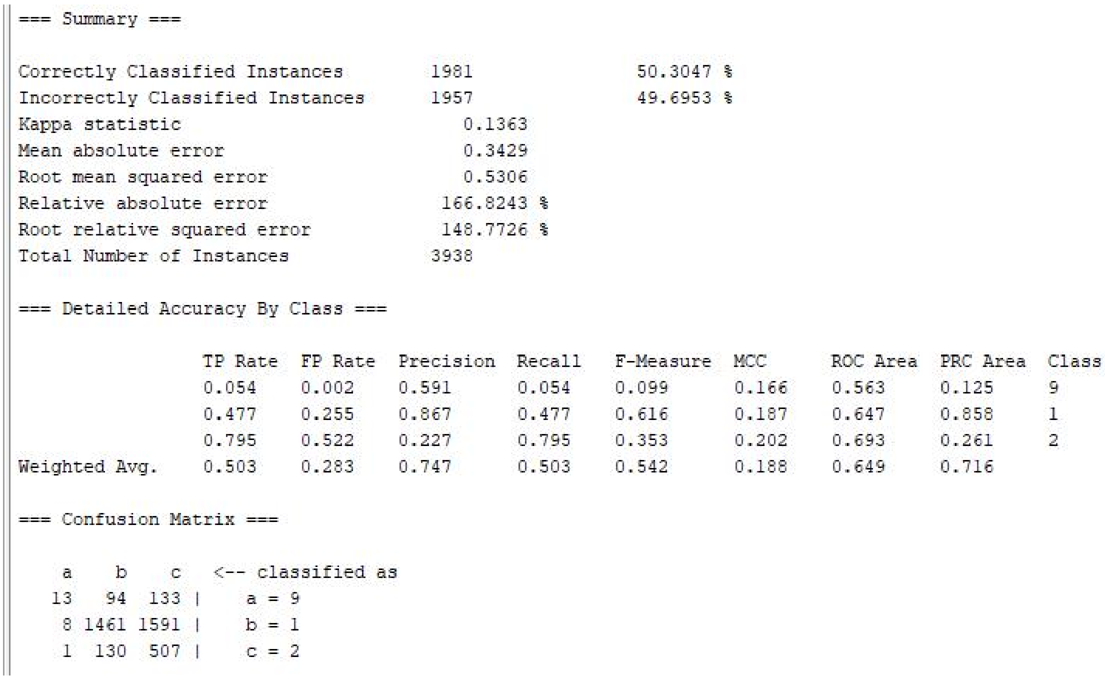


```
=== Summary ===

Correctly Classified Instances        1447               36.7445 %
Incorrectly Classified Instances      2491               63.2555 %
Kappa statistic                          0.0957
Mean absolute error                      0.2016
Root mean squared error                  0.3224
Relative absolute error                 92.2002 %
Root relative squared error             96.7776 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.501    0.373    0.428      0.501    0.461      0.124    0.605     0.420     2
                 0.540    0.432    0.352      0.540    0.426      0.100    0.553     0.347     3
                 0.202    0.074    0.221      0.202    0.211      0.133    0.736     0.218     4
                 0.015    0.007    0.382      0.015    0.029      0.037    0.669     0.326     1
                 0.112    0.015    0.162      0.112    0.133      0.116    0.840     0.128     5
                 0.000    0.002    0.000      0.000    0.000      -0.001   0.684     0.001     9
                 0.000    0.003    0.000      0.000    0.000      -0.002   0.461     0.003     7
Weighted Avg.    0.367    0.273    0.368      0.367    0.323      0.098    0.621     0.350

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 705 619  64   8   5   4   2 |   a = 2
 383 643 128  10  21   1   5 |   b = 3
  59 204  75   3  28   0   2 |   c = 4
 487 319  36  13   3   3   1 |   d = 1
  10  41  35   0  11   1   0 |   e = 5
   2   0   0   0   0   0   0 |   f = 9
   3   3   1   0   0   0   0 |   g = 7
```

