



# **Vidyavardhini's College of Engineering & Technology**

Department of Computer Engineering

Academic Year: 2025-26

---

Experiment No.9
Clustering, Classification and Association Data Mining using WEKA tool
Date of Performance: 24/9/25
Date of Submission: 01/10/25
Name:Sumit Metkari
Div/Roll no.: 2 / 32



# Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year: 2025-26

---

**Aim:** To implement clustering , classification and association data mining by using WEKA

**Objective:** Simulate K-Means Algorithm, Single Linkage Algorithm Decision tree induction and apriori algorithm by using WEKA

## Theory:

WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces.

### 1) K-Means Algorithm using WEKA

#### EXAMPLE:

Dataset:  $D = \{1, 2, 3, 8, 9, 10, 25\}$

1. Randomly assign means  $m1 = 3$  and  $m2 = 10$

$k1 = \{1, 2, 3\}$        $k2 = \{8, 9, 10, 25\}$

2.  $m1 = 2$  and  $m2 = 13$

$k1 = \{1, 2, 3\}$        $k2 = \{8, 9, 10, 25\}$

#### WEKA Code:

```
@RELATION iris
```

```
@ATTRIBUTE x NUMERIC
```

```
@DATA
```

```
1
```

```
2
```

```
3
```



8  
9  
10  
25

Weka Explorer

Preprocess Classify **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

**Cluster mode**

☒ Use training set  
☐ Supplied test set Set...  
☐ Percentage split % 66  
☐ Classes to clusters evaluation (Num) x  
☒ Store clusters for visualization

Ignore attributes

Start Stop

**Clusterer output**

Within cluster sum of squared errors: 0.3402777777777778

Initial starting points (random):

Cluster 0: 3  
Cluster 1: 1

Missing values globally replaced with mean/mode

Final cluster centroids:

	Cluster#		
Attribute	Full Data	0	1
	(7.0)	(4.0)	(3.0)

=====  
x

	0.2857	13	2
--	--------	----	---

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	4 ( 57%)
1	3 ( 43%)

**Result list (right-click for options)**

14:30:49 - SimpleKMeans



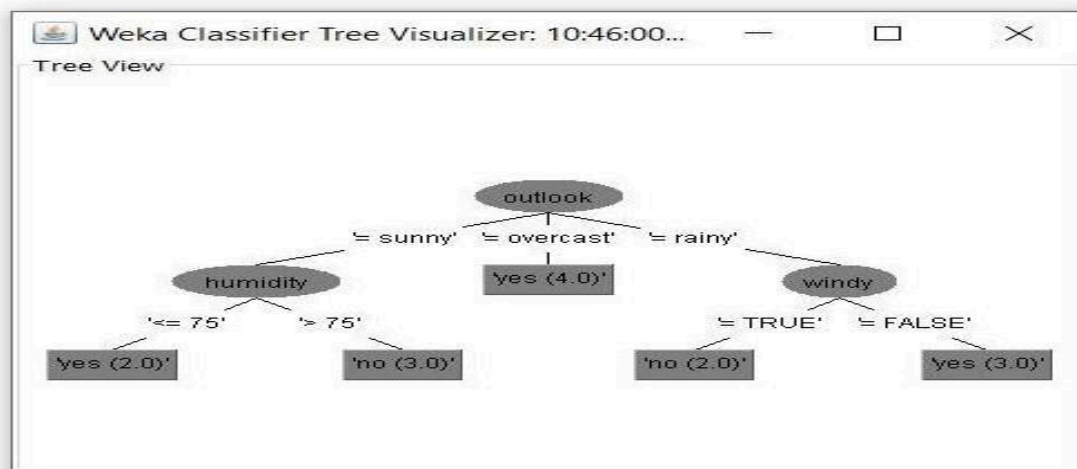
## 2) Decision Tree Induction using WEKA

A decision tree is a flowchart like tree structure, where each internal node(non-leaf node) denotes a test on an attribute,each branch represents an outcome of the test,and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node

Example:-

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

Output:-





### 3) Apriori Algorithm using WEKA

In this current world, globalization is the main feature of any environment. Everyone has to be update, fast and forward and information is the main element for it. For survival in this world it's the basic need to use and to store the information means to prepare a proper database or dataset to analyze. Using and storing the database is not an issue, but finding the relevant dataset or to analyze the meaningful dataset for a particular aspect, from the junkyard of the database is very big problem in analysis of a specific part of the database. To solve this problem the concept of data mining is used to abstracts the desirable information. Useful information from the large databases has been extracted in the form of the association rules. There are many algorithms have been developed to extract the association rules from the large databases. Apriori algorithm is the most popular algorithm to extract the association rules from the databases.

