



A.Y. 2022-2023

LAB EXPERIMENT NO. 01

Name: Parth Parekh

SAP ID:60004200006

Branch:Computer Engineering

Batch:A1

Aim: Perform data Pre-processing task using Weka data mining tool

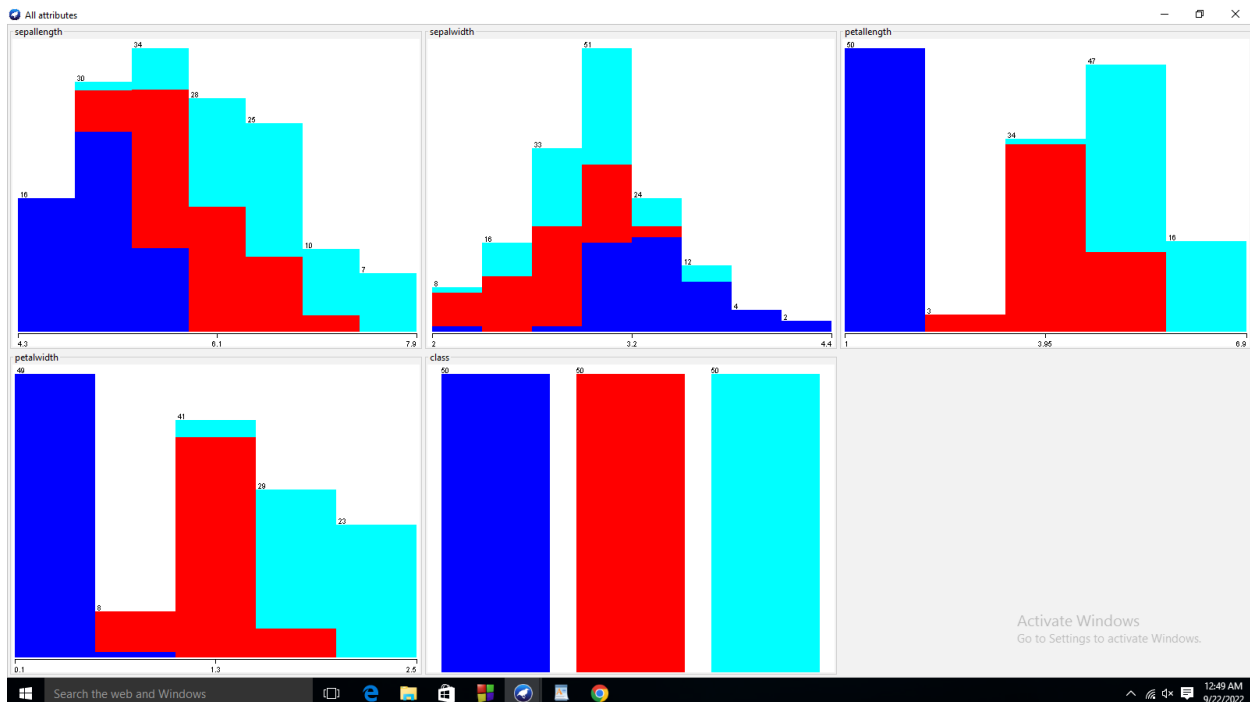
Theory:

WEKA - an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems

Tasks performed through Weka:

1. Preprocessing:

A. Visualize all:



This is histogram visualization of all attributes without any filter for iris dataset. We can infer that labels in sepal length and sepal width are closely related in terms of frequency distribution and

