Omkar Gurav
BE IT
8048

Camlin Page
Date / /

Assignment 1

- **Title:-** Study of platform for implementation of assignment using WEKA

- **Theory:-**

a. What is machine learning? Explain the ~~topics~~ types of Learning. Explain various applications of Machine Learning.

→ Machine learning is the study of algorithms that improve their performance P at some task T with experience E. It is study of computer systems that learn from data & experience.
Machine Learning has various application areas like medicines, advertising, Biometric Recognition etc.

There are 4 types of Machine Learning:-

① Supervised Learning:-

The task of grouping data with prior information in terms of labelled training data is known as supervised learning. In training data each instance is a pair of an input object & desired output value. Supervised learning problems can further separated into regression & classification problems.

In classification task the output variable is a label of class like "yes" or "no", "approved" or "not approved" or "red" or "blue".

In regression task the output variable is a real value, such as "dollars" or "weight".

② Unsupervised Learning:-

The algorithm starts to learn from the given data in unrestrained environment, after learning of the algorithm, the actual answers are not matched with the correct answers. The complete process has the unsupervised approach so it is known as unsupervised learning. Unsupervised learning problems can further separated into clustering & association problems.

The task of dividing the data into similar group is known as clustering.

The task of finding out rules that describe large portions of the data is known as association.

③ Semi Supervised Learning:-

It is a task that deals with a large amount of unlabelled data & small amount of labelled data. These problems initiate in between both supervised & unsupervised learning. Good example is web page classification where only some of the web pages are labelled & the majority are unlabelled.

④ Reinforcement Learning:-

It is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward.

Applications of Machine Learning:-

① Automatic Language Translation
② Medical Diagnosis
③ Stock Market Trading
④ Online Fraud Detection
⑤ Virtual Personal Assistant
⑥ Email Spam & Malware Filtering etc.

b. Explain the WEKA platform with respect to the following points
i. Introduction ii. Installation steps iii. Features
iv. Advantages v. Disadvantages vi. Applications

→ Introduction:-

WEKA is open source software under GNU General Public License. System is developed at the University of Waikato in New Zealand. "WEKA" stands for the Waikato Environment For knowledge Analysis. The system is written using object orientied language Java. Weka provides implementations of state-of-the-art data mining & machine learning algorithms. Weka contains modules for data preprocessing, classification, clustering & association rule extraction.

Installation Steps:-

① Download Weka from http://www.cs.waikato.ac.nz/ml/weka/
② Choose self extracting executable (including Java VM)
③ After Download is completed, run the self extracting

file to install Weka & use the default set-ups.

Features:-

1. 49 data pre processing tools
2. 79+ classification/regression algorithms
3. 3 graphical user interfaces
   - The Explorer (exploratory data analysis)
   - The Experimenter (experimental environment)
   - The Knowledge Flow (new process model inspired interface)
4. Platform independent
5. Easy to use even for non technical users

Advantages:-

1. Free availability
2. Supports standard data mining tasks
3. Easy to use GUI
4. Comparing different approaches is really easy
5. Constantly under development

Disadvantages:-

1. Memory bound
2. Not capable of multi relational data mining
3. Limited scaling
4. GUI does not implement all the possible options
5. Sequence modeling is not covered by algorithms included in Weka distribution.

