AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh

Assignment Title:	Project	Supervised
Assignment No: 2		Date of Submission: 15/05/2020
Course Title:		Data Warehousing and Data Mining
Course Code:		Section: A
Semester: Spring	20 19 - 20	Course Teacher: Rahman Mohammod Hafizur

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

** Student(s) must complete all details except the faculty use part.*

** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.:

No	Name	ID	Program	Signature
1	Sarkar, Christine Monisha	16-31255-1	CSE	

Faculty use only		
FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Project title: Supervised Learning

Problem statement: Compare between 5 classifier and choose the best -Understand the problem Definition -For solution, choose 5 different classifiers, you are free to choose your own -Build those 5 classifiers using “weka” -study them -Present a ROC graph-based comparison among the classifiers -Choose the best among them based on a scenario.

Introduction: In this project supervised learning is used to classify a dataset of a sea snails named Abalone using 5 classifiers in Weka. Supervised learning focuses on generating a required output by mapping the input (here the data on the dataset).

Dataset Information:

- The dataset Abalone [1] comes from an original study of a group of Australian researchers. It was obtained from UCI Machine Learning Repository.
- This dataset was mainly constructed to predict the age of abalone from its physical measurements, which are easier to obtain. Further information, such as weather patterns and location (hence food availability) may be required to accurately predict the age also.
- From the original data, examples with missing values were and the ranges of the continuous values have been scaled also.
- The number of rings is the value to predict to know the age of the abalone.

Attribute Information:

Name	Data Type	Measurement Unit	Description
Sex	nominal	--	M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer	--	+1.5 gives the age in years

Solution:

From this dataset I chose the attribute “Rings” as class attribute because adding +1.5 with the rings value gives the age of the Abalone.

Procedure:

- For this analysis the dataset was downloaded from the repository then converted into csv file manually.
- Then the file was opened in weka and the dataset was discretized with the filters option.
- After that, on the “Classify” tab the 5 classifiers were chosen one after another and they were run on the dataset orderly.
- Then the positive cases from each sets were taken to plot a ROC graph to compare the classifiers to determine the best performing classifier among them.

Classifiers used:

1. J48:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2292               54.8719 %
Incorrectly Classified Instances      1885               45.1281 %
Kappa statistic                          0.3102
Mean absolute error                      0.1165
Root mean squared error                  0.2448
Relative absolute error                 81.867  %
Root relative squared error             91.813  %
Total Number of Instances             4177

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.176     0.001     0.333       0.176   0.231      0.240    0.962     0.272     '(-inf-3.8]'
                 0.445     0.015     0.768       0.445   0.564      0.551    0.928     0.621     '(3.8-6.6]'
                 0.778     0.381     0.571       0.778   0.659      0.389    0.742     0.585     '(6.6-9.4]'
                 0.575     0.280     0.506       0.575   0.538      0.287    0.709     0.495     '(9.4-12.2]'
                 0.030     0.017     0.167       0.030   0.051      0.029    0.687     0.173     '(12.2-15]'
                 0.032     0.002     0.364       0.032   0.059      0.101    0.731     0.095     '(15-17.8]'
                 0.000     0.001     0.000       0.000   0.000      -0.005   0.730     0.059     '(17.8-20.6]'
                 0.000     0.000     0.000       0.000   0.000      -0.002   0.668     0.021     '(20.6-23.4]'
                 0.000     0.000     ?           0.000   ?          ?        0.416     0.001     '(23.4-26.2]'
                 0.000     0.000     ?           0.000   ?          ?        0.462     0.001     '(26.2-inf)'
Weighted Avg.    0.549     0.247     ?           0.549   ?          ?        0.744     0.483
```

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h    i    j   <-- classified as
   3   14    0    0    0    0    0    0    0    0 |    a = '(-inf-3.8]'
   6  192  229    4    0    0    0    0    0    0 |    b = '(3.8-6.6]'
   0   42 1282  313   11    0    0    0    0    0 |    c = '(6.6-9.4]'
   0    2  556  798   30    0    1    1    0    0 |    d = '(9.4-12.2]'
   0    0  137  277   13    2    3    0    0    0 |    e = '(12.2-15]'
   0    0   24   85   11    4    0    1    0    0 |    f = '(15-17.8]'
   0    0   13   75   10    2    0    0    0    0 |    g = '(17.8-20.6]'
   0    0    4   21    2    2    0    0    0    0 |    h = '(20.6-23.4]'
   0    0    0    3    1    0    0    0    0    0 |    i = '(23.4-26.2]'
   0    0    0    2    0    1    0    0    0    0 |    j = '(26.2-inf)'
```