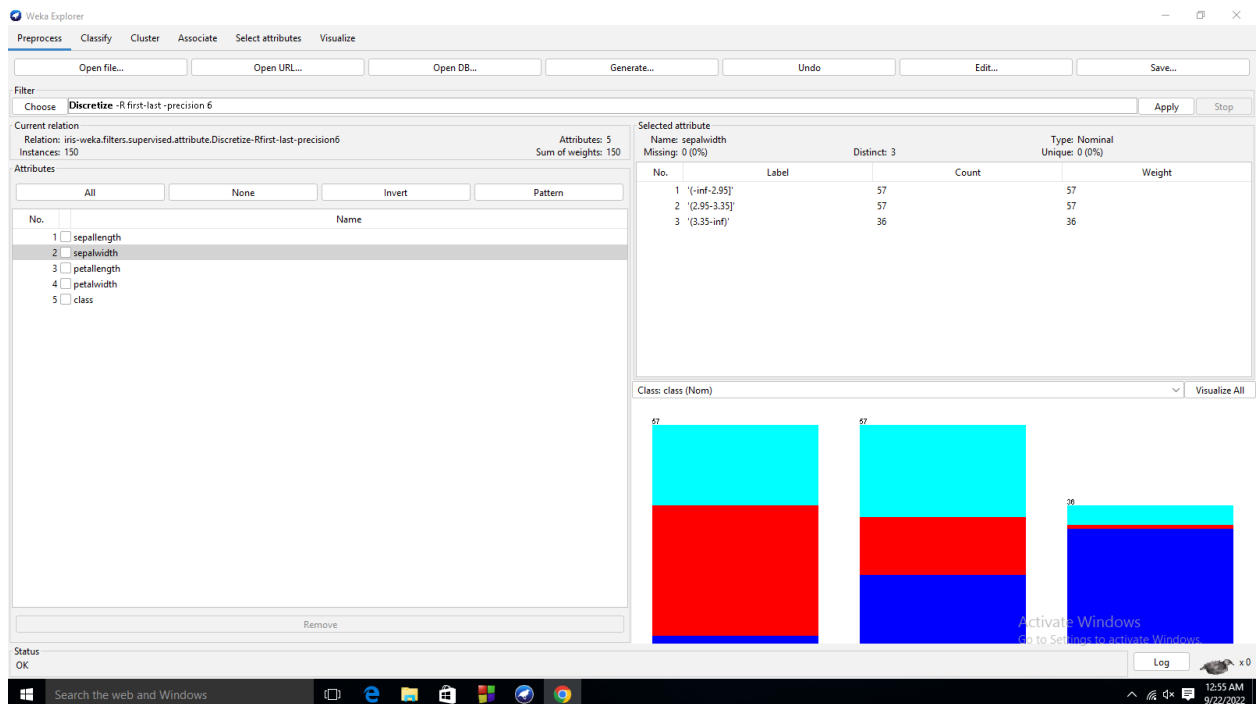


A.Y. 2022-2023

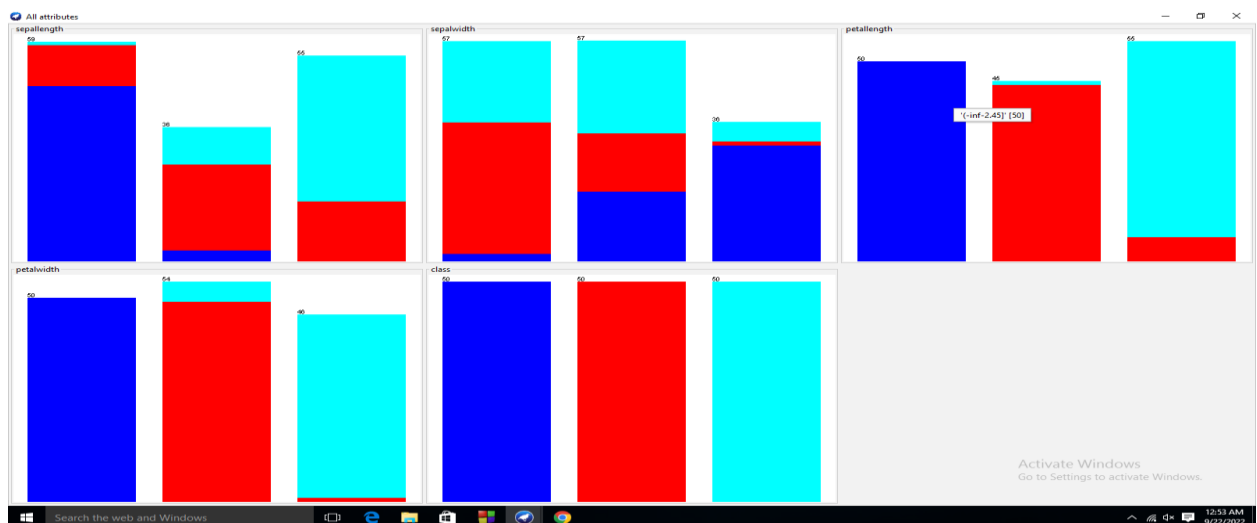
sepal length is skewed-right and sepal width is skewed-left whereas label in petal length and petal width are not so closely related in terms of frequency distribution.

B. Filter:

i) Discretization under supervised learning



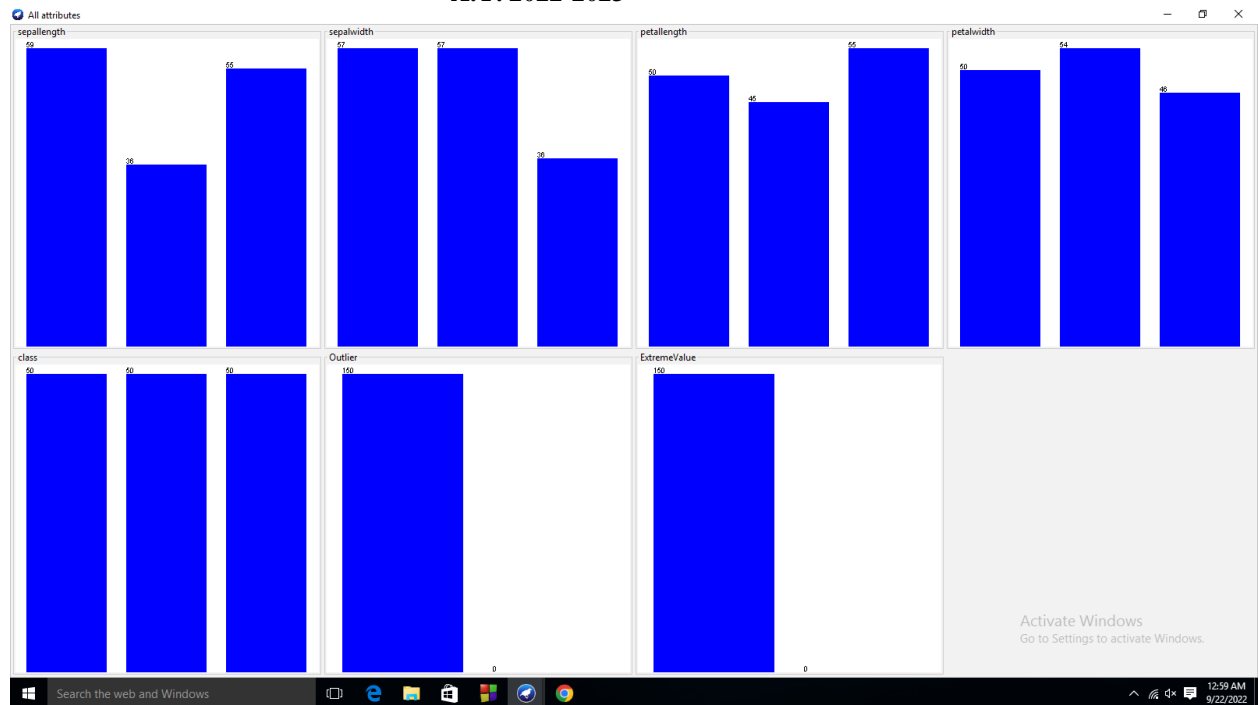
This is Preprocessing window of discrete supervised class (nominal datatype) where we can see it created three labels of different numerical intervals.



This is stacked column chart of discretized supervised learning

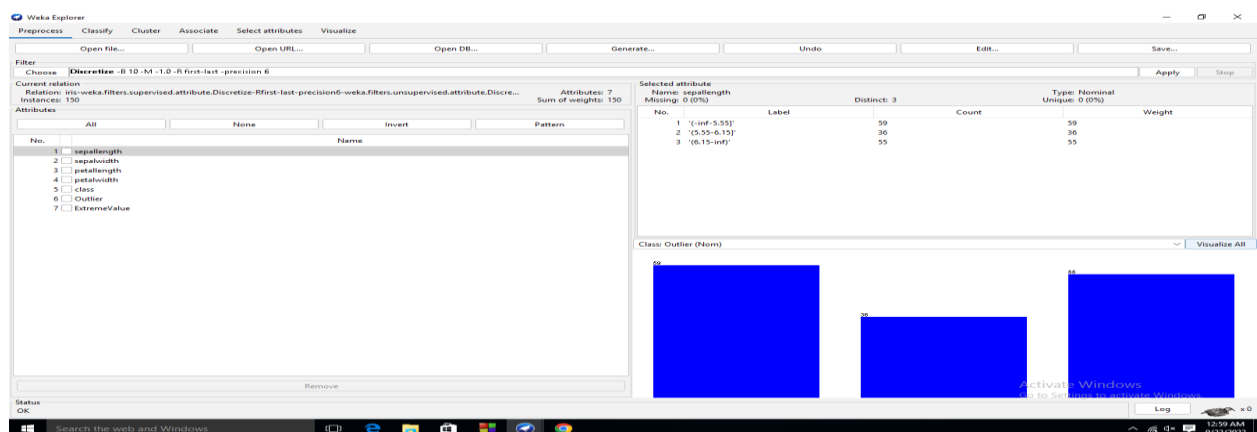


A.Y. 2022-2023



Above picture shows the extreme and outlier values of different attributes extreme values for discretize supervised learning . Insights from this is that extreme value for sepal length is 59 and two numerical intervals have same extreme value of 57 in sepal width.Outlier and Extreme values are 150.