5. SymmetricalUncertAttributeEval

1. NaiveBayes

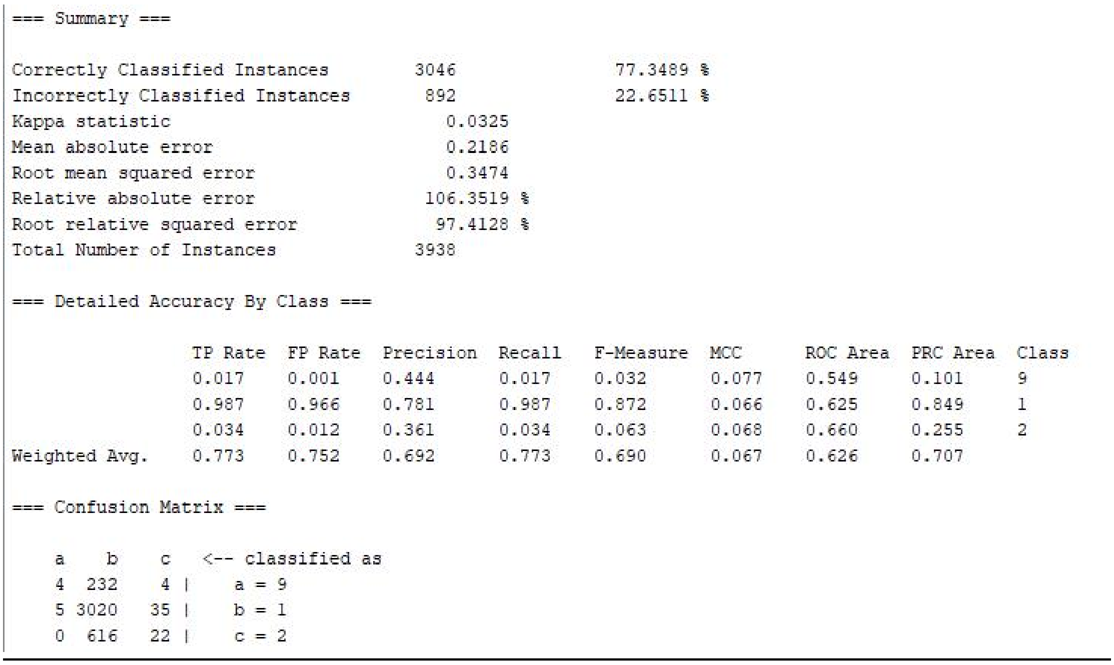
Accuracy: 50.3047%



```
=== Summary ===

Correctly Classified Instances        1981               50.3047 %
Incorrectly Classified Instances      1957               49.6953 %
Kappa statistic                          0.1363
Mean absolute error                      0.3429
Root mean squared error                  0.5306
Relative absolute error                166.8243 %
Root relative squared error            148.7726 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.054    0.002    0.591      0.054    0.099      0.166    0.563     0.125     9
                 0.477    0.255    0.867      0.477    0.616      0.187    0.647     0.858     1
                 0.795    0.522    0.227      0.795    0.353      0.202    0.693     0.261     2
Weighted Avg.    0.503    0.283    0.747      0.503    0.542      0.188    0.649     0.716

=== Confusion Matrix ===

    a    b    c   <-- classified as
   13   94  133 |    a = 9
    8 1461 1591 |    b = 1
    1  130  507 |    c = 2
```