2. Random Forest:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2176               52.0948 %
Incorrectly Classified Instances      2001               47.9052 %
Kappa statistic                          0.2978
Mean absolute error                      0.1103
Root mean squared error                  0.2491
Relative absolute error                 77.5136 %
Root relative squared error             93.415  %
Total Number of Instances             4177

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.235    0.000    0.667      0.235    0.348      0.395    0.964     0.317     '(-inf-3.8]'
                 0.552    0.036    0.642      0.552    0.594      0.553    0.926     0.650     '(3.8-6.6]'
                 0.658    0.325    0.569      0.658    0.610      0.327    0.748     0.628     '(6.6-9.4]'
                 0.552    0.274    0.501      0.552    0.525      0.272    0.723     0.518     '(9.4-12.2]'
                 0.155    0.050    0.265      0.155    0.196      0.135    0.692     0.191     '(12.2-15]'
                 0.088    0.011    0.200      0.088    0.122      0.115    0.769     0.117     '(15-17.8]'
                 0.060    0.010    0.133      0.060    0.083      0.075    0.757     0.086     '(17.8-20.6]'
                 0.000    0.003    0.000      0.000    0.000      -0.004   0.660     0.026     '(20.6-23.4]'
                 0.000    0.000    ?          0.000    ?          ?        0.605     0.003     '(23.4-26.2]'
                 0.000    0.000    ?          0.000    ?          ?        0.651     0.004     '(26.2-inf)'
Weighted Avg.    0.521    0.228    ?          0.521    ?          ?        0.753     0.514
```

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j   <-- classified as
    4   13    0    0    0    0    0    0    0    0 |    a = '(-inf-3.8]'
    2  238  184    7    0    0    0    0    0    0 |    b = '(3.8-6.6]'
    0  116 1084  412   27    6    3    0    0    0 |    c = '(6.6-9.4]'
    0    4  492  766   96   13   11    6    0    0 |    d = '(9.4-12.2]'
    0    0  110  225   67   12   15    3    0    0 |    e = '(12.2-15]'
    0    0   23   54   25   11    9    3    0    0 |    f = '(15-17.8]'
    0    0    9   48   29    8    6    0    0    0 |    g = '(17.8-20.6]'
    0    0    3   16    6    4    0    0    0    0 |    h = '(20.6-23.4]'
    0    0    0    1    2    1    0    0    0    0 |    i = '(23.4-26.2]'
    0    0    0    1    1    0    1    0    0    0 |    j = '(26.2-inf)'
```

3. Naïve Bayes:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2077               49.7247 %
Incorrectly Classified Instances      2100               50.2753 %
Kappa statistic                          0.2783
Mean absolute error                      0.1116
Root mean squared error                  0.2802
Relative absolute error                 78.4137 %
Root relative squared error            105.0789 %
Total Number of Instances             4177