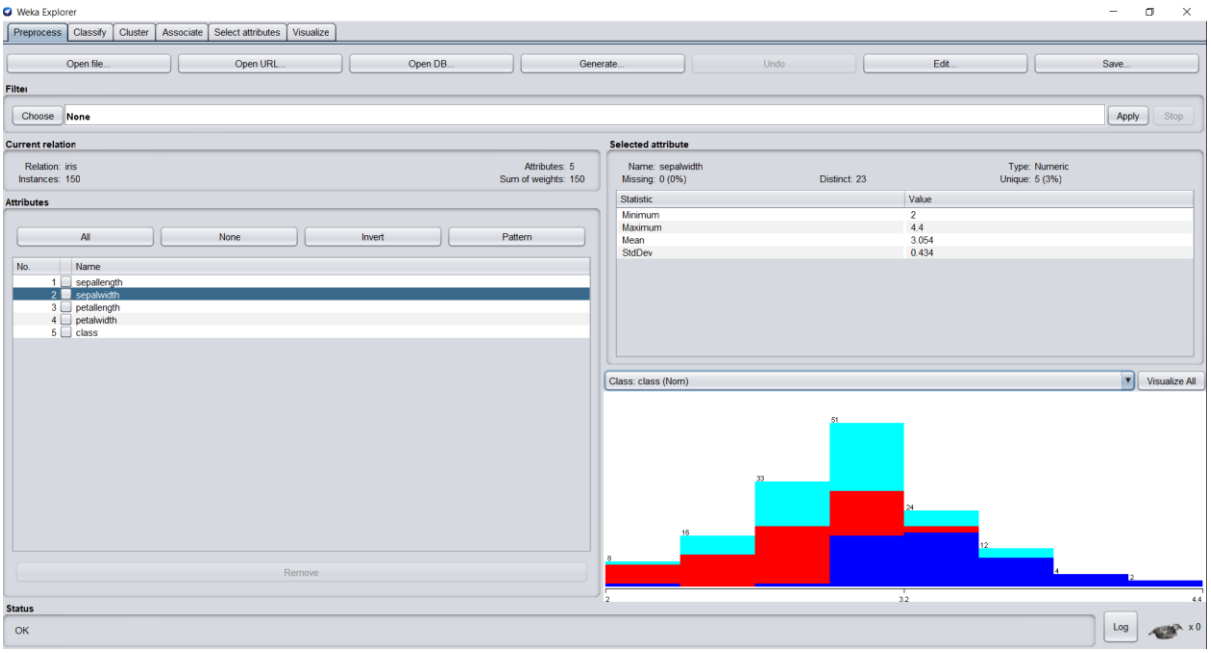
Camlin Page
Date / /

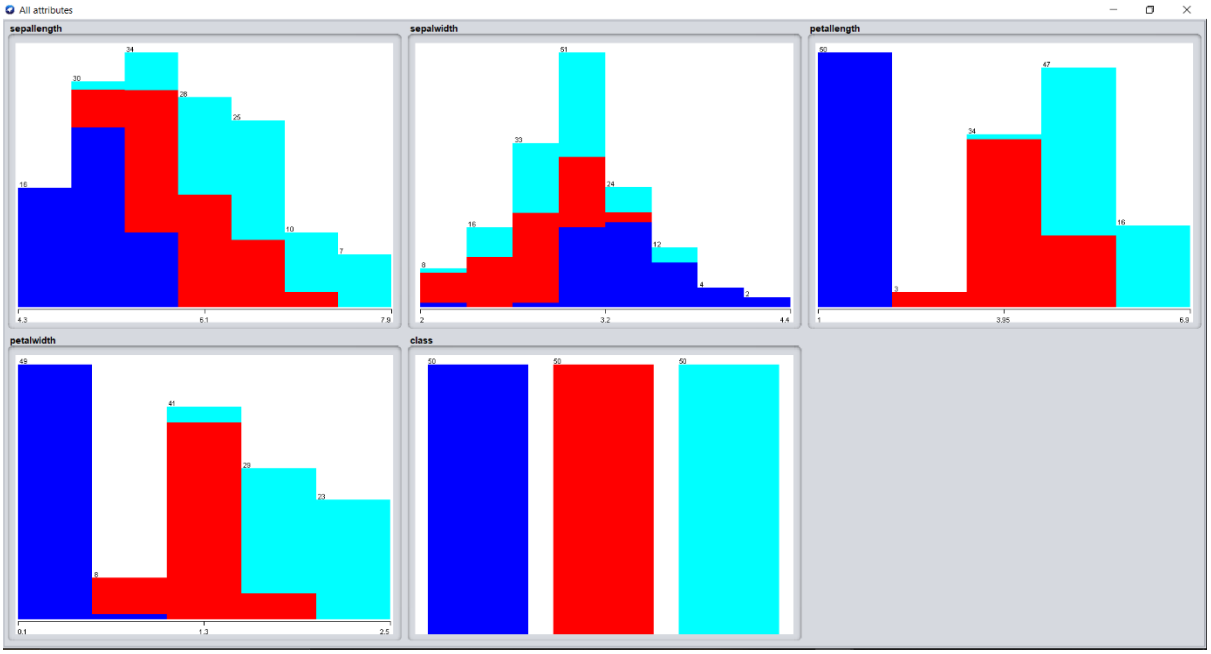
Applications:-

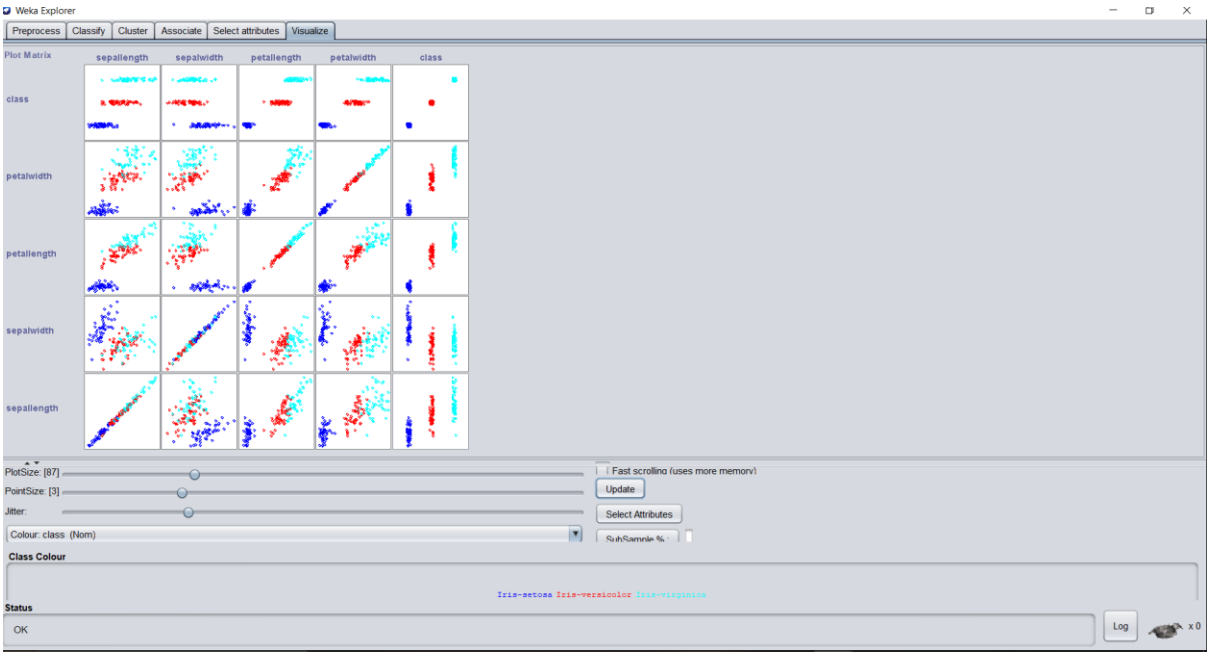
- The WEKA system has been successfully applied in a variety of areas including the areas of agriculture, machine learning research & education.

