
**S. B. JAIN INSTITUTE OF TECHNOLOGY, MANAGEMENT & RESEARCH, NAGPUR.**

# Practical No. 2

**Aim:** Demonstration on Discretize the attribute using Weka Tool on the training data set ionosphere and compare it with different binning values and equal frequency binning.

**Name of Student: Shrutika Pradeep Bagdi**

**Roll No.:** **CS22130**

**Semester/Year:** **6th / 3rd**

**Academic Session: 2024-2025**

**Date of Performance:**

**Date of Submission:**

**AIM:** Demonstration on Discretize the attribute using Weka Tool on the training data set ionosphere and compare it with different binning values and equal frequency binning.

**OBJECTIVE/EXPECTED LEARNING OUTCOME:**

The objectives and expected learning outcome of this practical are:

- The goal of discretization is to reduce the number of values a continuous variable assumes by grouping them into a number, b, of intervals or bins.
- In statistics and machine learning, discretization refers to the process of converting or partitioning continuous attributes, features or variables to discretized or nominal attributes /features/ variables/ intervals.

**HARDWARE AND SOFTWARE REQUIRMENTS:**

**Hardware Requirement:**

**Software Requirement:**

**THEORY:**

**Discretization in Data Mining:-**

Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization. Supervised discretization refers to a method in which the class data is used. Unsupervised discretization refers to a method depending upon the way which operation proceeds. It means it works on the top-down splitting strategy and bottom-up merging strategy.

Discretization is **the process through which we can transform continuous variables, models or functions into a discrete form**. We do this by creating a set of contiguous intervals (or bins) that go across the range of our desired variable/model/function. Continuous data is Measured, while Discrete data is Counted

Two key problems in association with discretization are how to select the number of intervals or bins and how to decide on their width. Discretization can be performed with or without taking class information, , into account. These are the supervised and unsupervised ways. If class labels were known in the training data, the discretization method ought to take advantage of it, especially if the subsequently used learning algorithm for model building is supervised.

A short description of two unsupervised techniques follows.

**Equal-width discretization**

The algorithm first finds the minimum and maximum values of every variable, Xi, and then divides this range into a number, mXi, of user-specified, equal-width intervals.

**Equal-frequency discretization**

The algorithm determines the minimum and maximum values of the variable, sorts all values in ascending order, and divides the range into a user-defined number of intervals, in such a way that every interval contains the equal number of sorted values.

**Techniques of data discretization:-**

**Binning**

Binning refers to a data smoothing technique that helps to group a huge number of continuous values into smaller values. For data discretization and the development of idea hierarchy, this technique can also be used.

**Histogram analysis**

Histogram refers to a plot used to represent the underlying frequency distribution of a continuous data set. Histogram assists the data inspection for data distribution. For example, Outliers, skewness representation, normal distribution representation, etc.

