About the exam

Høgskolen i Østfold

EXAMINATION

Course Code and name:
ITF301416 Big Data - Processing and Analysis

Date and duration :
13.12.2017, 4 hours

No support materials permitted.

Lecturers:
Cathrine Linnes
Edgar Bostrøm

Examination:
The final exam consists of 4 parts.

- You may give your answer in Norwegian or in English.
- You are able to view only one question at the time.
- You may go back to view each question before you submit your exam.
- You can have the exam question and appendix (vedlegg) open at the same time. You will find the appendix link on this page.

Part 1 (90 minutes)
Please refer to Appendix A while answering these questions.
Question 1-32

- 12 % - 12 True and false questions
- 20 % - 20 Multiple choice questions

Part 2 (30 minutes)
Question 33-34

- 18 % - Consists of 2 essays questions.

Part 3 (60 minutes)
Question 35-39

- 25 % - Consists of 5 essays questions.

Part 4 (60 minutes)
Question 40-43

- 25 % - Consists of 4 essays questions.

Deadline for submitting results: 11.1.2018

Appendix for Part 1

Appendix_A_BigData

Q1

Can data mining be applied in the following area: Image screening, e.g. detecting potential oil slicks in the sea from satellite data?

Select an alternative:

- True
- False

Q2

Can data mining be applied in the following area: Load forecasting, e.g. forecasting the demand for electricity for a particular hour, day, month, and year?

Select an alternative:

- True
- False

Q3

Data mining is about solving problems by analyzing data that is currently not available in the databases.

Select an alternative:

- True
- False

Q4

Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage e.g. an economic advantage.

Select an alternative:

- ○ True
- ○ False

5 **Q5**

The four interfaces in Weka are: Explorer, Experimenter, Knowledge Flow and Workbench?
Select an alternative:

- ○ True
- ○ False

6 **Q6**

ZeroR is the baseline classifier?
Select an alternative:

- ○ True
- ○ False

7 **Q7**

The default percentage split is set to 10%.
Select an alternative:

- ○ True
- ○ False

8 **Q8**

Weka accepts numeric files only.
Select an alternative:

- ○ True
- ○ False

9 **Q9**

A confusion matrix is also known as an error matrix.
Select an alternative:

- ○ True
- ○ False

10 **Q10**

Supervised files are the ones that use a class value for their operation.
Select an alternative:

- ○ True
- ○ False

11 **Q11**

In the visualization tab, the classifier errors are displayed using the symbol "square".
Select an alternative:

- ○ True
- ○ False

12 **Q12**

Looking at the segment-challenge dataset. With a training set percentage of 99% it gives a figure of 100% accuracy on the test set. Does this mean that this generates a perfect classifier?
Select an alternative:

- ○ True
- ○ False

13 **Q13**

How many instances are there in the IRIS data set?

Select an alternative:

- ○ 200
- ○ 150
- ○ 250
- ○ 100

14 Q14

How many attributes are there?
Select an alternative:

- ○ 10
- ○ 7
- ○ 4
- ○ 5

15 Q15

How many possible values does the class attributes have?
Select an alternative:

- ○ 50
- ○ 2
- ○ 3
- ○ 1

16 Q16

Weka data files are___files.
Select an alternative:

- ○ ARFF
- ○ CSV
- ○ ARC
- ○ DOC

17 **Q17**

Examine the IRIS file (header) and say when the dataset was first used by looking at the NotePad handout?

Select an alternative:

- ○ 1988
- ○ 1936
- ○ 1973
- ○ 1980

18 **Q18**

Choose the J48 tree classifier, and run it (with default parameters). How many instances are misclassified?

Select an alternative:

- ○ 2
- ○ 6
- ○ 4
- ○ 1

19 **Q19**

Now switch the classifier to Simple Logistic, which you will find in the functions category, and run it (with default parameters). How many instances are misclassified now?

Select an alternative:

- ○ 9
- ○ 6
- ○ 12
- ○ 3
- ○ 15

20 **Q20**

If you did the above experiment with 10 different random seeds rather than 5, how would you expect this to affect the mean and standard deviation?

Select an alternative:

- The mean would be about the same and the standard deviation would be a little smaller.
- Both the mean and standard deviation would be a bit smaller.
- The mean would be a bit bigger but the standard deviation would be about the same.
- They would both stay about the same

21 **Q21**

What should be the first row of a Comma Separated Values (.csv) format file that contains the nominal Weather data?

Select an alternative:

- Sunny, hot, high, TRUE, no
- Outlook, temperature, humidity, windy, play
- Sunny, hot, high, FALSE, no
- Rainy, mild, high, TRUE, no

22 **Q22**

The problem of finding hidden structure in unlabeled data is called

Select an alternative:

- None of the above
- Reinforcement learning
- Unsupervised learning
- Supervised learning

23 **Q23**

You are given data about seismic activity in Japan, and you want to predict a magnitude of the next earthquake, this is an example of

Select an alternative:

- Supervised learning
- All of the above
- Unsupervised learning
- Reduction learning

24 **Q24**

Algorithm is
Select an alternative:

- Computational procedure that takes some value as input and produces some value as output.
- Science of making machines perform tasks that would require intelligence when performed by humans.
- None of the above
- It uses machine-learning techniques. Here program can learn from past experience and adapt themselves to new situations.

25 **Q25**

Classification is
Select an alternative:

- The task of assigning a classification to a set of examples
- None of the above
- A subdivision of a set of examples into a number of classes
- A measure of the accuracy, of the classification of a concept that is given by a certain theory.

26 **Q26**

Self-organizing maps (type of artificial neural network) are an example of
Select an alternative:

- Missing data inputation
- Reinforcement learning
- Unsupervised learning
- Supervised learning

27 **Q27**

Bayesian classifiers is

Select an alternative:

- None of the above.
- A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.
- Any mechanism employed by learning system to constrain the search space of a hypothesis
- An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.

28 **Q28**

One of the most frequently used data mining algorithms is
Select an alternative:

- Evolution analysis
- Cluster analysis
- Outlier analysis
- Decision trees

29 **Q29**

Which of the following is not a data mining method?
Select an alternative:

- Classification and prediction
- A priori algorithms
- Dependency analysis
- Data description

30 **Q30**

Look at the glass dataset, go to the Classify panel, choose the J48 tree classifier, and run it (with default parameters). Using the confusion matrix to determine how many headlamps instances were misclassified as build wind float?

