D.W.M. Experiment No.

Data Preprocessing using WEKA

AIM:

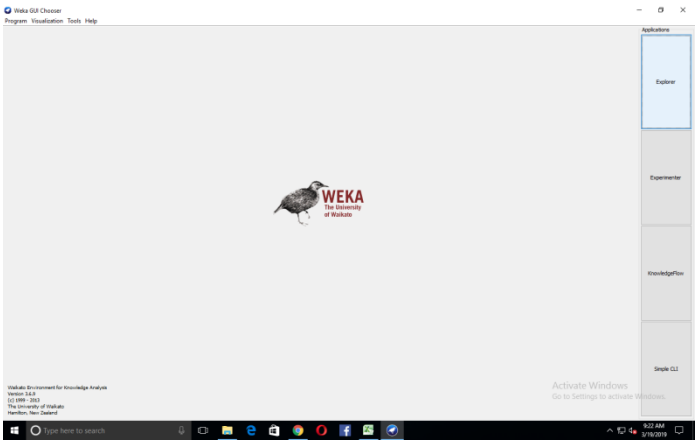
To perform data Preprocessing on a dataset using WEKA tool.

Theory:

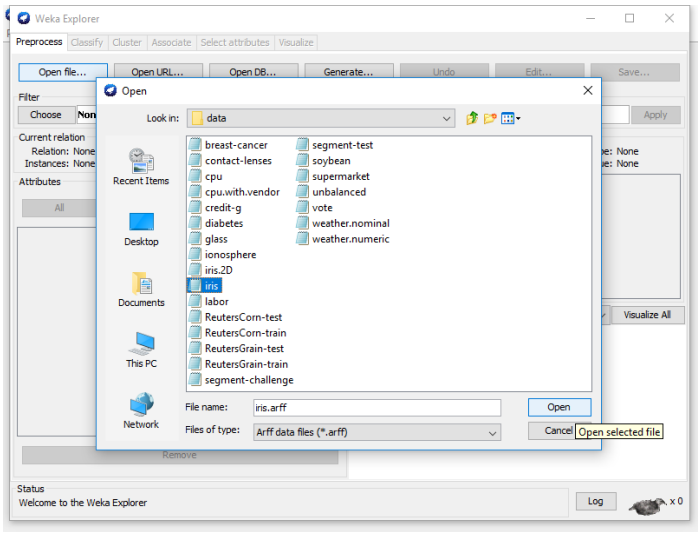
This exercise illustrates some of the basic data preprocessing operations that can be performed using WEKA. The sample data set used for this example is the "iris data".

