Learning Analytics Tools

Professor Ramkumar Rajendran

Educational Technology

Indian Institute of Technology, Bombay

WEKA - Data mining Tool

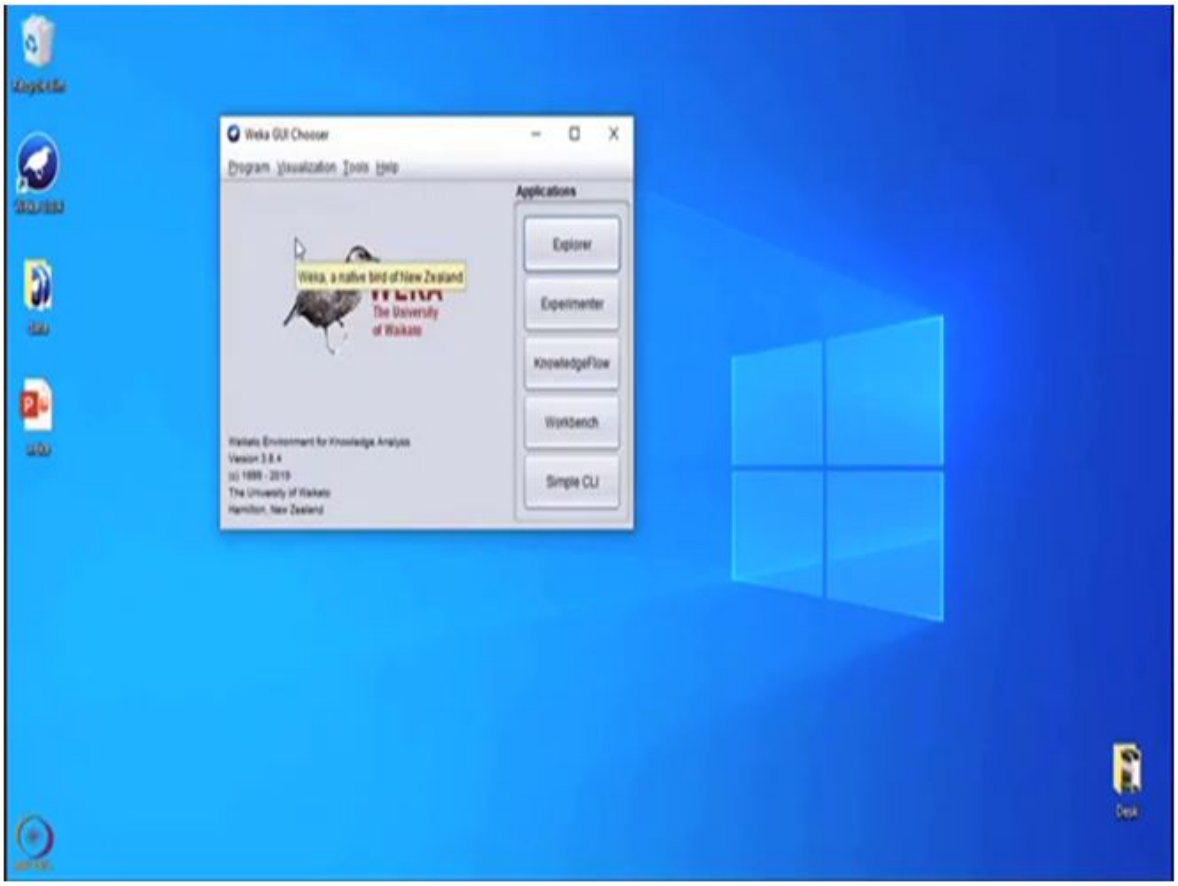
(Refer Slide Time: 00:18)

WEKA is an open-source data mining tool. It is developed by Waikato University, New Zealand.