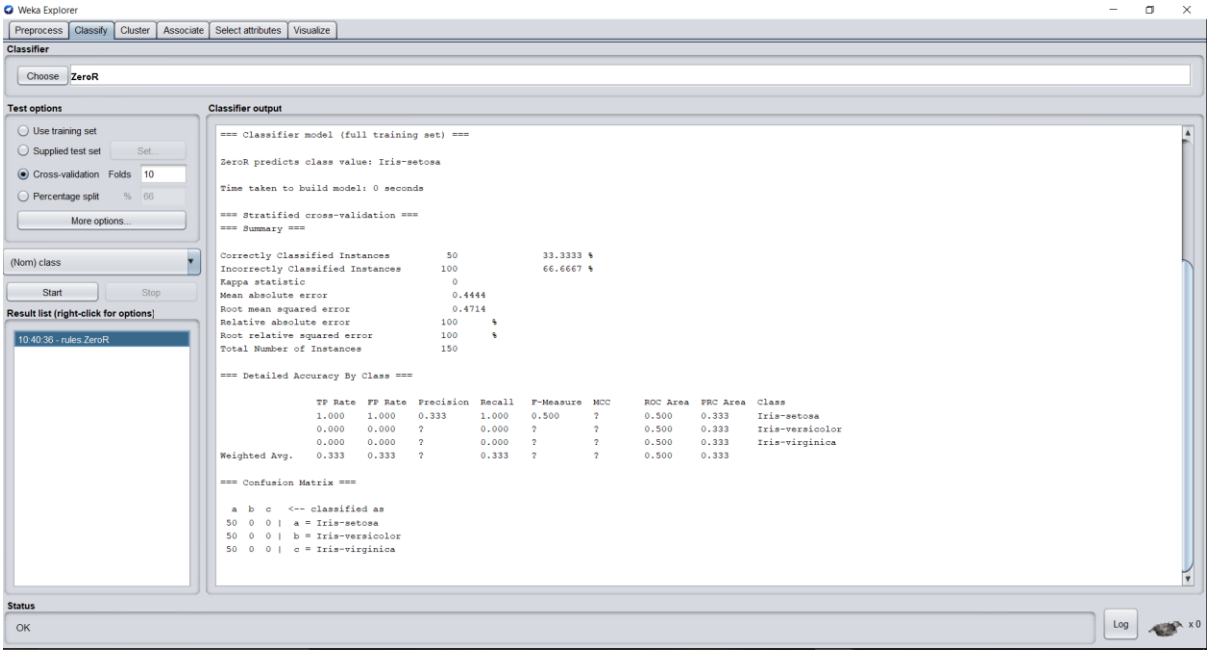
- Conclusion:- Study of WEKA platform is implemented.

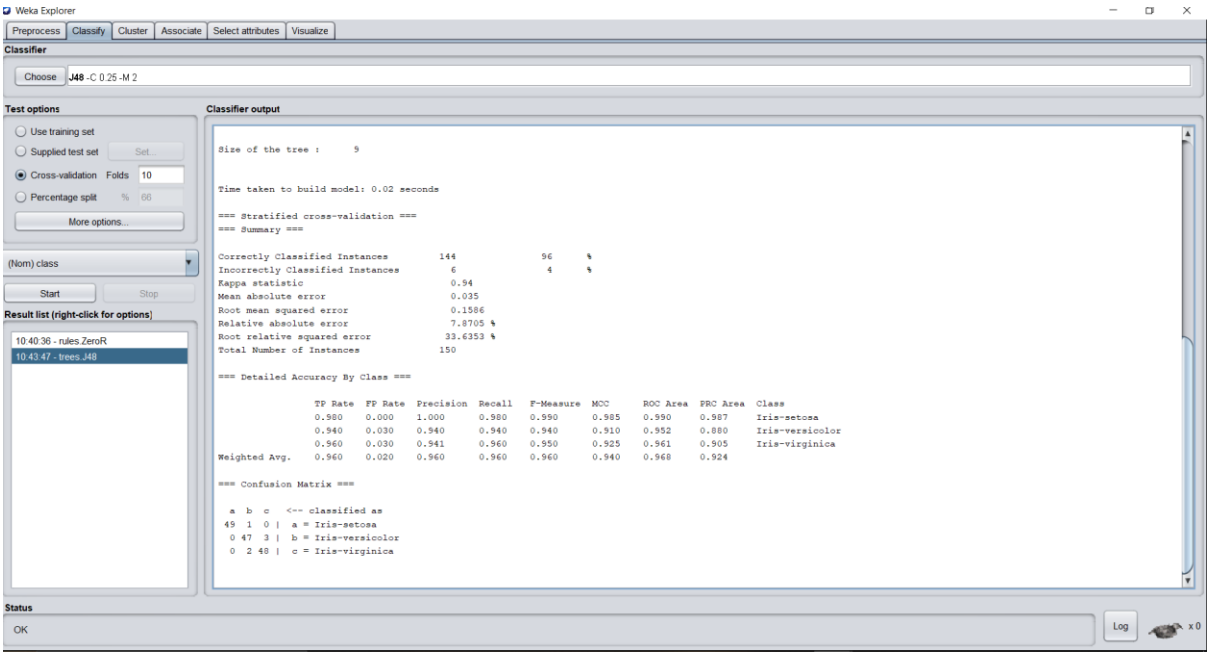
Screenshots :

