DATA MINING AND VISUALIZATION LABORATORY

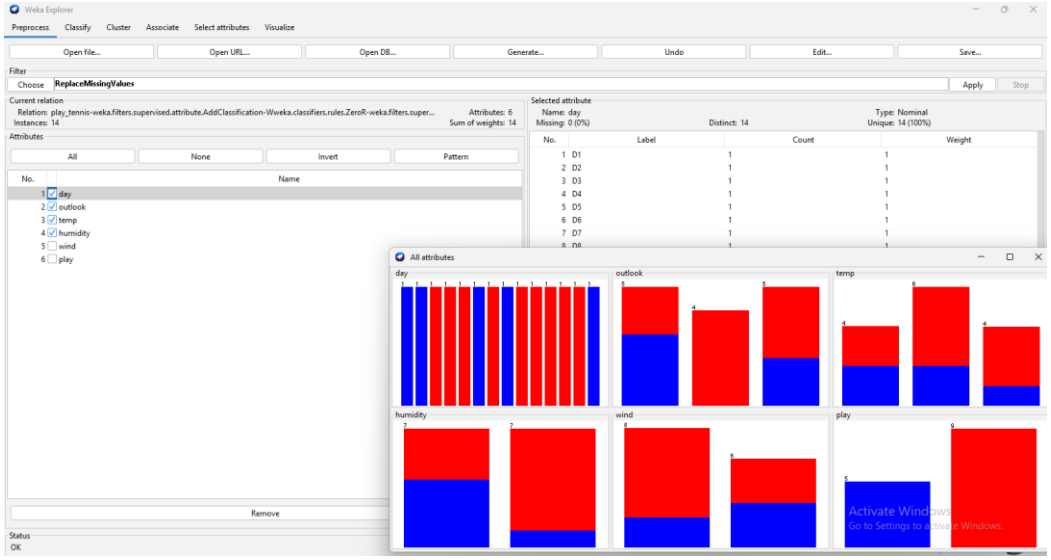
1. Experiment to be conducted using WEKA tool:

1	outlook	temperatu	humidity	windy	play
2					
3	sunny	85	85	FALSE	no
4	sunny	80	90	TRUE	no
5	overcast	83	86	FALSE	yes
6	rainy	70	96	FALSE	yes
7	rainy	68	80	FALSE	yes
8	rainy	65	70	TRUE	no
9	overcast	64	65	TRUE	yes
10	sunny	72	95	FALSE	no
11	sunny	69	70	FALSE	yes
12	rainy	75	80	FALSE	yes
13	sunny	75	70	TRUE	yes
14	overcast	72	90	TRUE	yes
15	overcast	81	75	FALSE	yes
16	rainy	71	91	TRUE	no

i). Preprocess(Data Cleaning, Data Integration, Data Transformation, Data Reduction) and Classify (Posteriori and Priori) panels. Analyze Input and Output Attributes.

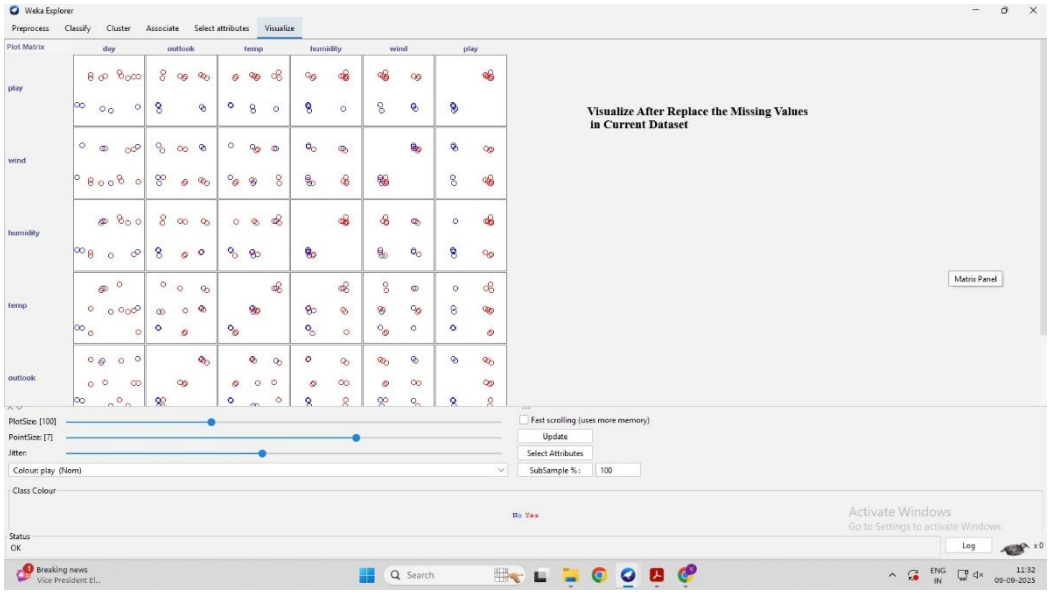
Preprocess

1. Open Explorer → Preprocess tab
2. Open file (load your dataset)
3. Clean data
   - Use filters: ReplaceMissingValues, RemoveDuplicates, Remove unwanted attributes
4. Integrate data
   - Use Append or MergeTwoFiles filter to combine datasets
5. Transform data
   - Use filters like NumericToNominal, NominalToBinary, Standardize, Normalize
6. Reduce data
   - Go to Select attributes tab → choose evaluator & search → Start