**Cluster Analysis:**

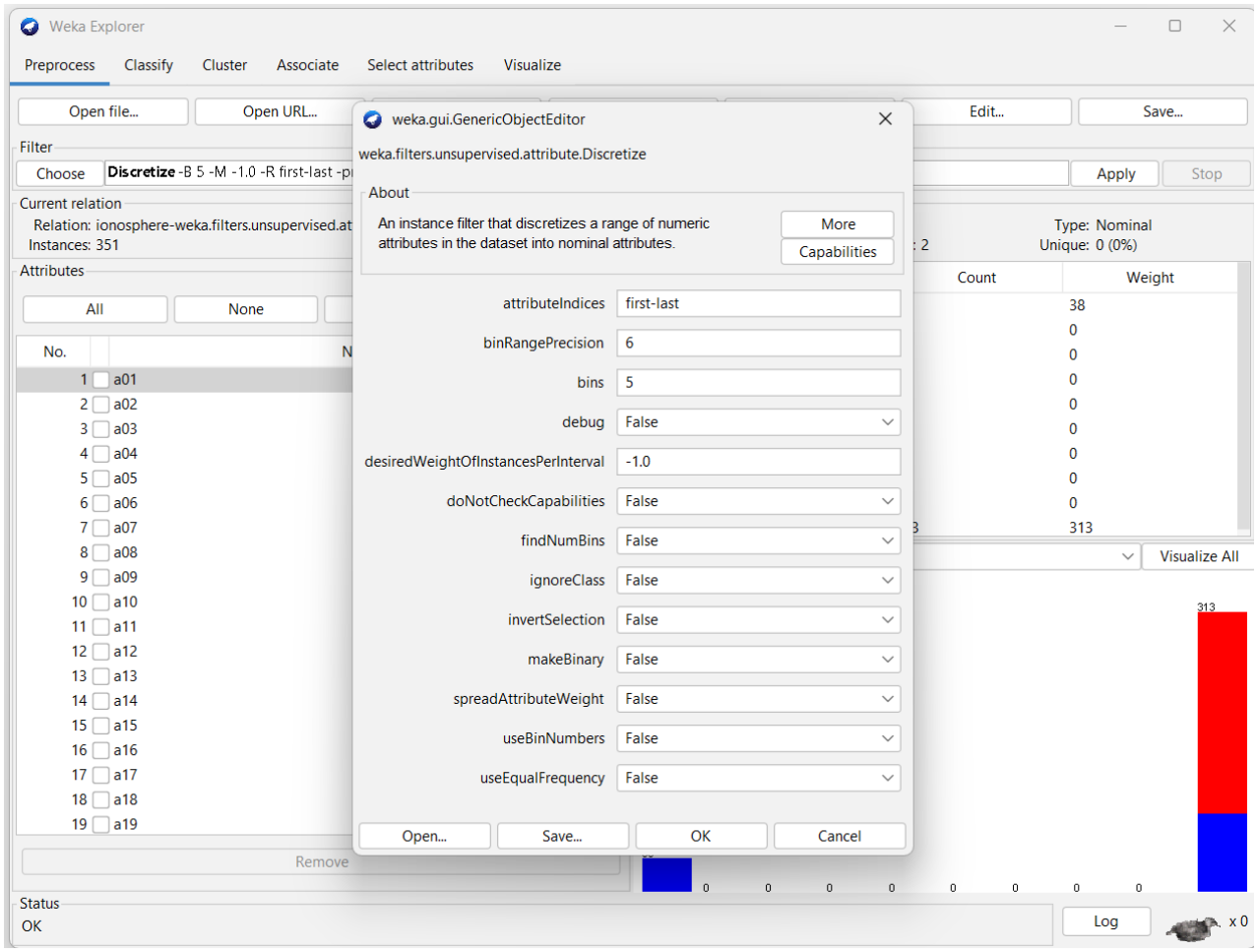
Cluster analysis is a form of data discretization. A clustering algorithm is executed by dividing the values of x numbers into clusters to isolate a computational feature of x.