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 50 15  3  0  0  1  1 |  a = build wind float
 16 47  6  0  2  3  2 |  b = build wind non-float
  5  5  6  0  0  1  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  2  0  0 10  0  1 |  e = containers
  1  1  0  0  0  7  0 |  f = tableware
  3  2  0  0  0  1 23 |  g = headlamps
```

Select an alternative:

- ○ 7
- ○ 6
- ○ 3
- ○ 1
- ○ 2

31 # Q31

Which of the following statements seems to be correct, based on your experiments?
Select an alternative:

- ○ The more test data, the greater the classifier's success rate.
- ○ The more training data, the greater the classifier's success rate.
- ○ Training set size has no influence on the classifier's success rate.

32 # Q32

When the percentage split option is used for evaluation, how good is the performance if (a) almost none of the data is used for testing; (b) almost all of the data is used for testing?
Select an alternative:

- ○ (a) has better performance than (b).
- ○ The performance for (a) and (b) are similar.
- ○ has poor performance, (b) has good performance.

i # Part 2

Part 2 - (30 minutes)

Consists of 2 essay question (Q33-Q34).

33 **Q33**

When a data set is loaded, IBM® Watson™ Analytics reads the data and assesses it for data quality. If the data quality score is low, you can improve its quality and usage so that your predictions, explorations, and views are more accurate. In class we worked on the following dataset, in a paragraph format what does a quality score of 63% mean?

IBM_HR_Training_2014-17

Sep 27, 2017 10:32 PM

63% Quality

Fill in your answer here

Format | B I U x_2 x^2 | I_x | | | | | | | | | | | | Σ |

Words: 0

34 **Q34**

In a paragraph, explain the term **cross validation**

Fill in your answer here

Format | B I U x_2 x^2 | I_x | ...

Words: 0

35 **Q35-39**

Part 3 - (60 minutes)

The questions to be answered in relational algebra uses the three relations below (from an earlier exercise). Use latin letters (in Inspera) or write the answers on a separate sheet. (Link to syntax found as .PDF file).

STUDENT(Stud#, StudName, Adress,..,PostalCode) **COURSE**(Course#, CourseName,.., Level)

EXAM (Stud#, Course#,Trial#, Date)

a. List Stud#, StudName for students who have taken Course# = 2110.

b. List all the information about students who have taken a course named «The Theory of Everything», preferably using semijoin to simplify the statement.

c. List Stud#, StudName for students who have NOT taken Course# = 2110.

d. List Stud#, StudName for students having taken all courses given at Level 3.

e. Explain some basic principles for optimizing a query.

Relational algebra – useful operators.

Set operators	*Notation, variant-1*	*Notation without special characters*
Union	$R \cup S$	R union S
Intersect / Snitt	$R \cap S$	R intersect S
Set difference / Mengdedifferanse	R - S, $R \setminus S$	R difference S
Set product / Mengdeprodukt, cartesian product ("all joined with all")	$R \times S$	R product S R times S
Relational operators		
Restriction[1]/Horizontal selection (sigma)	$\sigma_{<cond.>}(R)$	SIGMA $_{<cond.>}$ (R)
Projection/Vertical selection (pi)	$\pi_{<attribute\ list>}(R)$	PI$_{<attribute\ list>}$ (R)
Set division. (Given R[c,d] and S[d]. c is a member of the set R divided by S if c in R is found with all d-s in S.)	$R \div S$ R / S	R divideby S or R div S
Different types of product		
θ-join (product with some kind of condition on compatible attributes, e.g. >, <, =, !=, incl. combinations.)	$R \bowtie_{\theta <cond.>} S$	R THETAJOIN$_{<condition>}$ S
Equi-join (the θ-operation is =)	$R \bowtie_{=<cond.>} S$	R EQUIJOIN$_{<condition>}$ S
Natural join (equi-join where duplicate attributes are removed) ** most common join type **	$R \bowtie_{<cond.>} S$	R NATURALJOIN$_{<condition>}$ S, or just R JOIN$_{<condition>}$ S
Varieties for product[2]		
Outer join, normally left (all in R, plus all in S complying the join condition)	$R \overset{\supset}{\bowtie} S$	R LEFTOUTERJOIN$_{<cond.>}$ S
Right outer join (all in S, plus all in R complying the join condition)	$R \overset{\subset}{\bowtie} S$	R RIGHTOUTERJOIN$_{<cond.>}$ S
Full join (all in R, all in S, plus all complying with the join condition)	$R \overset{\supset\subset}{\bowtie} S$	R FULLJOIN$_{<cond.>}$ S
Semijoin (those in R complying with R join$_{<condition>}$ S)	$R \rhd_{<cond.>} S$[3]	R SEMIJOIN$_{<cond.>}$ S

Note: the operations are set-based, i.e. duplicates are deleted – corresponding to select distinct i SQL. Relational algebra is **set-based**, while SQL is **bag-based.**

[1] Originally, restriction was called selection (thereby sigma - σ). However, this makes it easy to be confused with SQL-seletion, whick in fact is a projection.

[2] If joining on primary/foreign key combination, we may drop <cond>.

[3] Some use this symbol for antijoin (not in), and use ⋉ for semijoin.

Fill in your answer here

Format

Words: 0

36 **Q40-43**

Part 4 - (60 minutes)

a. What is a distributed database system?