 Replace missing values **:**

- Under Filter **click** Choose → unsupervised → attribute → ReplaceMissingValues → click Apply**.**

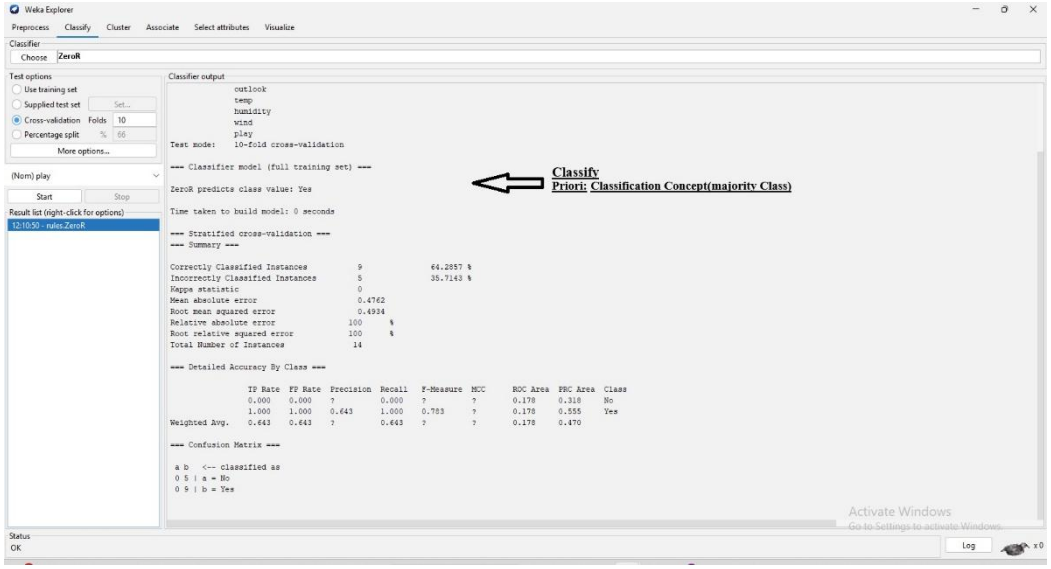


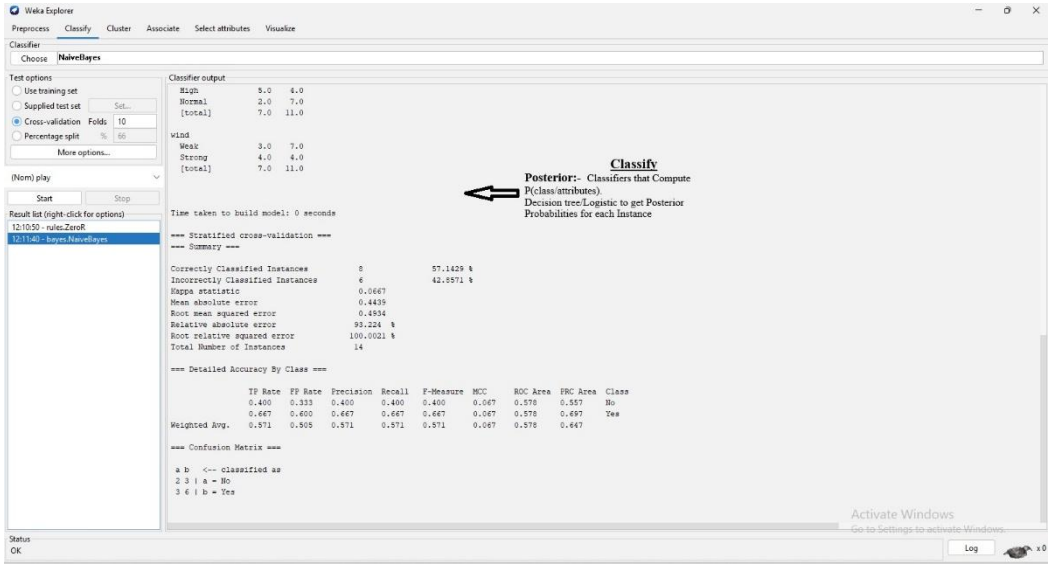
Normalize numeric attributes:

- Choose unsupervised → attribute → Normalize → Apply.

 Remove an attribute (reduce):

- Choose unsupervised → attribute → Remove.
- Click the filter name in the box to edit property attributeIndices.
  Example: to remove outlook and windy, set attributeIndices = 1,4. Then click Apply.