I.Loading the data

1.Start weka and select the explorer



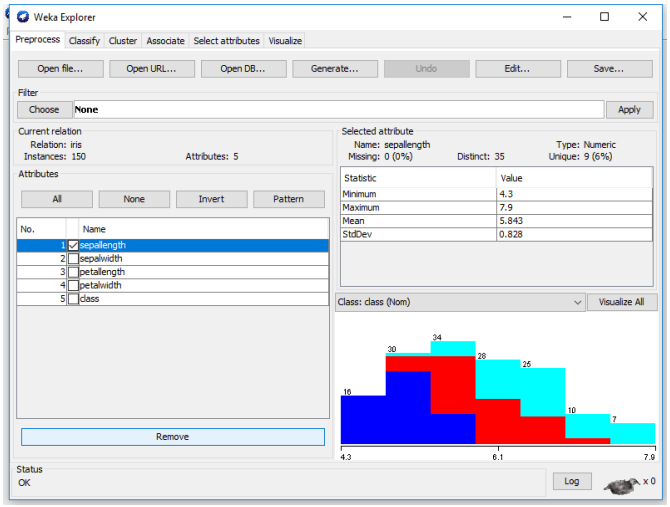
2.Open the dataset file

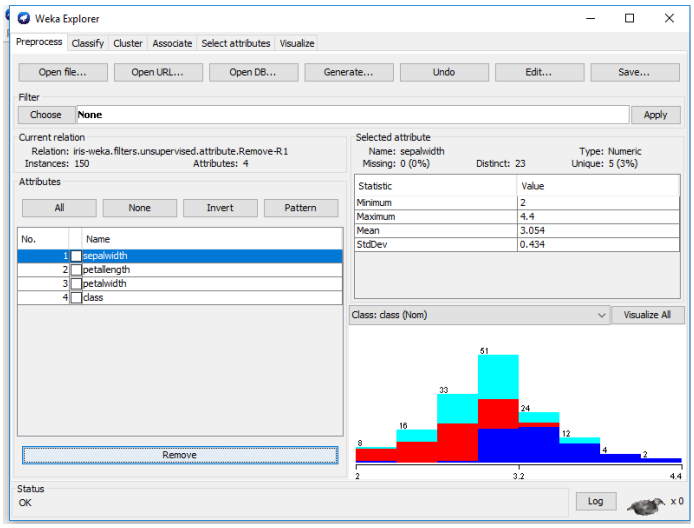


II.Selecting and filtering attributes

A.Deleting

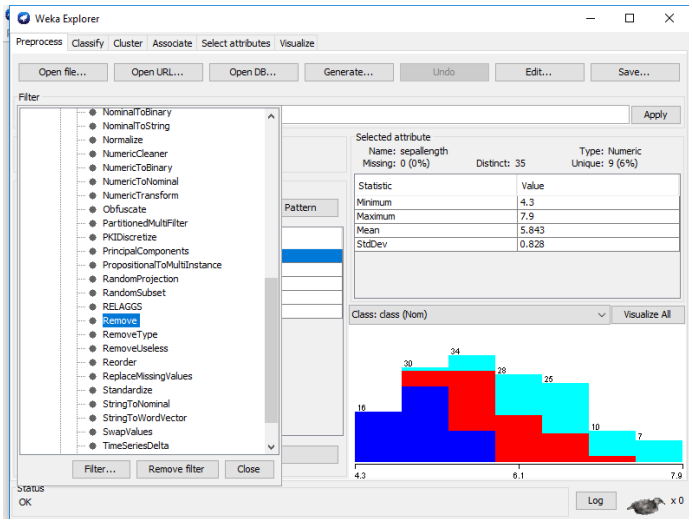
1.simply select the attribute and click on “Remove button” as shown in Figure



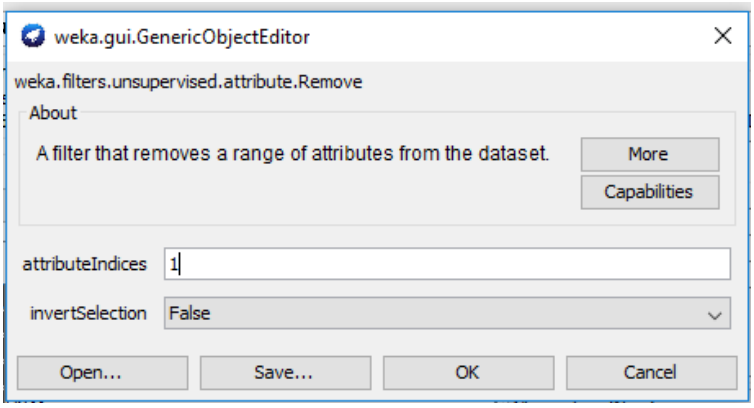


OR

1.Using the Attribute filters in WEKA. In the "Filter" panel, click on the "Choose" button. This will show a popup window with a list available filters. Scroll down the list and select the "weka.filters.unsupervised.attribute.Remove" filter as shown in Figure

