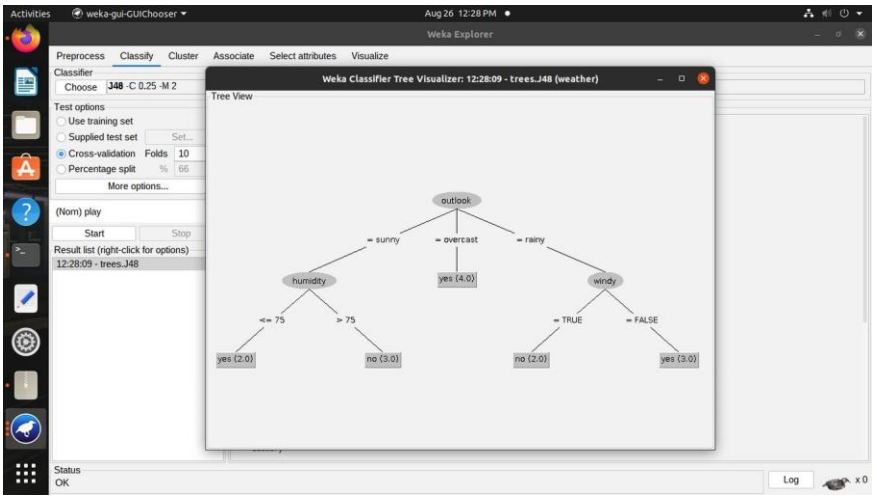
CSL503 Data Warehousing and Mining Lab
Sem V
Roll No.:15
Name – Adithya Menon