#### Example

TID	Items
1	A,B,C,D,G,H
2	A,B,C,D,E,F,H
3	B,C,D,E,H
4	B,E,G,H
5	A,B,D,E,G,H
6	A,C,F,G,H
7	B,D,E,G,H
8	A,C,D,E,G,H
9	B,C,D,E,H
10	A,C,E,F,H
11	C,E,H
12	A,D,E,F,H
13	B,C,E,F,H
14	A,B,C,F,H
15	A,B,E,F,H



# Vidyavardhini's College of Engineering & Technology

## Department of Computer Engineering

### Academic Year: 2025-26

## Output

```
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Associate
Choose Apriori H 10 -T 0 -C 0.9 -D 0.08 -M 1.0 -S 0.1 -B -1.0 -E -1
Start Stop
Result list (right-click)
10:35:30 - Apriori
Associate output
=== Run information ===
Scheme: weka.associations.Apriori -H 10 -T 0 -C 0.9 -D 0.08 -M 1.0 -S 0.1 -B -1.0 -E -1
Relation: TEST_DATA
Instances: 15
Attributes: 8
A
B
C
D
E
F
G
H
=== Associate model (Full training set) ===
Apriori
=====
Minimum supports: 0.5 (7 instances)
Minimum metric (confidence): 0.5
Number of cycles performed: 10
Generated sets of large itemsets:
Size of set of large itemsets L(1): 10
Size of set of large itemsets L(2): 12
Size of set of large itemsets L(3): 9
Best rules found:
1. A=TRUE => B=TRUE 11 conf:(1)
2. B=TRUE 10 => B=TRUE 10 conf:(1)
3. C=TRUE 10 => B=TRUE 10 conf:(1)
4. A=TRUE 9 => B=TRUE 9 conf:(1)
5. A=FALSE 9 => B=TRUE 9 conf:(1)
6. D=TRUE 9 => B=TRUE 9 conf:(1)
7. B=FALSE 9 => B=TRUE 9 conf:(1)
8. D=FALSE 7 => B=TRUE 7 conf:(1)
9. F=TRUE 7 => B=TRUE 7 conf:(1)
10. B=TRUE &=TRUE 7 => B=TRUE 7 conf:(1)
```



# Vidyavardhini's College of Engineering & Technology

## Department of Computer Engineering

### Academic Year: 2025-26

Output:

Clustering:

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Cluster 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
- ☐ Classes to clusters evaluation
- ☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

14:34:15 - SimpleKMeans

Cluster output

	80-89	80-89	80-89
germination	norm	abnorm	norm
plant-growth	norm	abnorm	abnorm
leaves	abnorm	abnorm	abnorm
leafspots-halo	no-yellow-halos	absent	no-yellow-halos
leafspots-marg	w-s-marg	dna	w-s-marg
leafspot-size	gt-l/8	dna	gt-l/8
leaf-shred	absent	absent	absent
leaf-malf	absent	absent	absent
leaf-mild	absent	absent	absent
stem	abnorm	abnorm	norm
lodging	yes	yes	yes
stem-cankers	absent	above-seed	absent
canker-lesion	dna	dk-brown-blt	dna
fruiting-bodies	absent	absent	absent
external-decay	absent	absent	absent
mycelium	absent	absent	absent
int-discolor	none	none	none
sclerotia	absent	absent	absent
fruit-pods	norm	norm	norm
fruit-spots	absent	absent	absent
seed	norm	norm	norm
seed-growth	absent	absent	absent
seed-discolor	absent	absent	absent
seed-size	norm	norm	norm
shriveling	absent	absent	absent
roots	norm	norm	norm
class	brown-spot	phytophthora-root	alternarialeaf-spot

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	261 ( 38%)
1	422 ( 62%)