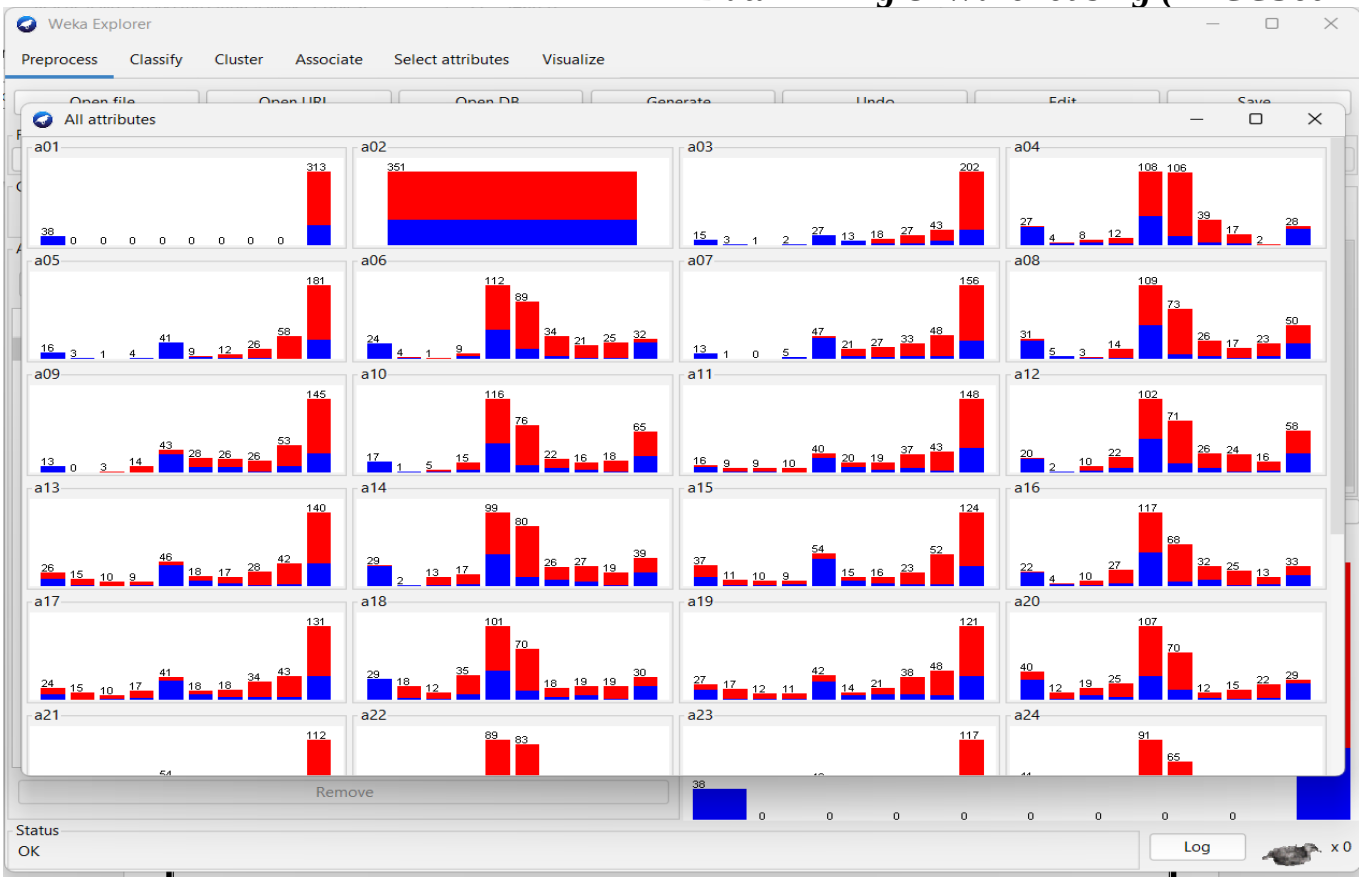
**Description:**

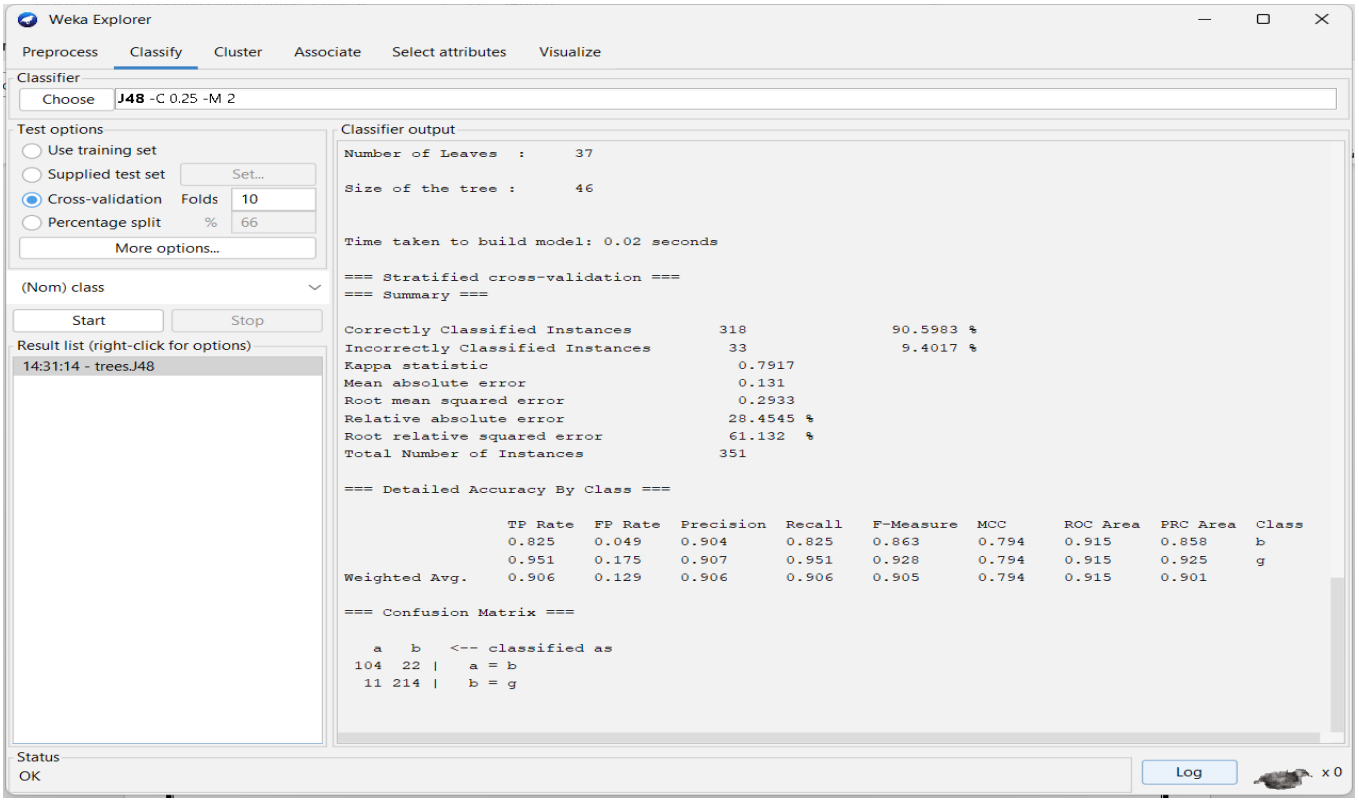
We need to use ionosphere data set which is already present in Weka training sets.

