DADS Experiment No: 1

Name: Abhishek S Waghchaure

PRN: **1032221714**

Dept: FY M Tech DSA(2022-24)

Aim:

Implement two pre-processing techniques on standardized data from UCI repository or any other as recommended by the instructor.

What is WEKA?:

WEKA - an open-source software provides tools for data pre-processing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world da It contains a Collection of visualization tools and algorithms for data analysis and predictive modelling coupled with graphical user interface.

Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regressing, visualization and feature selection.

Weka has the following advantages, such as: Free availability under the GNU General Public License. Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform. A comprehensive collection of data pre-processing and modelling techniques.

WEKA Interface:

Weka GUI Chooser

The entry point into the Weka interface is the Weka GUI Chooser.

It is an interface that lets you choose and launch a specific Weka environment.

1 Screenshot of the Weka GUI Chooser

Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

Weka Explorer

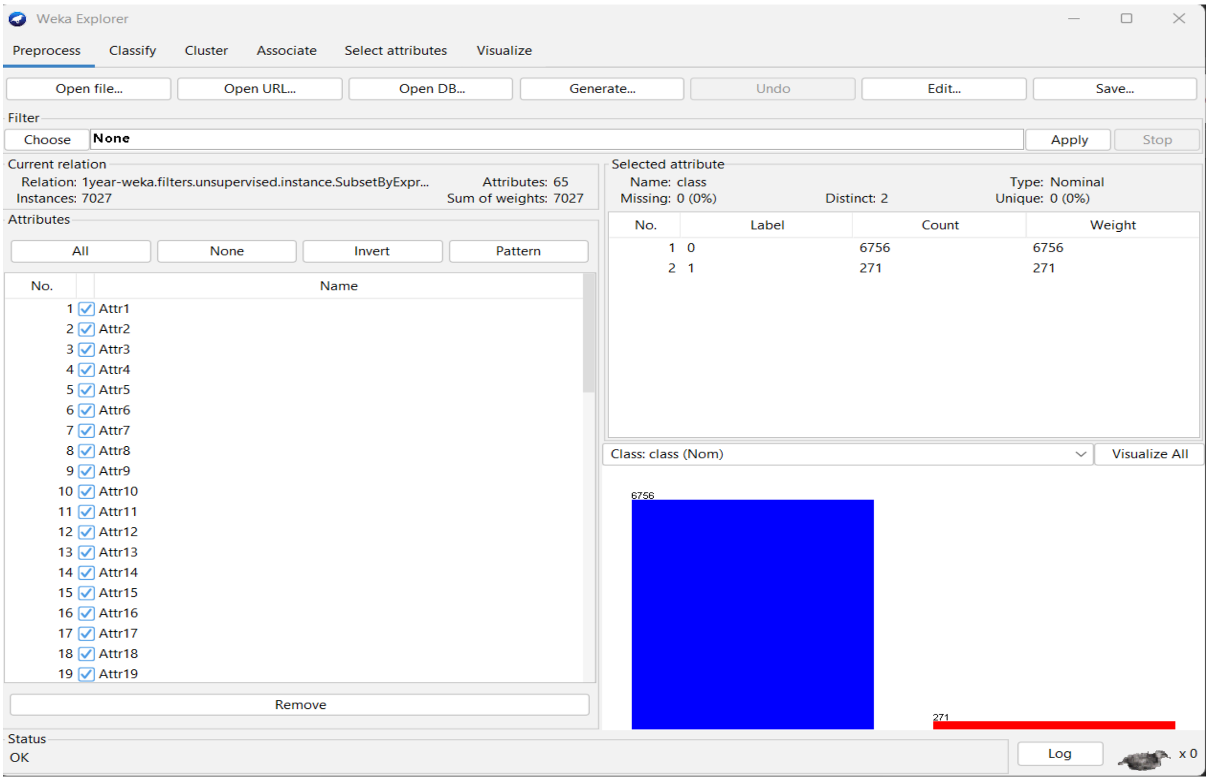
The Weka Explorer is designed to investigate your machine learning dataset.

It is useful when you are thinking about different data transforms and modelling algorithms that you could investigate with a controlled experiment later. It is excellent for getting ideas and playing what-if scenarios.