(Refer Slide Time: 00:27)

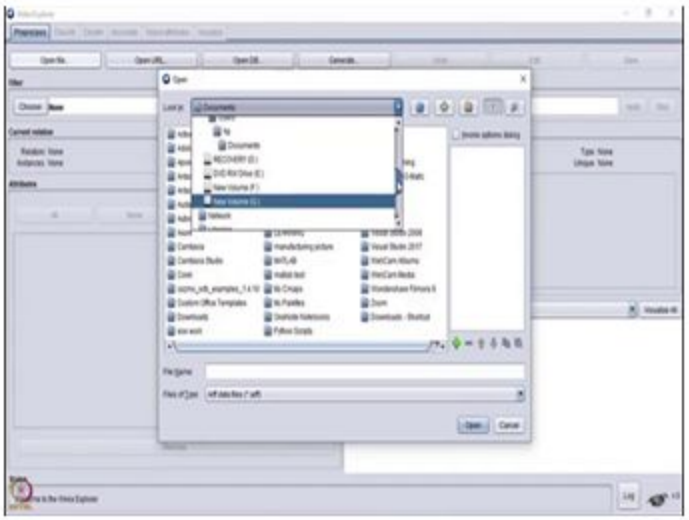
It can be downloaded from the following link and it runs on Windows, MAC operating system and Linux.

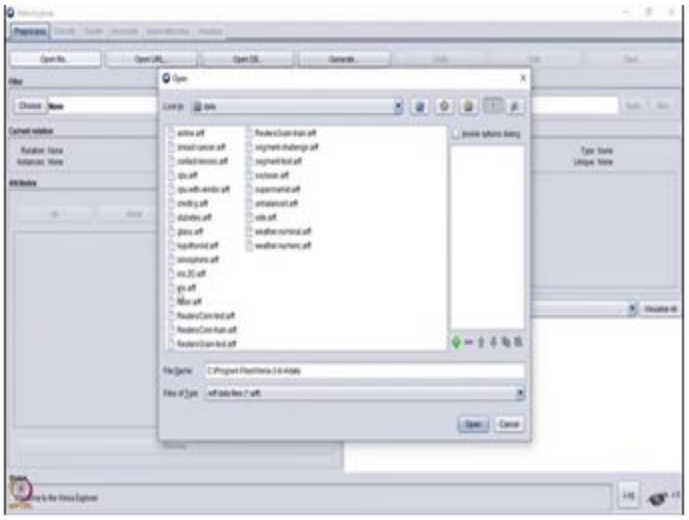
(Refer Slide Time: 00:37)

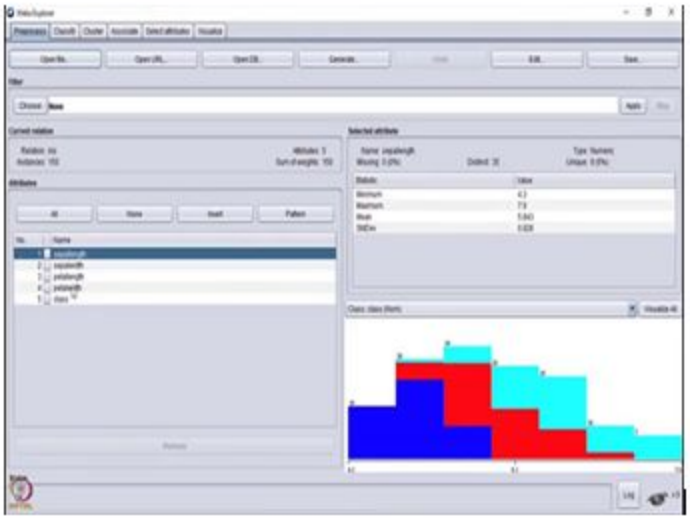


This is the WEKA GUI chooser. It has 5 applications: explorer, experimenter, knowledge flow, workbench and simple CLI. For most of our work, we will work on explorer.