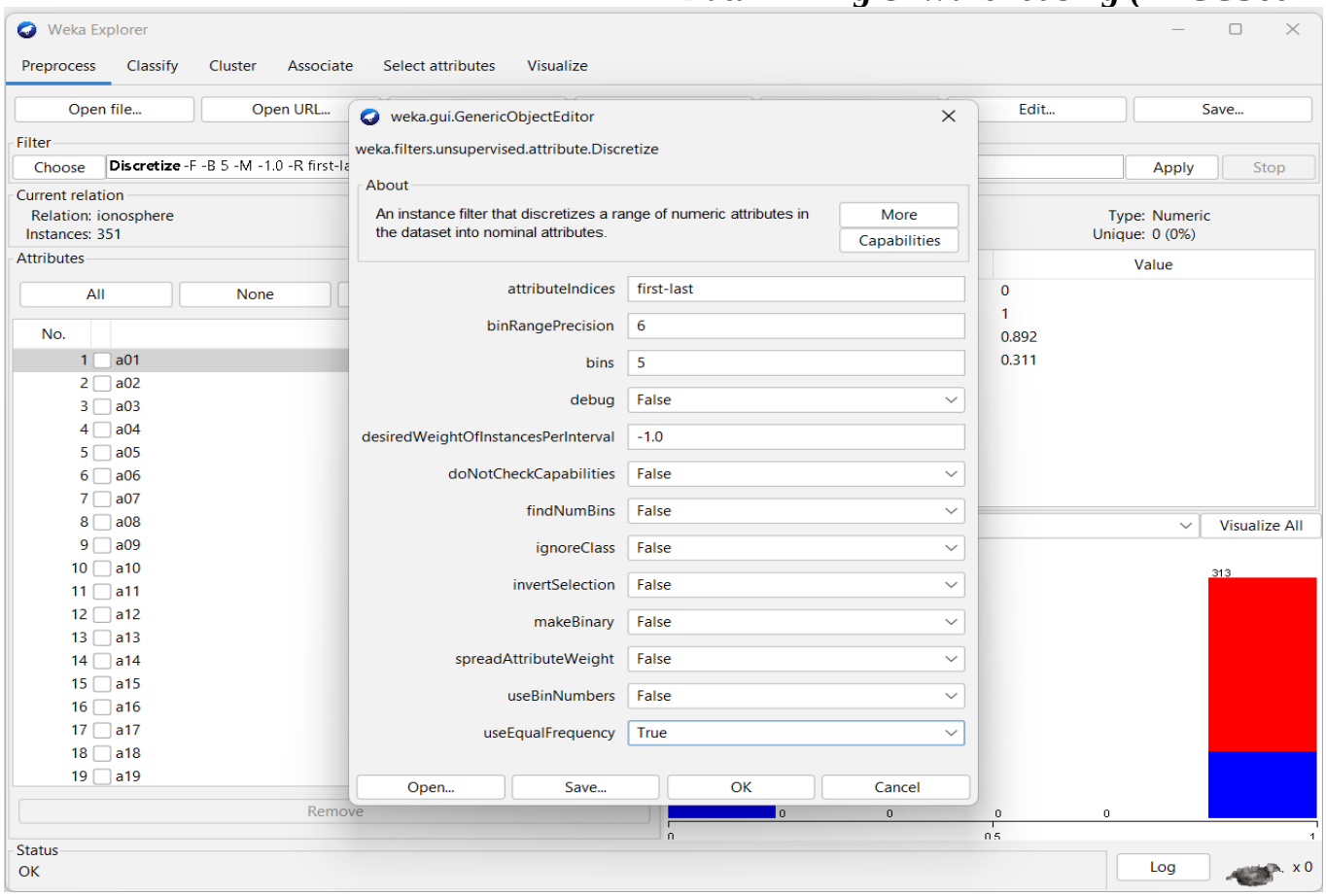
**Procedure:**

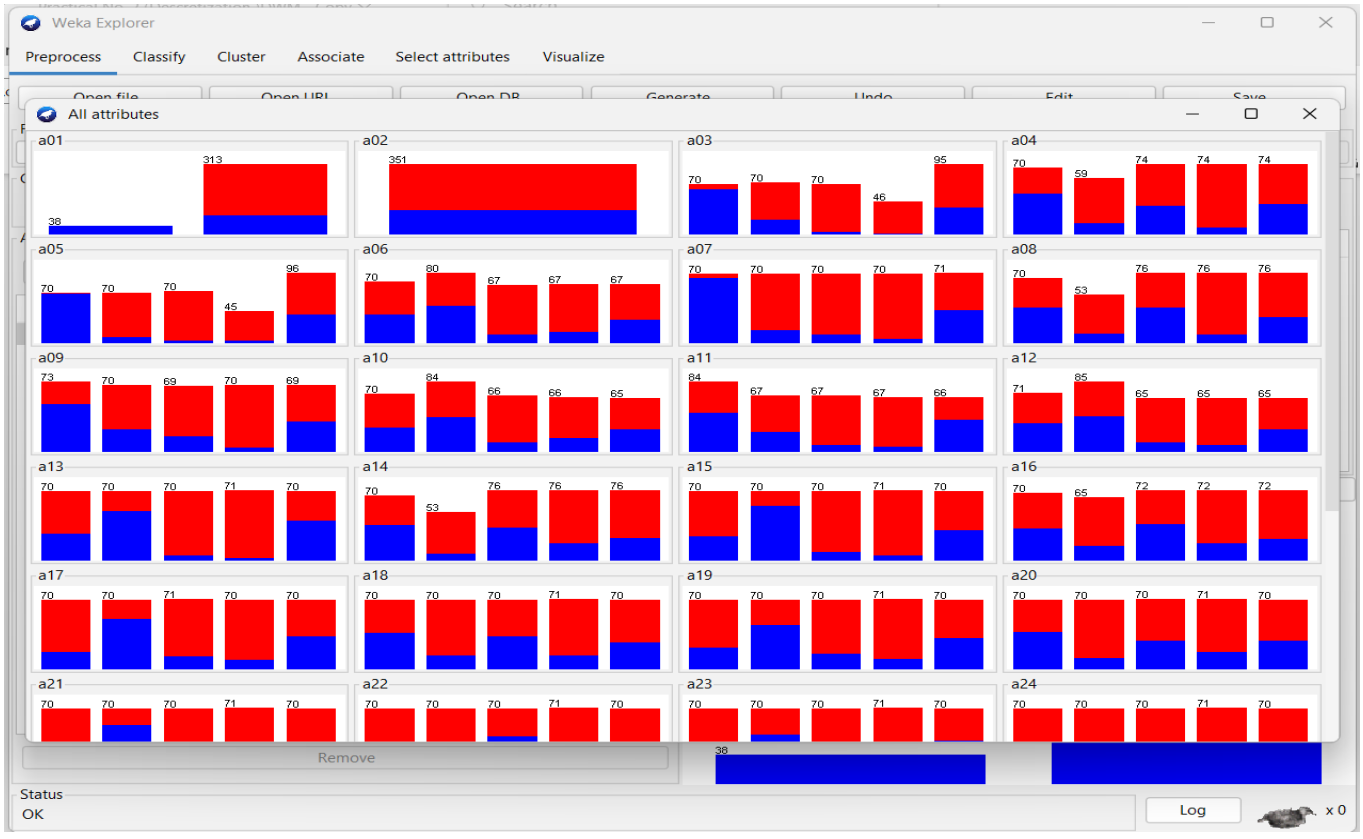
**OUTPUT (SCREENSHOTS):**

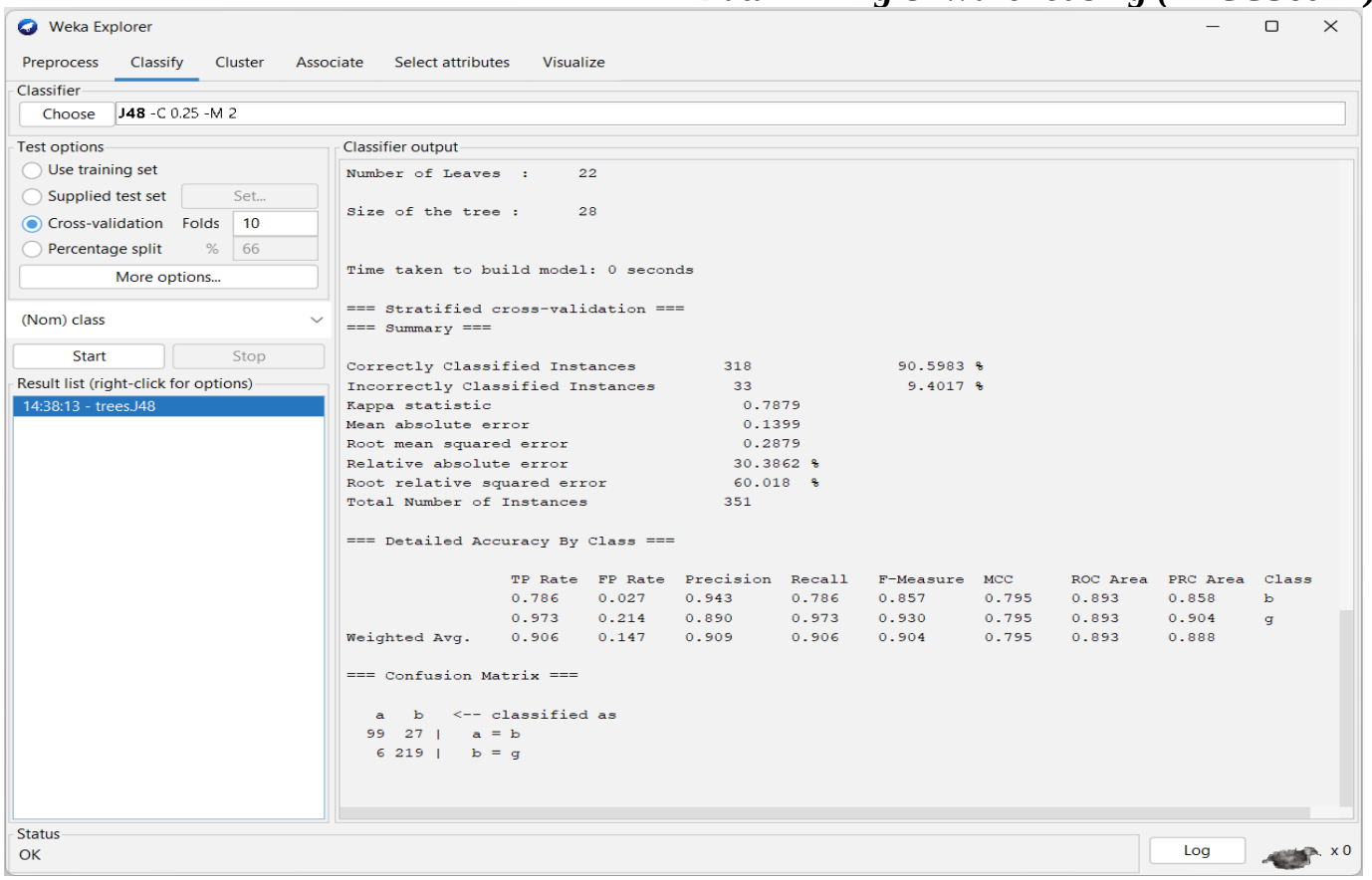



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
All attributes
a01
a02
a03
a04
a05
a06
a07
a08
a09
a10
a11
a12
a13
a14
a15
a16
a17
a18
a19
a20
a21
a22
a23
a24
Remove
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
14:31:14 - trees.J48
Classifier output
Number of Leaves : 37
Size of the tree : 46
Time taken to build model: 0.02 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 318 90.5983 %
Incorrectly Classified Instances 33 9.4017 %
Kappa statistic 0.7917
Mean absolute error 0.131
Root mean squared error 0.2933
Relative absolute error 28.4545 %
Root relative squared error 61.132 %
Total Number of Instances 351
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.825 0.049 0.904 0.825 0.863 0.794 0.915 0.858 b
0.951 0.175 0.907 0.951 0.928 0.794 0.915 0.925 g