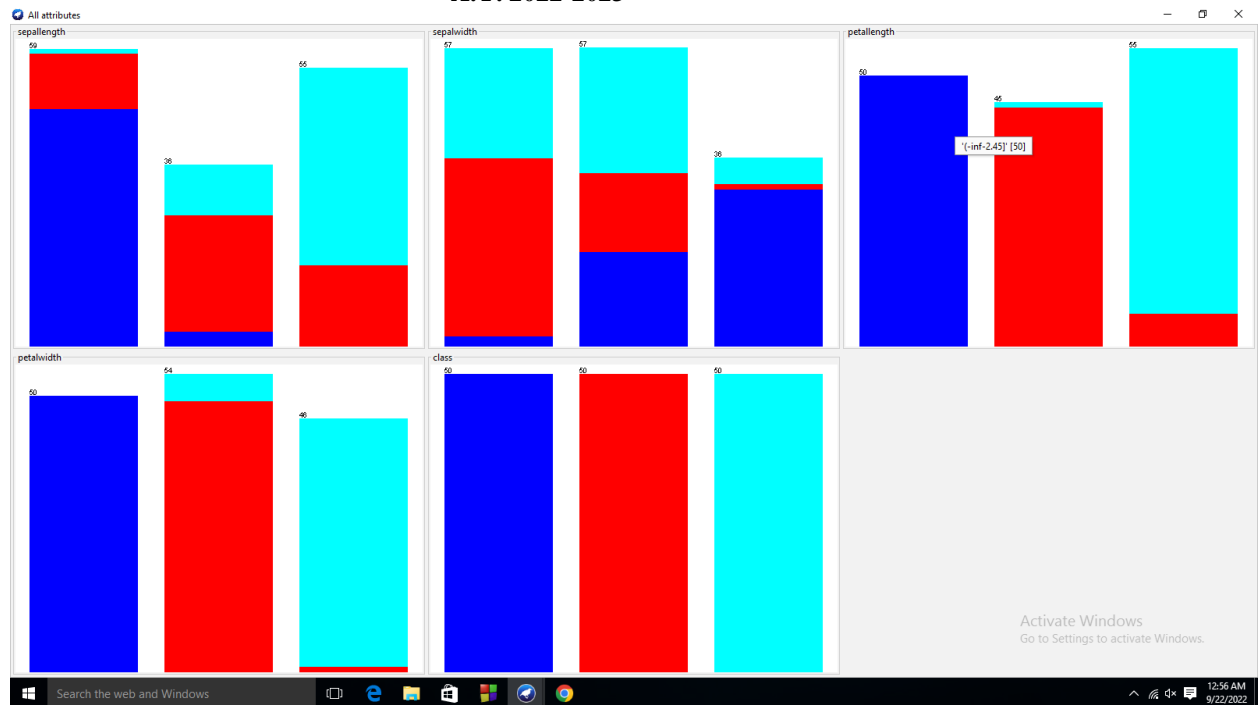
ii) Discretization under unsupervised learning



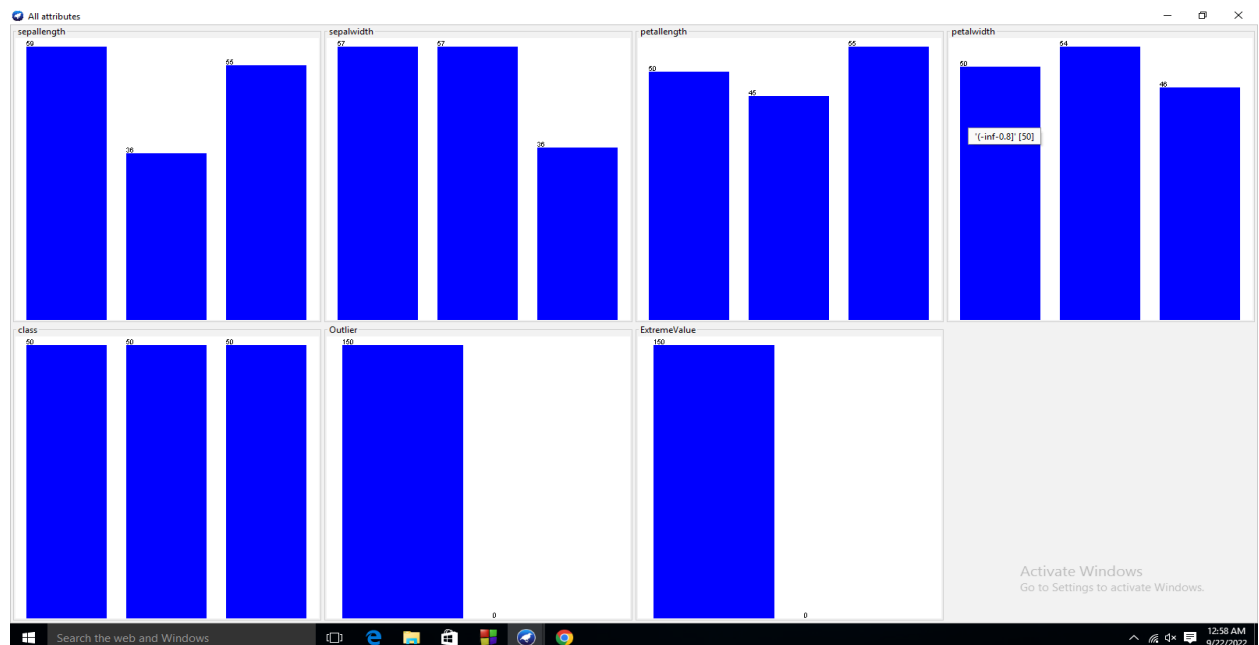
Preprocessing window of discretize unsupervised window where outlier for first bin is 59 which is the highest.



A.Y. 2022-2023



This is stacked column chart of discretized unsupervised learning

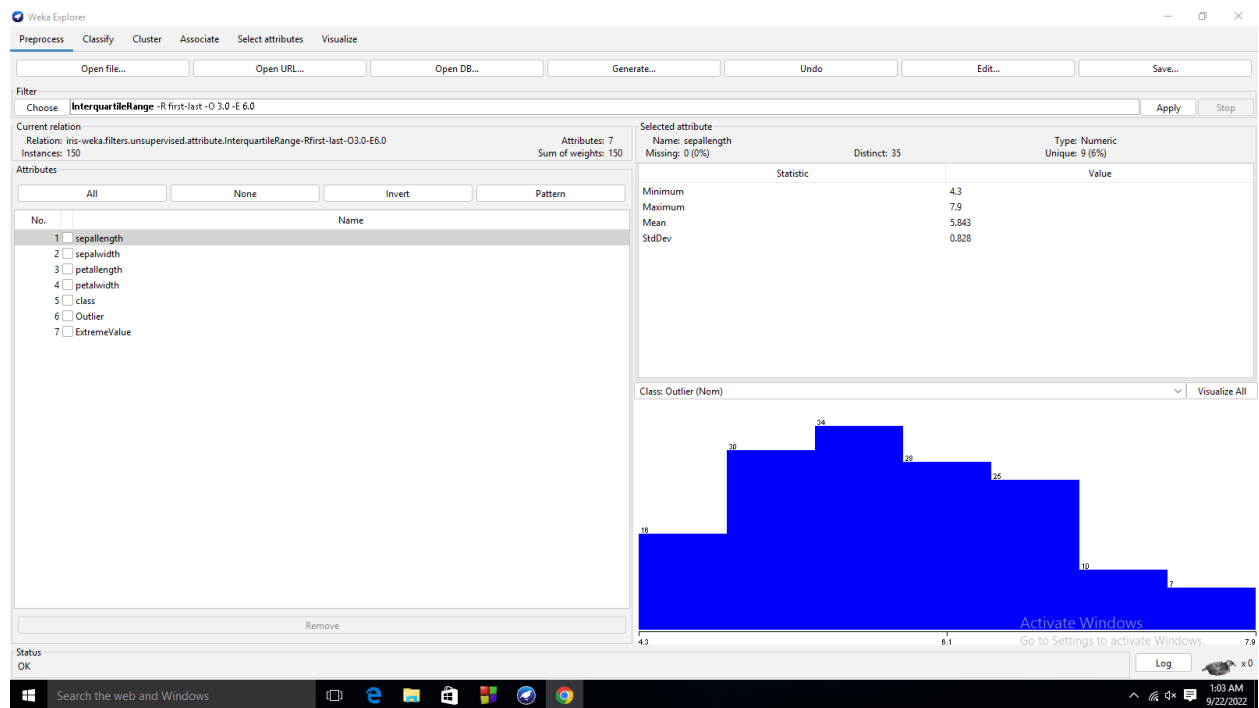


Above picture shows the extreme and outlier values of different attributes extreme values for discretize unsupervised learning . Insights from this is that extreme value for petal length is 55 and that of the petal width is 54.Outlier and Extreme values are 150