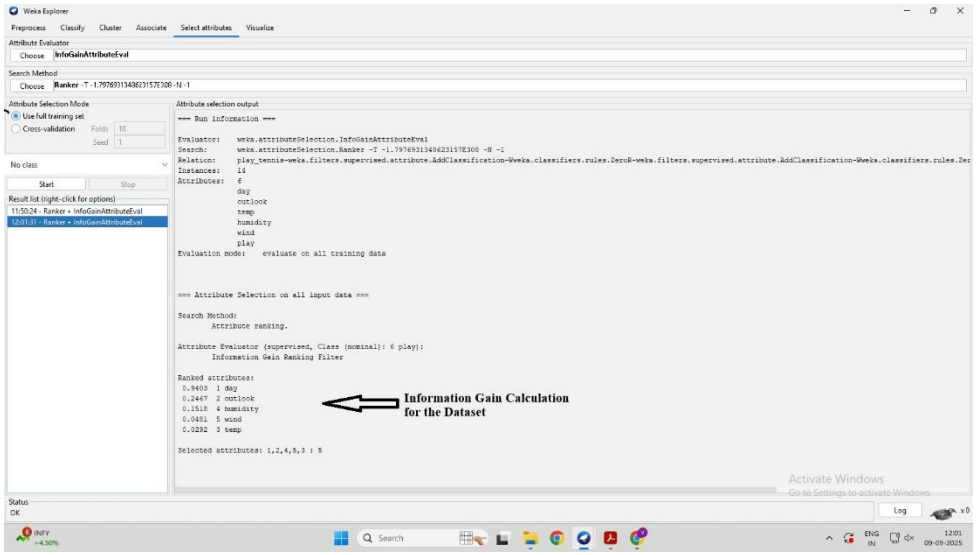


Edit dataset manually:

- Click the Edit button (bottom left) → you can change individual cell values, add rows, or delete rows.

ii). Calculate the information of the whole data set on the basis of whether play is held or not.

1. Open WEKA → Click on Explorer.
2. In the Preprocess tab, click Open file... and load your dataset (CSV or ARFF).
3. Once loaded, check at the top that the number of Instances is shown (e.g., 14).
4. In the left Attributes panel, click on the attribute play.
5. At the bottom-right, set Class to play.
6. In the Selected attribute panel (right side), note the counts of each class value (e.g., yes = 9, no = 5).
7. Compute the value to get the information (entropy) of the dataset.

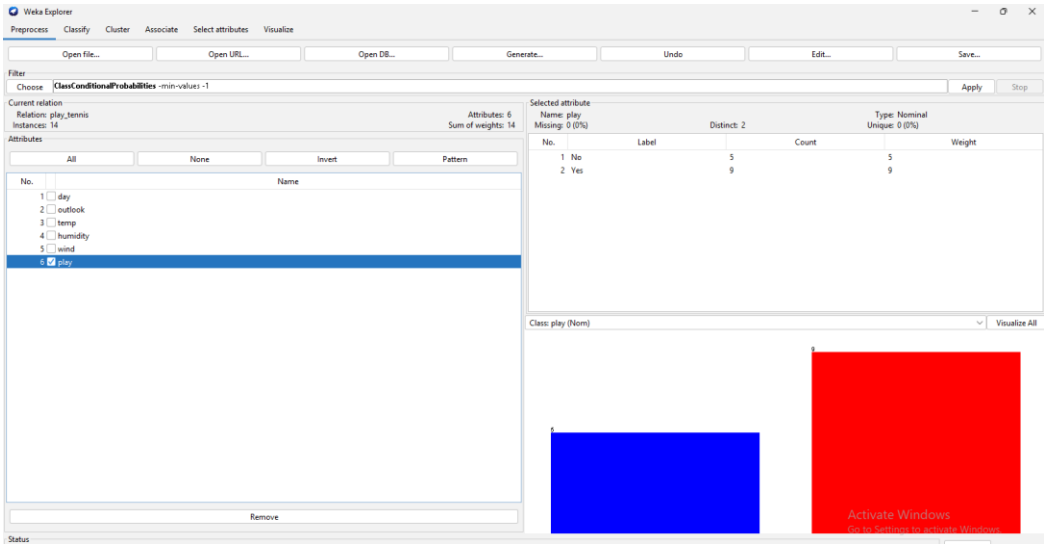


iii). Draw the histogram to show how the values of the play class occurs for each value of the outlook attribute .

Steps to draw the histogram in WEKA

1. Open **WEKA → Explorer**.
2. Go to the **Preprocess** tab.
3. Click **Open file…** and load your dataset (CSV or ARFF).
4. At the bottom-right, set **Class attribute = play**.
5. In the **Attributes** list (left panel), click on **outlook**.

6. Click the **Visualize** button (or double-click the `outlook` attribute).
7. A histogram window opens showing bars for each `outlook` value.
8. The bars will be **color-split based on** `play` **values (yes/no)**.
9. Hover over the bars to see exact counts for each class.