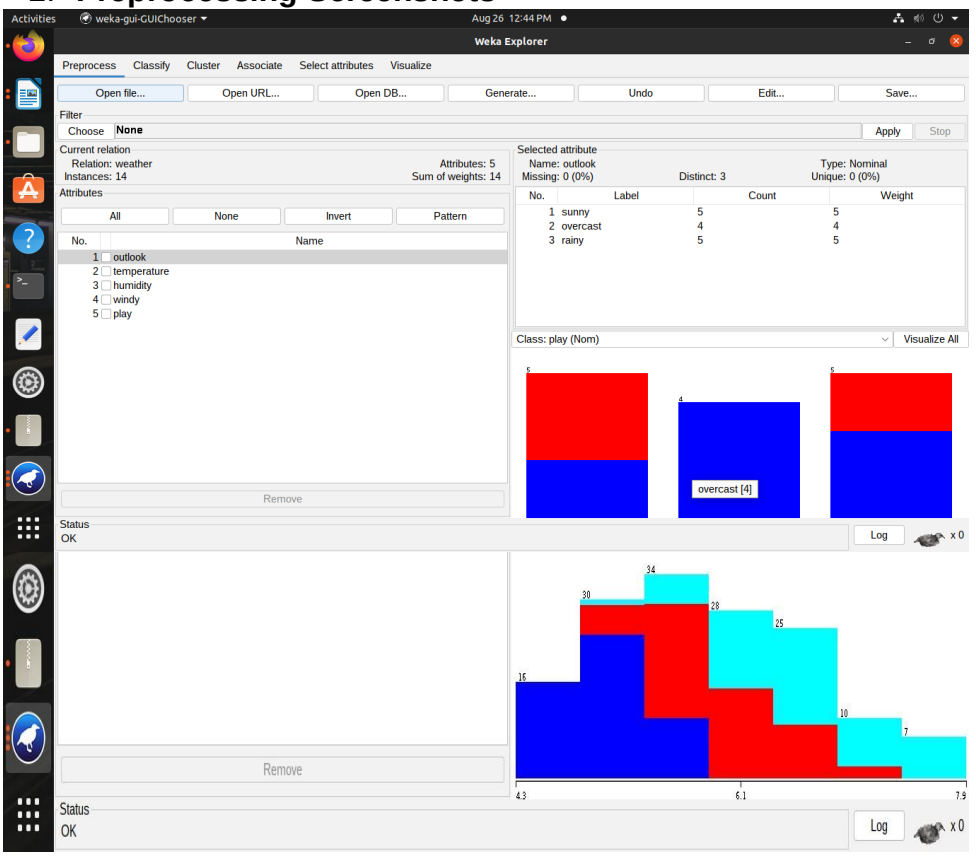
EXPERIMENT 5

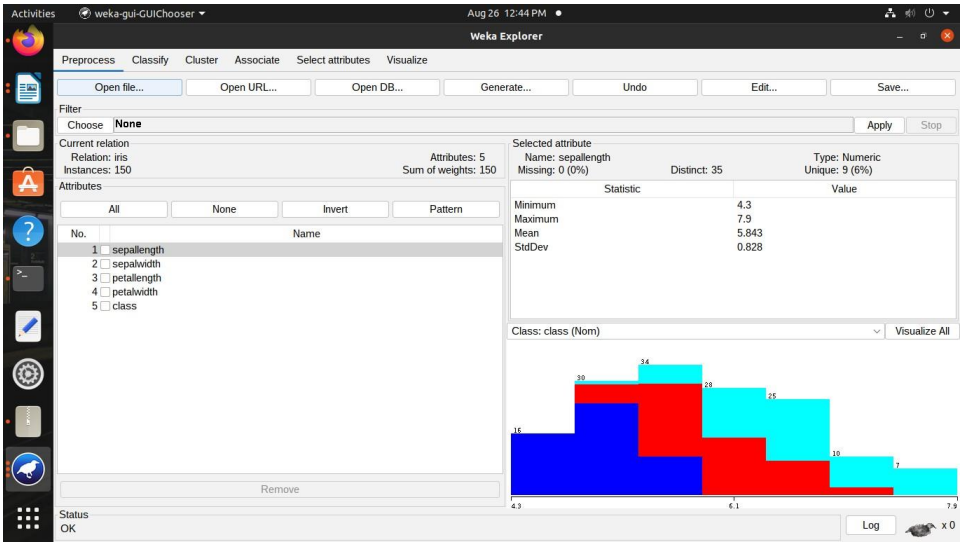
Title	Demonstrating Classification(J48 /ID3),clustering and association rule mining Algorithms using WEKA Tool
Pre requisite	Classification, clustering,association rule mining
Mapping with CO	To apply Data Mining algorithms on a given dataset for a real-time case study and evaluate their performance using Accuracy Measures. (CSL503.4)
Objective	To visualize results on real-time datasets using WEKA tool.
Outcome	To compare the working of various Data Mining algorithms on a given dataset using tools like WEKA ..
Instructions	- Screenshots must be readable - Annotate the important parts of screenshot using pen - Every screenshot must be explained
Delieverables	1. **weather.arff dataset explanation** This dataset is a classic machine learning dataset (often used in Weka). It models the problem of deciding whether to play a game (like tennis) based on weather conditions. • Independent variables (features): outlook, temperature, humidity, windy • Dependent variable (class label): play (yes/no) It's commonly used for decision tree learning (ID3, C4.5, J48, etc.) and other classification tasks.

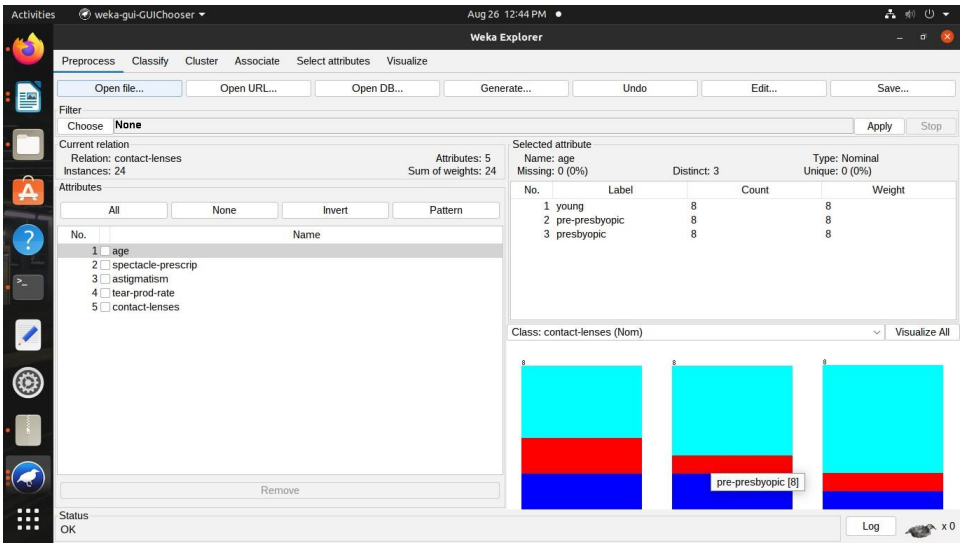
a. **Visualization screenshots of dataset**