b. Describe different principles of fragmentations as well as different kinds of fragmentation/replication.

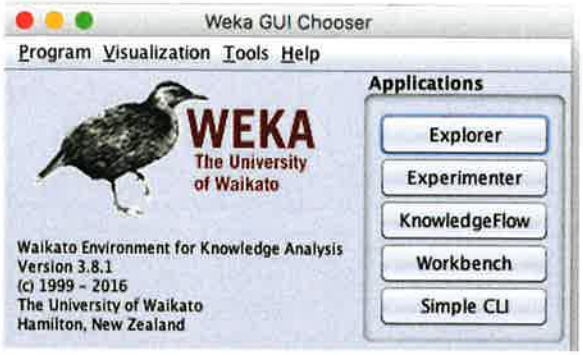
c. Describe the CAP theorem, and how this relates to relational DMBS and to «no-SQL» systems.

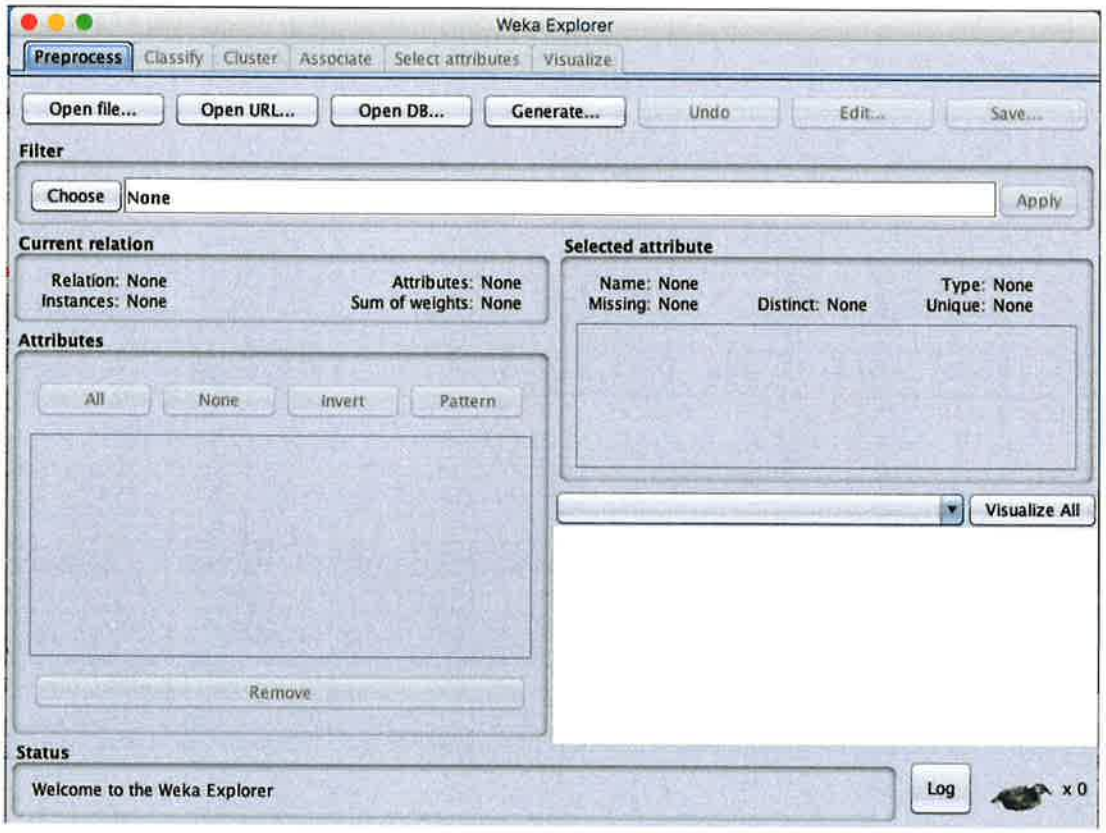
d. One of the most well-known no-SQL system is called Hadoop. Describe how this system works, where Hadoop is particulary useful, as well as what types of problems where it doesn't suit well.

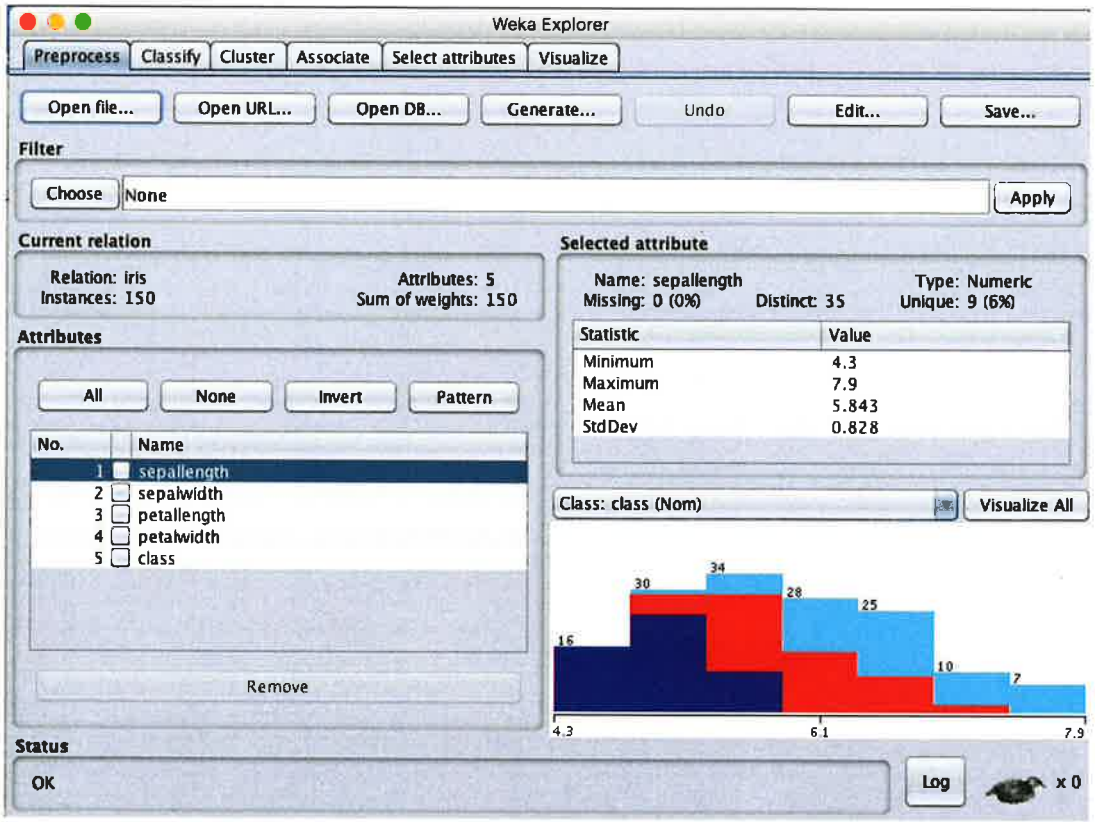
Big Data Final Exam

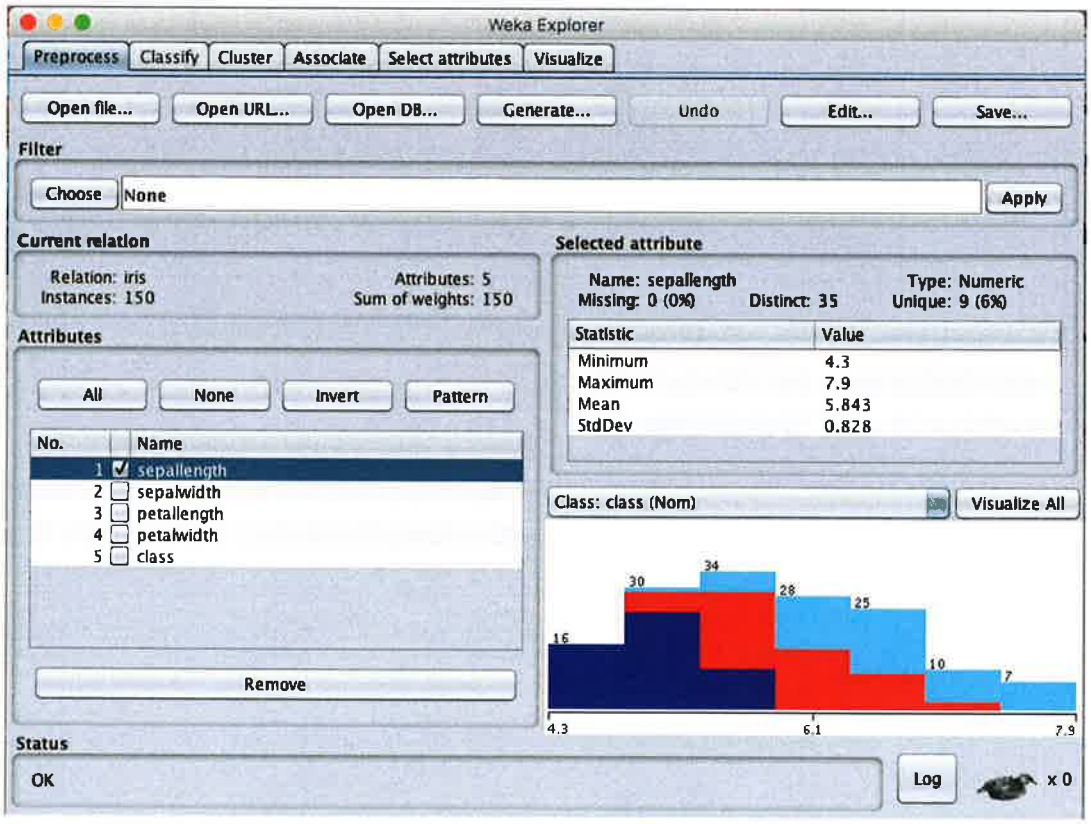
Appendix A

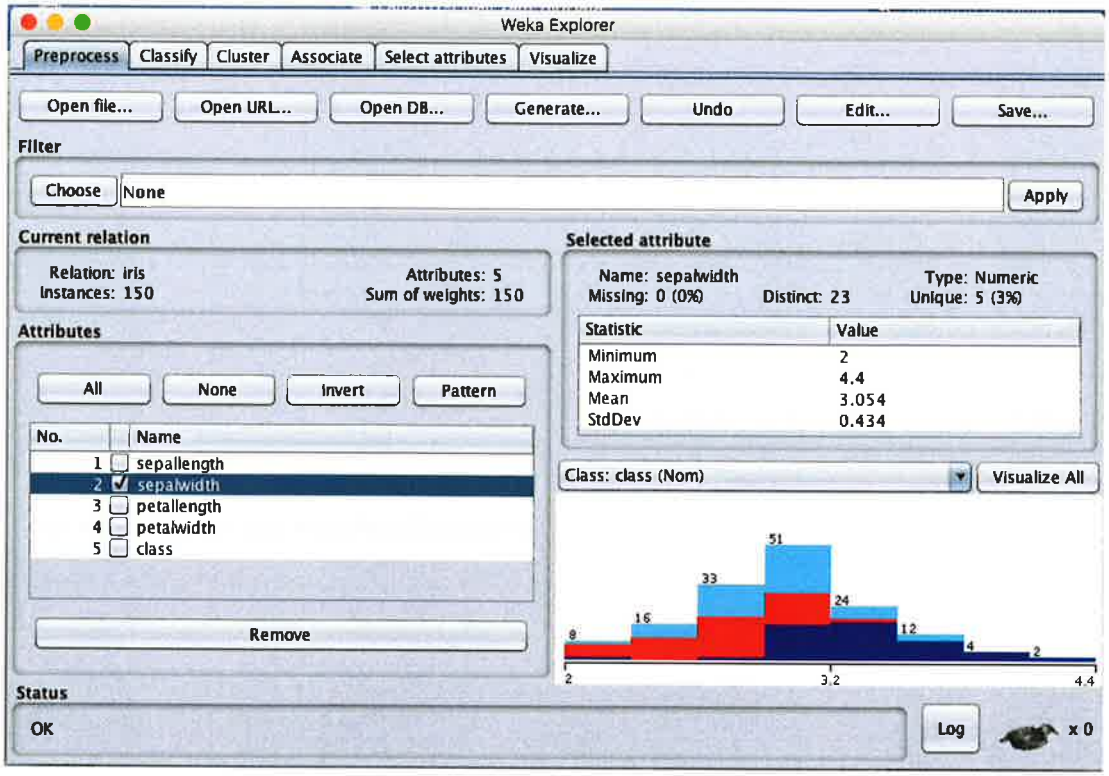
Appendix A: IRIS dataset

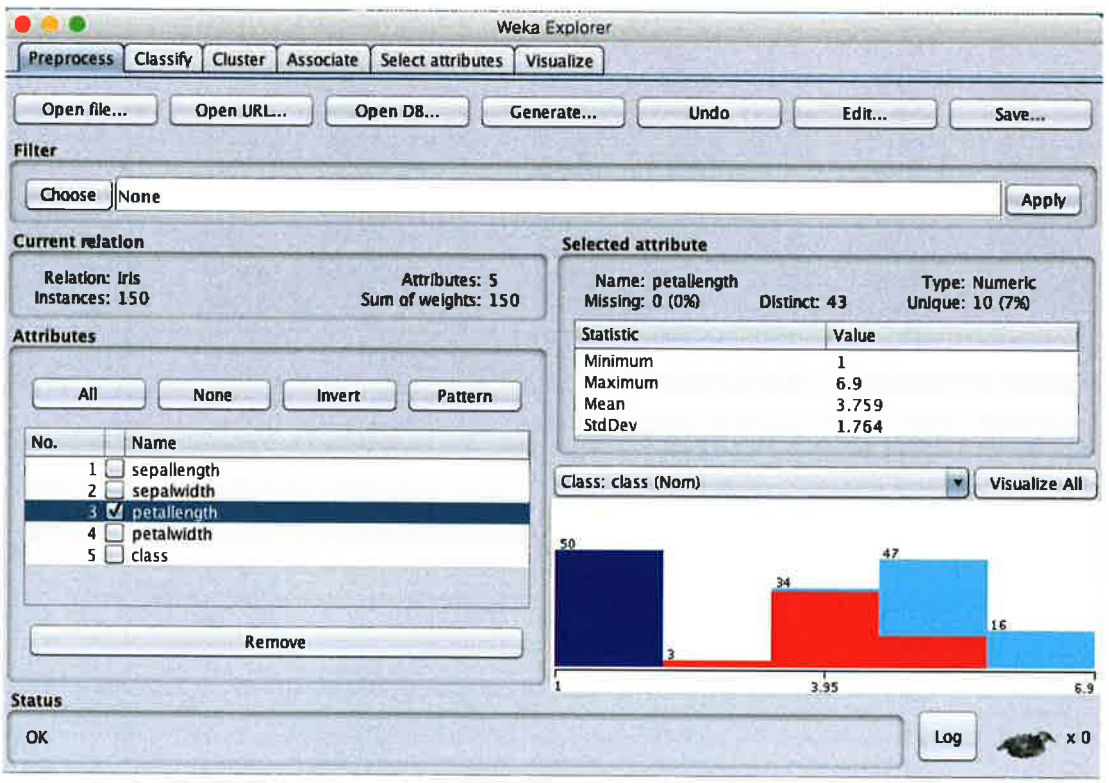


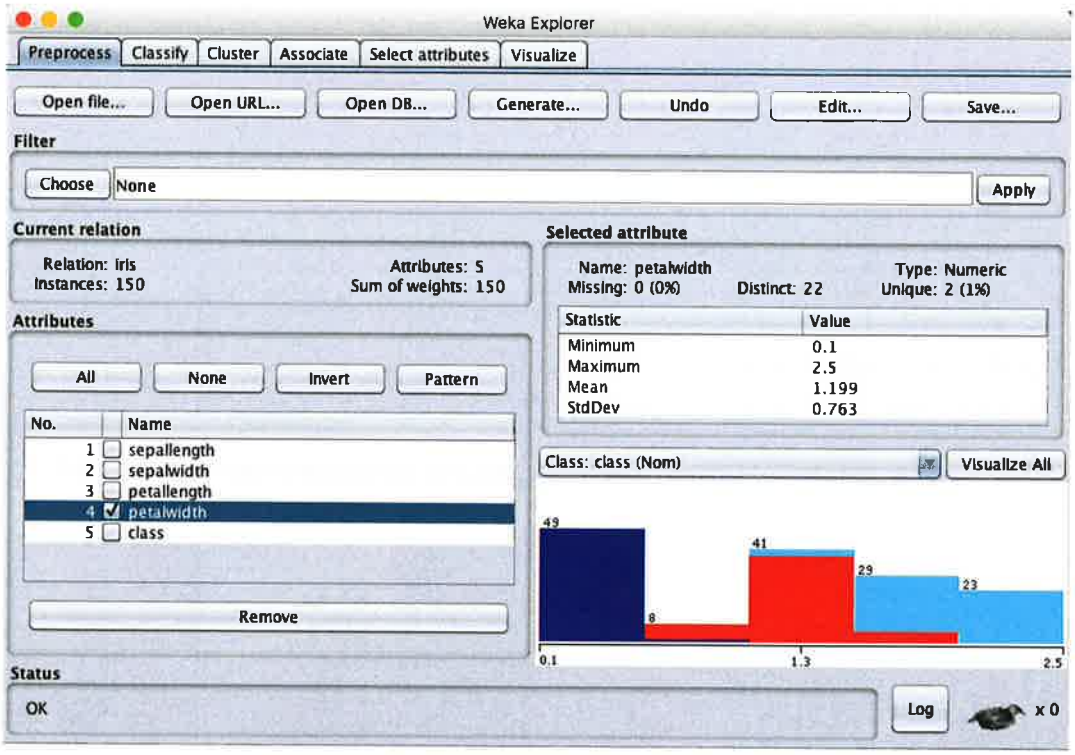


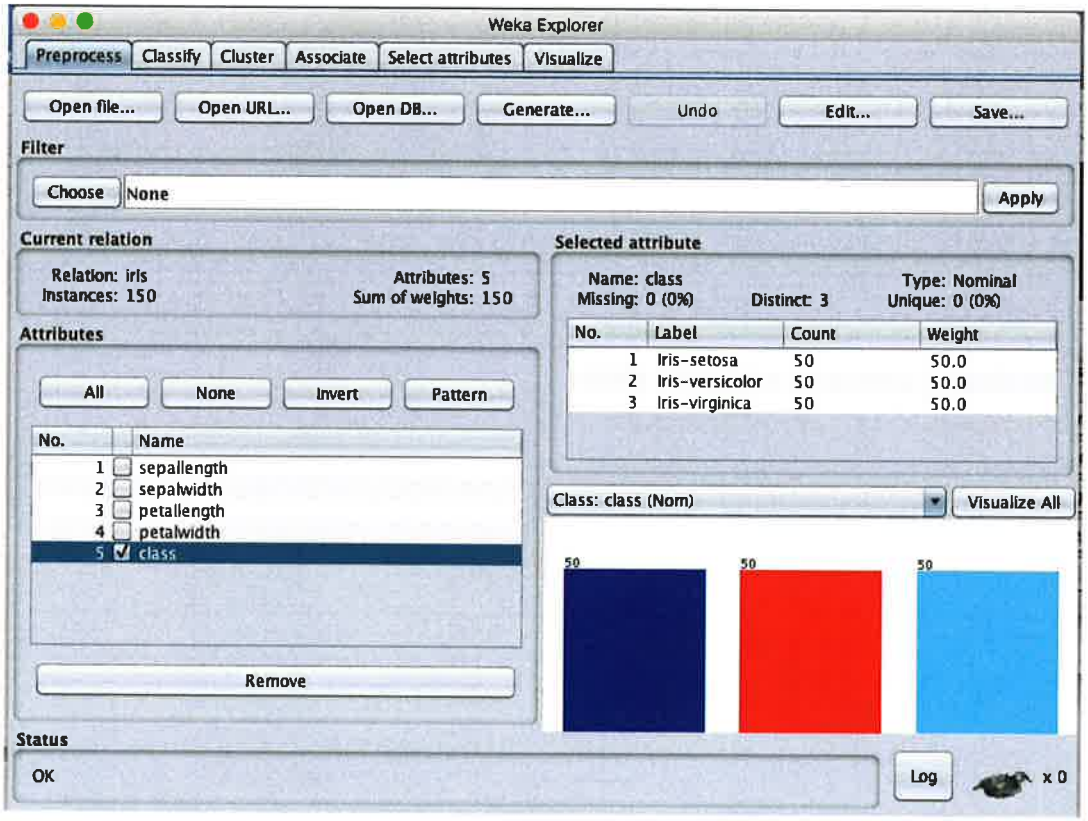

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Selected attribute
Name: sepallength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)
Statistic
Value
Minimum
4.3
Maximum
7.9
Mean
5.843
StdDev
0.828
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Class: class (Nom)
Visualize All
34
30
28
25
16
10
7
4.3
6.1
7.9
Status
OK
Log
x 0

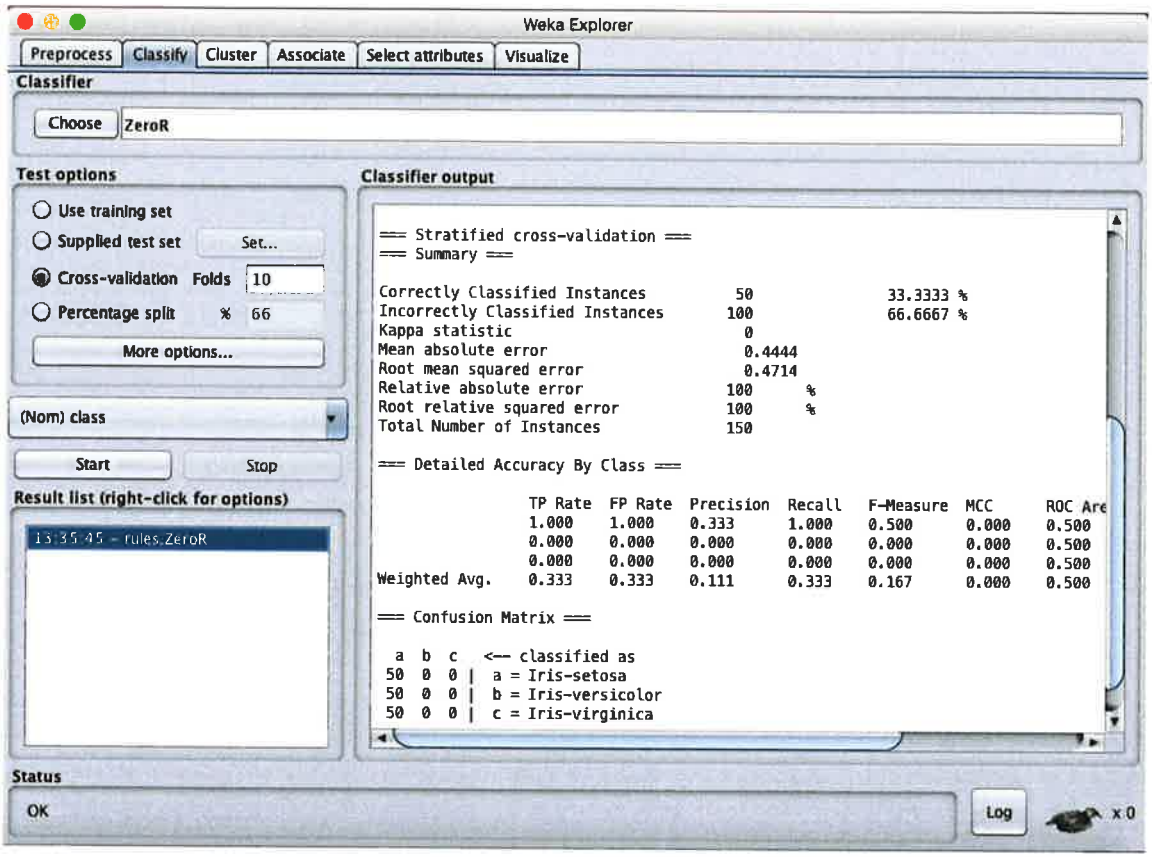

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Selected attribute
Name: sepallength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)
Statistic
Value
Minimum
4.3
Maximum
7.9
Mean
5.843
StdDev
0.828
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Class: class (Nom)
Visualize All
34
30
28
25
16
10
7
4.3
6.1
7.9
Status
OK
Log
x 0

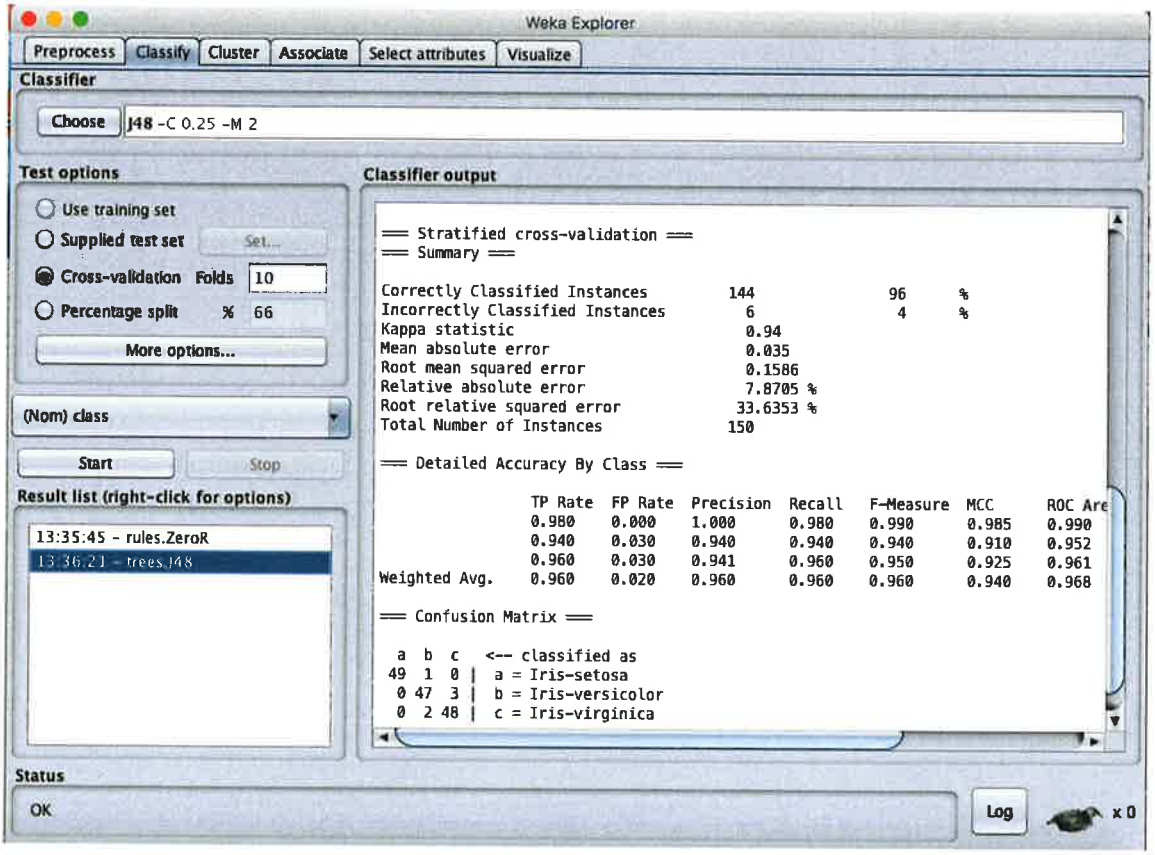

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Selected attribute
Name: sepalwidth
Missing: 0 (0%)
Distinct: 23
Type: Numeric
Unique: 5 (3%)
Statistic
Value
Minimum
2
Maximum
4.4
Mean
3.054
StdDev
0.434
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Class: class (Nom)
Visualize All
51
33
24
16
12
8
4
2
2
3.2
4.4
Status
OK
Log
x 0

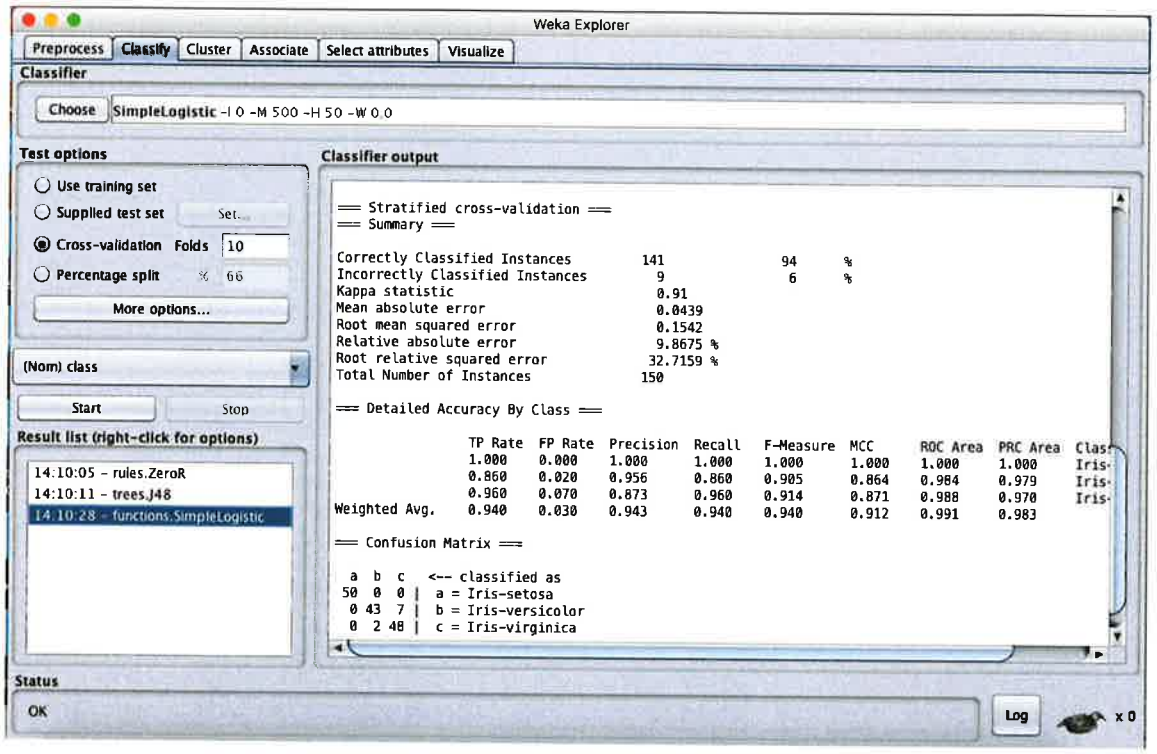

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Selected attribute
Name: petallength
Missing: 0 (0%)
Distinct: 43
Type: Numeric
Unique: 10 (7%)
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Class: class (Nom)
Visualize All
50
47
34
16
3
1
3.95
6.9
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Selected attribute
Name: petalwidth
Missing: 0 (0%)
Distinct: 22
Type: Numeric
Unique: 2 (1%)
Statistic
Value
Minimum
0.1
Maximum
2.5
Mean
1.199
StdDev
0.763
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Class: class (Nom)
Visualize All
49
8
41
29
23
0.1
1.3
2.5
Remove
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Selected attribute
Name: class
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 Iris-setosa 50 50.0
2 Iris-versicolor 50 50.0
3 Iris-virginica 50 50.0
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Class: class (Nom)
Visualize All
50
50
50
Remove
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)
13:35:45 - rules.ZeroR
Classifier output
=== 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 Are
1.000 1.000 0.333 1.000 0.500 0.000 0.500
0.000 0.000 0.000 0.000 0.000 0.000 0.500
0.000 0.000 0.000 0.000 0.000 0.000 0.500
Weighted Avg. 0.333 0.333 0.111 0.333 0.167 0.000 0.500
=== 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)
13:35:45 - rules.ZeroR
13:36:21 - trees.J48
Classifier output
=== 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 Are
0.980 0.000 1.000 0.980 0.990 0.985 0.990
0.940 0.030 0.940 0.940 0.940 0.910 0.952
0.960 0.030 0.941 0.960 0.950 0.925 0.961
Weighted Avg. 0.960 0.020 0.960 0.960 0.960 0.940 0.968
=== 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



The IRIS file has been opened in NotePad.

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
% 3. Past Usage:
%    - Publications: too many to mention!!!  Here are a few.
%    1. Fisher,R.A. "The use of multiple measurements in taxonomic problems"
%       Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions
%       to Mathematical Statistics" (John Wiley, NY, 1950).
%    2. Duda,R.O., & Hart,P.E. (1973) Pattern Classification and Scene Analysis.
%       (Q327.D83) John Wiley & Sons.  ISBN 0-471-22361-1.  See page 218.
%    3. Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New System
%       Structure and Classification Rule for Recognition in Partially Exposed
%       Environments".  IEEE Transactions on Pattern Analysis and Machine
%       Intelligence, Vol. PAMI-2, No. 1, 67-71.
%       -- Results:
%          -- very low misclassification rates (0% for the setosa class)
%    4. Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule".  IEEE
%       Transactions on Information Theory, May 1972, 431-433.
%       -- Results:
%          -- very low misclassification rates again
%    5. See also: 1988 MLC Proceedings, 54-64.  Cheeseman et al's AUTOCLASS II
%       conceptual clustering system finds 3 classes in the data.
%
% 4. Relevant Information:
%    --- This is perhaps the best known database to be found in the pattern
```