The interface is divided into 6 tabs, each with a specific function:

The pre-process tab is for loading your dataset and applying filters to transform the data into a form that better exposes the structure of the problem to the modelling processes. Also provides some summary statistics about loaded data.

Load a standard dataset in the data/ directory of your Weka installation, specifically ex. data/ 1year.arff. This is a binary classification problem that we will use on this tour.



Filter
Choose None

Current relation
Relation: 1year-weka.filters.unsupervised.instance.SubsetByExpr... Attributes: 65
Instances: 7027 Sum of weights: 7027

Selected attribute
Name: class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	6756	6756
2	1	271	271

Attributes

No.		Name
1	☑	Attr1
2	☑	Attr2
3	☑	Attr3
4	☑	Attr4
5	☑	Attr5
6	☑	Attr6
7	☑	Attr7
8	☑	Attr8
9	☑	Attr9
10	☑	Attr10
11	☑	Attr11
12	☑	Attr12
13	☑	Attr13
14	☑	Attr14
15	☑	Attr15
16	☑	Attr16
17	☑	Attr17
18	☑	Attr18
19	☑	Attr19

Class: class (Nom)

Status
OK

Dataset (website/sample data):

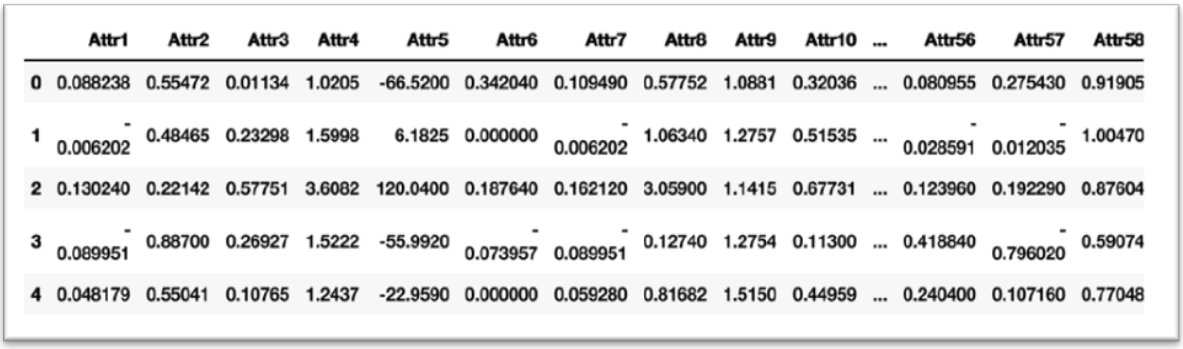
Website: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

Data Set Information:

The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service (EMIS, [Web Link]), which is a database containing information on emerging markets around the world. The bankrupt companies were analysed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.

Basing on the collected data depends on the forecasting period:- 1stYear the data contains financial rates from 1st year of the forecasting period and corresponding class label that indicates bankruptcy status after 5 years. The data contains 7027 instances (financial statements), 271 represents bankrupted companies, 6756 firms that did not bankrupt in the forecasting period.

Sample data:



	Attr1	Attr2	Attr3	Attr4	Attr5	Attr6	Attr7	Attr8	Attr9	Attr10	...	Attr56	Attr57	Attr58
0	0.088238	0.55472	0.01134	1.0205	-66.5200	0.342040	0.109490	0.57752	1.0881	0.32036	...	0.080955	0.275430	0.91905
1	-0.006202	0.48465	0.23298	1.5998	6.1825	0.000000	-0.006202	1.06340	1.2757	0.51535	...	-0.028591	-0.012035	1.00470
2	0.130240	0.22142	0.57751	3.6082	120.0400	0.187640	0.162120	3.05900	1.1415	0.67731	...	0.123960	0.192290	0.87604
3	-0.089951	0.88700	0.26927	1.5222	-55.9920	-0.073957	-0.089951	0.12740	1.2754	0.11300	...	0.418840	-0.796020	0.59074
4	0.048179	0.55041	0.10765	1.2437	-22.9590	0.000000	0.059280	0.81682	1.5150	0.44959	...	0.240400	0.107160	0.77048

