Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

Course Title : MACHINE LEARNING

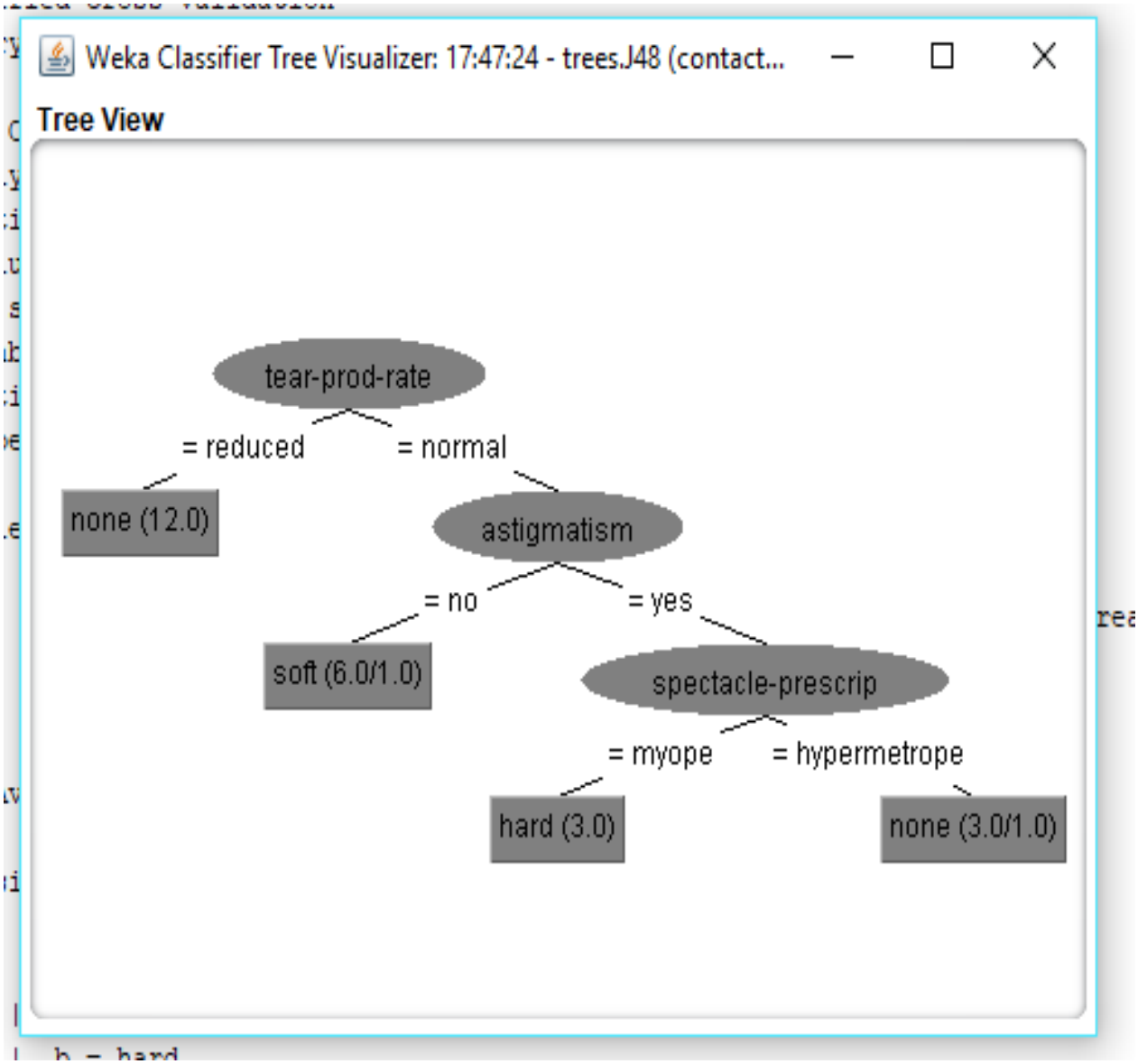
Download WEKA from https://www.cs.waikato.ac.nz/ml/weka/ and explore the different features in it.

1. Use the default values for building models using WEKA.

a) For dataset **contact-lenses** build a decision tree J4.8 in WEKA. [3+2]

1) Draw the decision tree.

Solution:



Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

2) Write down the confusion matrix for 10-fold cross validation.

Solution:

Selects Cross Validation by default with 10 folds. This means that the dataset is split into 10 parts:

- The first 9 are used to train the algorithm, and
- The 10th is used to assess the algorithm.

This process is repeated, allowing each of the 10 parts of the split dataset a chance to be the held-out test set.