2. *Bagging*

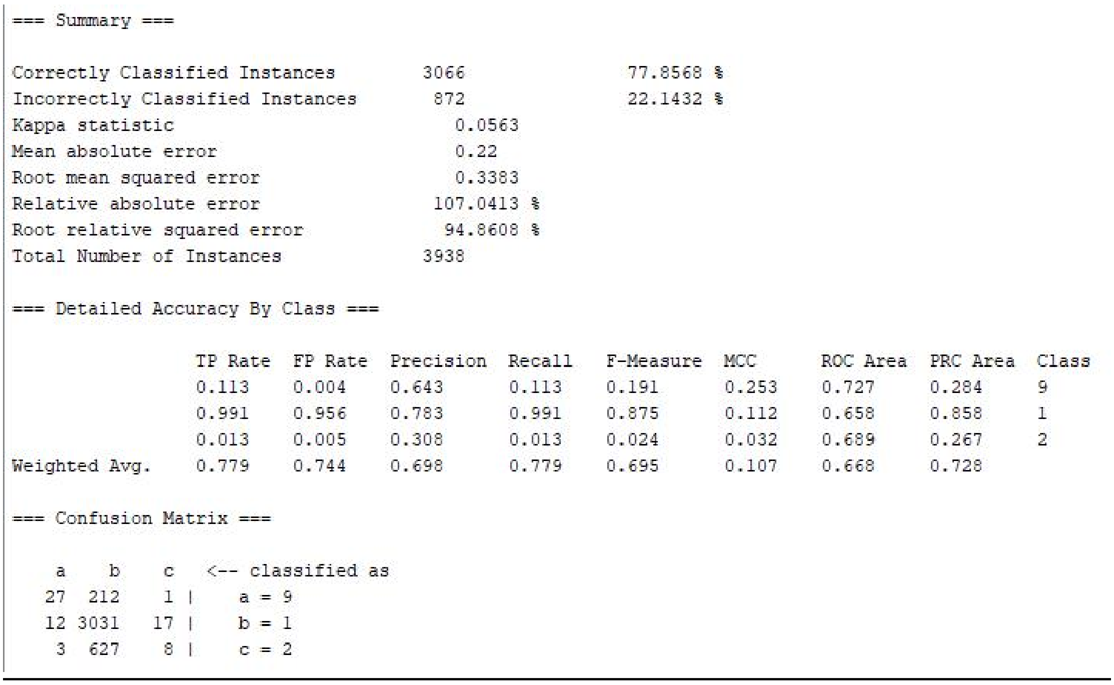
Accuracy: 77.3489%



```
=== Summary ===

Correctly Classified Instances        3046               77.3489 %
Incorrectly Classified Instances       892               22.6511 %
Kappa statistic                          0.0325
Mean absolute error                      0.2186
Root mean squared error                  0.3474
Relative absolute error                106.3519 %
Root relative squared error             97.4128 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.017    0.001    0.444      0.017    0.032      0.077    0.549     0.101     9
                 0.987    0.966    0.781      0.987    0.872      0.066    0.625     0.849     1
                 0.034    0.012    0.361      0.034    0.063      0.068    0.660     0.255     2
Weighted Avg.    0.773    0.752    0.692      0.773    0.690      0.067    0.626     0.707

=== Confusion Matrix ===

    a    b    c   <-- classified as
    4  232    4 |    a = 9
    5 3020   35 |    b = 1
    0  616   22 |    c = 2
```

3. *Logistic*

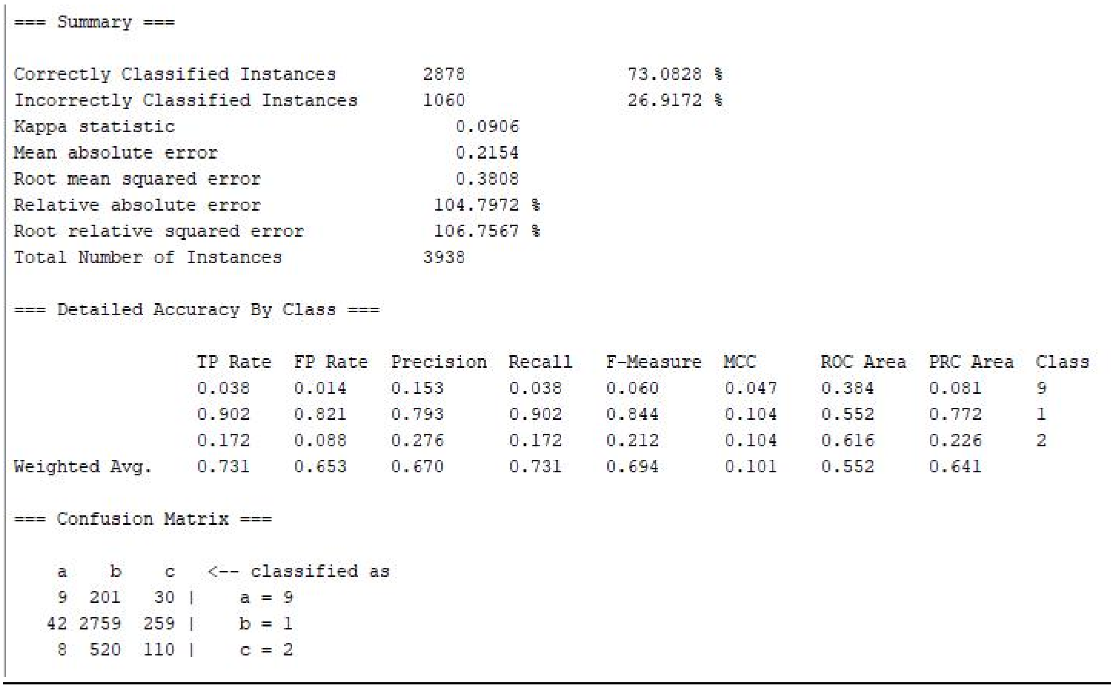
Accuracy: 77.8568%



```
=== Summary ===

Correctly Classified Instances        3066               77.8568 %
Incorrectly Classified Instances       872               22.1432 %
Kappa statistic                          0.0563
Mean absolute error                      0.22
Root mean squared error                  0.3383
Relative absolute error                107.0413 %
Root relative squared error             94.8608 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.113    0.004    0.643      0.113    0.191      0.253    0.727     0.284     9
                 0.991    0.956    0.783      0.991    0.875      0.112    0.658     0.858     1
                 0.013    0.005    0.308      0.013    0.024      0.032    0.689     0.267     2
Weighted Avg.    0.779    0.744    0.698      0.779    0.695      0.107    0.668     0.728

=== Confusion Matrix ===

    a    b    c   <-- classified as
   27  212    1 |    a = 9
   12 3031   17 |    b = 1
    3  627    8 |    c = 2
```