2. **Preprocessing Screenshots**







a. **Explain what is happening in pre processing (convert numeric to categorical values)**

The preprocessing step of converting numeric to categorical values is called discretization in Weka. It bins continuous attributes like temperature and humidity into ranges (e.g., hot/mild/cool, high/low). This makes the data suitable for algorithms that need categorical inputs.

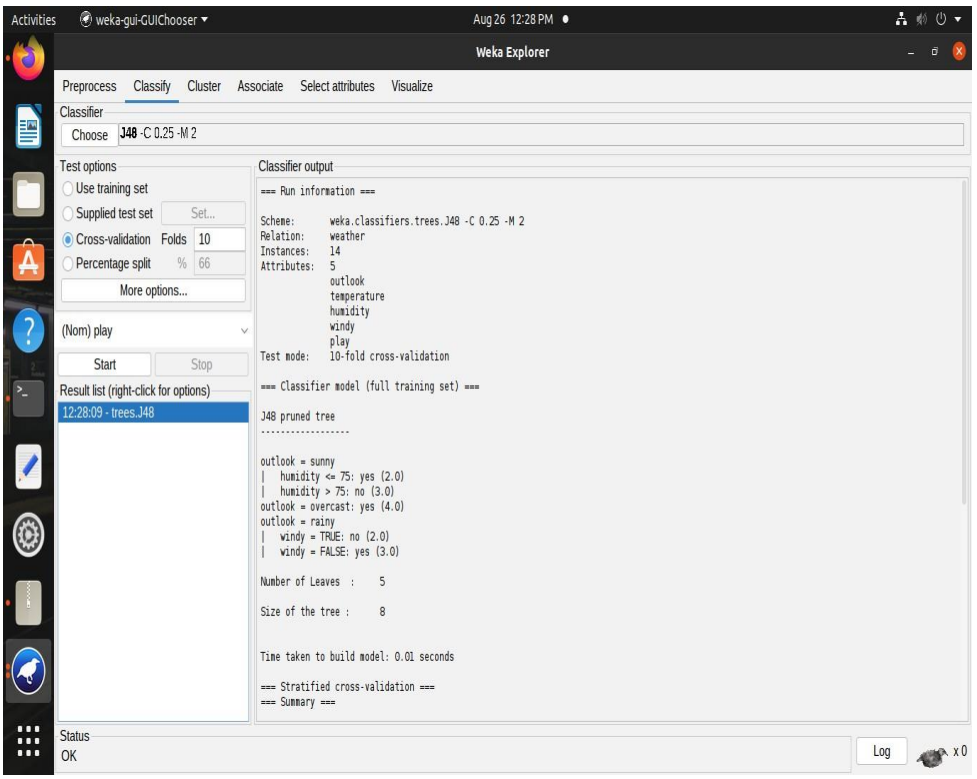
What happens:

- Weka automatically splits numeric values into intervals (bins). Example:
  - `temperature` → {low (≤70), medium (71–80), high (≥81)}
  - `humidity` → {low (≤75), high (>75)}