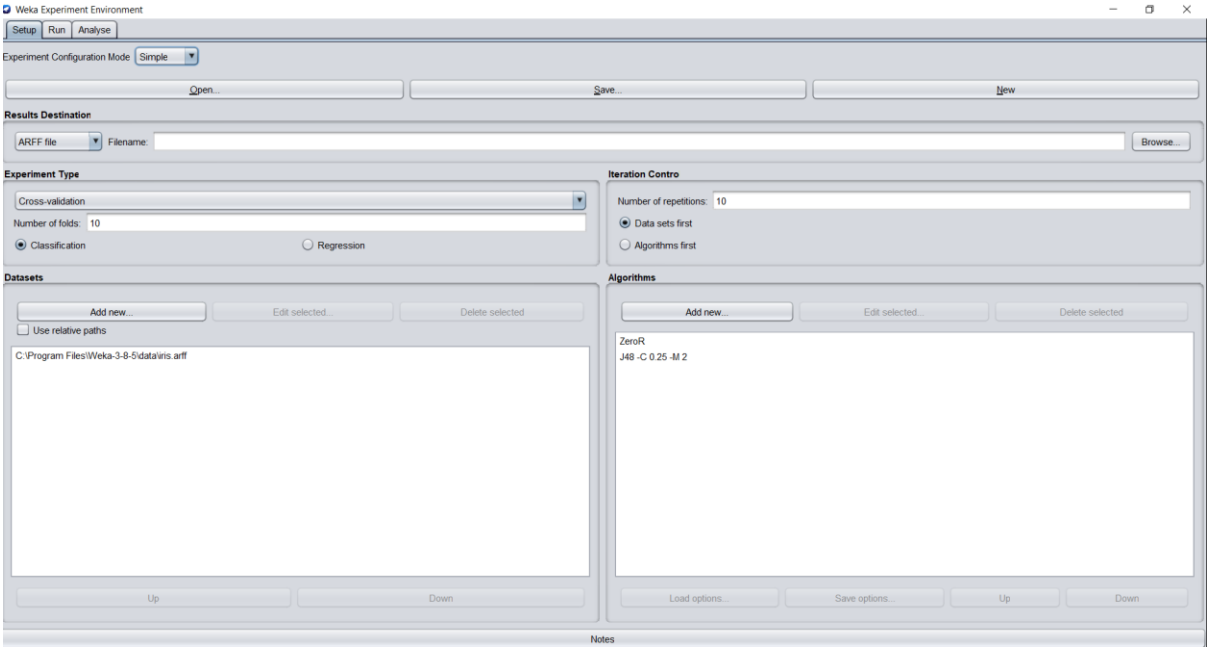


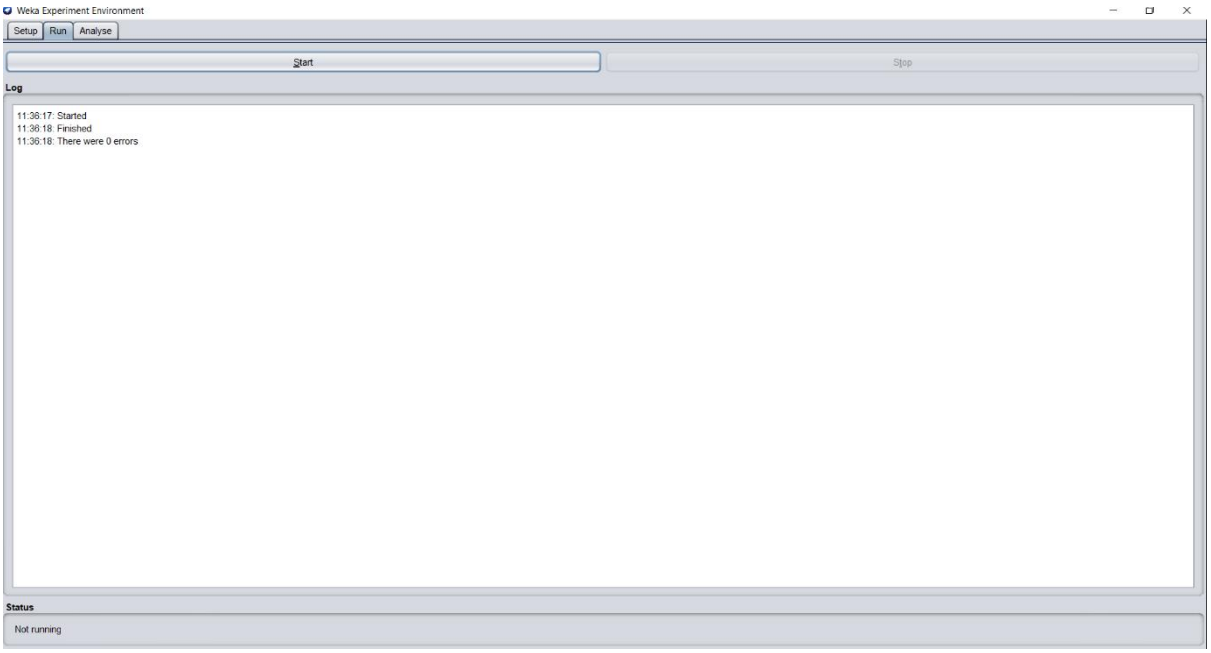


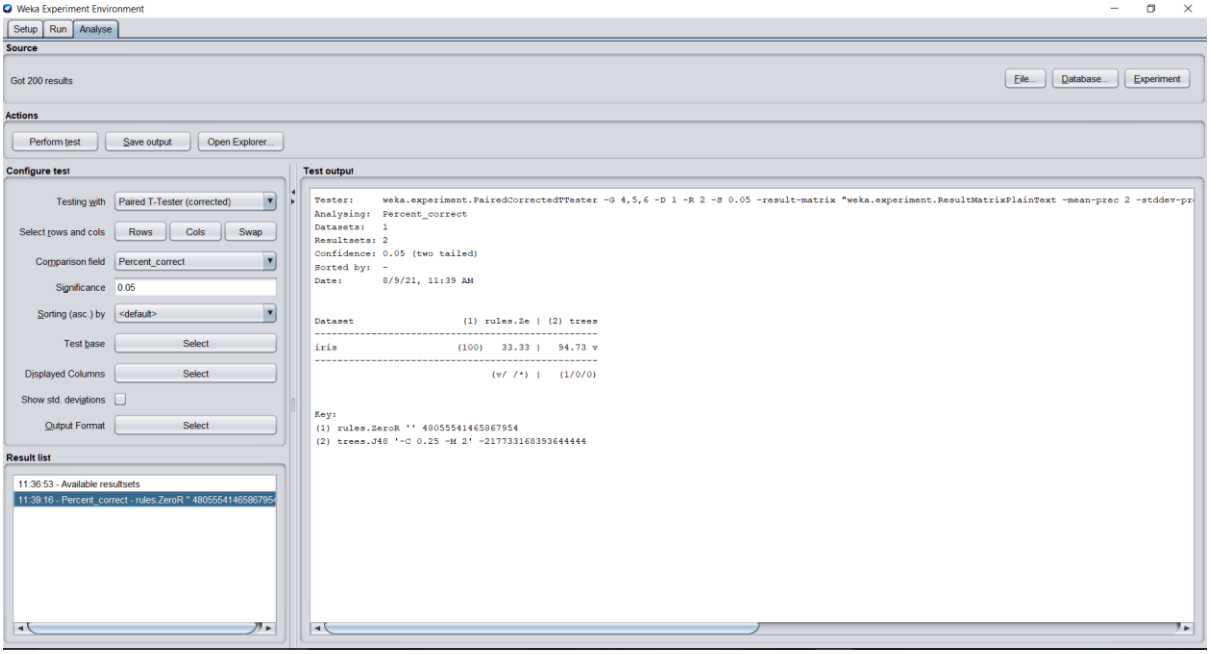

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
sepallength
sepalwidth
petallength
petalwidth
class
PlotSize: [87]
PointSize: [3]
Jitter:
Colour: class (Nom)
Fast scrolling (uses more memory)
Update
Select Attributes
Class Colour
Iris-setosa Iris-versicolor Iris-virginica
Status
OK
Log
x 0