(Refer Slide Time: 00:54)



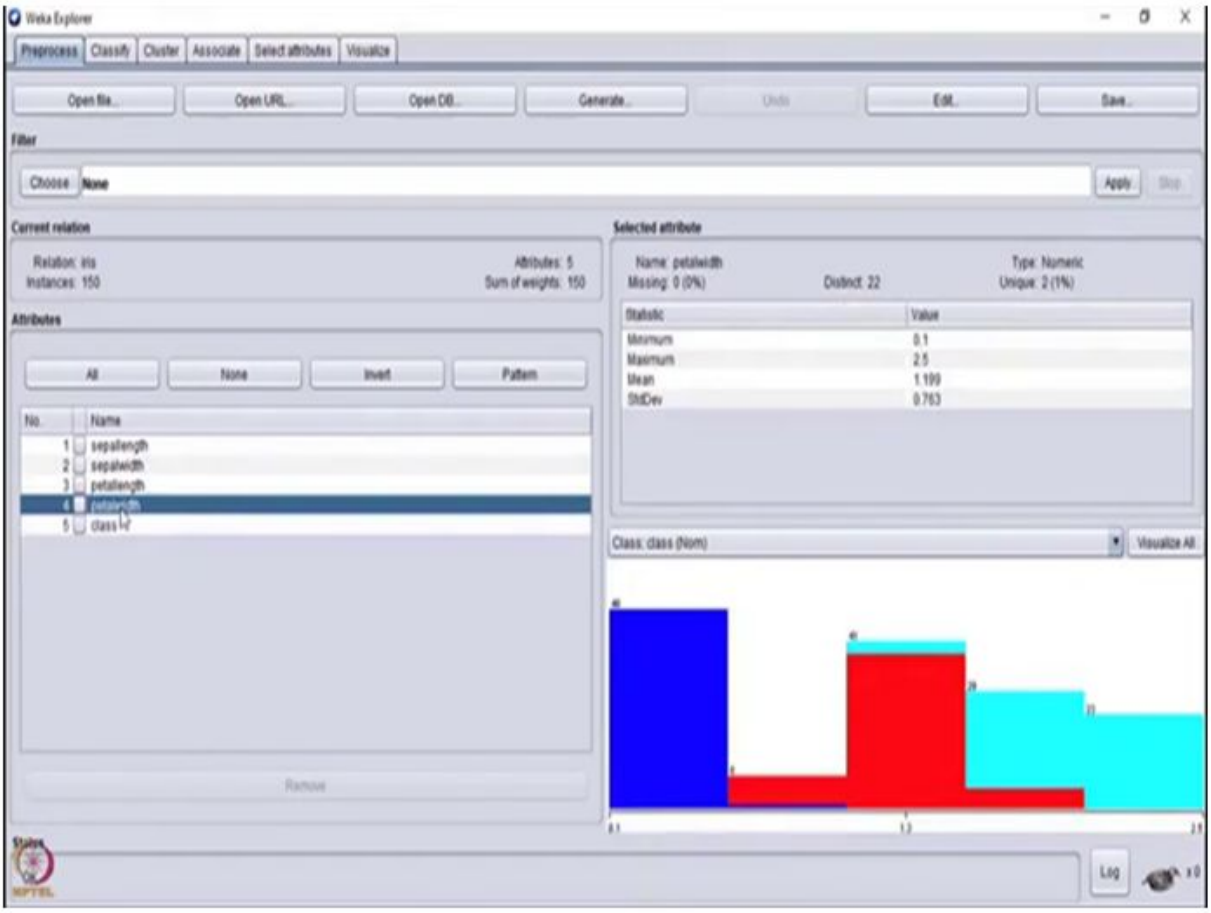


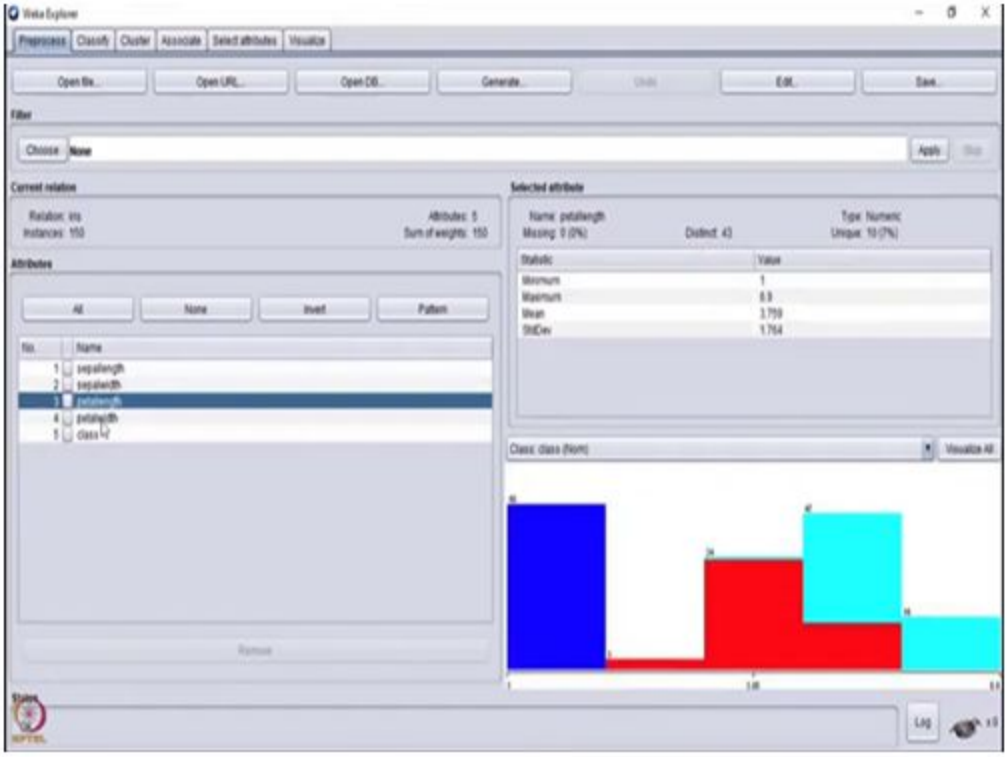


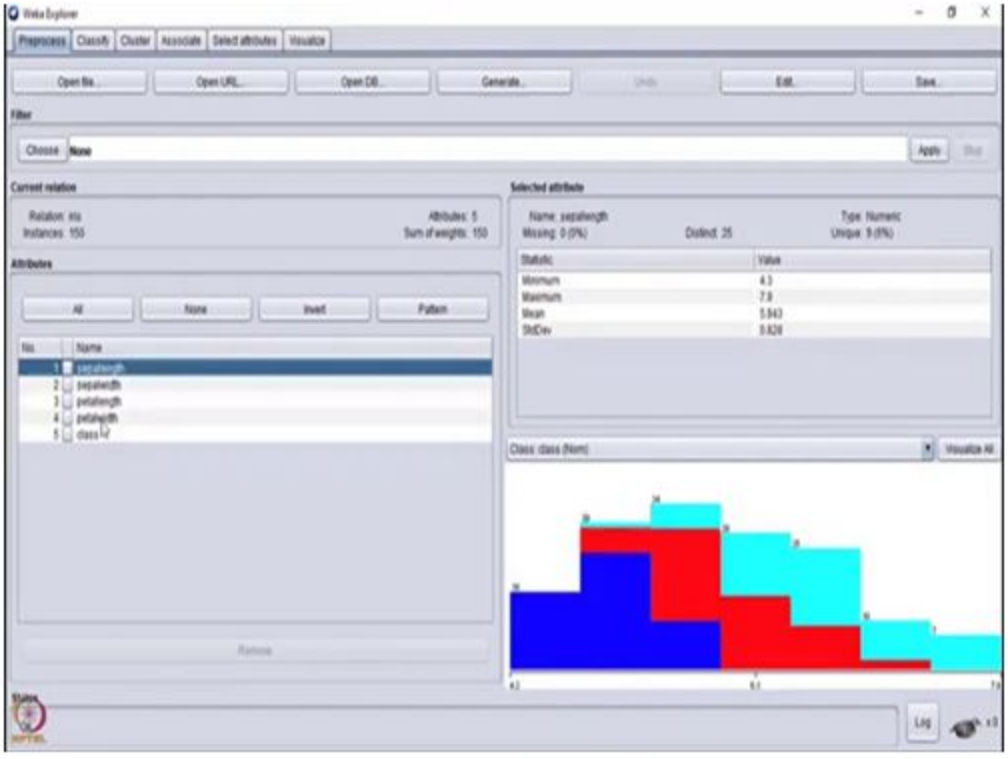
The explorer has the following menus: preprocess, classify, cluster and visualize. We can open the file either saved on the system or we can also use URL. For demonstration purposes, we will choose the file that you get when you download the WEKA software and this file can be located in C drive program files and WEKA. This is the sample data file.