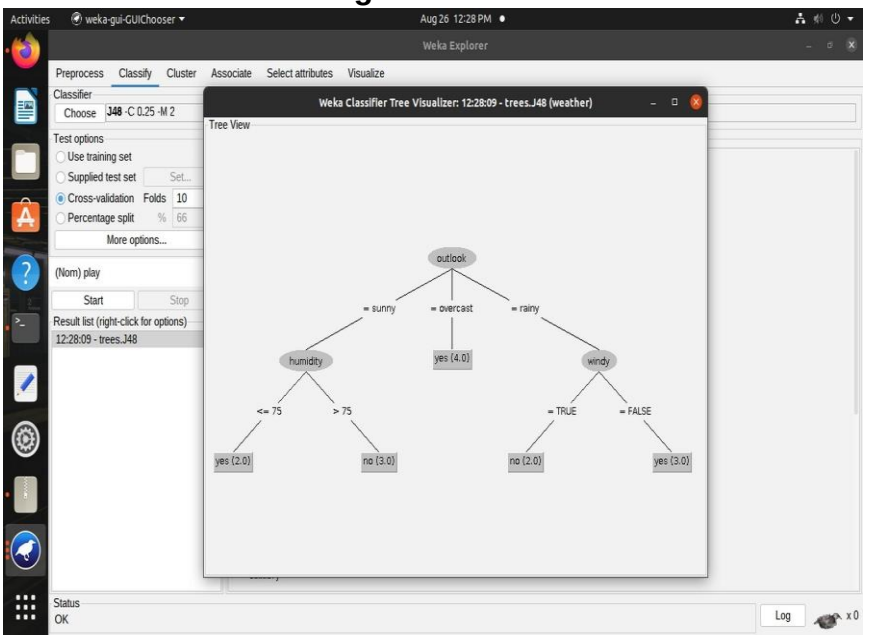
After conversion, these attributes are now nominal, and algorithms like ID3/J48 can handle them.

3. **Classification Algorithm Demonstration**

   a. **Attach screenshots of J48 demonstration in step-wise manner**



   i. **Explain the output, confusion matrix and show tree using visualization. Also visualize errors.**



Decision Tree Explanation

- Root Node = outlook
  The most important attribute (highest information gain) is chosen first.
    - outlook = sunny → check humidity
        - humidity ≤ 75 → play = yes (2 instances)
        - humidity > 75 → play = no (3 instances)
    - outlook = overcast → always play = yes (4 instances)
    - outlook = rainy → check windy
        - windy = TRUE → play = no (2 instances)
        - windy = FALSE → play = yes (3 instances)

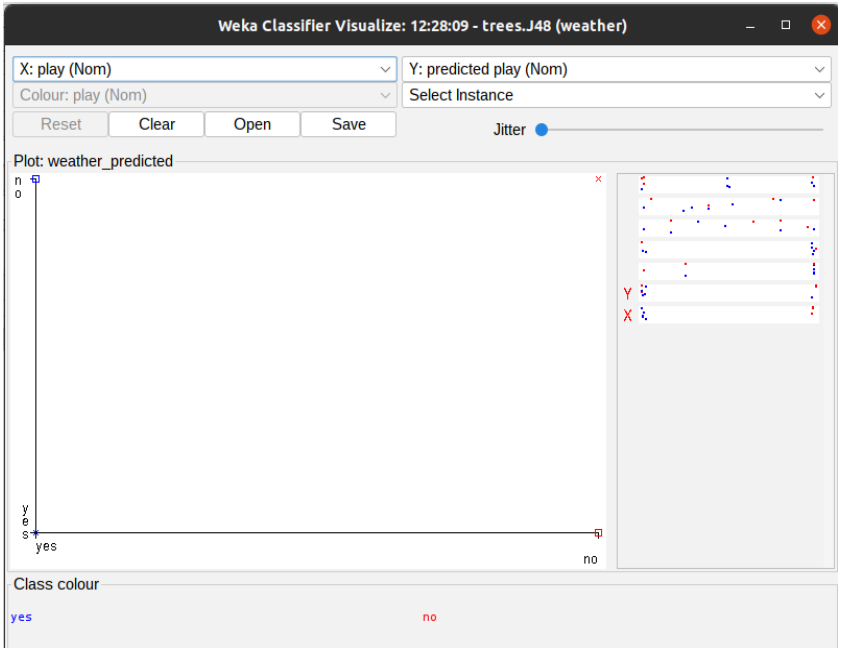
So the tree captures simple weather rules for playing.

Confusion Matrix (for 10-fold CV)

You would have seen something like:

	predicted yes	predicted no
actual yes	9	1
actual no	1	3

- Correctly classified = **12/14 (85.7%)**
- Misclassified = **2/14 (14.3%)**



b. **A brief one-line explanation under every screenshot**

Screenshot 1 – Preprocess Tab

"Preprocessing stage: dataset loaded with 5 attributes (2 numeric, 3 nominal) before discretization or filtering."