4. *Kstar*

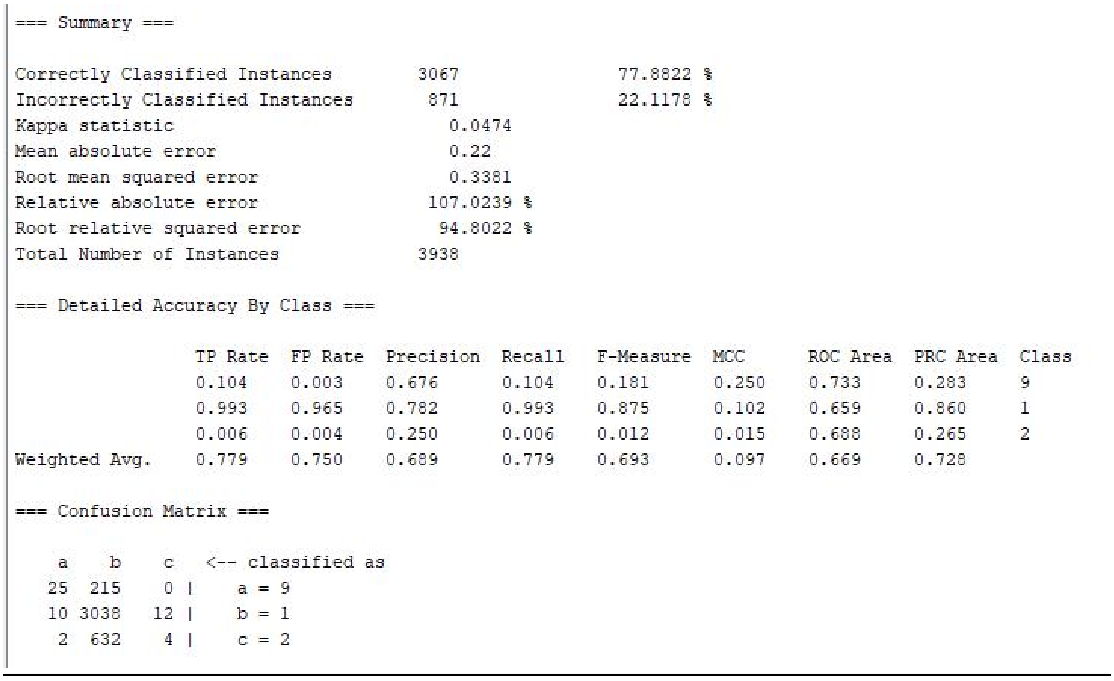
Accuracy: 73.0828%



```
=== Summary ===

Correctly Classified Instances        2878               73.0828 %
Incorrectly Classified Instances      1060               26.9172 %
Kappa statistic                          0.0906
Mean absolute error                      0.2154
Root mean squared error                  0.3808
Relative absolute error                104.7972 %
Root relative squared error            106.7567 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.038    0.014    0.153      0.038    0.060      0.047    0.384     0.081     9
                 0.902    0.821    0.793      0.902    0.844      0.104    0.552     0.772     1
                 0.172    0.088    0.276      0.172    0.212      0.104    0.616     0.226     2
Weighted Avg.    0.731    0.653    0.670      0.731    0.694      0.101    0.552     0.641

=== Confusion Matrix ===

    a    b    c   <-- classified as
    9  201   30 |    a = 9
   42 2759  259 |    b = 1
    8  520  110 |    c = 2
```