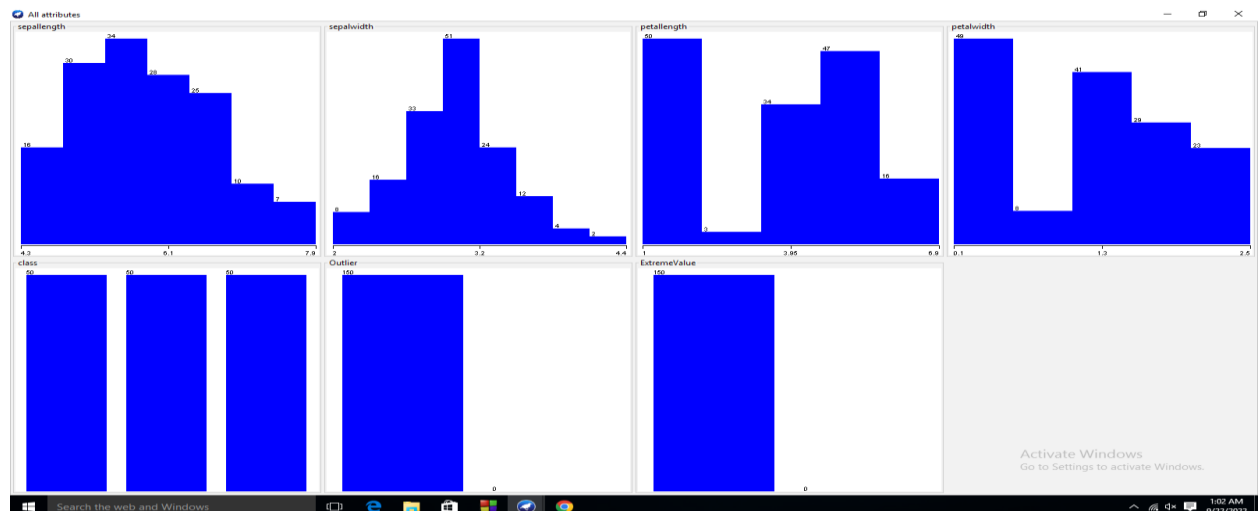


A.Y. 2022-2023

C. IQR:



This is the preprocessing window of IQR with sepal length as chosen attribute whose outlier histogram is skewed-right with mode 54 and it has minimum value of 4.3 and maximum value as 7.9 .



This is IQR outlier and extreme histogram visualization of all attributes with sepal length and sepal width being unimodal and petal length and petal width being multi modal.



A.Y. 2022-2023

2. Classification:

i. Naïve Bayes

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **NaiveBayes**

Test options:
☐ Use training set
☐ Supplied test set
☐ Cross-validation Folds: 10
☒ Percentage split % 70
More options...

(Nom) class: Start Stop

Result list (right-click for options):
01:04:30 - trees.J48
01:05:20 - trees.J48
01:06:27 - lazy.KStar
01:06:32 - lazy.KStar
01:07:36 - lazy.KStar
01:07:39 - lazy.KStar
01:08:17 - trees.J48
01:08:54 - trees.J48
01:09:56 - lazy.KStar
01:10:20 - lazy.KStar
01:11:09 - bayes.NaiveBayes

Classifier output:

ExtremeValue			
no	51.0	51.0	51.0
yes	1.0	1.0	1.0
[total]	52.0	52.0	52.0

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances	43	95.5556 %
Incorrectly Classified Instances	2	4.4444 %
Kappa statistic	0.9331	
Mean absolute error	0.0375	
Root mean squared error	0.158	
Relative absolute error	8.4241 %	
Root relative squared error	33.4979 %	
Total Number of Instances	45	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Iris-setosa
	1.000	0.069	0.889	1.000	0.941	0.910	0.987	0.976	Iris-versicolor
	0.867	0.000	1.000	0.867	0.929	0.901	0.987	0.979	Iris-virginica
Weighted Avg.	0.956	0.025	0.960	0.956	0.955	0.935	0.991	0.984	

=== Confusion Matrix ===

a	b	c	<-- classified as
14	0	0	a = Iris-setosa
0	16	0	b = Iris-versicolor
0	2	13	c = Iris-virginica

Status: OK

Search the web and Windows

Log 1:11 AM 9/22/2022

This is Naïve Bayes percentage split of 70% where correctly classified instances was 95.556
With mean absolute error as 0.0375

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **NaiveBayes**

Test options:
☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds: 10
☐ Percentage split % 70
More options...

(Nom) class: Start Stop

Result list (right-click for options):
01:04:30 - trees.J48
01:05:20 - trees.J48
01:06:27 - lazy.KStar
01:06:32 - lazy.KStar
01:07:36 - lazy.KStar
01:07:39 - lazy.KStar
01:08:17 - trees.J48
01:08:54 - trees.J48
01:09:56 - lazy.KStar
01:10:20 - lazy.KStar
01:11:09 - bayes.NaiveBayes
01:11:30 - bayes.NaiveBayes

Classifier output:

ExtremeValue			
yes	1.0	1.0	1.0
[total]	52.0	52.0	52.0

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	144	96 %
Incorrectly Classified Instances	6	4 %
Kappa statistic	0.94	
Mean absolute error	0.0342	
Root mean squared error	0.155	
Relative absolute error	7.6987 %	
Root relative squared error	32.8794 %	
Total Number of Instances	150	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Iris-setosa
	0.960	0.040	0.923	0.960	0.941	0.911	0.992	0.983	Iris-versicolor
	0.920	0.020	0.985	0.920	0.939	0.910	0.992	0.986	Iris-virginica
Weighted Avg.	0.960	0.020	0.960	0.960	0.960	0.940	0.994	0.989	

=== Confusion Matrix ===

a	b	c	<-- classified as
50	0	0	a = Iris-setosa
0	48	2	b = Iris-versicolor
0	4	46	c = Iris-virginica

Status: OK

Search the web and Windows

Log 1:11 AM 9/22/2022

This is naïve Bayes cross validation with 10 folds in which correctly classified instances has
increased to 96% with mean absolute error reduced to 0.342.