Screenshot 2 – Decision Tree Visualization

"Classification stage: J48 decision tree built where 'outlook' is the root and other attributes guide the play decision."

Screenshot 3 – Confusion Matrix Output

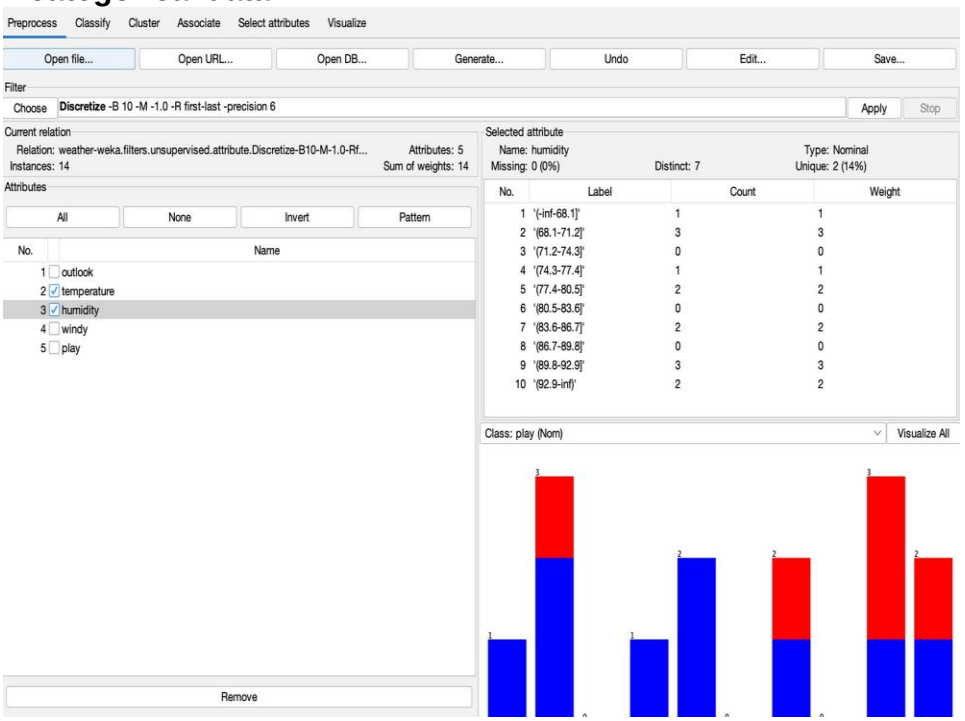
"Evaluation stage: confusion matrix shows correctly vs incorrectly classified instances with overall accuracy."

Screenshot 4 – Error Visualization

"Error analysis: misclassified instances highlighted in red, correctly classified in blue for visual inspection."

4. **Classification Exercise**

Use ID3 algorithm to classify weather data from the "weather.arff" file. Perform initial preprocessing and create a version of the initial dataset in which all numeric attributes should be converted to categorical data.



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) play

Start | Stop

Result list (right-click for options)

- 19:50:02 - trees.J48
- 19:50:48 - trees.J48
- 19:52:41 - trees.J48
- 19:55:57 - trees.J48
- 19:57:07 - trees.J48
- 19:57:21 - trees.J48