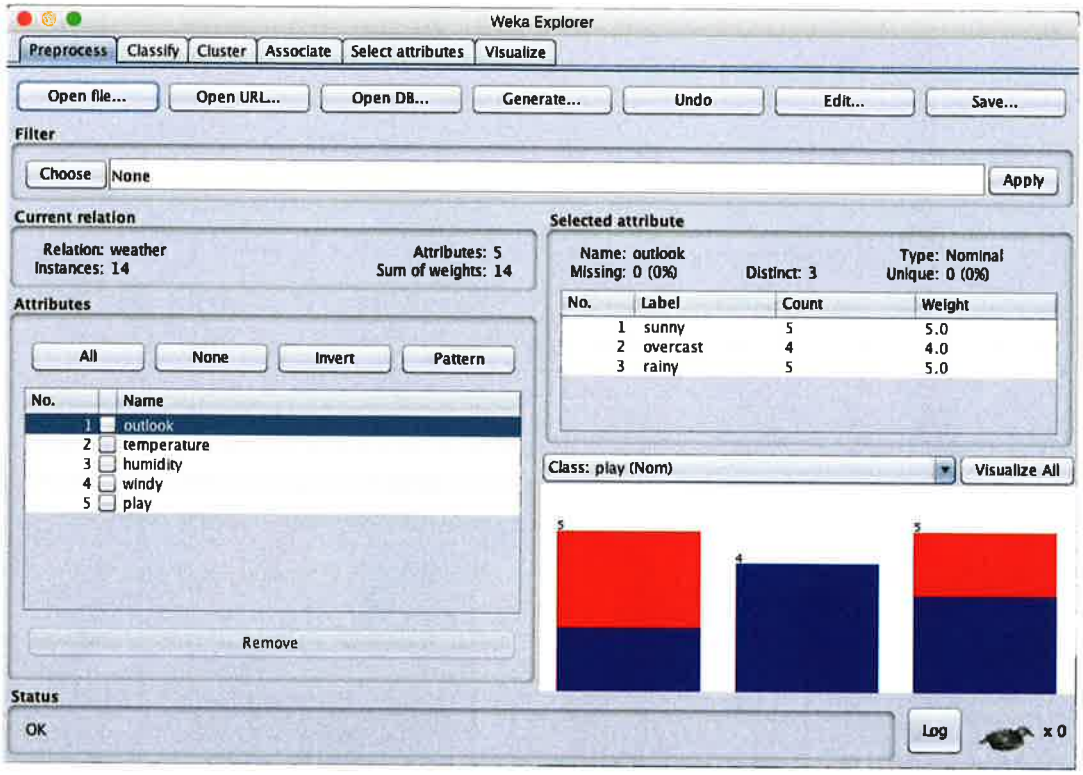
```
%       recognition literature.  Fisher's paper is a classic in the field
%       and is referenced frequently to this day.  (See Duda & Hart, for
%       example.)  The data set contains 3 classes of 50 instances each,
%       where each class refers to a type of iris plant.  One class is
%       linearly separable from the other 2; the latter are NOT linearly
%       separable from each other.
%    --- Predicted attribute: class of iris plant.
%    --- This is an exceedingly simple domain.
%
% 5. Number of Instances: 150 (50 in each of three classes)
%
% 6. Number of Attributes: 4 numeric, predictive attributes and the class
%
% 7. Attribute Information:
%    1. sepal length in cm
%    2. sepal width in cm
%    3. petal length in cm
%    4. petal width in cm
%    5. class:
%       -- Iris Setosa
%       -- Iris Versicolour
%       -- Iris Virginica
%
% 8. Missing Attribute Values: None
%
% Summary Statistics:
%                Min  Max   Mean    SD   Class Correlation
%    sepal length: 4.3  7.9   5.84  0.83    0.7826
%     sepal width: 2.0  4.4   3.05  0.43   -0.4194
%    petal length: 1.0  6.9   3.76  1.76    0.9490  (high!)
%     petal width: 0.1  2.5   1.20  0.76    0.9565  (high!)
%
% 9. Class Distribution: 33.3% for each of 3 classes.

@RELATION iris

@ATTRIBUTE sepallength       REAL
@ATTRIBUTE sepalwidth        REAL
@ATTRIBUTE petallength       REAL
@ATTRIBUTE petalwidth  REAL
@ATTRIBUTE class       {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
```