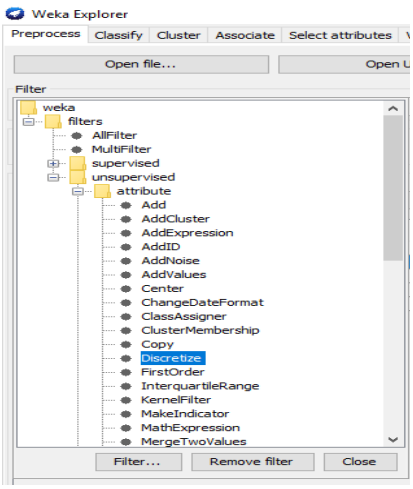


2.Click on text box immediately to the right of the "Choose" button. In the resulting dialog box enter the index of the attribute to be filtered out.Then click "OK". Now, in the filter box you will see "Remove R 1" .Click the "Apply" button to apply this filter to the data. This will remove the attribute and create a new working relation

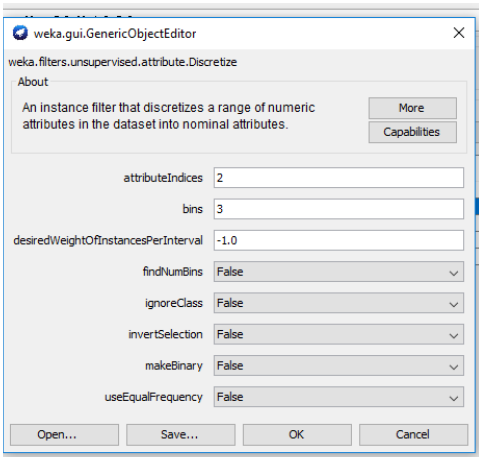


III.Discretization

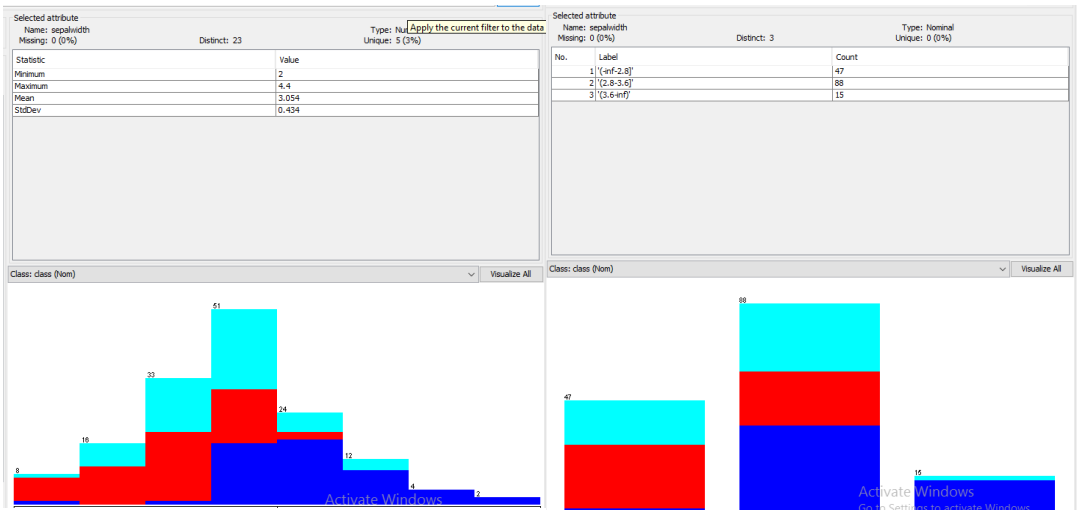
1.We will rely on WEKA to perform discretization on the attributes. In this example, we divide each of these into 3 bins (intervals). The WEKA discretization filter, can divide the ranges blindly, or used various statistical techniques to automatically determine the best way of partitioning the data.We activate the Filter dialog box, but this time, we will select "weka.filters.unsupervised.attribute.Discretize"



2.Next, to change the defaults for this filter, click on the box immediately to the right of the "Choose" button. This will open the Discretize Filter dialog box. We enter the index for the attributes to be discretized. We also enter 3 as the number of bins. Since we are doing simple binning, all of the other available options are set to "false".



3.Click "Apply" in the Filter panel. This will result in a new working relation with the selected attribute partitioned into 3 bins



III.Missing Values

1.Check if there is any missing values in any attribute.