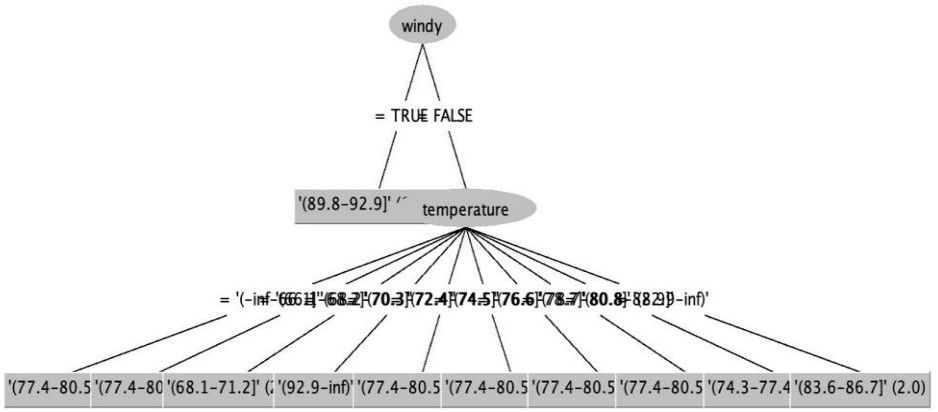
Classifier output

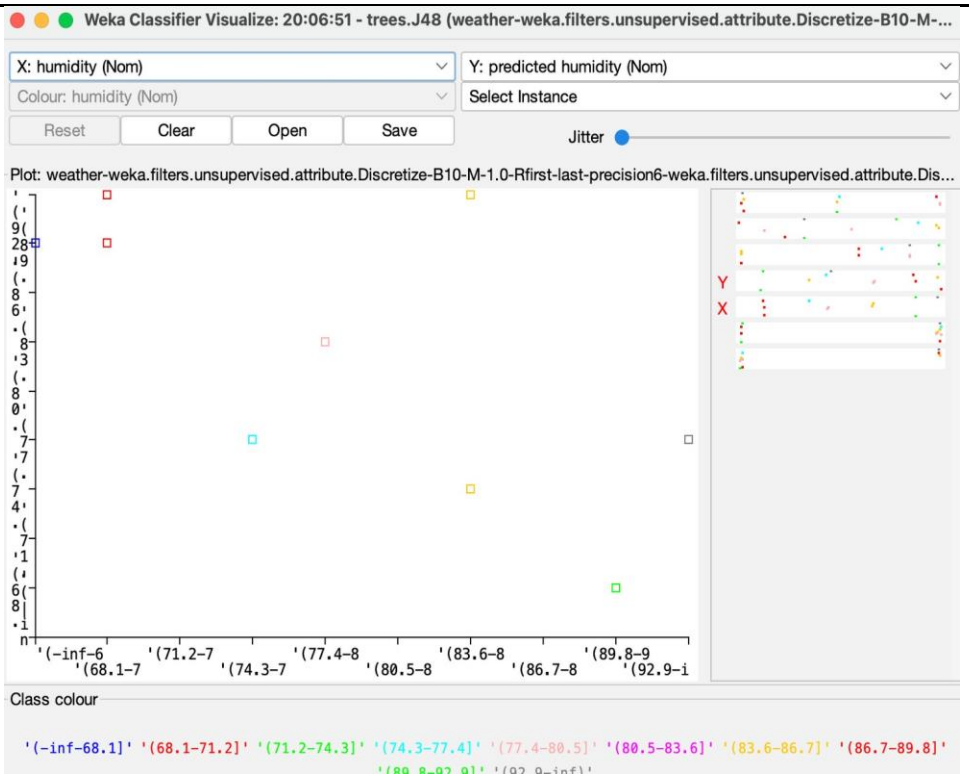
```
            windy
            play
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------
: yes (14.0/5.0)

Number of Leaves  :     1

Size of the tree :     1


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           7               50      %
Incorrectly Classified Instances         7               50      %
Kappa statistic                         -0.2564
Mean absolute error                      0.4794
Root mean squared error                  0.5431
Relative absolute error                100.6731 %
Root relative squared error            110.0798 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.778    1.000    0.583      0.778   0.667      -0.304   0.322     0.661     yes
               0.000    0.222    0.000      0.000   0.000      -0.304   0.322     0.336     no
Weighted Avg.  0.500    0.722    0.375      0.500   0.429      -0.304   0.322     0.545

=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 5 0 | b = no
```