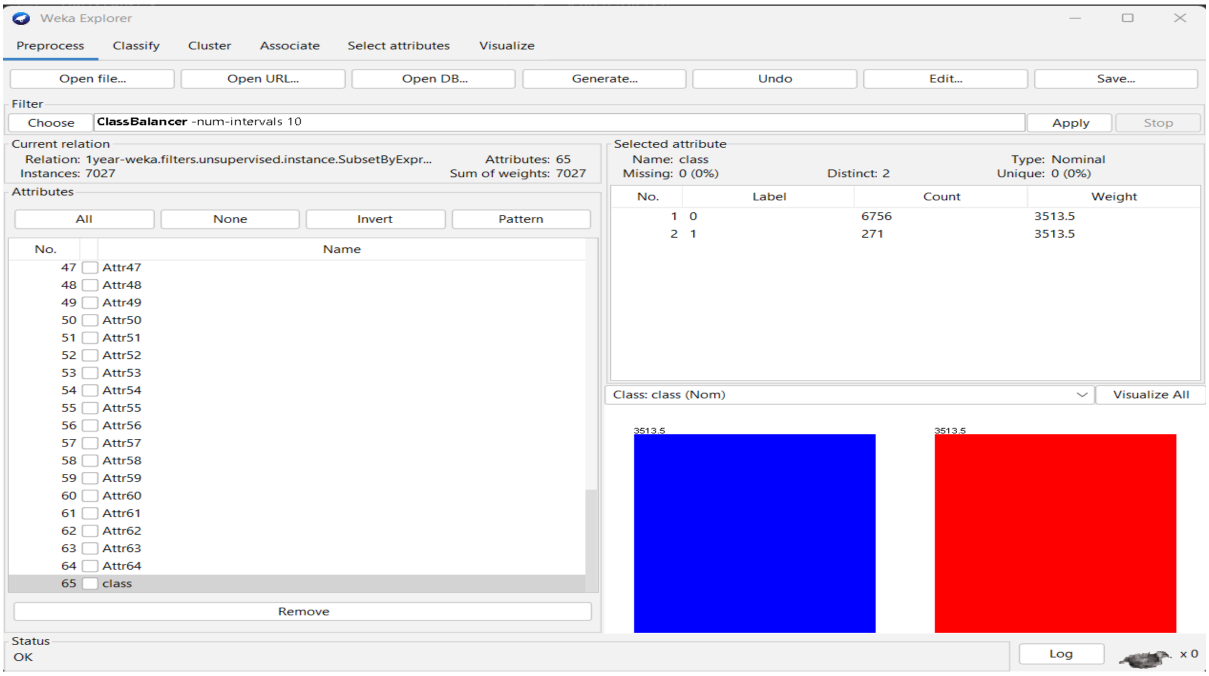
Steps involved in WEKA:

1. Class balancer:

A frequent question of Weka users is how to implement oversampling or under sampling, which are two common strategies for dealing with imbalanced classes in classification problems. This post provides some explanation.

When a binary classification problem has a lot less data in one class than in the other one, e.g., when learning from results of a medical test for which the vast majority of instances have a negative outcome and only a few returns positive, some machine learning algorithms will simply learn to ignore the minority class and classify all cases into the majority class because this will trivially yield high classification accuracy. This kind of classification model is clearly not useful. Two common methods for combating this problem are under sampling of the majority class and oversampling of the minority class respectively.

In our case we use Oversampling for a given dataset of Bankruptcy.



Filter
ClassBalancer -num-intervals 10

Current relation
Relation: 1year-weka.filters.unsupervised.instance.SubsetByExpr... Attributes: 65
Instances: 7027 Sum of weights: 7027

Attributes

No.	Name
47	Attr47
48	Attr48
49	Attr49
50	Attr50
51	Attr51
52	Attr52
53	Attr53
54	Attr54
55	Attr55
56	Attr56
57	Attr57
58	Attr58
59	Attr59
60	Attr60
61	Attr61
62	Attr62
63	Attr63
64	Attr64
65	class

Selected attribute
Name: class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