A.Y. 2022-2023

ii. J48 (Example of Decision Tree):

Classifier
Choose J48 -C 0.25 -M 2

Test options
☐ Use training set
☐ Supplied test set
☐ Cross-validation Folds 10
☒ Percentage split % 70

(Nom) class
Start Stop

Result list (right-click for options)
01:04:30 - trees.J48
01:05:20 - trees.J48
01:06:27 - lazy.KStar
01:06:32 - lazy.KStar
01:07:36 - lazy.KStar
01:07:39 - lazy.KStar
01:08:17 - trees.J48

Classifier output
petalwidth > 1.7: Iris-virginica (46.0/1.0)
Number of Leaves : 5
Size of the tree : 9
Time taken to build model: 0 seconds
Time taken to test model on test split: 0 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 43 95.5556 %
Incorrectly Classified Instances 2 4.4444 %
Kappa statistic 0.9331
Mean absolute error 0.0416
Root mean squared error 0.1652
Relative absolute error 9.3466 %
Root relative squared error 35.6559 %
Total Number of Instances 45
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 Iris-setosa
1.000 0.069 0.889 1.000 0.941 0.910 0.966 0.889 Iris-versicolor
0.867 0.000 1.000 0.867 0.929 0.901 0.964 0.931 Iris-virginica
Weighted Avg. 0.956 0.025 0.960 0.956 0.955 0.935 0.976 0.938
=== Confusion Matrix ===
a b c <-- classified as
14 0 0 | a = Iris-setosa
0 16 0 | b = Iris-versicolor
0 2 13 | c = Iris-virginica

This is J48 performance split of 70/30 with correctly classified instances of 95.556 and mean absolute error of 0.0416

Classifier
Choose J48 -C 0.25 -M 2

Test options
☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds 10
☐ Percentage split % 70

(Nom) class
Start Stop

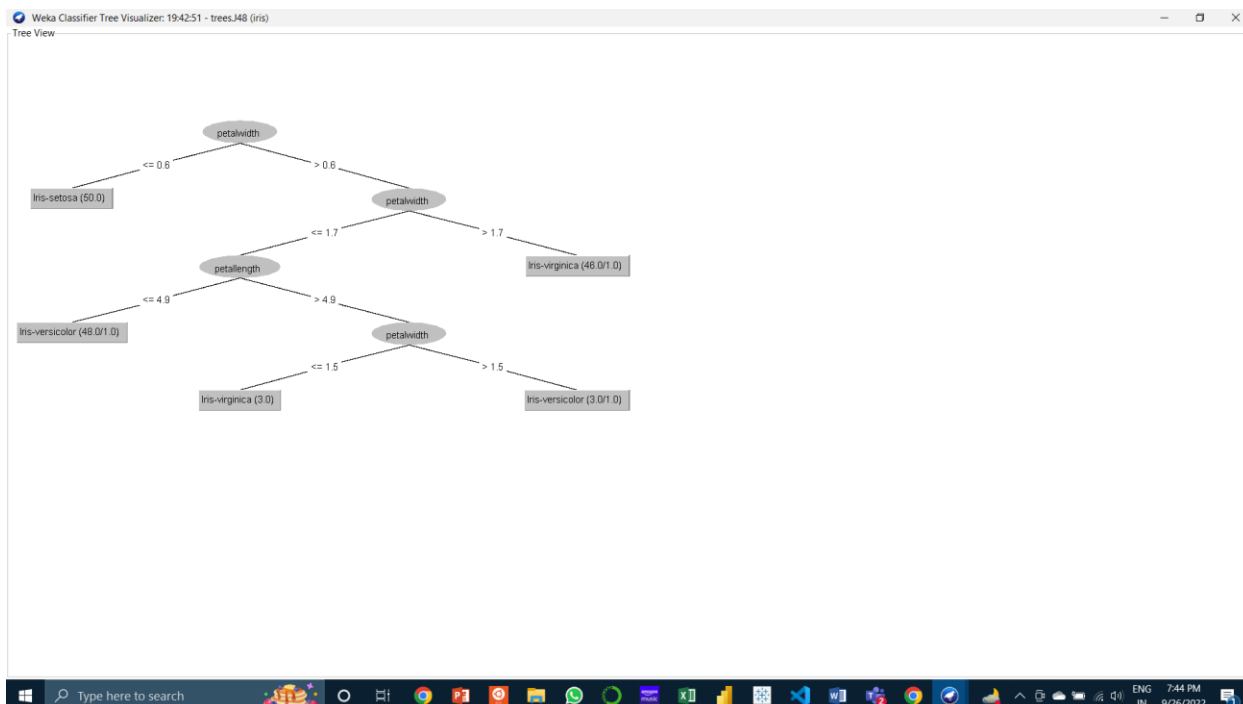
Result list (right-click for options)
01:04:30 - trees.J48
01:05:20 - trees.J48
01:06:27 - lazy.KStar
01:06:32 - lazy.KStar
01:07:36 - lazy.KStar
01:07:39 - lazy.KStar
01:08:17 - trees.J48
01:08:54 - trees.J48

Classifier output
petallength > 4.9
| petalwidth <= 1.5: Iris-virginica (3.0)
| petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| petalwidth > 1.7: Iris-virginica (46.0/1.0)
Number of Leaves : 5
Size of the tree : 9
Time taken to build model: 0 seconds
Time taken to test model on test split: 0 seconds
=== 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 Area PRC Area Class
0.980 0.000 1.000 0.980 0.990 0.985 0.990 0.987 Iris-setosa
0.940 0.030 0.940 0.940 0.940 0.910 0.952 0.880 Iris-versicolor
0.960 0.030 0.941 0.960 0.950 0.925 0.961 0.905 Iris-virginica
Weighted Avg. 0.960 0.020 0.960 0.960 0.960 0.940 0.968 0.924
=== 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

This is J48 cross validation with 10 folds and correctly classified instances at 96% and mean absolute error as 0.035