Viewer

Relation: iris

No.	sepallength Numeric	sepalwidth Numeric	petallength Numeric	petalwidth Numeric	class Nominal
10	4.9	3.1	1.5	0.1	Iris-se...
11	5.4	3.7	1.5	0.2	Iris-se...
12	4.8	3.4	1.6	0.2	Iris-se...
13	4.8	3.0	1.4	0.1	Iris-se...
14	4.3	3.0	1.1	0.1	Iris-se...
15	5.8	4.0		0.2	Iris-se...
16	5.7	4.4	1.5	0.4	Iris-se...
17	5.4	3.9	1.3	0.4	Iris-se...
18	5.1	3.5	1.4	0.3	
19	5.7	3.8	1.7	0.3	Iris-se...
20	5.1	3.8	1.5	0.3	Iris-se...
21	5.4		1.7	0.2	Iris-se...
22	5.1	3.7	1.5	0.4	Iris-se...
23	4.6	3.6	1.0	0.2	Iris-se...
24	5.1	3.3	1.7	0.5	Iris-se...
25	4.8	3.4	1.9	0.2	Iris-se...
26	5.0		1.6	0.2	Iris-se...
27	5.0	3.4	1.6	0.4	Iris-se...
28	5.2	3.5	1.5	0.2	Iris-se...
29	5.2	3.4	1.4	0.2	Iris-se...
30	4.7	3.2	1.6	0.2	Iris-se...
31	4.8	3.1	1.6	0.2	Iris-se...
32	5.4	3.4	1.5	0.4	Iris-se...
33	5.2	4.1	1.5	0.1	Iris-se...
34	5.5	4.2	1.4	0.2	Iris-se...
35	4.9	3.1	1.5	0.1	Iris-se...

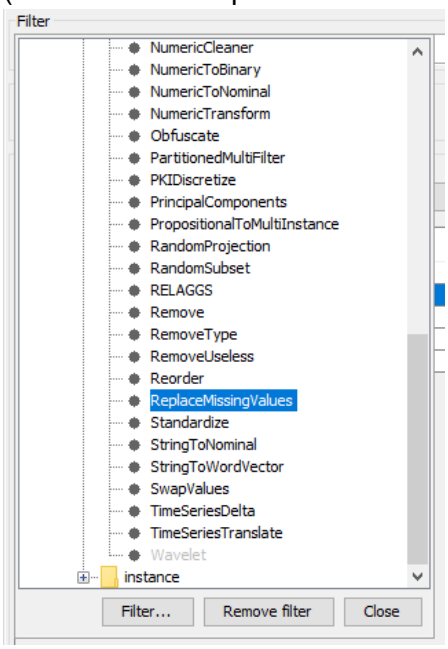
Selected attribute

Name: sepalwidth
Missing: 2 (1%)
Distinct: 23
Type: Numeric
Unique: 5 (3%)

Statistic	Value
Minimum	2
Maximum	4.4
Mean	3.052
StdDev	0.436

2.Choose “ReplaceMissingValues” filter

(weka.filters.unsupervised.attribute.ReplaceMissingValues). Then, click on Apply button.



Relation: iris-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filte

No.	sepallength Numeric	sepalwidth Numeric	petallength Numeric	petalwidth Numeric	class Nominal
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5.0	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3.0	1.4	0.1	Iris-setosa
14	4.3	3.0	1.1	0.1	Iris-setosa
15	5.8	4.0	3.775838...	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa
20	5.1	3.8	1.5	0.3	Iris-setosa
21	5.4	3.052027...	1.7	0.2	Iris-setosa
22	5.1	3.7	1.5	0.4	Iris-setosa
23	4.6	3.6	1.0	0.2	Iris-setosa
24	5.1	3.3	1.7	0.5	Iris-setosa

Selected attribute

Name: sepalwidth Type: Numeric
Missing: 0 (0%) Distinct: 24 Unique: 5 (3%)

Statistic	Value
Minimum	2
Maximum	4.4
Mean	3.052
StdDev	0.433