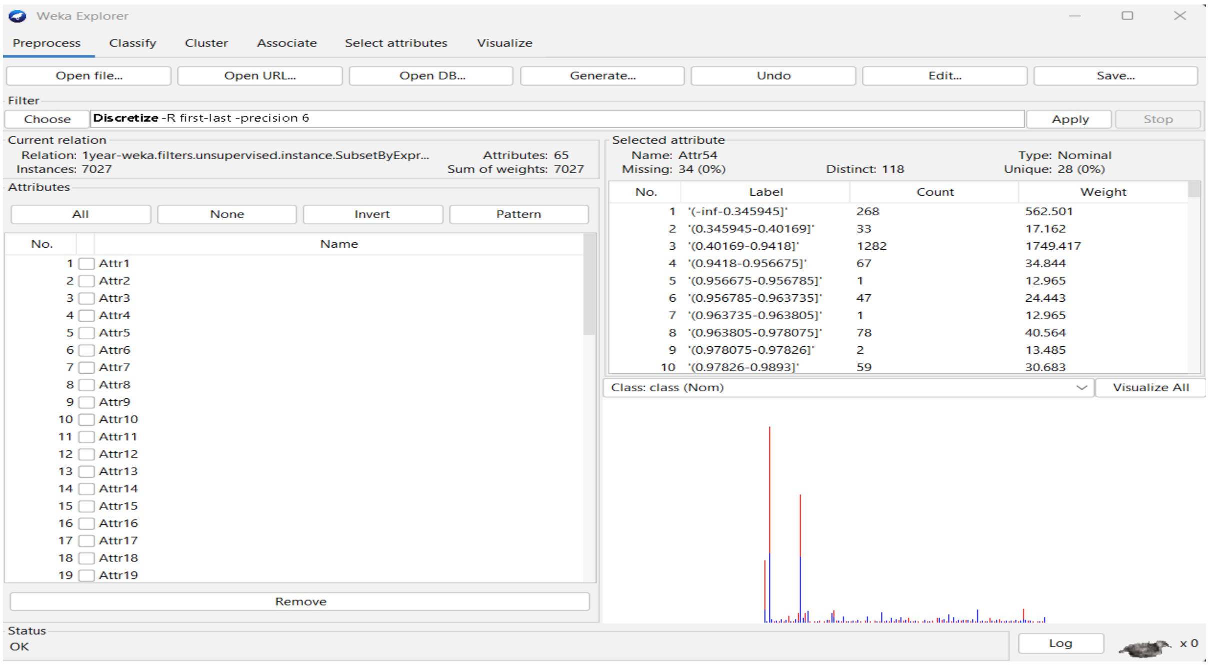
No.	Label	Count	Weight
1	0	6756	3513.5
2	1	271	3513.5

Class: class (Nom)

Status
OK

2. Discretize

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes. Discretization is by simple binning. Skips the class attribute if set.



Filter
Discretize -R first-last -precision 6

Current relation
Relation: 1year-weka.filters.unsupervised.instance.SubsetByExpr... Attributes: 65
Instances: 7027 Sum of weights: 7027

Attributes

No.	Name
1	Attr1
2	Attr2
3	Attr3
4	Attr4
5	Attr5
6	Attr6
7	Attr7
8	Attr8
9	Attr9
10	Attr10
11	Attr11
12	Attr12
13	Attr13
14	Attr14
15	Attr15
16	Attr16
17	Attr17
18	Attr18
19	Attr19

Selected attribute
Name: Attr54 Type: Nominal
Missing: 34 (0%) Distinct: 118 Unique: 28 (0%)