For demonstration, we will use the iris data set. This is the specifications of the given iris data set. It has 5 attributes i.e. 5 columns and it stores 150 instances. It has 3 classes: iris setosa, iris versicolour, iris virginica is having 50 counts.

(Refer Slide Time: 02:09)

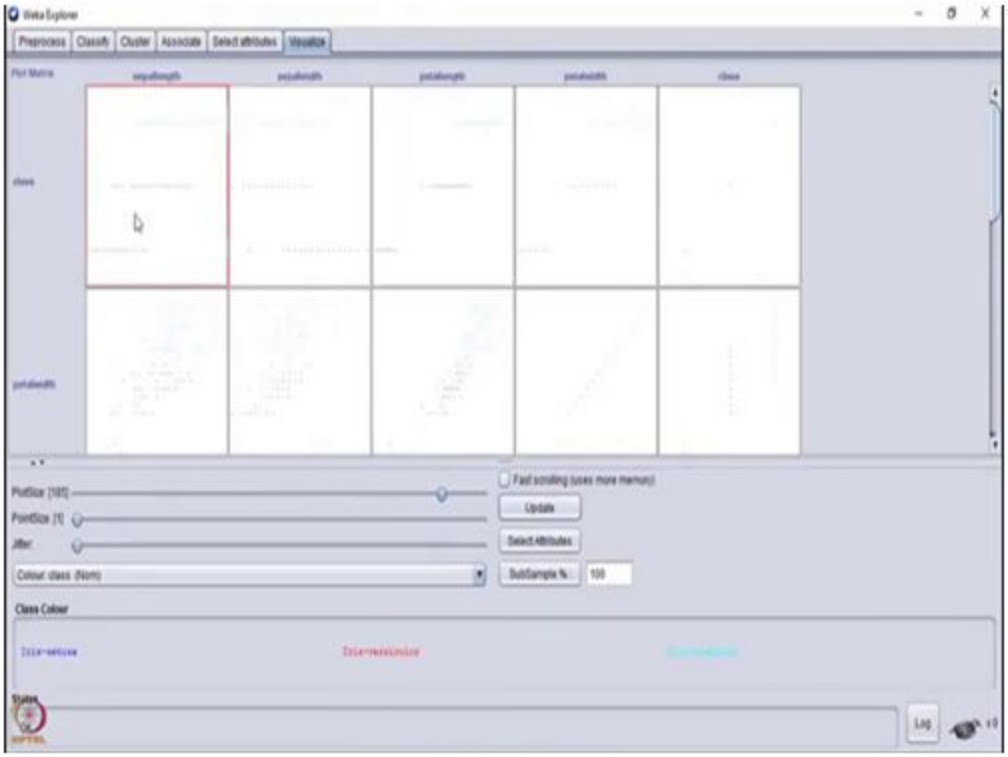


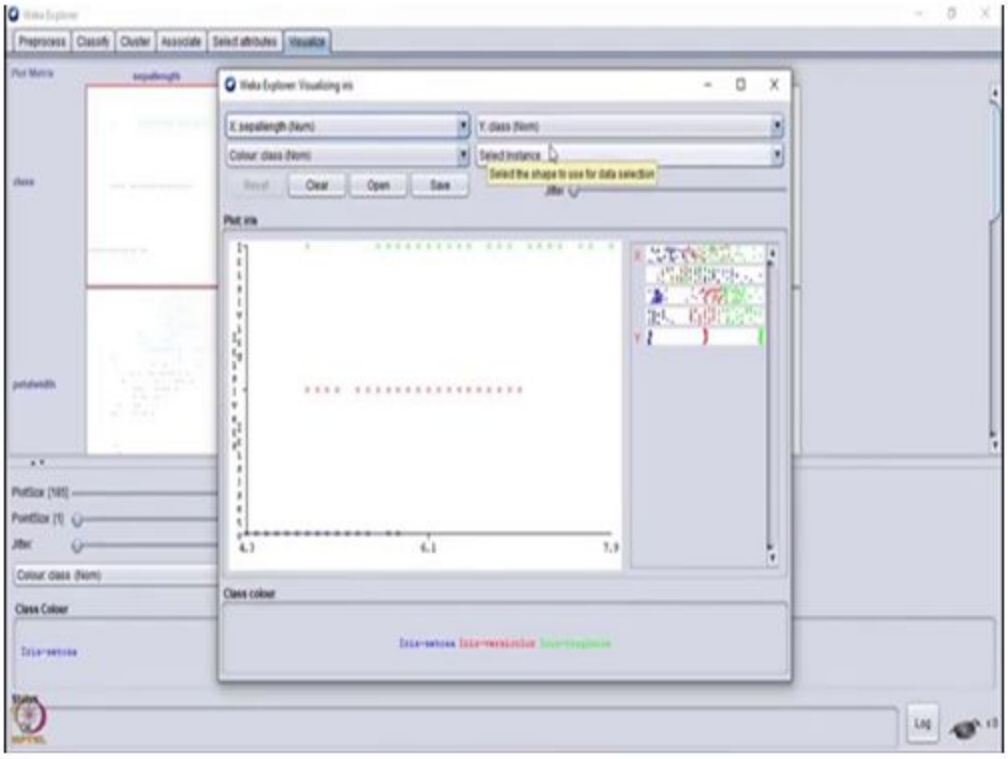

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
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
Class: class (Nom)
Visualize All
Status
OK
Log
x 0