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.294    0.001    0.500      0.294    0.370      0.382    0.989     0.385     '(-inf-3.8]'
                 0.828    0.102    0.483      0.828    0.610      0.579    0.928     0.516     '(3.8-6.6]'
                 0.488    0.248    0.562      0.488    0.523      0.247    0.694     0.590     '(6.6-9.4]'
                 0.649    0.342    0.485      0.649    0.555      0.291    0.722     0.506     '(9.4-12.2]'
                 0.019    0.020    0.098      0.019    0.031      -0.003   0.663     0.152     '(12.2-15]'
                 0.000    0.001    0.000      0.000    0.000      -0.006   0.735     0.067     '(15-17.8]'
                 0.010    0.001    0.143      0.010    0.019      0.032    0.752     0.061     '(17.8-20.6]'
                 0.000    0.002    0.000      0.000    0.000      -0.004   0.744     0.018     '(20.6-23.4]'
                 0.000    0.004    0.000      0.000    0.000      -0.002   0.787     0.004     '(23.4-26.2]'
                 0.000    0.005    0.000      0.000    0.000      -0.002   0.834     0.005     '(26.2-inf)'
Weighted Avg.    0.497    0.224    0.448      0.497    0.459      0.256    0.729     0.475
```

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   <-- classified as
   5  12   0   0   0   0   0   0   0   0 |   a = '(-inf-3.8]'
   5 357  66   3   0   0   0   0   0   0 |   b = '(3.8-6.6]'
   0 313 805 495  33   0   0   1   0   1 |   c = '(6.6-9.4]'
   0  53 379 901  28   2   1   3  12   9 |   d = '(9.4-12.2]'
   0   4 129 279   8   1   1   3   4   3 |   e = '(12.2-15]'
   0   0  30  83   5   0   1   3   0   3 |   f = '(15-17.8]'
   0   0  16  73   6   2   1   0   0   2 |   g = '(17.8-20.6]'
   0   0   8  18   1   0   1   0   0   1 |   h = '(20.6-23.4]'
   0   0   0   3   0   0   1   0   0   0 |   i = '(23.4-26.2]'
   0   0   0   1   1   0   1   0   0   0 |   j = '(26.2-inf)'
```

4. LMT

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2270               54.3452 %
Incorrectly Classified Instances      1907               45.6548 %
Kappa statistic                          0.306
Mean absolute error                      0.1147
Root mean squared error                  0.2398
Relative absolute error                 80.6045 %
Root relative squared error             89.9399 %
Total Number of Instances             4177

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.294     0.001     0.500       0.294   0.370      0.382    0.982     0.307     '(-inf-3.8]'
                 0.490     0.026     0.685       0.490   0.571      0.540    0.943     0.668     '(3.8-6.6]'
                 0.732     0.365     0.566       0.732   0.639      0.359    0.763     0.647     '(6.6-9.4]'
                 0.599     0.298     0.500       0.599   0.545      0.289    0.743     0.555     '(9.4-12.2]'
                 0.032     0.008     0.311       0.032   0.059      0.071    0.747     0.231     '(12.2-15]'
                 0.024     0.003     0.200       0.024   0.043      0.060    0.809     0.134     '(15-17.8]'
                 0.000     0.000     0.000       0.000   0.000      -0.002   0.806     0.086     '(17.8-20.6]'
                 0.000     0.000     0.000       0.000   0.000      -0.002   0.772     0.028     '(20.6-23.4]'
                 0.000     0.001     0.000       0.000   0.000      -0.001   0.458     0.001     '(23.4-26.2]'
                 0.000     0.000     0.000       0.000   0.000      -0.001   0.143     0.001     '(26.2-inf)'
Weighted Avg.    0.543     0.247     0.501       0.543   0.501      0.304    0.776     0.540
```

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h    i    j   <-- classified as
   5   12    0    0    0    0    0    0    0    0 |    a = '(-inf-3.8]'
   5  211  212    3    0    0    0    0    0    0 |    b = '(3.8-6.6]'
   0   80 1206  356    6    0    0    0    0    0 |    c = '(6.6-9.4]'
   0    5  537  831   13    1    0    0    0    1 |    d = '(9.4-12.2]'
   0    0  131  281   14    4    0    1    1    0 |    e = '(12.2-15]'
   0    0   29   83    9    3    1    0    0    0 |    f = '(15-17.8]'
   0    0   10   82    2    3    0    0    2    1 |    g = '(17.8-20.6]'
   0    0    4   23    0    2    0    0    0    0 |    h = '(20.6-23.4]'
   0    0    0    2    1    0    0    1    0    0 |    i = '(23.4-26.2]'
   0    0    0    1    0    2    0    0    0    0 |    j = '(26.2-inf)'
```