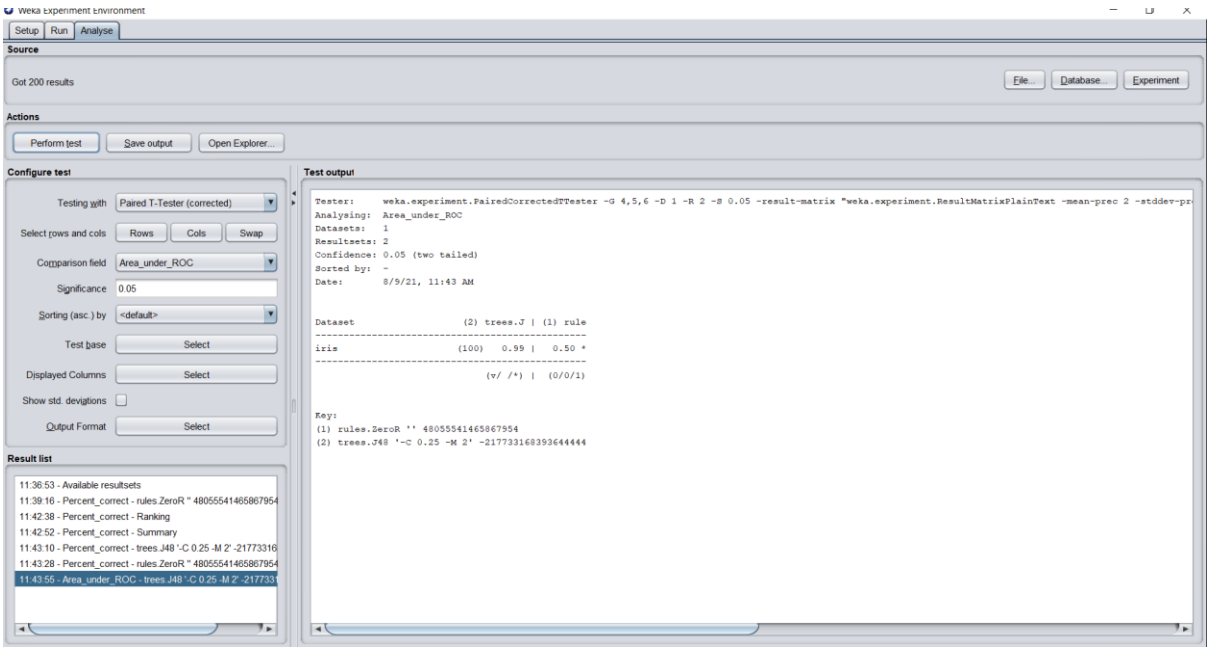

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
10:40:36 - rules.ZeroR
Classifier output
=== Classifier model (full training set) ===
ZeroR predicts class value: Iris-setosa
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 50 33.3333 %
Incorrectly Classified Instances 100 66.6667 %
Kappa statistic 0
Mean absolute error 0.4444
Root mean squared error 0.4714
Relative absolute error 100 %
Root relative squared error 100 %
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 1.000 0.333 1.000 0.500 ? 0.500 0.333 Iris-setosa
0.000 0.000 ? 0.000 ? ? 0.500 0.333 Iris-versicolor
0.000 0.000 ? 0.000 ? ? 0.500 0.333 Iris-virginica
Weighted Avg. 0.333 0.333 ? 0.333 ? ? 0.500 0.333
=== Confusion Matrix ===
a b c <-- classified as
50 0 0 | a = Iris-setosa
50 0 0 | b = Iris-versicolor
50 0 0 | c = Iris-virginica
Status
OK
Log
x 0

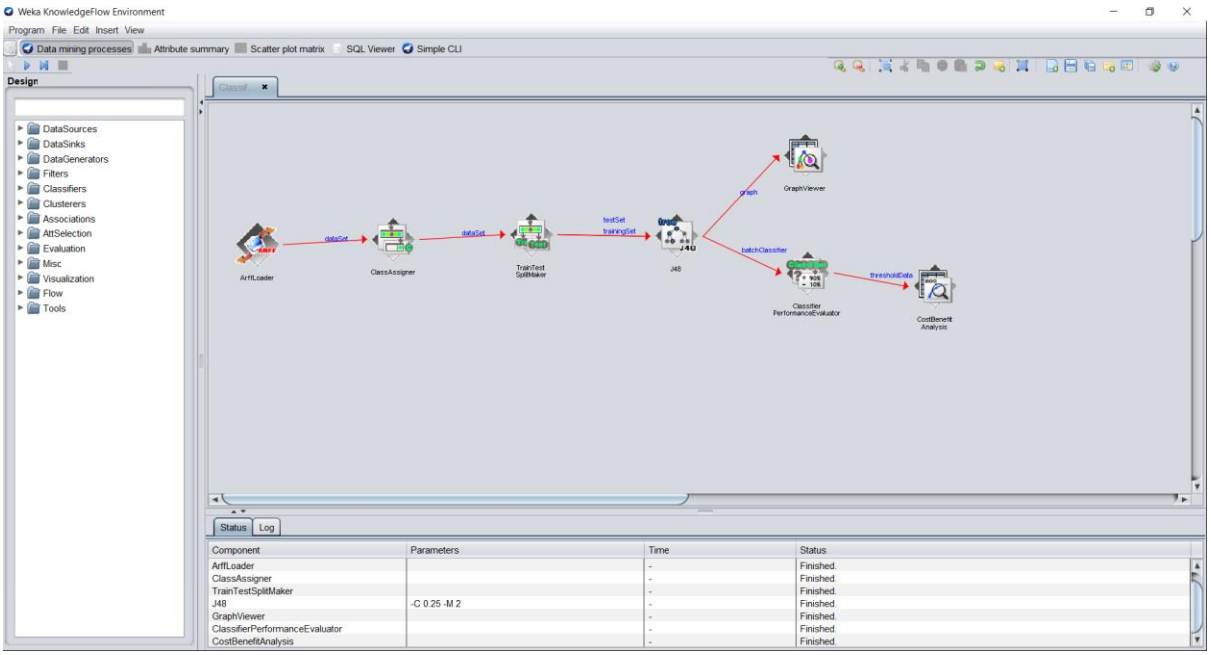

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
10:40:36 - rules.ZeroR
10:43:47 - trees.J48
Classifier output
Size of the tree : 9
Time taken to build model: 0.02 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 144 96 %
Incorrectly Classified Instances 6 4 %
Kappa statistic 0.94
Mean absolute error 0.035
Root mean squared error 0.1586
Relative absolute error 7.8705 %
Root relative squared error 33.6353 %
Total Number of Instances 150
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.980 0.000 1.000 0.980 0.990 0.985 0.990 0.987 Iris-setosa
0.940 0.030 0.940 0.940 0.940 0.910 0.952 0.880 Iris-versicolor
0.960 0.030 0.941 0.960 0.950 0.925 0.961 0.905 Iris-virginica
Weighted Avg. 0.960 0.020 0.960 0.960 0.960 0.940 0.968 0.924
=== Confusion Matrix ===
a b c <-- classified as
49 1 0 | a = Iris-setosa
0 47 3 | b = Iris-versicolor
0 2 48 | c = Iris-virginica
Status
OK
Log
x 0