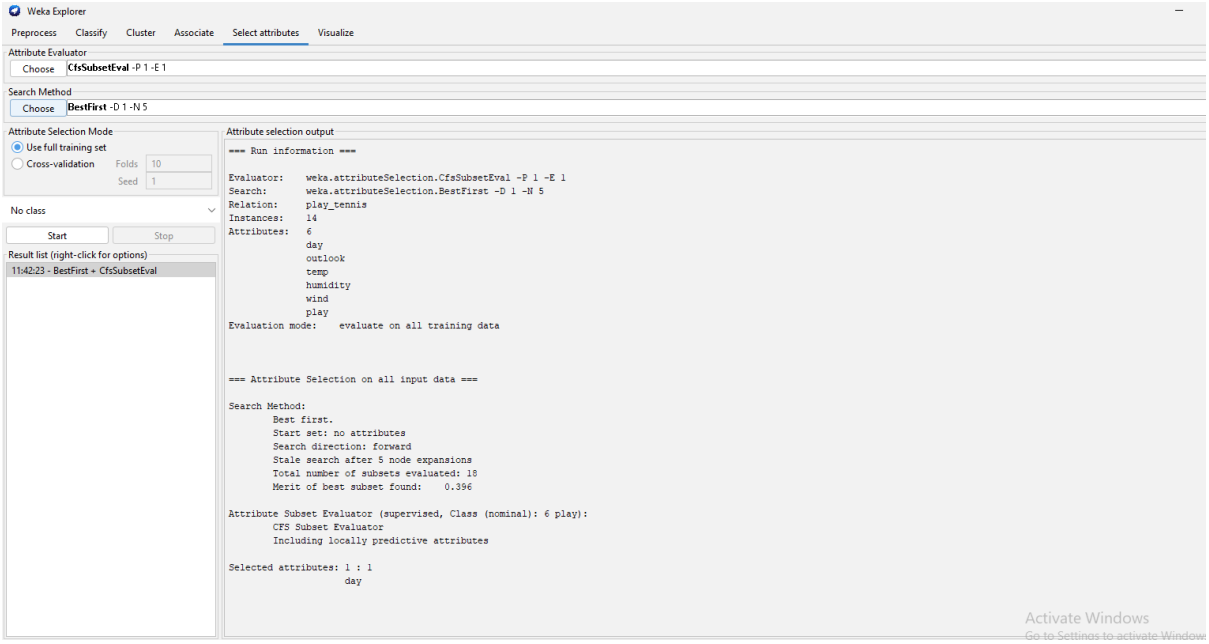


iv). Derive minimum and maximum values, mean, and standard deviation.

- Open Weka → Explorer → Preprocess tab
- Open file (load your dataset)
- In the Attributes panel, click on the attribute name.

On the right side (Selected attribute) you will see:

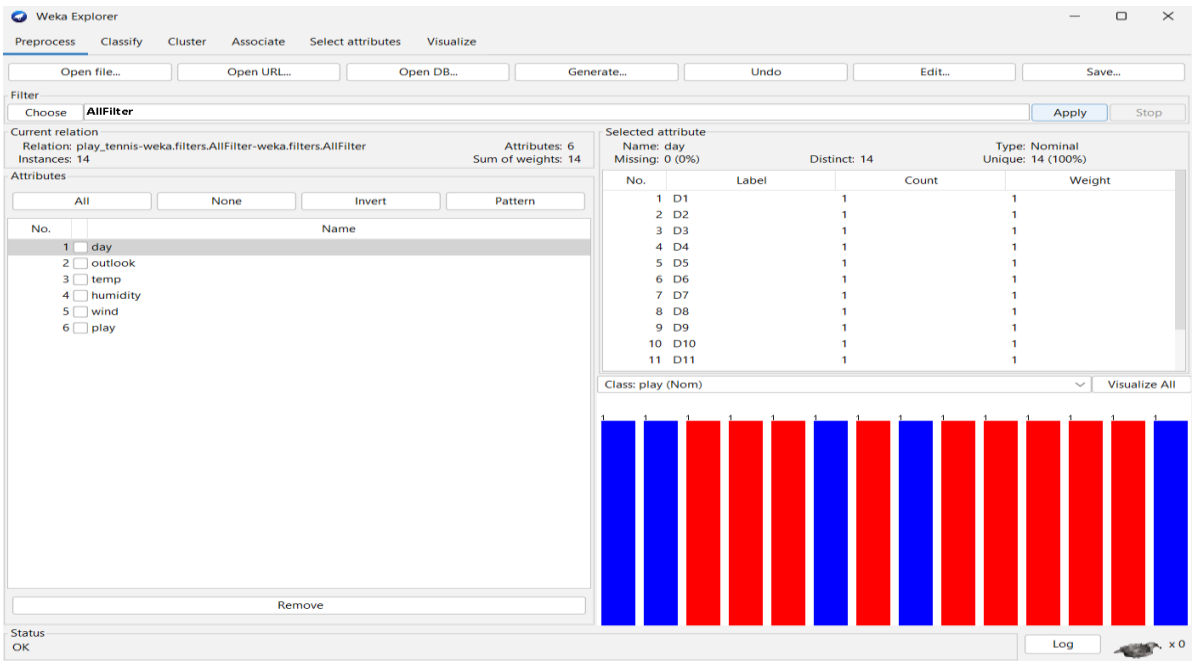
- Minimum
- Maximum
- Mean
- Standard deviation



v). Perform operations such as filter, delete, invert, Pattern, Undo, Edit, search, Select, Conversions etc.