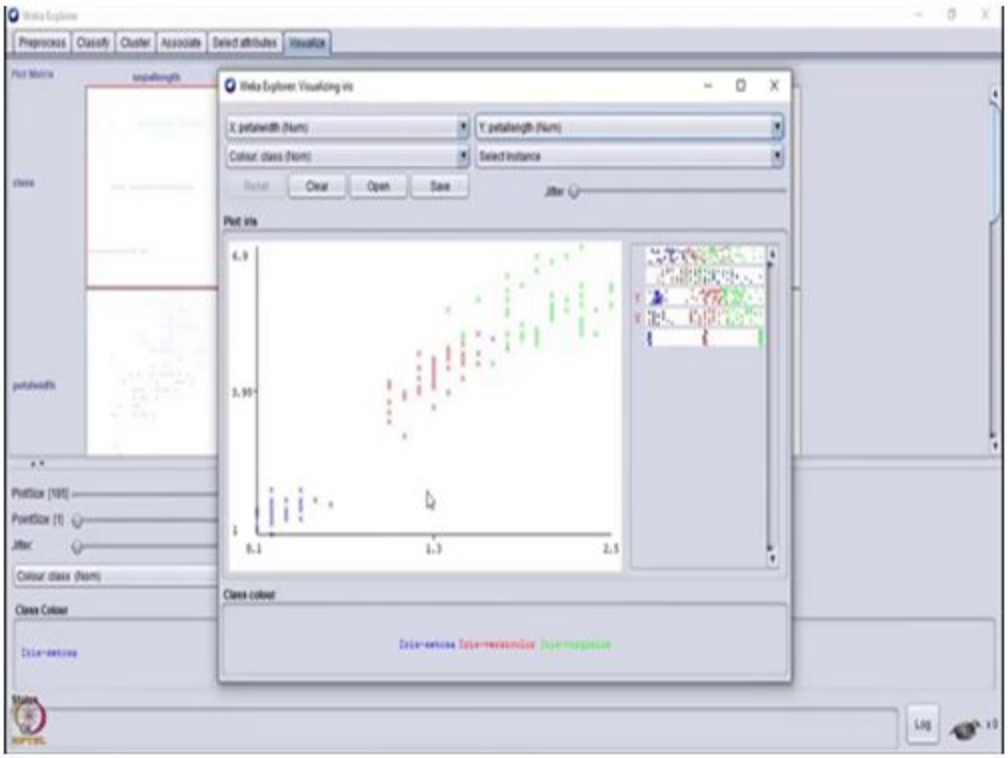

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
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
Class: class (Nom)
Visualize All
Status
OK
Log
x 0

It also shows the petal width, petal length, sample width and sample length variation across the different attributes or classes. We can also visualize the given data set by choosing the visualize menus.

(Refer Slide Time: 02:28)


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
sepallength
petalwidth
petallength
class
PlotSize [100]
PointSize [1]
Jitter
Colour: class (Nom)
Fast scrolling (uses more memory)
Update
Select Attributes
SubSample %
100
Class Colour
Iris-setosa
Iris-versicolor
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Weka Explorer Visualizing iris
X: sepallength (Num)
Y: class (Nom)
Colour: class (Nom)
Select Instance
Select the shape to use for data selection
Reset
Clear
Open
Save
Jitter
Plot iris
4.3
6.1
7.9
Class colour
Iris-setosa
Iris-versicolor
Iris-virginica
Status
OK
Log
x 0



Here we can visualize our given data set. In this example, the sample length is plotted along the x-axis and this class level is on the y-axis. This plot shows the variation of class with respect to sample length or we can also choose the different parameter.

Along the x-axis suppose we choose petal width and along the y axis we can choose petal length. This plot shows the variations of different classes along petal width and petal length.

(Refer Slide Time: 03:19)

Now let us learn about classification in the WEKA tool. For classification also we choose the explorer application and we choose the downloaded data sets from our system which is stored in the C file. In WEKA we have different classifier. First, we choose the classified menu and choose the appropriate classifier.

In WEKA we have different classifier. We have the Naive-Bayes classifier. We have the IBk classifier. We have zero-classifier. Zero-classifier is also known as trivial classifier or baseline classifier.

We also have the J48 classifier. For our analysis purpose, we will choose the J48 classifier. Before analyzing the data we create a model. First, we divide the data into 2 sets: the training set and testing set. The training set is used to create a classifier model and the test data set can be used to check the performance of the model created.

For most of the application, two-third of the given data set is selected for training the model and the remaining one third data set is used for testing the model and now we run. This is the classifier output and we can obtain the summary as follows. We are having 51 as the total number of instances. This belongs to the testing set. Out of which 49 instances are classified correctly and only 2 instances are classified incorrectly.

We can also see the kappa statistic, mean absolute error values, root mean square error value, relative absolute error values and etc. Here we can see the detailed accuracy of the classes. We have the value of the precision which equals to 0.965. We have the value of recall. We have the average value of the f measures, yes

$F1 = 0.961$

We also have the values of ROC area and on the left-hand side at the bottom, you can see the confusion matrix.

Here we can see that for the iris setosa, all the instances are classified correctly. No instance is classified as an incorrect one. Same is the case for the iris Versicolor. Here also all the instances are classified as correct and no instance is classified as incorrect. But for Iris virginica out of 17 instances, 15 instances are classified as correct and 2 instances are the classified as incorrect.