Weighted Avg. 0.906 0.129 0.906 0.906 0.905 0.794 0.915 0.901
=== Confusion Matrix ===
a b <-- classified as
104 22 | a = b
11 214 | b = g
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Edit...
Save...
Filter
Choose
Discretize -F -B 5 -M -1.0 -R first-la
Apply
Stop
Current relation
Relation: ionosphere
Instances: 351
Attributes
All
None
No.
1 a01
2 a02
3 a03
4 a04
5 a05
6 a06
7 a07
8 a08
9 a09
10 a10
11 a11
12 a12
13 a13
14 a14
15 a15
16 a16
17 a17
18 a18
19 a19
Remove
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Discretize
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
Capabilities
attributeIndices first-last
binRangePrecision 6
bins 5
debug False
desiredWeightOfInstancesPerInterval -1.0
doNotCheckCapabilities False
findNumBins False
ignoreClass False
invertSelection False
makeBinary False
spreadAttributeWeight False
useBinNumbers False
useEqualFrequency True
Open...
Save...
OK
Cancel
Type: Numeric
Unique: 0 (0%)
Value
0
1
0.892
0.311
Visualize All
313
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
All attributes
a01
a02
a03
a04
a05
a06
a07
a08
a09
a10
a11
a12
a13
a14
a15
a16
a17
a18
a19
a20
a21
a22
a23
a24
Remove
Status
OK
Log
x 0




**CONCLUSION:**

**DISCUSSION AND VIVA VOCE:**

- What is Classification?
- What is Classification Accuracy?
- What is data discretization in data mining?
- Why is discretization needed in data mining?
- What is binning?
- What is the J48 Classifier?

**REFERENCE:**


- http://www.cs.waikato.ac.nz/ml/weka/documentation.html
- http://ai.fon.bg.ac.rs/wp-content/uploads/2015/04/ML-Attribute-Discretisation-and-Selection-

Clustering-2014_eng.pdf

- https://weka.wikispaces.com/Discretizing+datasets

Data Mining – Concepts and Techniques, Jiawei Han & Micheline Kamber, Morgan Kauf-mann Publishers, Elsevier, 2nd Edition, 2006.

| Observation book: (3) | Viva-Voce (3) | `Quality of Submission and timely Evaluation (4) |
|---|---|---|
| | | |
| **Total:** | | **Sign with date:** |