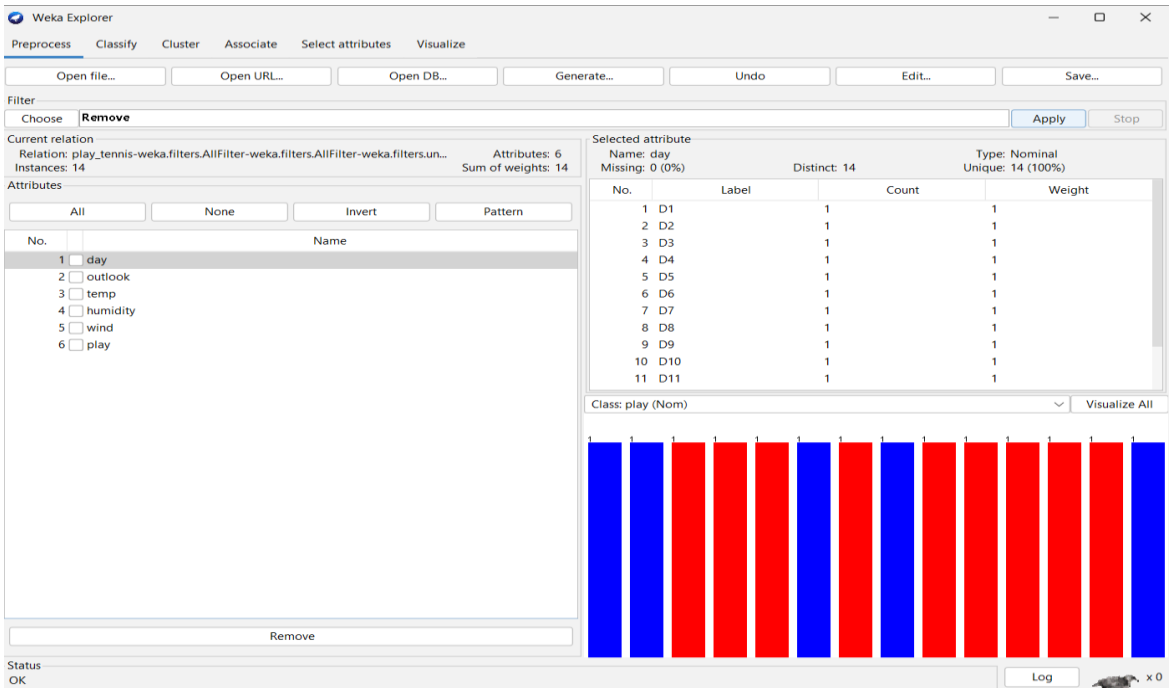
Filter:

1. Preprocess → Filter → Choose → select filter → set options → Apply.



Delete Attribute/Instance:

1. Tick attribute(s) → Remove
2. Preprocess → Edit → select row(s) → Delete rows → Close

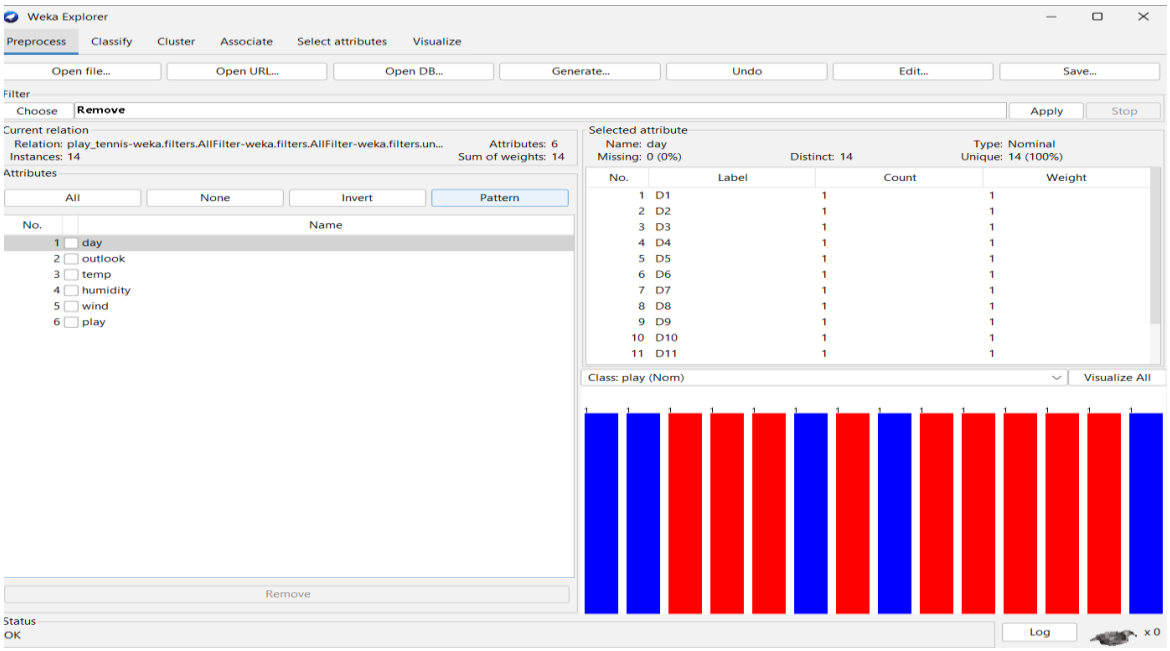


Invert Selection:

1. Click Invert above attributes list

Pattern Selection:

1. Pattern box → type regex → Enter



Undo:

1. Click Undo (top-right)

Edit Values:

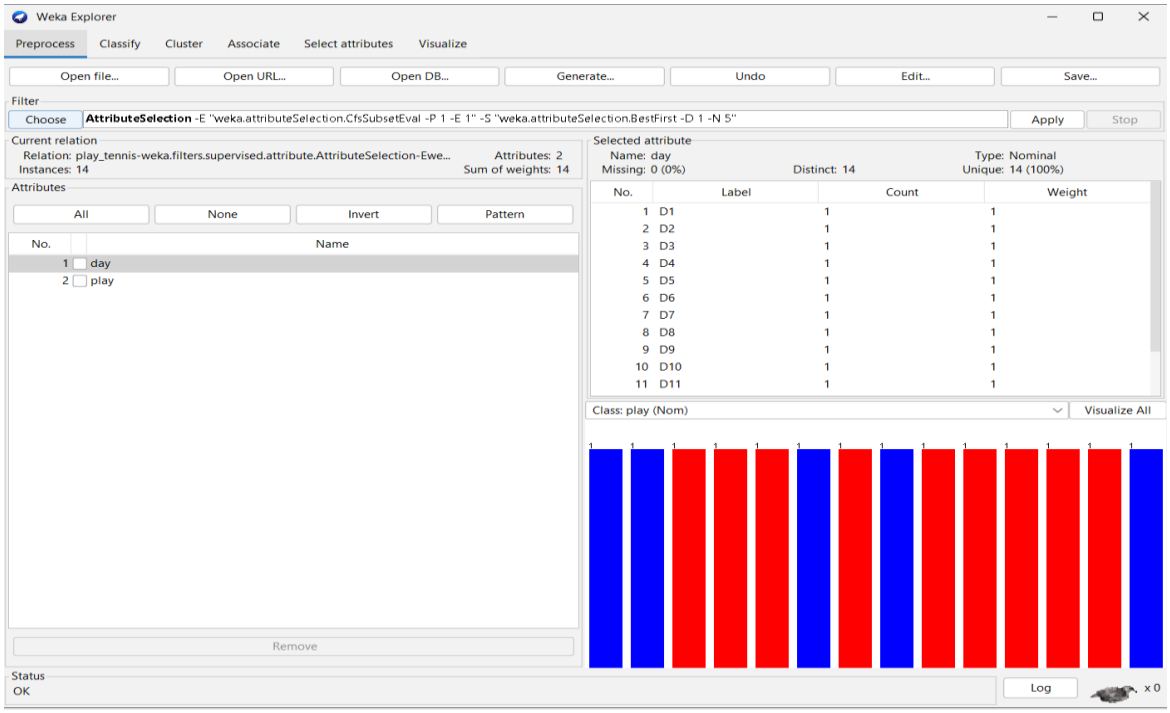
1. Preprocess → Edit → double-click cell → Close → Save

Search Values:

1. Preprocess → Edit → scan column or use filter → Apply

Select Attributes/Instances:

1. Tick attributes → or Visualize → select points/rectangle



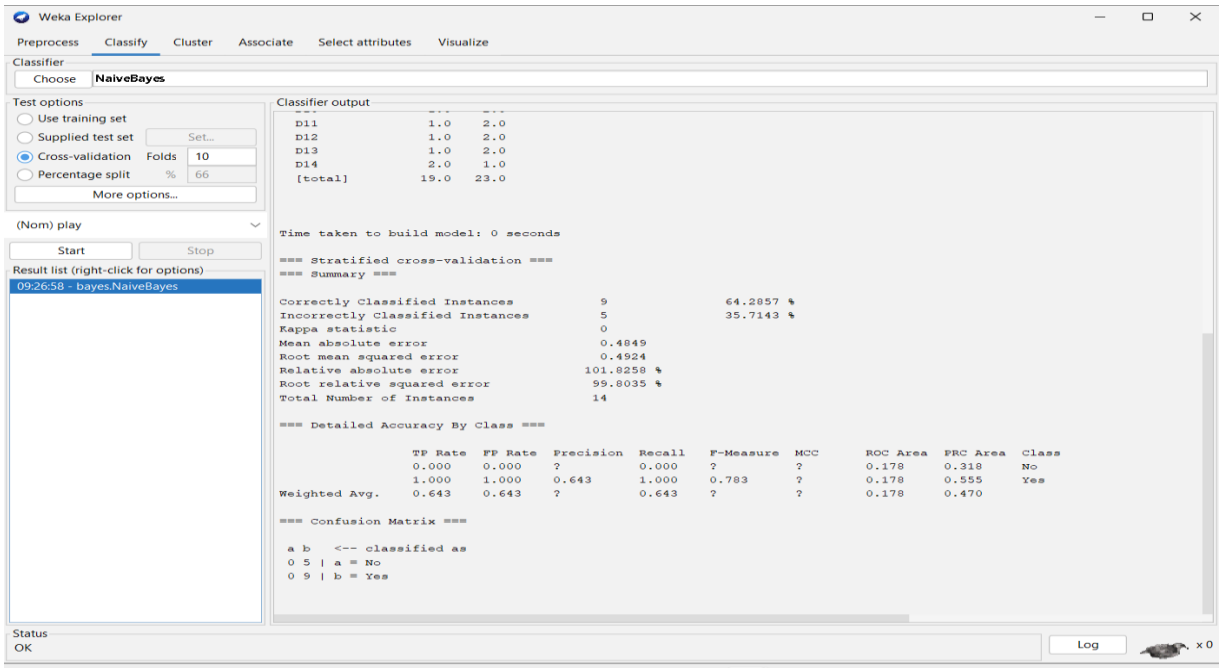
Conversions (type changes):