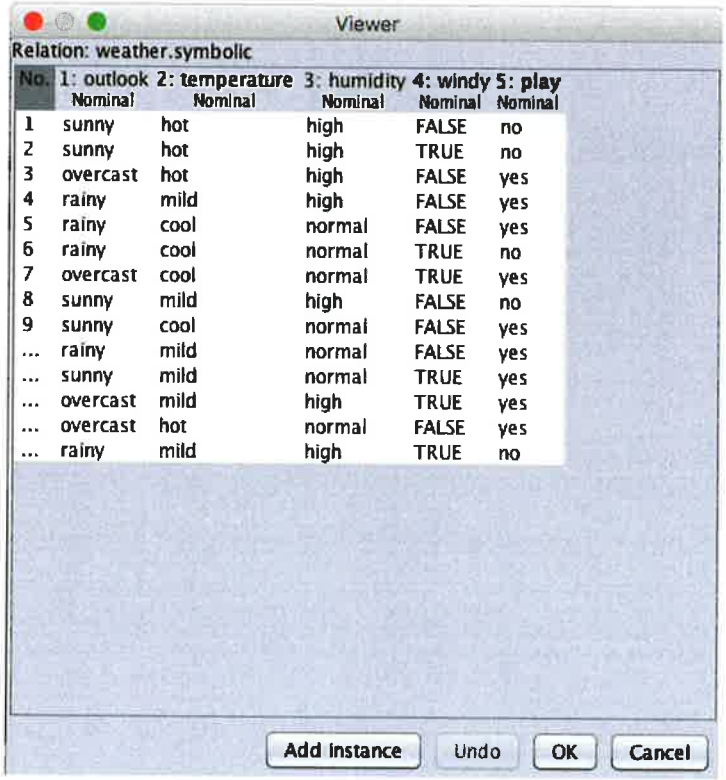
```
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
5.0,3.4,1.6,0.4,Iris-setosa
5.2,3.5,1.5,0.2,Iris-setosa
5.2,3.4,1.4,0.2,Iris-setosa
4.7,3.2,1.6,0.2,Iris-setosa
4.8,3.1,1.6,0.2,Iris-setosa
5.4,3.4,1.5,0.4,Iris-setosa
5.2,4.1,1.5,0.1,Iris-setosa
5.5,4.2,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.0,3.2,1.2,0.2,Iris-setosa
5.5,3.5,1.3,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
4.4,3.0,1.3,0.2,Iris-setosa
5.1,3.4,1.5,0.2,Iris-setosa
5.0,3.5,1.3,0.3,Iris-setosa
4.5,2.3,1.3,0.3,Iris-setosa
4.4,3.2,1.3,0.2,Iris-setosa
5.0,3.5,1.6,0.6,Iris-setosa
5.1,3.8,1.9,0.4,Iris-setosa
4.8,3.0,1.4,0.3,Iris-setosa
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
6.3,3.3,4.7,1.6,Iris-versicolor
4.9,2.4,3.3,1.0,Iris-versicolor
6.6,2.9,4.6,1.3,Iris-versicolor
5.2,2.7,3.9,1.4,Iris-versicolor
5.0,2.0,3.5,1.0,Iris-versicolor
```

```
5.9,3.0,4.2,1.5,Iris-versicolor
6.0,2.2,4.0,1.0,Iris-versicolor
6.1,2.9,4.7,1.4,Iris-versicolor
5.6,2.9,3.6,1.3,Iris-versicolor
6.7,3.1,4.4,1.4,Iris-versicolor
5.6,3.0,4.5,1.5,Iris-versicolor
5.8,2.7,4.1,1.0,Iris-versicolor
6.2,2.2,4.5,1.5,Iris-versicolor
5.6,2.5,3.9,1.1,Iris-versicolor
5.9,3.2,4.8,1.8,Iris-versicolor
6.1,2.8,4.0,1.3,Iris-versicolor
6.3,2.5,4.9,1.5,Iris-versicolor
6.1,2.8,4.7,1.2,Iris-versicolor
6.4,2.9,4.3,1.3,Iris-versicolor
6.6,3.0,4.4,1.4,Iris-versicolor
6.8,2.8,4.8,1.4,Iris-versicolor
6.7,3.0,5.0,1.7,Iris-versicolor
6.0,2.9,4.5,1.5,Iris-versicolor
5.7,2.6,3.5,1.0,Iris-versicolor
5.5,2.4,3.8,1.1,Iris-versicolor
5.5,2.4,3.7,1.0,Iris-versicolor
5.8,2.7,3.9,1.2,Iris-versicolor
6.0,2.7,5.1,1.6,Iris-versicolor
5.4,3.0,4.5,1.5,Iris-versicolor
6.0,3.4,4.5,1.6,Iris-versicolor
6.7,3.1,4.7,1.5,Iris-versicolor
6.3,2.3,4.4,1.3,Iris-versicolor
5.6,3.0,4.1,1.3,Iris-versicolor
5.5,2.5,4.0,1.3,Iris-versicolor
5.5,2.6,4.4,1.2,Iris-versicolor
6.1,3.0,4.6,1.4,Iris-versicolor
5.8,2.6,4.0,1.2,Iris-versicolor
5.0,2.3,3.3,1.0,Iris-versicolor
5.6,2.7,4.2,1.3,Iris-versicolor
5.7,3.0,4.2,1.2,Iris-versicolor
5.7,2.9,4.2,1.3,Iris-versicolor
6.2,2.9,4.3,1.3,Iris-versicolor
5.1,2.5,3.0,1.1,Iris-versicolor
5.7,2.8,4.1,1.3,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
6.3,2.9,5.6,1.8,Iris-virginica
6.5,3.0,5.8,2.2,Iris-virginica
7.6,3.0,6.6,2.1,Iris-virginica
4.9,2.5,4.5,1.7,Iris-virginica
7.3,2.9,6.3,1.8,Iris-virginica
6.7,2.5,5.8,1.8,Iris-virginica
7.2,3.6,6.1,2.5,Iris-virginica
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
```