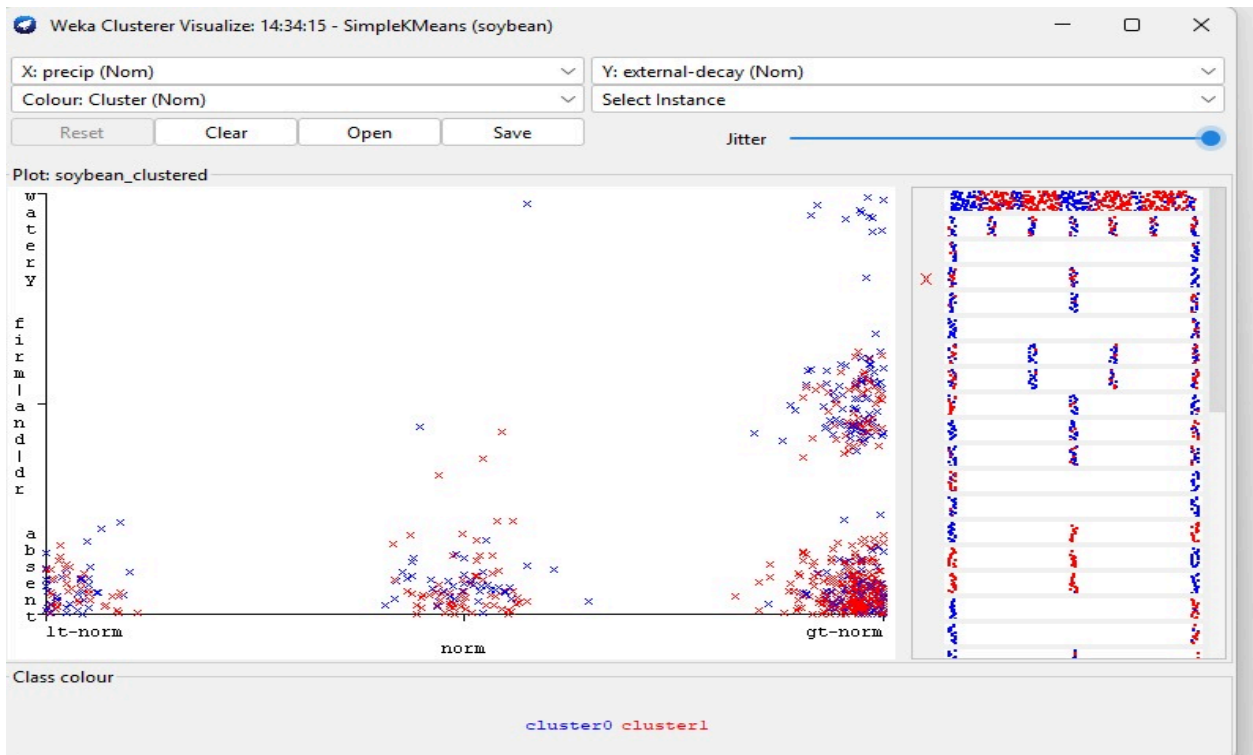
Status OK

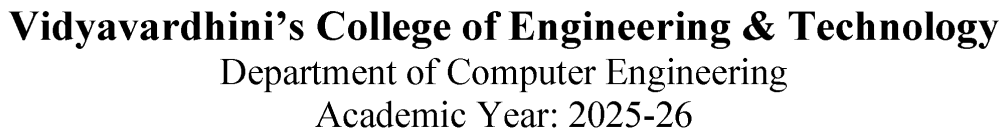
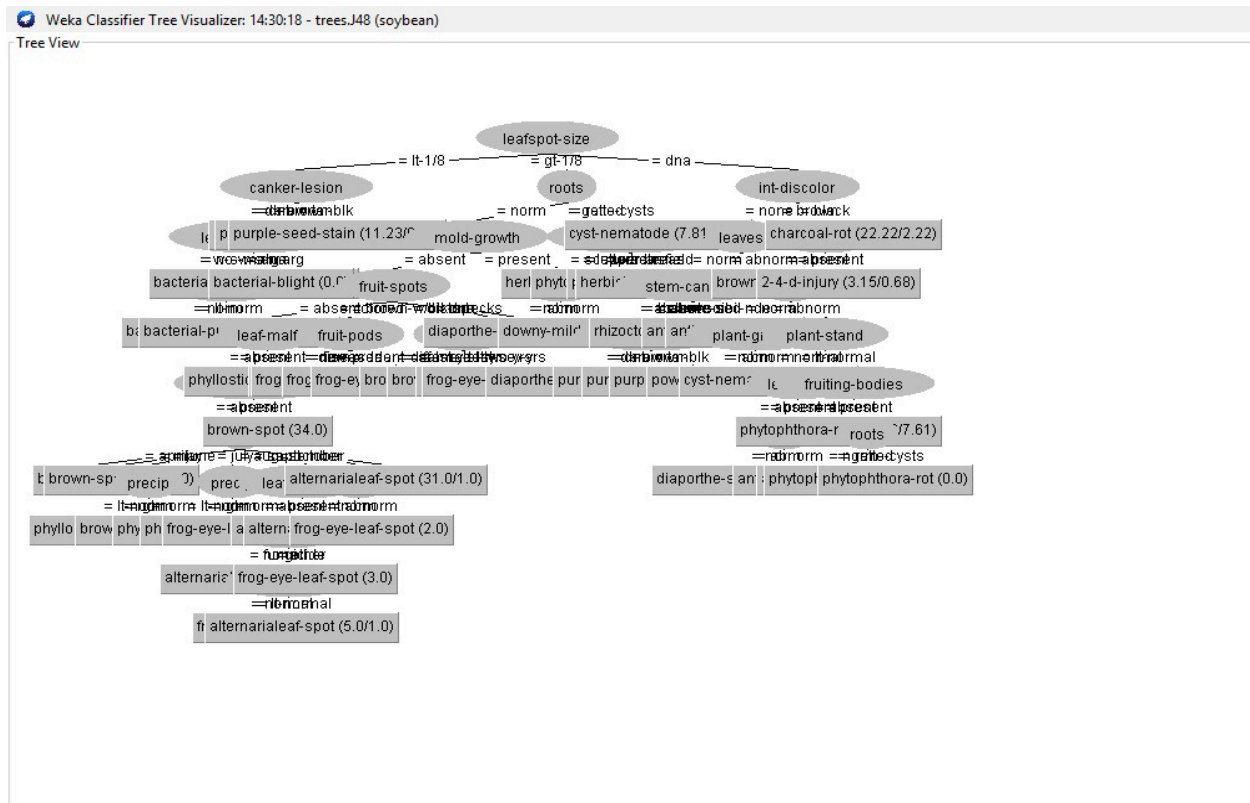
Log