5. OneR

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2252               53.9143 %
Incorrectly Classified Instances      1925               46.0857 %
Kappa statistic                          0.2823
Mean absolute error                      0.0922
Root mean squared error                  0.3036
Relative absolute error                 64.7689 %
Root relative squared error            113.8546 %
Total Number of Instances             4177

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.294    0.001    0.455      0.294    0.357      0.364    0.646     0.137     '(-inf-3.8]'
                 0.448    0.015    0.778      0.448    0.568      0.558    0.717     0.405     '(3.8-6.6]'
                 0.835    0.518    0.512      0.835    0.635      0.323    0.658     0.493     '(6.6-9.4]'
                 0.488    0.199    0.550      0.488    0.518      0.299    0.645     0.439     '(9.4-12.2]'
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.103     '(12.2-15]'
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.030     '(15-17.8]'
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.024     '(17.8-20.6]'
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.007     '(20.6-23.4]'
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.001     '(23.4-26.2]'
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.001     '(26.2-inf)'
Weighted Avg.    0.539    0.272    ?          0.539    ?          ?        0.634     0.395
```

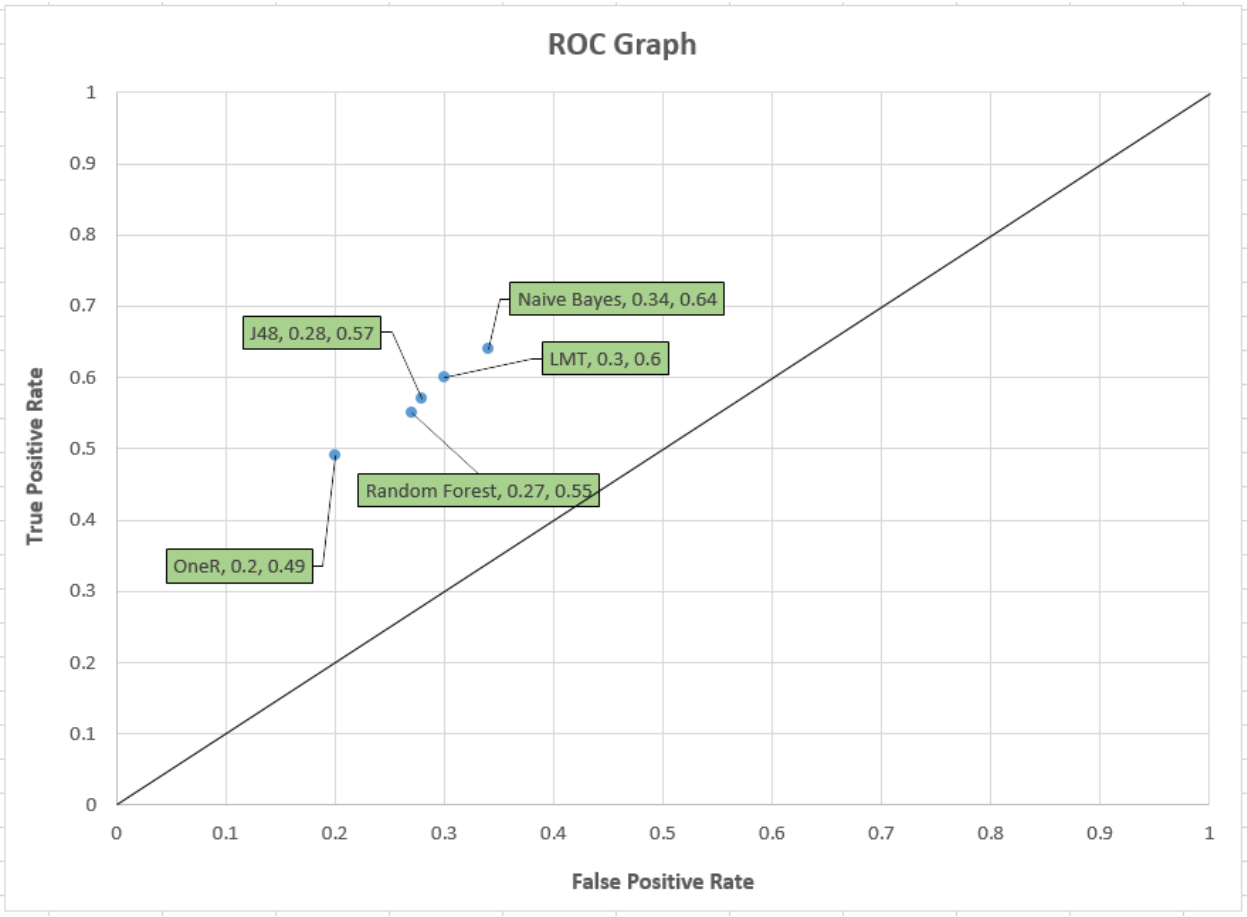
```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j   <-- classified as
    5   12    0    0    0    0    0    0    0    0 |    a = '(-inf-3.8]'
    6  193  230    2    0    0    0    0    0    0 |    b = '(3.8-6.6]'
    0   42 1376  230    0    0    0    0    0    0 |    c = '(6.6-9.4]'
    0    1  709  678    0    0    0    0    0    0 |    d = '(9.4-12.2]'
    0    0  248  184    0    0    0    0    0    0 |    e = '(12.2-15]'