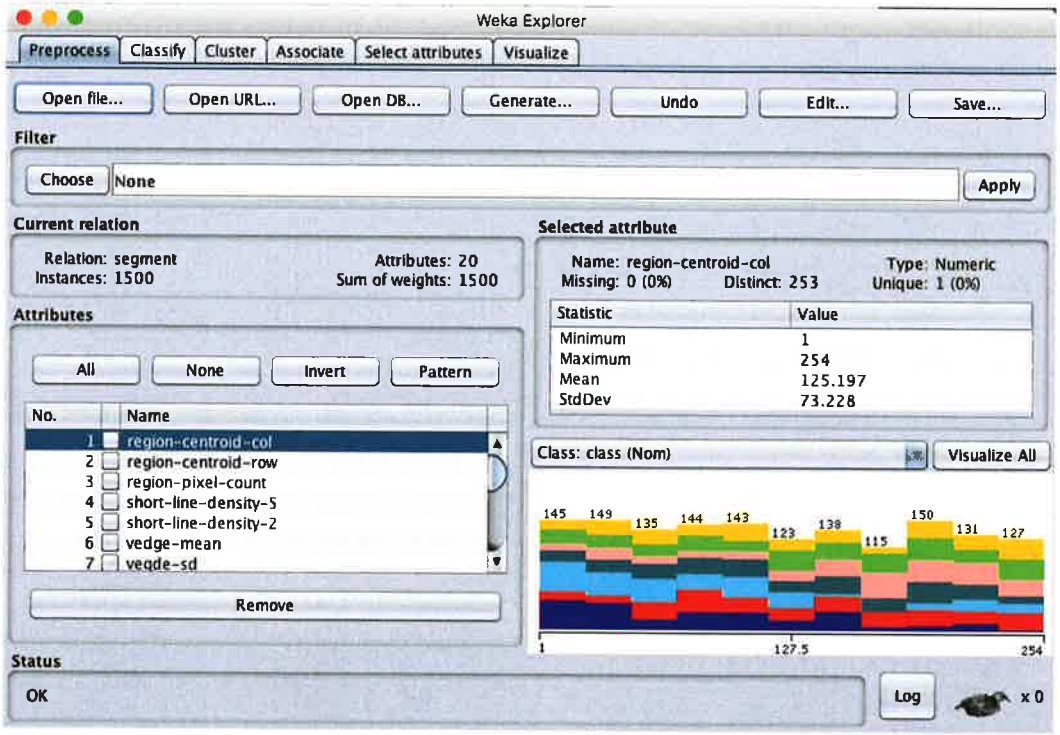
```
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
6.4,3.2,5.3,2.3,Iris-virginica
6.5,3.0,5.5,1.8,Iris-virginica
7.7,3.8,6.7,2.2,Iris-virginica
7.7,2.6,6.9,2.3,Iris-virginica
6.0,2.2,5.0,1.5,Iris-virginica
6.9,3.2,5.7,2.3,Iris-virginica
5.6,2.8,4.9,2.0,Iris-virginica
7.7,2.8,6.7,2.0,Iris-virginica
6.3,2.7,4.9,1.8,Iris-virginica
6.7,3.3,5.7,2.1,Iris-virginica
7.2,3.2,6.0,1.8,Iris-virginica
6.2,2.8,4.8,1.8,Iris-virginica
6.1,3.0,4.9,1.8,Iris-virginica
6.4,2.8,5.6,2.1,Iris-virginica
7.2,3.0,5.8,1.6,Iris-virginica
7.4,2.8,6.1,1.9,Iris-virginica
7.9,3.8,6.4,2.0,Iris-virginica
6.4,2.8,5.6,2.2,Iris-virginica
6.3,2.8,5.1,1.5,Iris-virginica
6.1,2.6,5.6,1.4,Iris-virginica
7.7,3.0,6.1,2.3,Iris-virginica
6.3,3.4,5.6,2.4,Iris-virginica
6.4,3.1,5.5,1.8,Iris-virginica
6.0,3.0,4.8,1.8,Iris-virginica
6.9,3.1,5.4,2.1,Iris-virginica
6.7,3.1,5.6,2.4,Iris-virginica
6.9,3.1,5.1,2.3,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
6.8,3.2,5.9,2.3,Iris-virginica
6.7,3.3,5.7,2.5,Iris-virginica
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
%
%
%
```