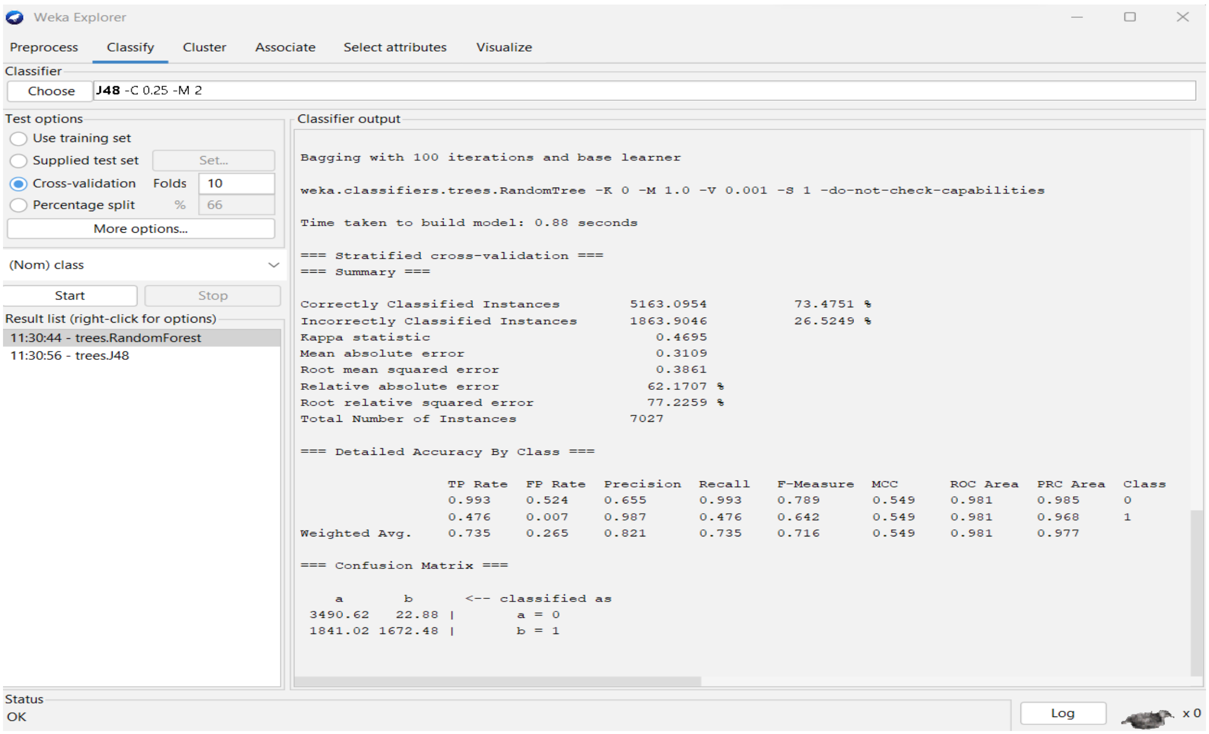
No.	Label	Count	Weight
1	'(-inf-0.345945]'	268	562.501
2	'(0.345945-0.40169]'	33	17.162
3	'(0.40169-0.9418]'	1282	1749.417
4	'(0.9418-0.956675]'	67	34.844
5	'(0.956675-0.956785]'	1	12.965
6	'(0.956785-0.963735]'	47	24.443
7	'(0.963735-0.963805]'	1	12.965
8	'(0.963805-0.978075]'	78	40.564
9	'(0.978075-0.97826]'	2	13.485
10	'(0.97826-0.9893]'	59	30.683

Class: class (Nom)

Status
OK

Results obtained:

Random Forest gives 73% accuracy.