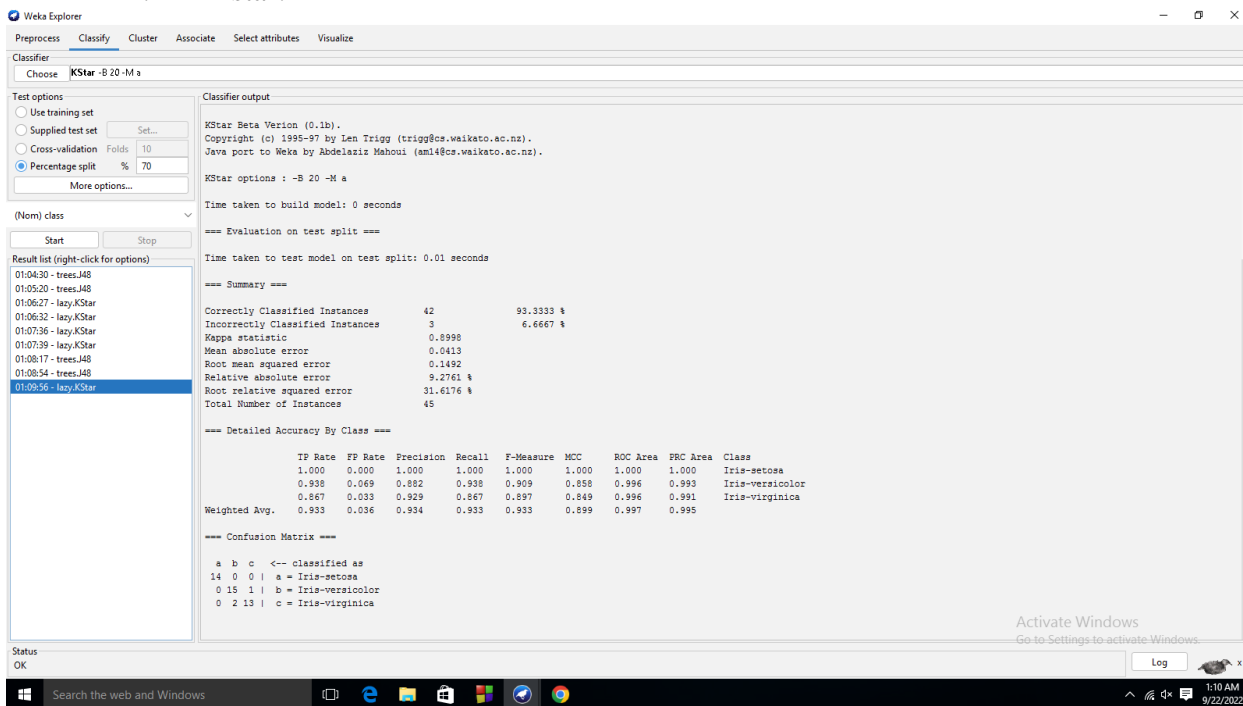


A.Y. 2022-2023



This is the visualization of J48 tree

iii. K-star:



This is K-Star percentage split 70/30 with correctly classified instances at 93.33% with mean absolute error being 0.0413



A.Y. 2022-2023

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose **KStar -B 20 -M a**

Test options:
☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds: 10
☐ Percentage split % 70
More options...

(Nom) class: Start Stop

Result list (right-click for options):
01:04:30 - treesJ48
01:05:20 - treesJ48
01:06:27 - lazyKStar
01:06:32 - lazyKStar
01:07:36 - lazyKStar
01:07:39 - lazyKStar
01:08:17 - treesJ48
01:08:54 - treesJ48
01:09:56 - lazyKStar
01:10:20 - lazyKStar

Classifier output:
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
KStar Beta Version (0.1b).
Copyright (c) 1995-97 by Len Trigg (trigg@cs.waikato.ac.nz).
Java port to Weka by Abdelaziz Mahoui (am14@cs.waikato.ac.nz).
KStar options: -B 20 -M a
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 142 94.6667 %
Incorrectly Classified Instances 8 5.3333 %
Kappa statistic 0.92
Mean absolute error 0.0429
Root mean squared error 0.1555
Relative absolute error 9.658 %
Root relative squared error 32.9823 %
Total Number of Instances 150
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Iris-setosa
	0.920	0.040	0.920	0.920	0.920	0.880	0.994	0.989	Iris-versicolor
	0.920	0.040	0.920	0.920	0.920	0.880	0.994	0.988	Iris-virginica
Weighted Avg.	0.947	0.027	0.947	0.947	0.947	0.920	0.996	0.992	

=== Confusion Matrix ===
a b c <-- classified as
50 0 0 | a = Iris-setosa
0 46 4 | b = Iris-versicolor
0 4 46 | c = Iris-virginica

Status: OK

This is the K-Star cross-validation with 10 folds having correctly classified instances at 94.67% and mean absolute error being 0.0429

3. Clustering:

i. K-Means

Weka Explorer

Preprocess **Cluster** Associate Select attributes Visualize

Clusterer: Choose **SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance" -R first-last -I 500 -num-slots 1 -S 10**

Cluster mode:
☒ Use training set
☐ Supplied test set
☐ Percentage split % 66
☐ Classes to clusters evaluation
(Nom) ExtremeValue
☒ Store clusters for visualization

Ignore attributes: Start Stop

Result list (right-click for options):
01:13:12 - SimpleKMeans

Clusterer output:
KMeans
=====

Number of iterations: 7
Within cluster sum of squared errors: 62.1436882815797
Initial starting points (random):
Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor,no,no
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor,no,no
Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data (150.0)	Cluster# 0 (100.0)	Cluster# 1 (50.0)
sepal.length	5.8433	6.262	5.006
sepal.width	3.054	2.872	3.418
petal.length	3.7587	4.906	1.464
petal.width	1.1987	1.676	0.244
class	Iris-setosa Iris-versicolor Iris-setosa		
Outlier	no	no	no
ExtremeValue	no	no	no

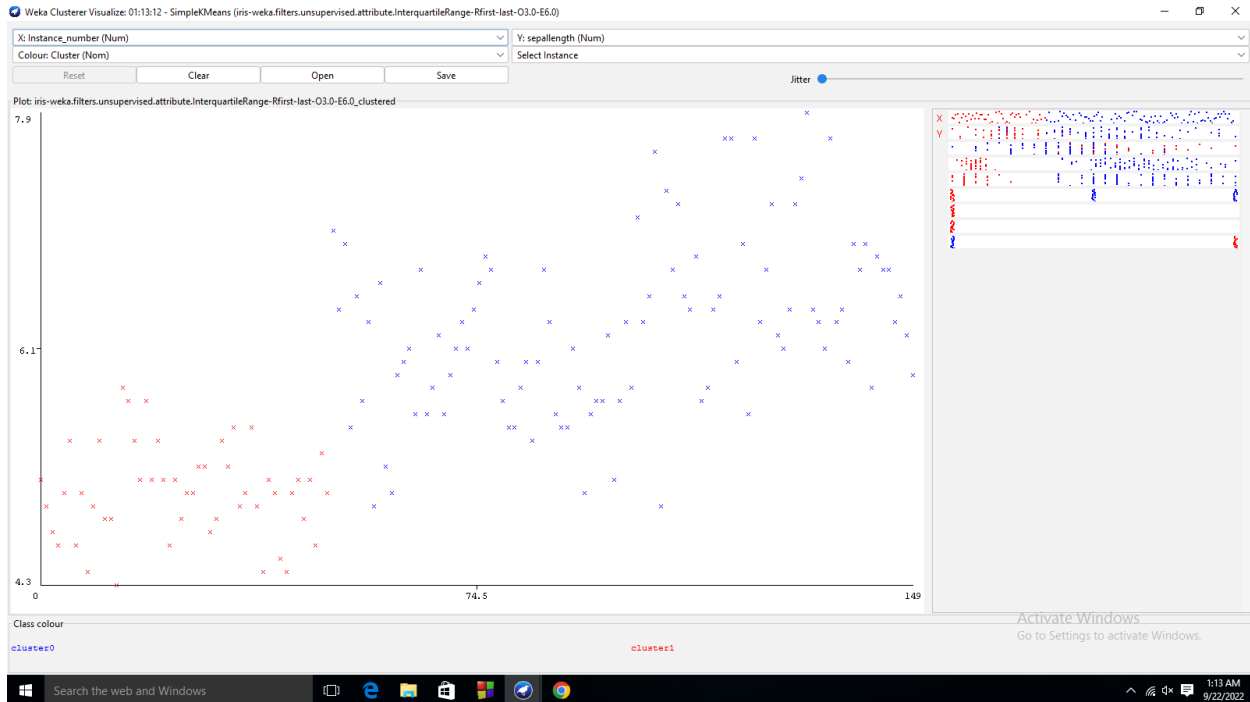
Time taken to build model (full training data): 0 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 100 (67%)
1 50 (33%)

Status: OK

This is output of the model of K-Means with cluster 0 having 67% and cluster 1 having 33%