29°C

Search

2:41 PM



[illegible]





# Vidyavardhini's College of Engineering & Technology

## Department of Computer Engineering

### Academic Year: 2025-26

#### Associate - Apriori:

```
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize

Associate
Choose: Apriori-N10-T0-C0.9-D0.05-U1.0-M0.1-S1.0-c-1

Start Stop
Result list (right-click for op...
14:11:59 - FilteredAssociate
14:19:06 - Apriori

Associate output
Attributes: 41
[list of attributes omitted]
=== Associate model (full training set) ===

Apriori
=====

Minimum support: 0.15 (694 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 44
Size of set of large itemsets L(2): 380
Size of set of large itemsets L(3): 910
Size of set of large itemsets L(4): 633
Size of set of large itemsets L(5): 105
Size of set of large itemsets L(6): 1

Best rules found:

1. biscuits=t frozen foods=t fruit=t total-high 788 ==> bread and cake=t 723 <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)
2. baking needs=t biscuits=t fruit=t total-high 760 ==> bread and cake=t 696 <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)
3. baking needs=t frozen foods=t fruit=t total-high 770 ==> bread and cake=t 705 <conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)
4. biscuits=t fruit=t vegetables=t total-high 815 ==> bread and cake=t 746 <conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)
5. party snack foods=t fruit=t total-high 854 ==> bread and cake=t 779 <conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)
6. biscuits=t frozen foods=t vegetables=t total-high 797 ==> bread and cake=t 725 <conf:(0.91)> lift:(1.26) lev:(0.03) [151] conv:(3.06)
7. baking needs=t biscuits=t vegetables=t total-high 772 ==> bread and cake=t 701 <conf:(0.91)> lift:(1.26) lev:(0.03) [145] conv:(3.01)
8. biscuits=t fruit=t total-high 954 ==> bread and cake=t 866 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(3)
9. frozen foods=t fruit=t vegetables=t total-high 834 ==> bread and cake=t 757 <conf:(0.91)> lift:(1.26) lev:(0.03) [156] conv:(3)
10. frozen foods=t fruit=t total-high 969 ==> bread and cake=t 877 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92)

Status
OK
Log x0
```

#### Conclusion:

Through clustering, classification, and association rule mining, WEKA effectively demonstrates how different data mining techniques can be applied to extract meaningful patterns and insights from datasets. The K-Means algorithm groups data points into clusters based on similarity, decision tree induction provides an easy-to-interpret model for classification, and the Apriori algorithm generates useful association rules from large datasets. WEKA simplifies these processes with its intuitive graphical interface, pre-built algorithms, and automated steps, allowing users to focus on analyzing results rather than complex coding. The generated outputs clearly illustrate hidden patterns, decision rules, and associations, making WEKA a powerful and user-friendly tool for practical data mining applications.