Classifier: J48 -C 0.25 -M 2

● Cross-validation Folds 10

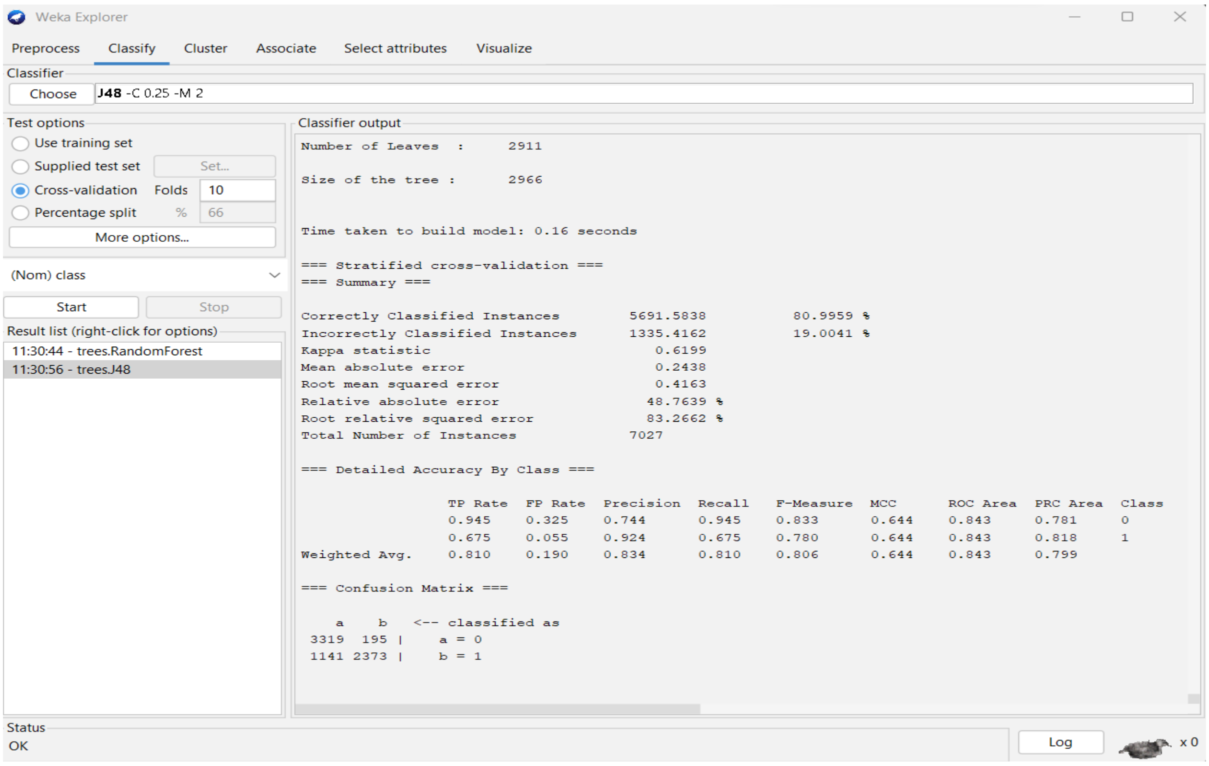
(Nom) class

11:30:44 - trees.RandomForest
11:30:56 - trees.J48

```
Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.88 seconds

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

Correctly Classified Instances        5163.0954          73.4751 %
Incorrectly Classified Instances      1863.9046          26.5249 %
Kappa statistic                          0.4695
Mean absolute error                      0.3109
Root mean squared error                  0.3861
Relative absolute error                 62.1707 %
Root relative squared error             77.2259 %
Total Number of Instances             7027

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.993    0.524    0.655      0.993    0.789      0.549    0.981     0.985     0
                 0.476    0.007    0.987      0.476    0.642      0.549    0.981     0.968     1
Weighted Avg.    0.735    0.265    0.821      0.735    0.716      0.549    0.981     0.977

=== Confusion Matrix ===

     a       b     <-- classified as
 3490.62   22.88 |      a = 0
 1841.02 1672.48 |      b = 1
```

Status: OK

Next, J48 model gives 73% accuracy which is more than random Forest model.



Classifier: J48 -C 0.25 -M 2

● Cross-validation Folds 10

(Nom) class

11:30:44 - trees.RandomForest
11:30:56 - trees.J48

```
Number of Leaves  :     2911

Size of the tree :      2966


Time taken to build model: 0.16 seconds

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

Correctly Classified Instances        5691.5838          80.9959 %
Incorrectly Classified Instances      1335.4162          19.0041 %
Kappa statistic                          0.6199
Mean absolute error                      0.2438
Root mean squared error                  0.4163
Relative absolute error                 48.7639 %
Root relative squared error             83.2662 %
Total Number of Instances             7027

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.945    0.325    0.744      0.945    0.833      0.644    0.843     0.781     0
                 0.675    0.055    0.924      0.675    0.780      0.644    0.843     0.818     1
Weighted Avg.    0.810    0.190    0.834      0.810    0.806      0.644    0.843     0.799

=== Confusion Matrix ===

    a    b   <-- classified as
 3319  195 |    a = 0
 1141 2373 |    b = 1
```

Status: OK

Conclusion:

WEKA is a powerful tool for developing machine learning models. It provides implementation of several most widely used ML algorithms. Before these algorithms are applied to your dataset, it also allows you to pre-process the data. The types of algorithms that are supported are classified under Classify, Cluster, Associate, and Select attributes. The result at various stages of processing can be visualized with a beautiful and powerful visual representation. This makes it easier for a Data Scientist to quickly apply the various machine learning techniques on his dataset, compare the results and create the best model for the final use.

Reference:

- https://machinelearningmastery.com/tour-weka-machine-learning-workbench/
- https://www.tutorialspoint.com/weka/weka_quick_guide.htm
- https://weka.sourceforge.io/doc.dev/weka/filters/unsupervised/attribute/Discretize.html
- https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/