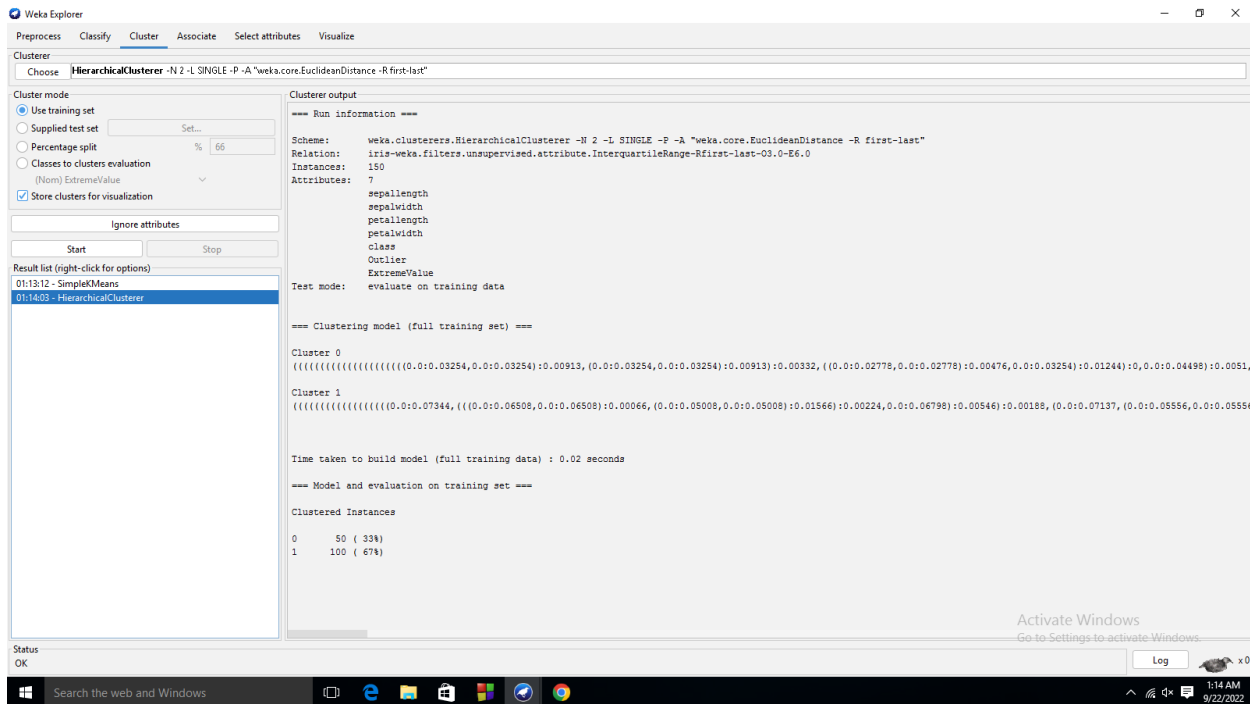


A.Y. 2022-2023



This is the scatter plot for K-Means cluster assignment where we can conclude that there are two clusters one at the beginning and second in the middle also there is a upward trend.

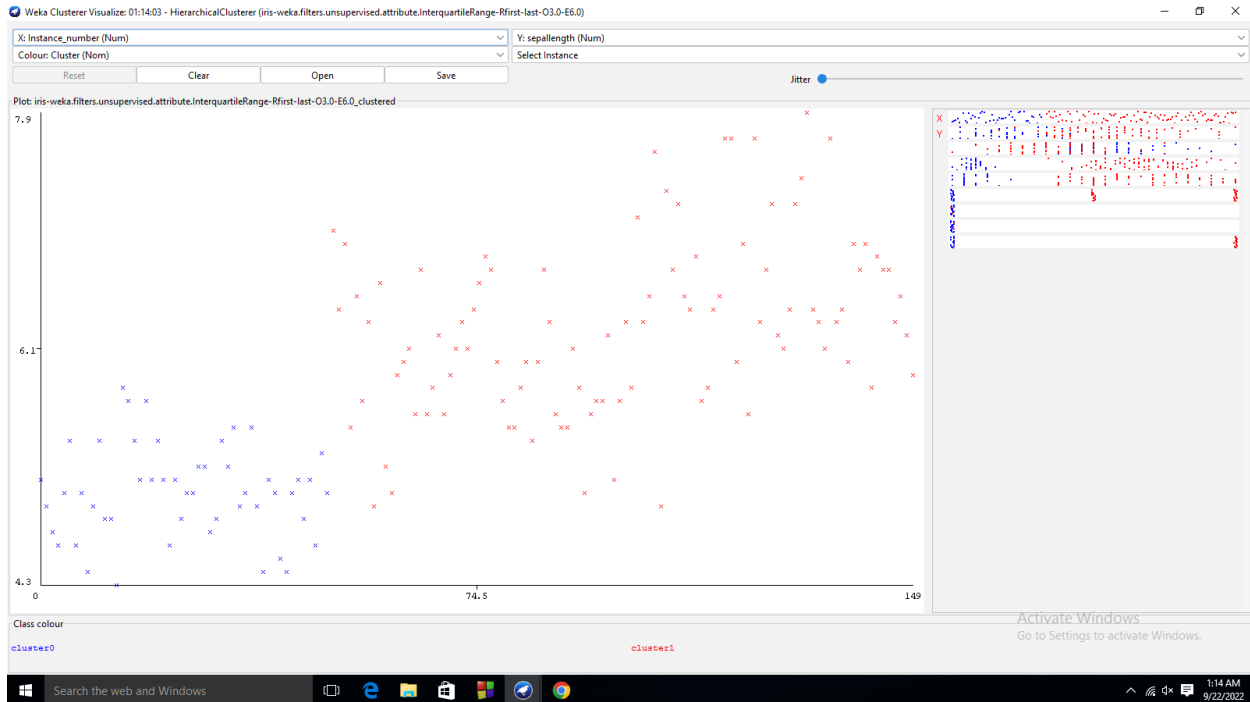
ii. Hierarchical Cluster:



The percentage split of cluster is same as above.