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: contact-lenses

Instances: 24

Attributes: 5

age

spectacle-prescrip

astigmatism

tear-prod-rate

contact-lenses

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

---

tear-prod-rate = reduced: none (12.0)

tear-prod-rate = normal

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

| astigmatism = no: soft (6.0/1.0)

| astigmatism = yes

| | spectacle-prescrip = myope: hard (3.0)

| | spectacle-prescrip = hypermetrope: none (3.0/1.0)

Number of Leaves : 4

Size of the tree : 7

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	20	83.3333 %
Incorrectly Classified Instances	4	16.6667 %
Kappa statistic	0.71	
Mean absolute error	0.15	
Root mean squared error	0.3249	
Relative absolute error	39.7059 %	
Root relative squared error	74.3898 %	
Total Number of Instances	24	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.053	0.833	1.000	0.909	0.889	0.947	0.833	soft
	0.750	0.100	0.600	0.750	0.667	0.596	0.813	0.592	hard
	0.800	0.111	0.923	0.800	0.857	0.669	0.811	0.865	none
Weighted Avg.	0.833	0.097	0.851	0.833	0.836	0.703	0.840	0.813	

=== Confusion Matrix ===

a	b	c	<-- classified as
5	0	0	a = soft

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

0 3 1 | b = hard

1 2 12 | c = none

b) In your class you have learnt following algorithms. For dataset **iris.2D** and **supermarket** find out classification accuracy for each of the algorithms and fill the following table. [3+2]

Solution :

Algorithms	Accuracy iris.2D	Accuracy supermarket	In which lecture and by what name you learnt this algorithm
BayesNet	96%	63.713%	Lecture: Arun Chauhan Name: Lecture L6
NaiveBayes	96%	63.713%	Lecture: Arun Chauhan Name: Lecture L4
Logistic Regression	96%	63.713%	Lecture: Arun Chauhan Name: Lecture L4
Multilayer Perceptron	96.6667%	63.713%	Lecture: Arun Chauhan Name: Lecture L10
SMO	96%	63.713%	Lecture: Arun Chauhan Name: Lecture L15
IBk	99.3333%	89.8422%	Lecture: Arun Chauhan Name: Lecture L12
LWL	96%	73.468%	Lecture: Arun Chauhan Name: Lecture L12
J4.8	98%	63.713%	Lecture: Arun Chauhan Name: Lecture L7

<<Please find the snapshots of the Weka for the above algorithms in the subsequent pages>>

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

Weka SnapShots in support of the above table:

DataSet iris.2D

Algorithm : BayesNet :

```
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.0304
Root mean squared error                  0.1368
Relative absolute error                  6.8301 %
Root relative squared error             29.0144 %
Total Number of Instances              150

=== 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     Iris-setosa
                 0.980    0.050    0.907      0.980    0.942      0.913    0.987     0.961     Iris-versicolor
                 0.900    0.010    0.978      0.900    0.938      0.910    0.987     0.962     Iris-virginica
Weighted Avg.    0.960    0.020    0.962      0.960    0.960      0.941    0.991     0.974

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 49  1 |  b = Iris-versicolor
  0  5 45 |  c = Iris-virginica
```

DataSet iris.2D

Algorithm : Naïve Bayes

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

```
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.0265
Root mean squared error                  0.1294
Relative absolute error                  5.9721 %
Root relative squared error             27.443  %
Total Number of Instances              150

=== 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     Iris-setosa
                 0.940    0.030    0.940      0.940    0.940      0.910    0.998     0.995     Iris-versicolor
                 0.940    0.030    0.940      0.940    0.940      0.910    0.998     0.995     Iris-virginica
Weighted Avg.    0.960    0.020    0.960      0.960    0.960      0.940    0.998     0.997

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  3 47 |  c = Iris-virginica
```

DataSet iris.2D

Algorithm : Logistic Regression

```
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.0289
Root mean squared error                  0.1244
Relative absolute error                  6.4963 %
Root relative squared error             26.381  %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000
                 0.940    0.030    0.940      0.940    0.940      0.910    0.997
                 0.940    0.030    0.940      0.940    0.940      0.910    0.997
Weighted Avg.    0.960    0.020    0.960      0.960    0.960      0.940    0.998

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  3 47 |  c = Iris-virginica
```

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

DataSet iris.2D

Algorithms : Multilayer Perceptron

```
=== Summary ===

Correctly Classified Instances         145               96.6667 %
Incorrectly Classified Instances         5                3.3333 %
Kappa statistic                          0.95
Mean absolute error                      0.0437
Root mean squared error                  0.1263
Relative absolute error                  9.8223 %
Root relative squared error             26.7935 %
Total Number of Instances              150

=== 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     Iris-setosa
                 0.940    0.020    0.959      0.940    0.949      0.925    0.998     0.995     Iris-versicolor
                 0.960    0.030    0.941      0.960    0.950      0.925    0.998     0.995     Iris-virginica
Weighted Avg.    0.967    0.017    0.967      0.967    0.967      0.950    0.998     0.997

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

DataSet iris.2D

Algorithm : SMO

