Data Warehousing & Data Mining LAB - G2
EXPERIMENT 3b

- ASHISH KUMAR
- 2K18/SE/041

Aim:- Generate Box Plot for a sample data in WEKA.

Theory:-

A Box and Whisker Plot (or Box Plot) is a convenient way of visually displaying the data distribution through their quartiles. The lines extending parallel from the boxes are known as the "whiskers", which are used to indicate variability outside the upper and lower quartiles. Outliers are sometimes plotted as individual dots that are in-line with whiskers. Box Plots can be drawn either vertically or horizontally. Although Box Plots may seem primitive in comparison to a Histogram or Density Plot, they have the advantage of taking up less space, which is useful when comparing distributions between many groups or datasets.

I have chosen iris dataset.

Iris Dataset
This is perhaps the best known database to be found in the pattern recognition literature. The dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant and 5 attributes and these are:

- sepal length in cm
- sepal width in cm
- petal length in cm
- petal width in cm
- class:
  -- Iris Setosa
  -- Iris Versicolour
  -- Iris Virginica

Procedure:

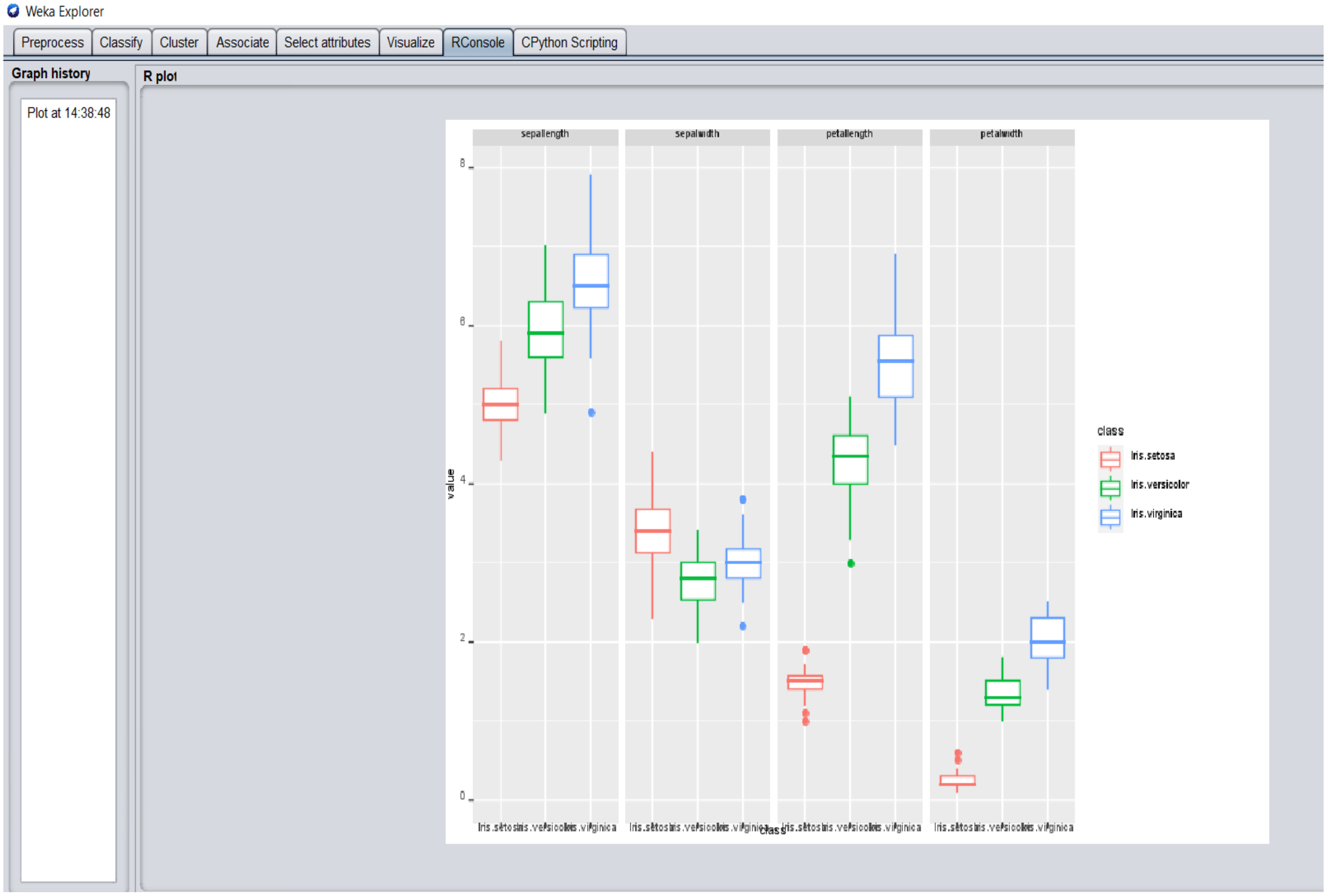
Plotting Box-Plots

- Open Weka software 3.8.5.
- Install R software on your computer. Install rJava package using the command install.packages(“rJava”) in the R application console.
- Using package manager in WEKA tools, install RPlugin to use the R console with WEKA.
- Go to weka explorer.
- Choose dataset by going into open files option and choose iris.arff.
- Using the following commands(in Rconsole), plot the box plot of the given dataset:

```
>>> install.packages("ggplot2")

>>> library(ggplot2)

>>> install.packages("reshape2")

>>> library(reshape2)

>>> ndata=melt(rdata)
        Using class as id variables

>>> ggplot(ndata, aes(y = value, x = class, color = class)) + geom_boxplot() +
     facet_grid(.~variable)
```

Output:-

The following Box plot have been obtained for the dataset iris.arff consisting 4 box plots , one for each attribute - sepal width , sepal length , petal width and petal length. Boxplot is used to see how the categorical feature “Species” is distributed with all other four input variables.



Findings and Learning:

Box plots are important tools to give an appropriate graphical representation of the raw data and hence are extensively used in data visualization. Box plot can tell you about your outliers and what their values are. WEKA Software provides a good set of functions to plot box plots with various combinations of attributes. We have successfully plotted boxplots in WEKA.