5. *MultipleClassClassifier*

Accuracy: 77.8822%



```
=== Summary ===

Correctly Classified Instances        3067               77.8822 %
Incorrectly Classified Instances       871               22.1178 %
Kappa statistic                          0.0474
Mean absolute error                      0.22
Root mean squared error                  0.3381
Relative absolute error                107.0239 %
Root relative squared error             94.8022 %
Total Number of Instances             3938

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.104    0.003    0.676      0.104    0.181      0.250    0.733     0.283     9
                 0.993    0.965    0.782      0.993    0.875      0.102    0.659     0.860     1
                 0.006    0.004    0.250      0.006    0.012      0.015    0.688     0.265     2
Weighted Avg.    0.779    0.750    0.689      0.779    0.693      0.097    0.669     0.728

=== Confusion Matrix ===

    a    b    c   <-- classified as
   25  215    0 |    a = 9
   10 3038   12 |    b = 1
    2  632    4 |    c = 2
```

Better Model

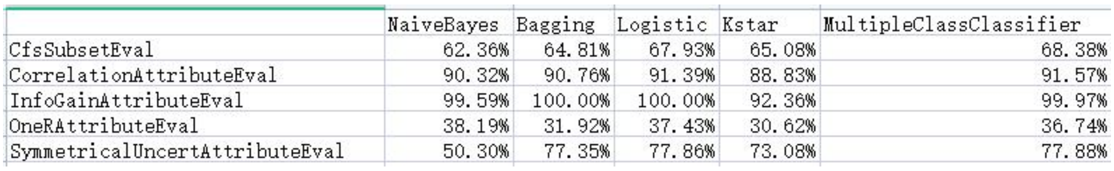
Criteria

1. *Accuracy:* This is the most straightforward criterion. The higher the accuracy, the better the model.

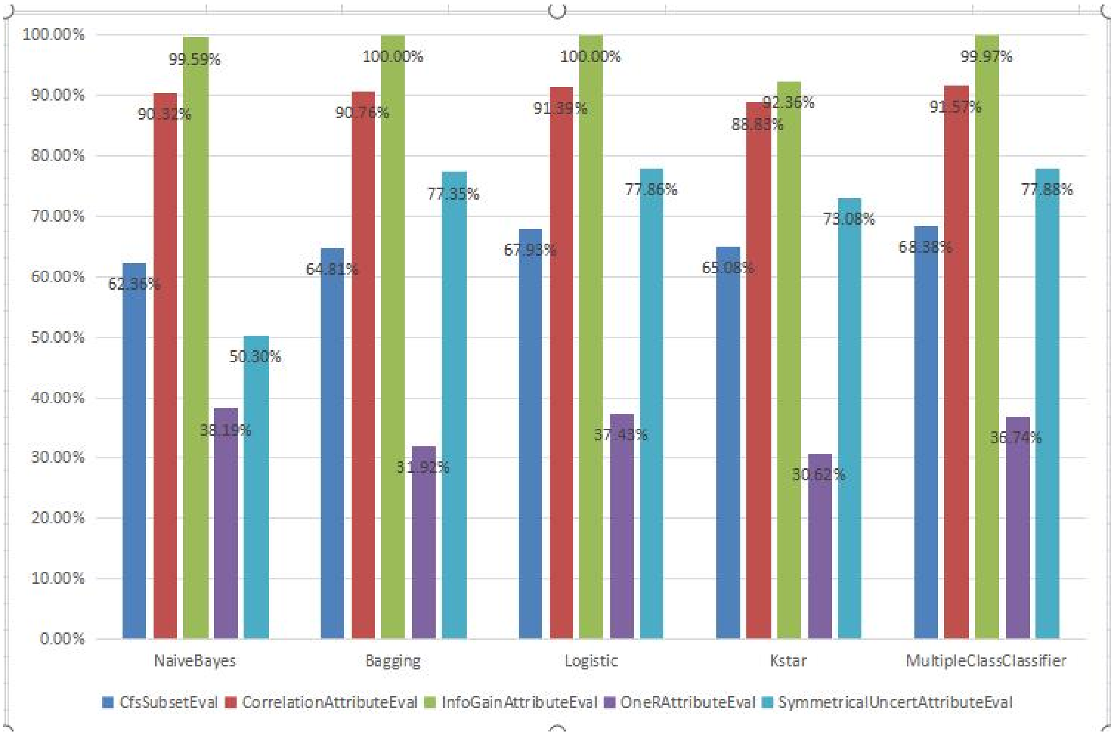
2. *Consistency across different evaluation methods:* A good model should perform well not just in one evaluation method but across different methods. If a model performs well in one method but poorly in another, it might be overfitting to a specific type of data distribution.