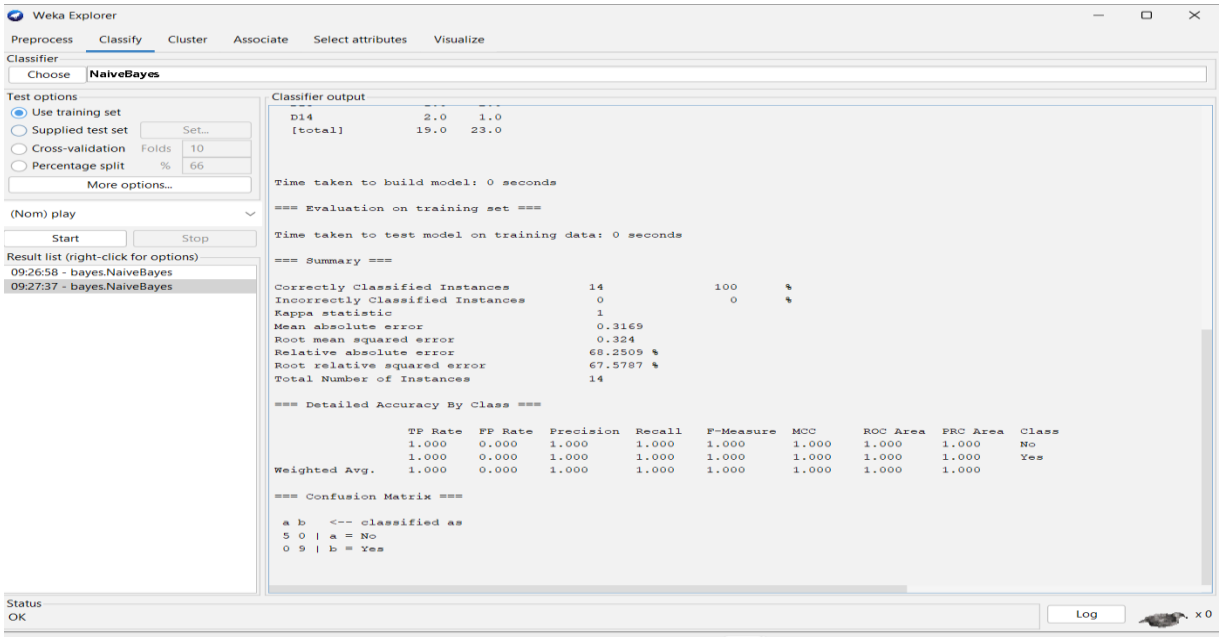
1. Preprocess → Filter → Choose → conversion filter → set attributes → Apply.

vi). Examine the Output , classification error and Kappa statistics.

1. Open Explorer → Classify tab
2. Choose Classifier → e.g., J48, NaiveBayes, etc. → set options if needed
3. Select Test options → Use training set / Supplied test set / Cross-validation

4. Click Start → wait for classifier to run
5. Look at Classifier output panel:
    - % Correct / % Incorrect → classification error
    - Kappa statistic → directly shown in output panel
6. Optional: Scroll down to Detailed Accuracy. By Class for per-class metrics.