    0    0   63   62    0    0    0    0    0    0 |    f = '(15-17.8]'
    0    0   46   54    0    0    0    0    0    0 |    g = '(17.8-20.6]'
    0    0   13   16    0    0    0    0    0    0 |    h = '(20.6-23.4]'
    0    0    0    4    0    0    0    0    0    0 |    i = '(23.4-26.2]'
    0    0    1    2    0    0    0    0    0    0 |    j = '(26.2-inf)'
```

ROC graph:

To plot the ROC graph first from all 5 confusion matrix a positive case was considered, which was class d= '(9.4-12.2)' and then the TPR and FPR of all the classifiers were computed and plotted.

Classifiers	FPR	TPR
Naïve Bayes	0.34	0.64
J48	0.28	0.57
Random Forest	0.27	0.55
OneR	0.2	0.49
LMT	0.3	0.6



Analysis:

From the ROC graph, we can measure the performance of the classifiers by measuring the distances of the points to the best possible classifier value (0, 1).

We know,

$$\text{Euc} = \sqrt{(FPR)^2 + (1 - TPR)^2}$$

1. J48:

$\text{Euc} = \sqrt{(0.28)^2 + (1 - 0.57)^2}$ = 0.51

2. Random Forest:

$\text{Euc} = \sqrt{(0.27)^2 + (1 - 0.55)^2}$ = 0.52

3. Naïve Bayes:

$\text{Euc} = \sqrt{(0.34)^2 + (1 - 0.64)^2}$ = 0.49

4. LMT:

$\text{Euc} = \sqrt{(0.30)^2 + (1 - 0.60)^2}$ = 0.50

5. OneR:

$\text{Euc} = \sqrt{(0.20)^2 + (1 - 0.49)^2}$ = 0.54

So, for the classifiers we can say that the Naïve Bayes classifier outperforms others in regards of accuracy, because it has the lowest distance from (0,1) the best point.

Conclusion:

From this dataset to determine the age of an Abalone 5 classifiers were performed and the classifiers J48 and LMT provided the best result in terms of determining the age. By considering the positive case as d= '(9.4-12.2)', the ROC graph was plotted and Naïve Bayes determines the age most efficiently than other classifiers used.

Reference:

[1] Abalone, UCI Machine Learning Repository,

http://archive.ics.uci.edu/ml/datasets/Abalone