Appendix A: Weather dataset

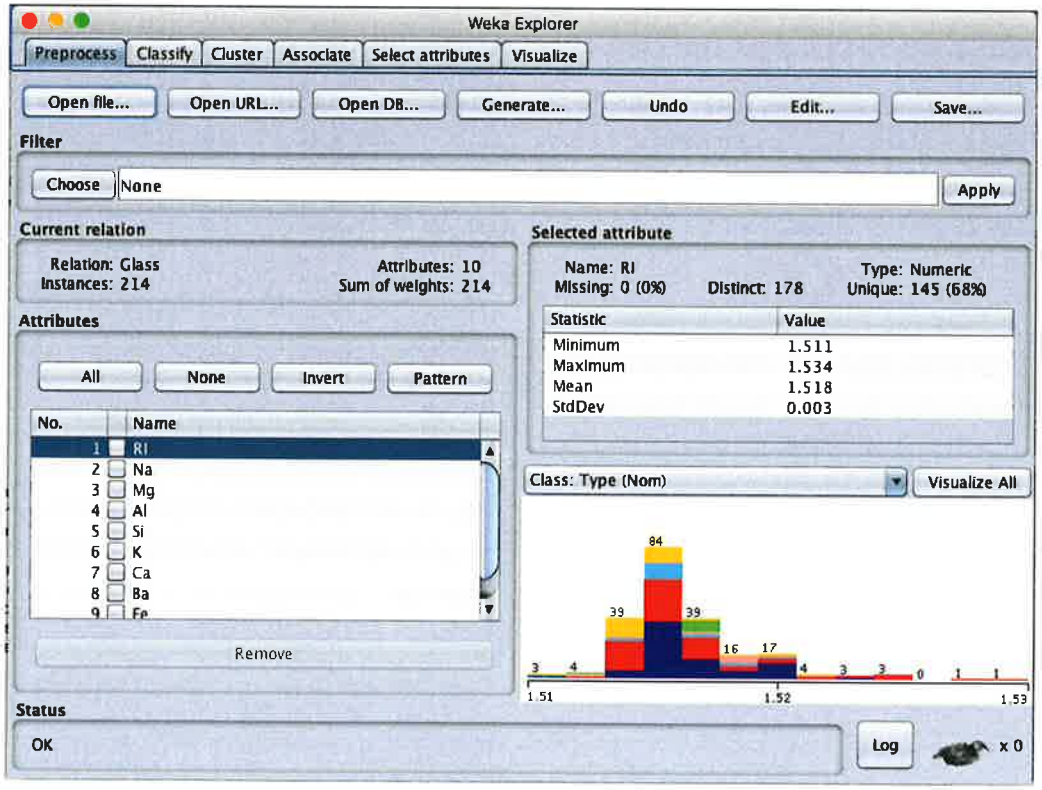


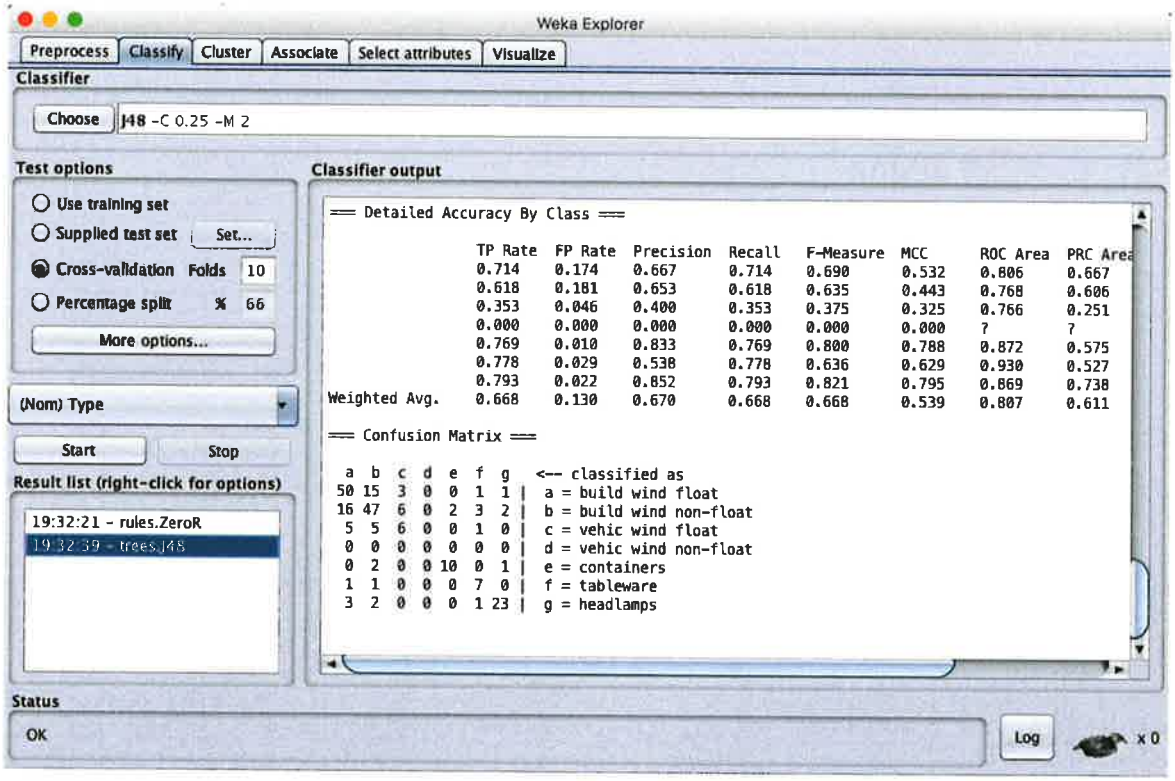


Appendix A: Segment.challenge dataset



Appendix A: glass dataset





Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Glass
Instances:    214
Attributes:   10
              RI
              Na
              Mg
              Al
              Si
              K
              Ca
              Ba
              Fe
              Type
Test mode:    10-fold cross-validation