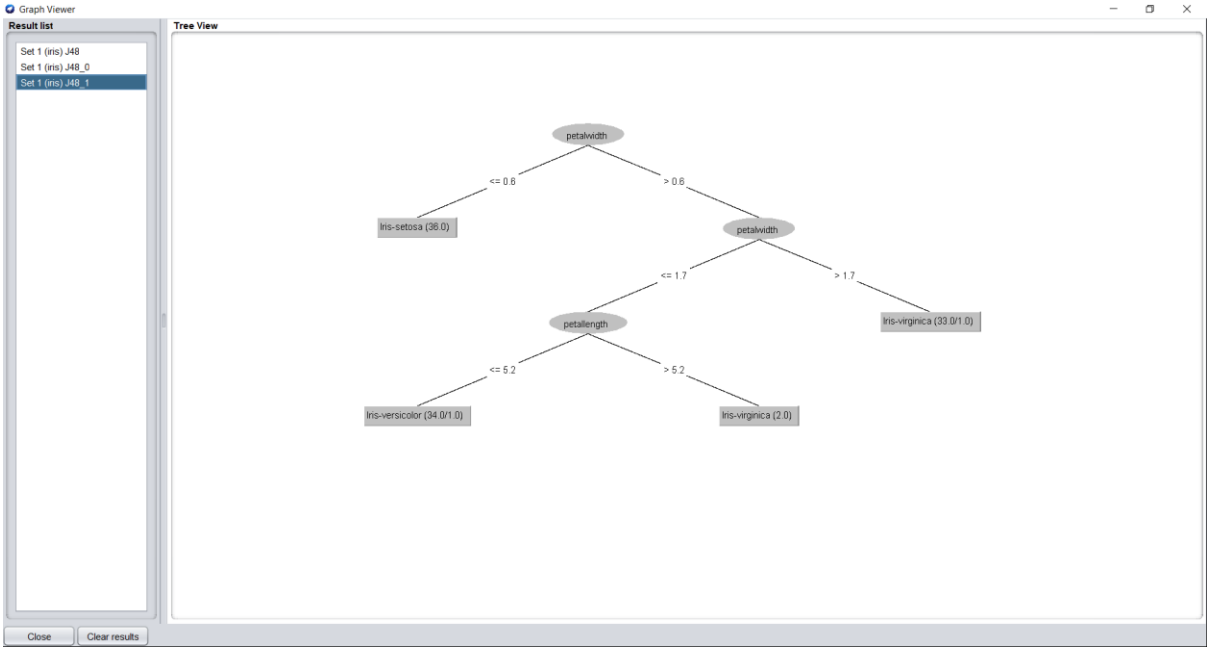

Weka Experiment Environment
Setup
Run
Analyse
Experiment Configuration Mode
Simple
Open...
Save...
New
Results Destination
ARFF file
Filename:
Browse...
Experiment Type
Cross-validation
Number of folds:
10
Classification
Regression
Iteration Contro
Number of repetitions:
10
Data sets first
Algorithms first
Datasets
Add new...
Edit selected...
Delete selected
Use relative paths
C:\Program Files\Weka-3-8-5\data\iris.arff
Up
Down
Algorithms
Add new...
Edit selected...
Delete selected
ZeroR
J48 -C 0.25 -M 2
Load options...
Save options...
Up
Down
Notes