Status: OK | Log | x 0

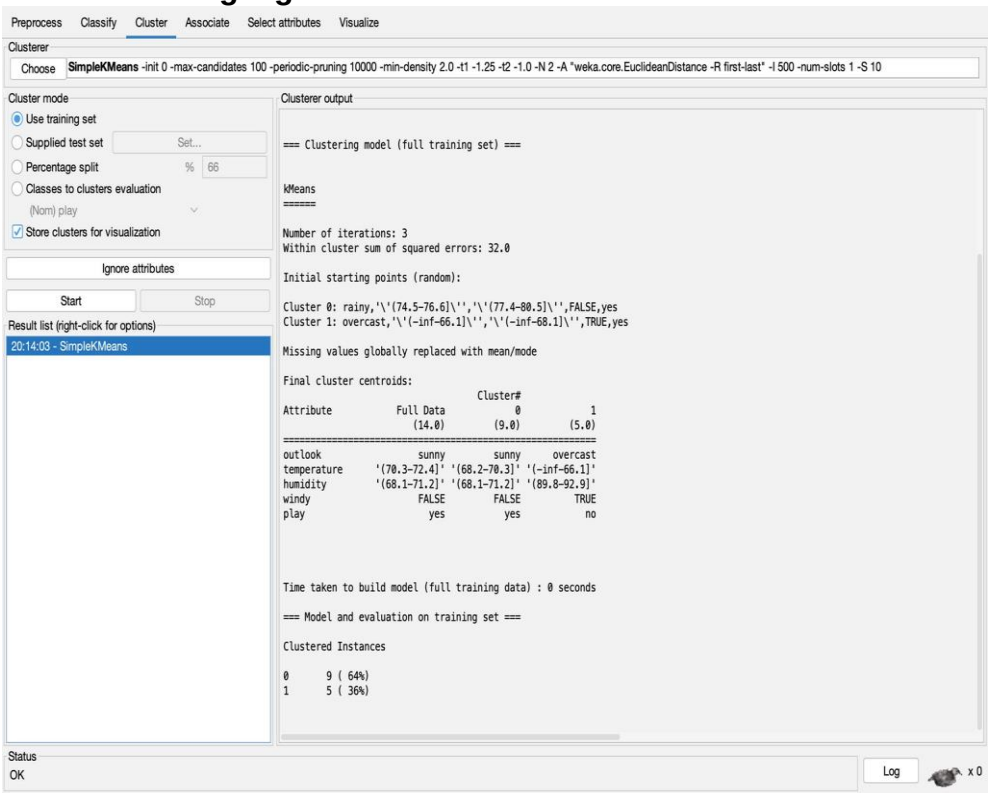
Weka Classifier Tree Visualizer: 20:06:51 - trees.J48 (weather-weka.filters.unsupervised.attribute.Discretize-B1...