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

J48 pruned tree
------------------

Ba <= 0.27
|   Mg <= 2.41
|   |   K <= 0.03
|   |   |   Na <= 13.75: build wind non-float (3.0)
|   |   |   Na > 13.75: tableware (9.0)
|   |   K > 0.03
|   |   |   Na <= 13.49
|   |   |   |   RI <= 1.5241: containers (13.0/1.0)
|   |   |   |   RI > 1.5241: build wind non-float (3.0)
|   |   |   Na > 13.49: build wind non-float (7.0/1.0)
|   Mg > 2.41
|   |   Al <= 1.41
|   |   |   RI <= 1.51707
|   |   |   |   RI <= 1.51596: build wind float (3.0)
|   |   |   |   RI > 1.51596
|   |   |   |   |   Fe <= 0.12
|   |   |   |   |   |   Mg <= 3.54: vehic wind float (5.0)
```

```
|   |   |   |   |   |   Mg > 3.54
|   |   |   |   |   |   |   RI <= 1.51667: build wind non-float (2.0)
|   |   |   |   |   |   |   RI > 1.51667: vehic wind float (2.0)
|   |   |   |   |   |  Fe > 0.12: build wind non-float (2.0)
|   |   |  RI > 1.51707
|   |   |   |  K <= 0.23
|   |   |   |   |  Mg <= 3.34: build wind non-float (2.0)
|   |   |   |   |  Mg > 3.34
|   |   |   |   |   |  Si <= 72.64
|   |   |   |   |   |   |  Na <= 14.01: build wind float (14.0)
|   |   |   |   |   |   |  Na > 14.01
|   |   |   |   |   |   |   |  RI <= 1.52211
|   |   |   |   |   |   |   |   |  Na <= 14.32: vehic wind float (3.0)
|   |   |   |   |   |   |   |   |  Na > 14.32: build wind float (2.0)
|   |   |   |   |   |   |   |  RI > 1.52211: build wind float (3.0)
|   |   |   |   |   |   |  Si > 72.64: vehic wind float (3.0)
|   |   |   |   |  K > 0.23
|   |   |   |   |   |  Mg <= 3.75
|   |   |   |   |   |   |  Fe <= 0.14
|   |   |   |   |   |   |   |  RI <= 1.52043: build wind float (36.0)
|   |   |   |   |   |   |   |  RI > 1.52043: build wind non-float (2.0/1.0)
|   |   |   |   |   |   |  Fe > 0.14
|   |   |   |   |   |   |   |  Al <= 1.17: build wind non-float (5.0)
|   |   |   |   |   |   |   |  Al > 1.17: build wind float (6.0/1.0)
|   |   |   |   |   |  Mg > 3.75: build wind non-float (10.0)
|   |   |  Al > 1.41
|   |   |   |  Si <= 72.49
|   |   |   |   |  Ca <= 8.28: build wind non-float (6.0)
|   |   |   |   |  Ca > 8.28: vehic wind float (5.0/1.0)
|   |   |   |  Si > 72.49
|   |   |   |   |  RI <= 1.51732
|   |   |   |   |   |  Fe <= 0.22: build wind non-float (30.0/1.0)
|   |   |   |   |   |  Fe > 0.22
|   |   |   |   |   |   |  RI <= 1.51629: build wind float (2.0)
|   |   |   |   |   |   |  RI > 1.51629: build wind non-float (2.0)
|   |   |   |   |  RI > 1.51732
|   |   |   |   |   |  RI <= 1.51789: build wind float (3.0)
|   |   |   |   |   |  RI > 1.51789: build wind non-float (2.0)
Ba > 0.27
|   Si <= 70.16: build wind non-float (2.0/1.0)
|   Si > 70.16: headlamps (27.0/1.0)
```

```
Number of Leaves  :     30

Size of the tree :      59


Time taken to build model: 0.07 seconds

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

Correctly Classified Instances         143               66.8224 %
Incorrectly Classified Instances        71               33.1776 %
Kappa statistic                          0.55
Mean absolute error                      0.1026
Root mean squared error                  0.2897
Relative absolute error                 48.4507 %
Root relative squared error             89.2727 %
Total Number of Instances              214

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.714    0.174    0.667      0.714    0.690      0.532    0.806     0.667     build wind float
               0.618    0.181    0.653      0.618    0.635      0.443    0.768     0.606     build wind non-float
               0.353    0.046    0.400      0.353    0.375      0.325    0.766     0.251     vehic wind float
               0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         vehic wind non-float
               0.769    0.010    0.833      0.769    0.800      0.788    0.872     0.575     containers
               0.778    0.029    0.538      0.778    0.636      0.629    0.930     0.527     tableware
               0.793    0.022    0.852      0.793    0.821      0.795    0.869     0.738     headlamps
Weighted Avg.  0.668    0.130    0.670      0.668    0.668      0.539    0.807     0.611

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 50 15  3  0  0  1  1 |  a = build wind float
 16 47  6  0  2  3  2 |  b = build wind non-float
  5  5  6  0  0  1  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  2  0  0 10  0  1 |  e = containers
  1  1  0  0  0  7  0 |  f = tableware
  3  2  0  0  0  1 23 |  g = headlamps
```

Appendix A: Self-organizing maps

Self-organizing map

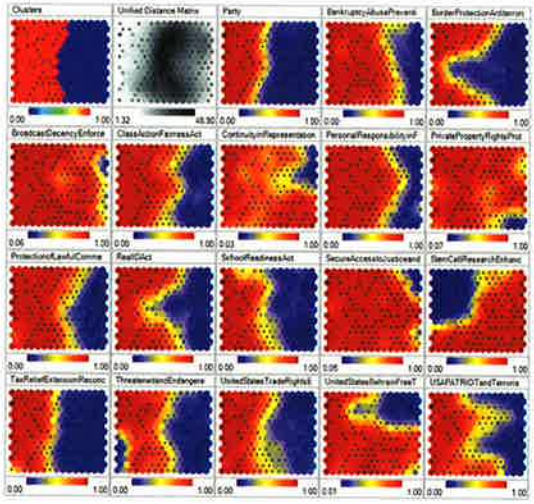
A **self-organizing map** (**SOM**) or **self-organizing feature map** (**SOFM**) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a **map**, and is therefore a method to do dimensionality reduction. Self-organizing maps differ from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as backpropagation with gradient descent), and in the sense that they use a neighborhood function to preserve the topological properties of the input space.

This makes SOMs useful for visualizing low-dimensional views of high-dimensional data, akin to multidimensional scaling. The artificial neural network introduced by the Finnish professor Teuvo Kohonen in the 1980s is sometimes called a **Kohonen map** or **network**.[1][2] The Kohonen net is a computationally convenient abstraction building on biological models of neural systems from the 1970s[3] and morphogenesis models dating back to Alan Turing in the 1950s.[4]

Like most artificial neural networks, SOMs operate in two modes: training and mapping. "Training" builds the map using input examples (a competitive process, also called vector quantization), while "mapping" automatically classifies a new input vector.

A self-organizing map consists of components called nodes or neurons. Associated with each node are a weight vector of the same dimension as the input data vectors, and a position in the map space. The usual arrangement of nodes is a two-dimensional regular spacing in a hexagonal or rectangular grid. The self-organizing map describes a mapping from a higher-dimensional input space to a lower-dimensional map space. The procedure for placing a vector from data space onto the map is to find the node with the closest (smallest distance metric) weight vector to the data space vector.

While it is typical to consider this type of network structure as related to feedforward networks where the nodes are visualized as being attached, this type of architecture is fundamentally different in arrangement and motivation.



A self-organizing map showing U.S. Congress voting patterns. The input data was a table with a row for each member of Congress, and columns for certain votes containing each member's yes/no/abstain vote. The SOM algorithm arranged these members in a two-dimensional grid placing similar members closer together. **The first plot** shows the grouping when the data are split into two clusters. **The second plot** shows average distance to neighbours: larger distances are darker. **The third plot** predicts Republican (red) or Democratic (blue) party membership. **The other plots** each overlay the resulting map with predicted values on an input dimension: red means a predicted 'yes' vote on that bill, blue means a 'no' vote. The plot was created in Synapse.