Data Warehousing & Data Mining LAB - G2
EXPERIMENT 3a

- ASHISH KUMAR
- 2K18/SE/041

Aim:- Generate Histogram for a sample data in WEKA.

Theory:-

A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data. This allows the inspection of the data for its underlying distribution (e.g., normal distribution), outliers, skewness, etc. A Histogram visualises the distribution of data over a continuous interval or certain time period. Each bar in a histogram represents the tabulated frequency at each interval/bin. Histograms help give an estimate as to where values are concentrated, what the extremes are and whether there are any gaps or unusual values. They are also useful for giving a rough view of the probability distribution.

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

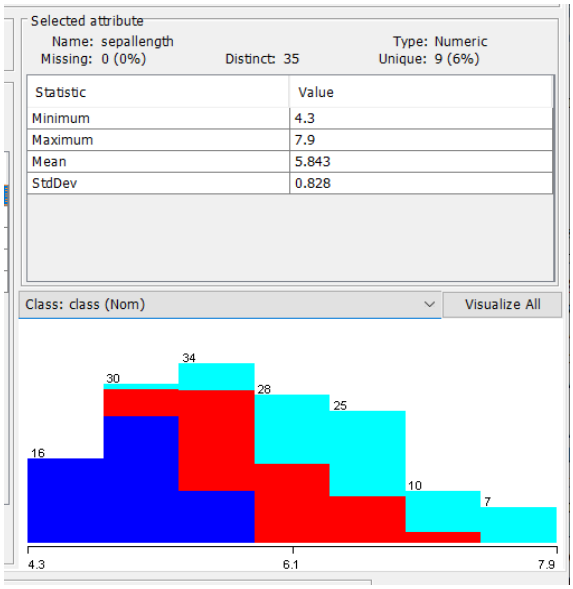
Procedure:

Plotting histograms

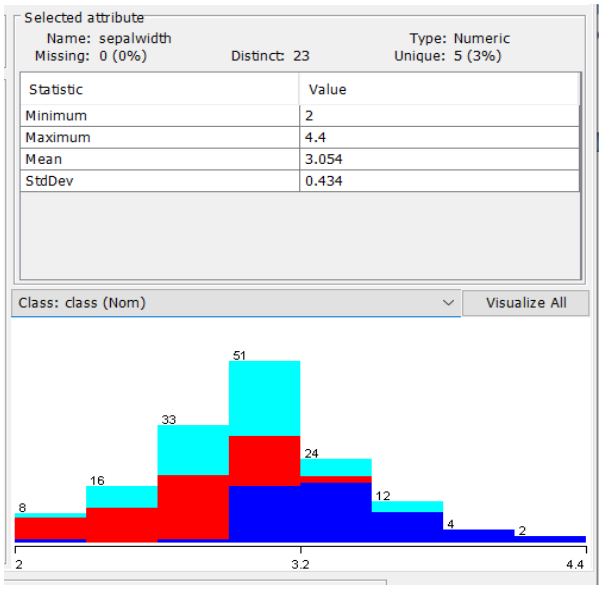
- Open Weka software 3.8.5.
- Go to weka explorer.
- Choose dataset by going into open files and choose iris.arff
- Above histogram, click visualize all.

Output:-

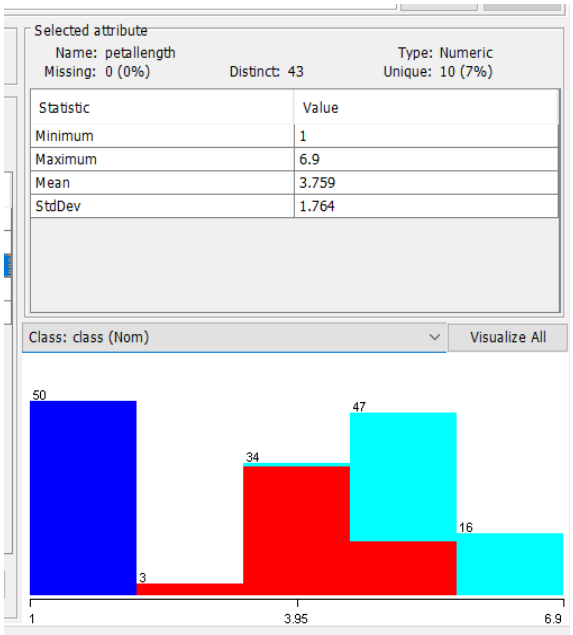
The following plots have been obtained for the dataset iris.arff
Histogram



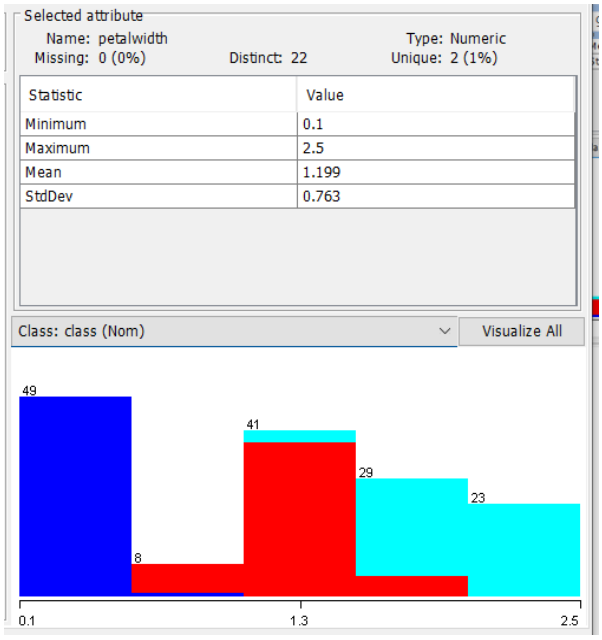
Sepallength



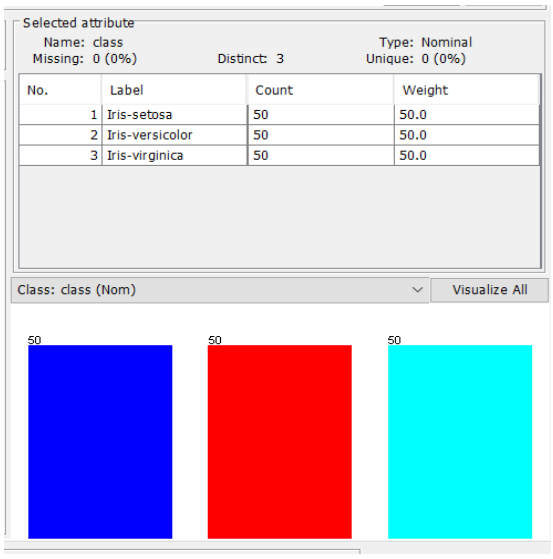
Sepalwidth



Petallength



Petalwidth



Class

FINDINGS & LEARNING:-

Through this experiment we learnt about the importance of histogram in data visualization. It can provide information on the degree of variation of the data and show the distribution pattern of the data by bar graphing the number of units in each class or category. We also learned about how histograms can be generated using WEKA Data Mining Tool.