3. *Trade-off between simplicity and accuracy:* Simpler models (like NaiveBayes) are easier to understand and interpret, and they also tend to generalize better to new data. So if the accuracy difference is not too big, one might prefer a simpler model.

Comparation Table



	NaiveBayes	Bagging	Logistic	Kstar	MultipleClassClassifier
CfsSubsetEval	62.36%	64.81%	67.93%	65.08%	68.38%
CorrelationAttributeEval	90.32%	90.76%	91.39%	88.83%	91.57%
InfoGainAttributeEval	99.59%	100.00%	100.00%	92.36%	99.97%
OneRAttributeEval	38.19%	31.92%	37.43%	30.62%	36.74%
SymmetricalUncertAttributeEval	50.30%	77.35%	77.86%	73.08%	77.88%



Summary

Based on the table and the criteria, the *Logistic model* with *InfoGainAttributeEval* method seems to be the best choice as it offers the highest accuracy (100%) and is also relatively consistent across different evaluation methods. However, the final decision should also consider other factors such as the specific application and data characteristics, the computational cost of the model, and the trade-off between model complexity and interpretability.

Reflections

5 attribute I think most relevant

x.age80: This attribute appears in the top five for three out of the five feature selection methods (CorrelationAttributeEval, InfoGainAttributeEval, SymmetricalUncertAttributeEval). This suggests that it is a highly relevant attribute for the class.

diffwalk: This attribute appears in the top five for three out of the five feature selection methods (CfsSubsetEval, CorrelationAttributeEval, SymmetricalUncertAttributeEval). This indicates that it is a significant attribute for the class.

x.ageg5yr: This attribute appears in the top five for two out of the five feature selection methods (CorrelationAttributeEval, SymmetricalUncertAttributeEval). It also appears in the top 20 for InfoGainAttributeEval, suggesting its relevance.

employ1: This attribute appears in the top five for two out of the five feature selection methods

(CfsSubsetEval, SymmetricalUncertAttributeEval). It also appears in the top 20 for InfoGainAttributeEval and OneRAttributeEval, indicating its importance.

checkup1: This attribute appears in the top five for two out of the five feature selection methods (CorrelationAttributeEval, SymmetricalUncertAttributeEval). It also appears in the top 20 for InfoGainAttributeEval, suggesting its relevance.

These attributes are considered relevant based on their frequency of appearance in the top five across different feature selection methods. However, the specific relevance of these attributes may also depend on the context of your data and the specific class attribute you are predicting.

I learned from this project

The project involved the use of various machine learning techniques and algorithms, including

Naive Bayes, Bagging, and Logistic Regression. These algorithms were applied in different contexts, demonstrating their strengths and weaknesses. For example, Naive Bayes showed high accuracy in some cases, but lower accuracy in others.

The project also involved feature selection methods like InfoGainAttributeEval and OneRAttributeEval. These methods were used to rank attributes based on their predictive power, which is crucial in improving the performance of the models.

The project demonstrated the importance of preprocessing in a machine learning pipeline, including feature selection and dimensionality reduction.

Other Observations

The accuracy of the models varied significantly depending on the feature selection method and the classification algorithm used. This highlights the importance of choosing the right combination of feature selection method and classifier for a given dataset.

The project demonstrated the use of various feature selection methods and classifiers, providing a comprehensive learning experience in applying machine learning techniques.