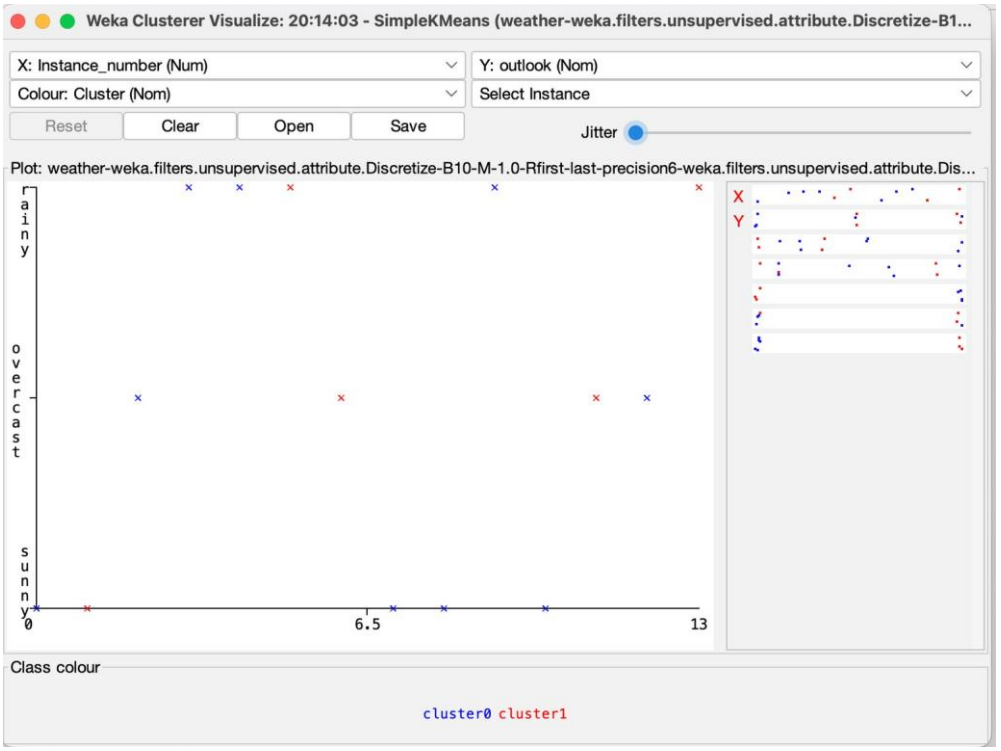
Tree View



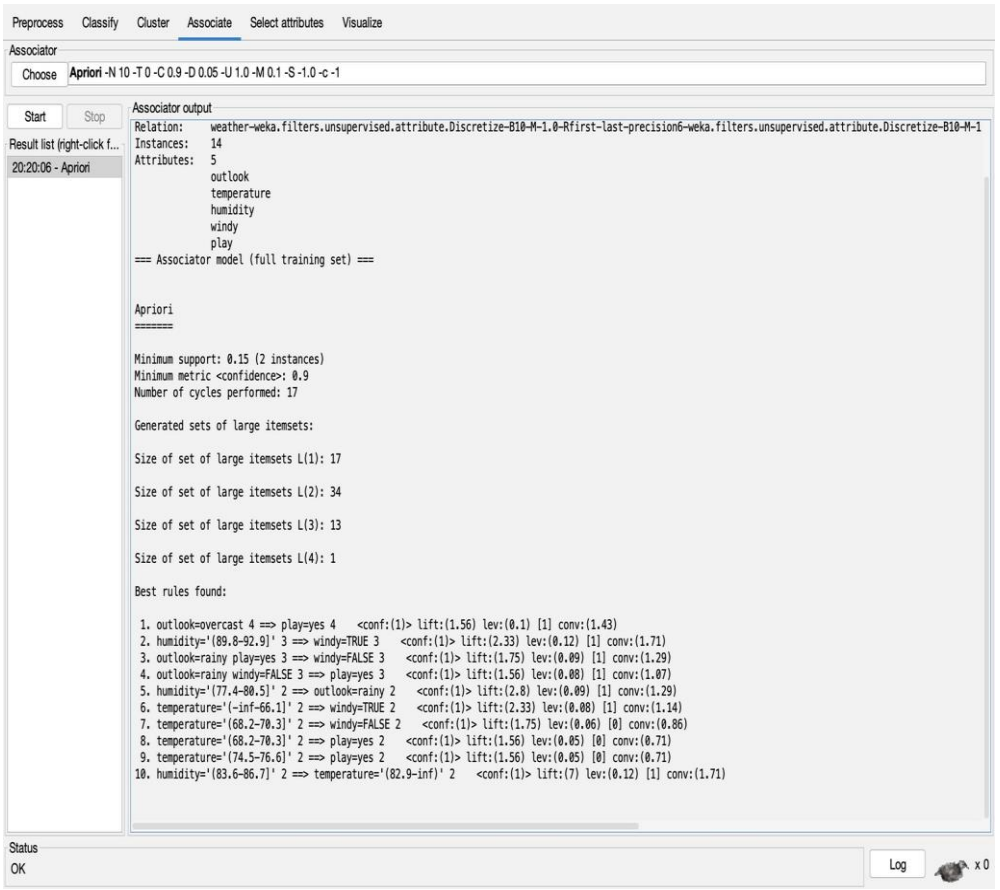


5. **Clustering algorithm screenshots**





6.Association rule mining algorithm screenshots



Conclusion	Preprocessing the weather dataset allowed ID3 to accurately classify instances, while clustering revealed natural groupings and patterns. Association rule mining uncovered strong relationships among attributes, providing valuable insights. Together, these techniques demonstrate how data mining can predict outcomes and discover hidden patterns effectively.
References	Follow the instruction manual given in this experiment.Also find the datasets in the same experiment. https://docs.google.com/document/d/1v6kit0FREEMuA-VH441r4HEIeNnuPi4X/edit?usp=sharing&ouid=115483059404226605817&rtpof=true&sd=true