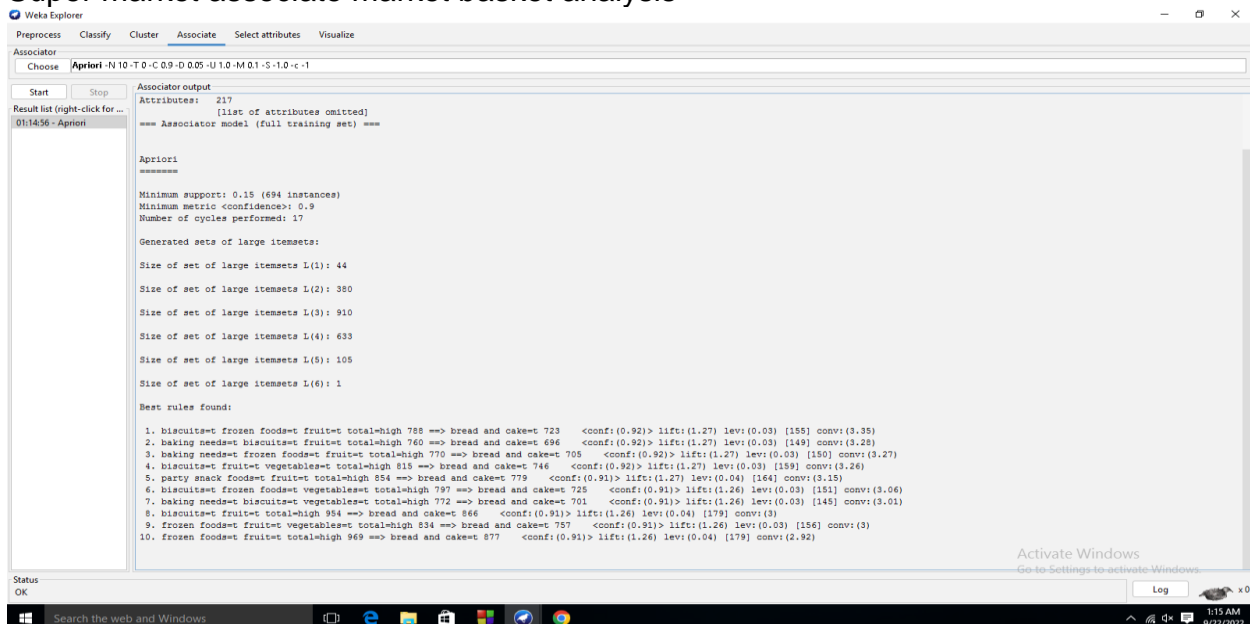
A.Y. 2022-2023



This is scatter plot of hierarichal cluster assignment . According to me, data is evenly distributed with upward trend and two cluster forming one at the start and other at middle.

4. Association Rule:

Super market associate market basket analysis



Confidence level of top 10 rules is either 0.92 or 0.91 .We found that there is high relation between biscuits,frozen foods and fruits and many more.



A.Y. 2022-2023

5. Select Attributes:

i. Gain Ratio:

The screenshot shows the Weka Explorer interface with the 'Select attributes' tab active. The 'Attribute Evaluator' is set to 'GainRatioAttributeEval' and the 'Search Method' is 'Ranker -T -1.7976931348623157E300 -N -1'. The 'Attribute Selection Mode' is 'Use full training set'. The 'Result list' shows '00:50:50 - Ranker - GainRatioAttributeEval'. The 'Attribute selection output' pane displays the following information:

```
=== Run information ===
Evaluator: weka.attributeSelection.GainRatioAttributeEval
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E300 -N -1
Relation: iris
Instances: 150
Attributes: 5
  sepalwidth
  sepalwidth
  petalwidth
  petalwidth
  class
Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===
Search Method:
  Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 5 class):
  Gain Ratio feature evaluator

Ranked attributes:
0.871 4 petalwidth
0.734 3 petalwidth
0.381 1 sepalwidth
0.242 2 sepalwidth
Selected attributes: 4,3,1,2 : 4
```

Petal width has the highest rank with gain ratio 0.871 and sepal width has the lowest rank with gain ratio 0.242

ii. Info Gain Ratio

The screenshot shows the Weka Explorer interface with the 'Select attributes' tab active. The 'Attribute Evaluator' is set to 'InfoGainAttributeEval' and the 'Search Method' is 'Ranker -T -1.7976931348623157E300 -N -1'. The 'Attribute Selection Mode' is 'Use full training set'. The 'Result list' shows '00:52:02 - Ranker - InfoGainAttributeEval'. The 'Attribute selection output' pane displays the following information:

```
=== Run information ===
Evaluator: weka.attributeSelection.InfoGainAttributeEval
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E300 -N -1
Relation: iris
Instances: 150
Attributes: 5
  sepalwidth
  sepalwidth
  petalwidth
  petalwidth
  class
Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===
Search Method:
  Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 5 class):
  Information Gain Ranking Filter

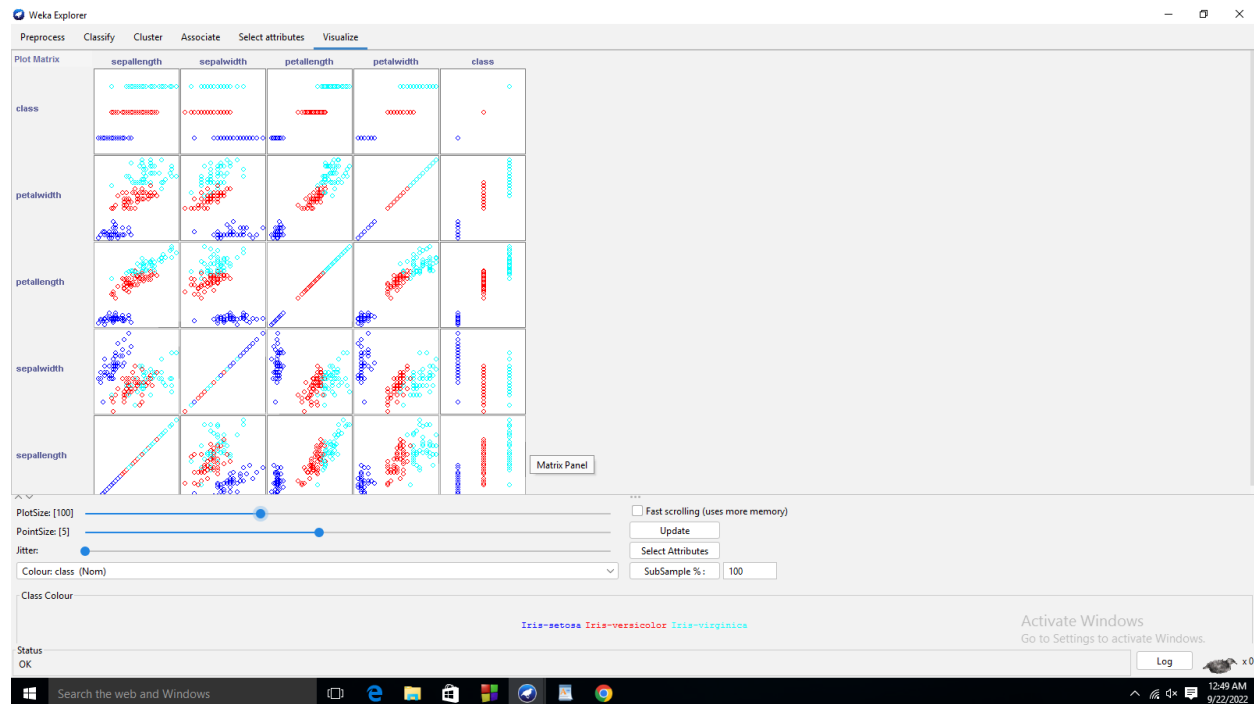
Ranked attributes:
1.418 3 petalwidth
1.378 4 petalwidth
0.696 1 sepalwidth
0.376 2 sepalwidth
Selected attributes: 3,4,1,2 : 4
```

Petal length has the highest info gain with 1.418 and sepal width with least of 0.376



A.Y. 2022-2023

6. Visualization:



Petal width and Petal length are related by upward trendline

For lower value of petal length we find values of sepal width to be more and we find it to be downward line.

Sepal length and Sepal Width are closely related with all the clusters in close proximity.

Conclusion:

J48 ,Naïve Bayes had highest correct classified instances with 96%.

Confidence level of top rules was around 0.92