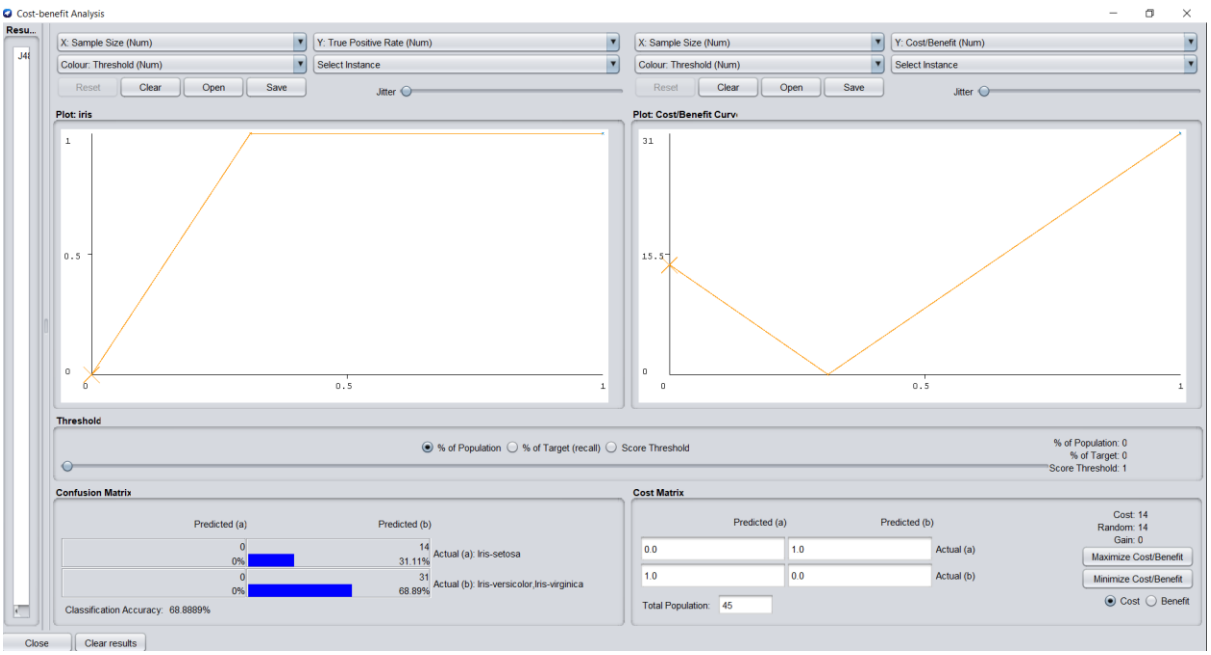

Weka Experiment Environment
Setup
Run
Analyse
Start
Stop
Log
11:36:17: Started
11:36:18: Finished
11:36:18: There were 0 errors
Status
Not running


Weka Experiment Environment
Setup
Run
Analyse
Source
Got 200 results
File...
Database...
Experiment
Actions
Perform test
Save output
Open Explorer...
Configure test
Testing with
Paired T-Tester (corrected)
Select rows and cols
Rows
Cols
Swap
Comparison field
Percent_correct
Significance
0.05
Sorting (asc.) by
<default>
Test base
Select
Displayed Columns
Select
Show std. deviations
Output Format
Select
Result list
11:36:53 - Available resultsets
11:39:16 - Percent_correct - rules.ZeroR '' 48055541465867954
Test output
Tester: weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-pr
Analysing: Percent_correct
Datasets: 1
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by: -
Date: 8/9/21, 11:39 AM
Dataset (1) rules.Ze | (2) trees
iris (100) 33.33 | 94.73 v
(v/ /*) | (1/0/0)
Key:
(1) rules.ZeroR '' 48055541465867954
(2) trees.J48 '-C 0.25 -M 2' -2177331683936444444


Weka Experiment Environment
Setup
Run
Analyse
Source
Got 200 results
File...
Database...
Experiment
Actions
Perform test
Save output
Open Explorer...
Configure test
Testing with
Paired T-Tester (corrected)
Select rows and cols
Rows
Cols
Swap
Comparison field
Area_under_ROC
Significance
0.05
Sorting (asc.) by
<default>
Test base
Select
Displayed Columns
Select
Show std. deviations
Output Format
Select
Result list
11:36:53 - Available resultsets
11:39:16 - Percent_correct - rules.ZeroR '' 48055541465867954
11:42:38 - Percent_correct - Ranking
11:42:52 - Percent_correct - Summary
11:43:10 - Percent_correct - trees.J48 '-C 0.25 -M 2' -21773316
11:43:28 - Percent_correct - rules.ZeroR '' 48055541465867954
11:43:55 - Area_under_ROC - trees.J48 '-C 0.25 -M 2' -217733
Test output
Tester: weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-pr
Analysing: Area_under_ROC
Datasets: 1
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by: -
Date: 8/9/21, 11:43 AM
Dataset (2) trees.J | (1) rule
iris (100) 0.99 | 0.50 *
(v/ /*) | (0/0/1)
Key:
(1) rules.ZeroR '' 48055541465867954
(2) trees.J48 '-C 0.25 -M 2' -2177331683936444444


Weka KnowledgeFlow Environment
Program File Edit Insert View
Data mining processes
Attribute summary
Scatter plot matrix
SQL Viewer
Simple CLI
Design
DataSources
DataSinks
DataGenerators
Filters
Classifiers
Clusterers
Associations
AttSelection
Evaluation
Misc
Visualization
Flow
Tools
ArffLoader
ClassAssigner
TrainTestSplitMaker
J48
GraphViewer
ClassifierPerformanceEvaluator
CostBenefitAnalysis
dataSet
testSet
trainingSet
graph
batchClassifier
thresholdData
Status
Log
Component
Parameters
Time
Status
ArffLoader
ClassAssigner
TrainTestSplitMaker
J48
-C 0.25 -M 2
GraphViewer
ClassifierPerformanceEvaluator
CostBenefitAnalysis
Finished.


Graph Viewer
Result list
Set 1 (iris) J48
Set 1 (iris) J48_0
Set 1 (iris) J48_1
Tree View
petalwidth
<= 0.6
> 0.6
Iris-setosa (36.0)
petalwidth
<= 1.7
> 1.7
petallength
Iris-virginica (33.0/1.0)
<= 5.2
> 5.2
Iris-versicolor (34.0/1.0)
Iris-virginica (2.0)
Close
Clear results


Cost-benefit Analysis
X: Sample Size (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: iris
Y: Cost/Benefit (Num)
Plot: Cost/Benefit Curv
Threshold
% of Population
% of Target (recall)
Score Threshold
% of Population: 0
% of Target: 0
Score Threshold: 1
Confusion Matrix
Predicted (a)
Predicted (b)
Actual (a): Iris-setosa
Actual (b): Iris-versicolor,Iris-virginica
Classification Accuracy: 68.8889%
Cost Matrix
Actual (a)
Actual (b)
Total Population:
Cost: 14
Random: 14
Gain: 0
Maximize Cost/Benefit
Minimize Cost/Benefit
Cost
Benefit
Close
Clear results