```
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.2311
Root mean squared error                  0.288
Relative absolute error                 52      %
Root relative squared error             61.101   %
Total Number of Instances              150

=== 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     Iris-setosa
                 0.960    0.040    0.923      0.960    0.941      0.911    0.960     0.899     Iris-versicolor
                 0.920    0.020    0.958      0.920    0.939      0.910    0.971     0.923     Iris-virginica
Weighted Avg.    0.960    0.020    0.960      0.960    0.960      0.940    0.977     0.941

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 48  2 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

DataSet iris.2D

Algorithm : IBk

```
=== Summary ===

Correctly Classified Instances         149               99.3333 %
Incorrectly Classified Instances         1                0.6667 %
Kappa statistic                          0.99
Mean absolute error                      0.0118
Root mean squared error                  0.0549
Relative absolute error                  2.6616 %
Root relative squared error             11.6437 %
Total Number of Instances              150

=== 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     Iris-setosa
                 0.980    0.000    1.000      0.980    0.990      0.985    1.000     0.999     Iris-versicolor
                 1.000    0.010    0.980      1.000    0.990      0.985    1.000     0.999     Iris-virginica
Weighted Avg.    0.993    0.003    0.993      0.993    0.993      0.990    1.000     0.999

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 49  1 |  b = Iris-versicolor
  0  0 50 |  c = Iris-virginica
```

Algorithm : LWL

```
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.0712
Root mean squared error                  0.1693
Relative absolute error                 16.0148 %
Root relative squared error             35.9179 %
Total Number of Instances              150

=== 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     Iris-setosa
                 0.980    0.050    0.907      0.980    0.942      0.913    0.987     0.973     Iris-versicolor
                 0.900    0.010    0.978      0.900    0.938      0.910    0.986     0.977     Iris-virginica
Weighted Avg.    0.960    0.020    0.962      0.960    0.960      0.941    0.991     0.983

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 49  1 |  b = Iris-versicolor
  0  5 45 |  c = Iris-virginica
```

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

DataSet iris.2D

Algorithm : J4.8

```
=== Summary ===

Correctly Classified Instances         147               98      %
Incorrectly Classified Instances         3                2      %
Kappa statistic                          0.97
Mean absolute error                      0.0233
Root mean squared error                  0.108
Relative absolute error                  5.2482 %
Root relative squared error             22.9089 %
Total Number of Instances              150

=== 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     Iris-setosa
                 0.980    0.020    0.961      0.980    0.970      0.955    0.990     0.969     Iris-versicolor
                 0.960    0.010    0.980      0.960    0.970      0.955    0.990     0.970     Iris-virginica
Weighted Avg.    0.980    0.010    0.980      0.980    0.980      0.970    0.993     0.980

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 49  1 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

DataSet : Supermarket

Algorithm : BayesNet

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

```
=== Summary ===

Correctly Classified Instances        2948               63.713  %
Incorrectly Classified Instances      1679               36.287  %
Kappa statistic                          0
Mean absolute error                      0.4624
Root mean squared error                  0.4808
Relative absolute error                 99.9982 %
Root relative squared error            100      %
Total Number of Instances             4627

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.637      1.000    0.778      0.000    0.500     0.637     low
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.363     high
Weighted Avg.    0.637    0.637    0.406      0.637    0.496      0.000    0.500     0.538

=== Confusion Matrix ===

    a    b   <-- classified as
 2948    0 |    a = low
 1679    0 |    b = high
```

DataSet : Supermarket

Algorithm : Naives Bayes

```
=== Summary ===

Correctly Classified Instances        2948               63.713  %
Incorrectly Classified Instances      1679               36.287  %
Kappa statistic                          0
Mean absolute error                      0.4624
Root mean squared error                  0.4808
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             4627

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.637      1.000    0.778      0.000    0.500     0.637     low
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.363     high
Weighted Avg.    0.637    0.637    0.406      0.637    0.496      0.000    0.500     0.538

=== Confusion Matrix ===

    a    b   <-- classified as
 2948    0 |    a = low
 1679    0 |    b = high
```

DataSet : Supermarket

Algorithm : Logistic Regression

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

```
=== Summary ===

Correctly Classified Instances        2948               63.713  %
Incorrectly Classified Instances      1679               36.287  %
Kappa statistic                          0
Mean absolute error                      0.4624
Root mean squared error                  0.4808
Relative absolute error                 99.9965 %
Root relative squared error            100      %
Total Number of Instances             4627

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.637      1.000    0.778      0.000    0.500     0.637     low
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.363     high
Weighted Avg.    0.637    0.637    0.406      0.637    0.496      0.000    0.500     0.538

=== Confusion Matrix ===

    a    b   <-- classified as
 2948    0 |    a = low
 1679    0 |    b = high
```

DataSet : Supermarket

Algorithm : MultiLayer Perception

```
Time taken to test model on training data: 1.36 seconds

=== Summary ===

Correctly Classified Instances        2948               63.713  %
Incorrectly Classified Instances      1679               36.287  %
Kappa statistic                          0
Mean absolute error                      0.4627
Root mean squared error                  0.4808
Relative absolute error                100.0727 %
Root relative squared error            100.0004 %
Total Number of Instances             4627

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.637      1.000    0.778      0.000    0.500     0.637     low
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.363     high
Weighted Avg.    0.637    0.637    0.406      0.637    0.496      0.000    0.500     0.538

=== Confusion Matrix ===

    a    b   <-- classified as
 2948    0 |    a = low
 1679    0 |    b = high
```

DataSet : Supermarket

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

Algorithm : LWL

```
=== Summary ===

Correctly Classified Instances        3398               73.4385 %
Incorrectly Classified Instances      1229               26.5615 %
Kappa statistic                          0.4177
Mean absolute error                      0.4105
Root mean squared error                  0.439
Relative absolute error                 88.7727 %
Root relative squared error             91.3005 %
Total Number of Instances             4627

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.809    0.397      0.782     0.809     0.795     0.418     0.784     0.842     low
                 0.603    0.191      0.643     0.603     0.622     0.418     0.784     0.715     high
Weighted Avg.    0.734    0.322      0.731     0.734     0.732     0.418     0.784     0.796

=== Confusion Matrix ===

    a    b   <-- classified as
 2386  562 |    a = low
  667 1012 |    b = high
```

DataSet : Supermarket

Algorithm :SMO

```
=== Summary ===

Correctly Classified Instances        2948               63.713  %
Incorrectly Classified Instances      1679               36.287  %
Kappa statistic                          0
Mean absolute error                      0.3629
Root mean squared error                  0.6024
Relative absolute error                 78.4742 %
Root relative squared error            125.2812 %
Total Number of Instances             4627

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000      0.637     1.000     0.778     0.000     0.500     0.637     low
                 0.000    0.000      0.000     0.000     0.000     0.000     0.500     0.363     high
Weighted Avg.    0.637    0.637      0.406     0.637     0.496     0.000     0.500     0.538

=== Confusion Matrix ===

    a    b   <-- classified as
 2948    0 |    a = low
 1679    0 |    b = high
```

DataSet : Supermarket

Algorithm : iBk

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

```
=== Summary ===

Correctly Classified Instances        4157               89.8422 %
Incorrectly Classified Instances       470               10.1578 %
Kappa statistic                          0.7925
Mean absolute error                      0.1093
Root mean squared error                  0.2772
Relative absolute error                 23.644  %
Root relative squared error             57.6468 %
Total Number of Instances             4627

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.843     0.004     0.998       0.843     0.914      0.809     0.998     0.999     low
               0.996     0.157     0.783       0.996     0.877      0.809     0.998     0.997     high
Weighted Avg.  0.898     0.059     0.920       0.898     0.900      0.809     0.998     0.998

=== Confusion Matrix ===

    a    b   <-- classified as
 2484  464 |    a = low
    6 1673 |    b = high
```

DataSet : Supermarket

Algorithm : J.48

```
=== Summary ===

Correctly Classified Instances        2948               63.713  %
Incorrectly Classified Instances      1679               36.287  %
Kappa statistic                          0
Mean absolute error                      0.4624
Root mean squared error                  0.4808
Relative absolute error                 99.9965 %
Root relative squared error            100      %
Total Number of Instances             4627

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000     1.000     0.637       1.000     0.778      0.000     0.500     0.637     low
               0.000     0.000     0.000       0.000     0.000      0.000     0.500     0.363     high
Weighted Avg.  0.637     0.637     0.406       0.637     0.496      0.000     0.500     0.538

=== Confusion Matrix ===

    a    b   <-- classified as
 2948    0 |    a = low
 1679    0 |    b = high
```

Which algorithm is giving best accuracy and why?

Solution:

Name : Amit Goswami
Enrollment : 2014HT15501
Subject : Machine Learning

Data Set Name	Variable Type	#Instances	#Attributes
Iris.2d	Real	150	3
Supermarket	Real	4627	217

Considering the above data sets its no clear co-relation between the result of the classification and type of variables ,cardinality of instances, number of attributes or values of the target class. It seems that dataset with the better accuracy (Iris) is one with the lesser number of attributes.

It seems IBk Algorithm provide the best accuracy and its evident from the result from the Weka tool which is use for the analysis of the data sets and enclosed the detailed snapshots for your reference.

----------End----------