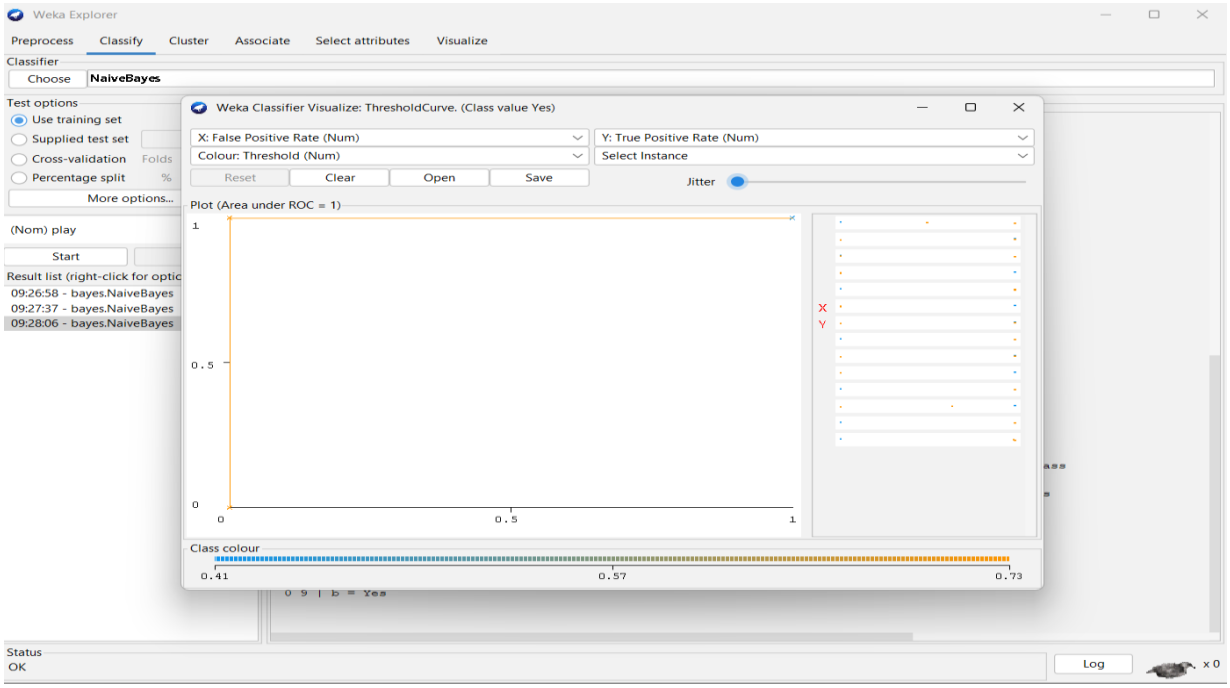




vii). Visualize threshold curve.

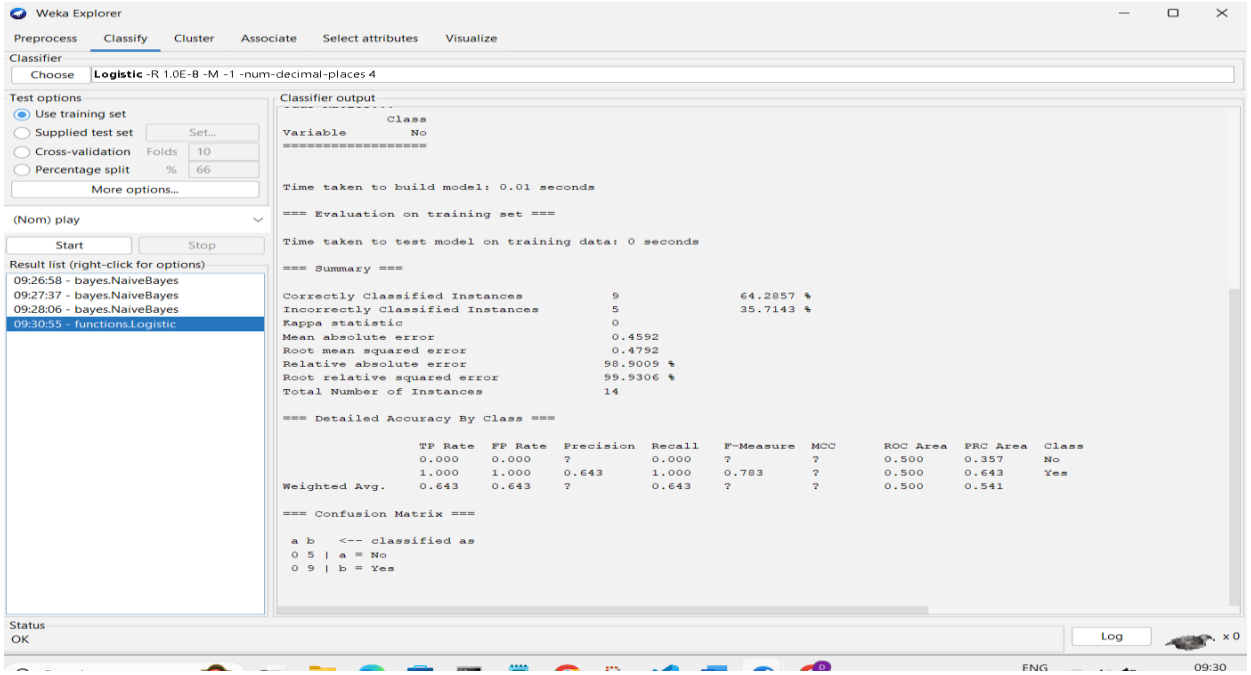
1. Open Explorer → Classify tab

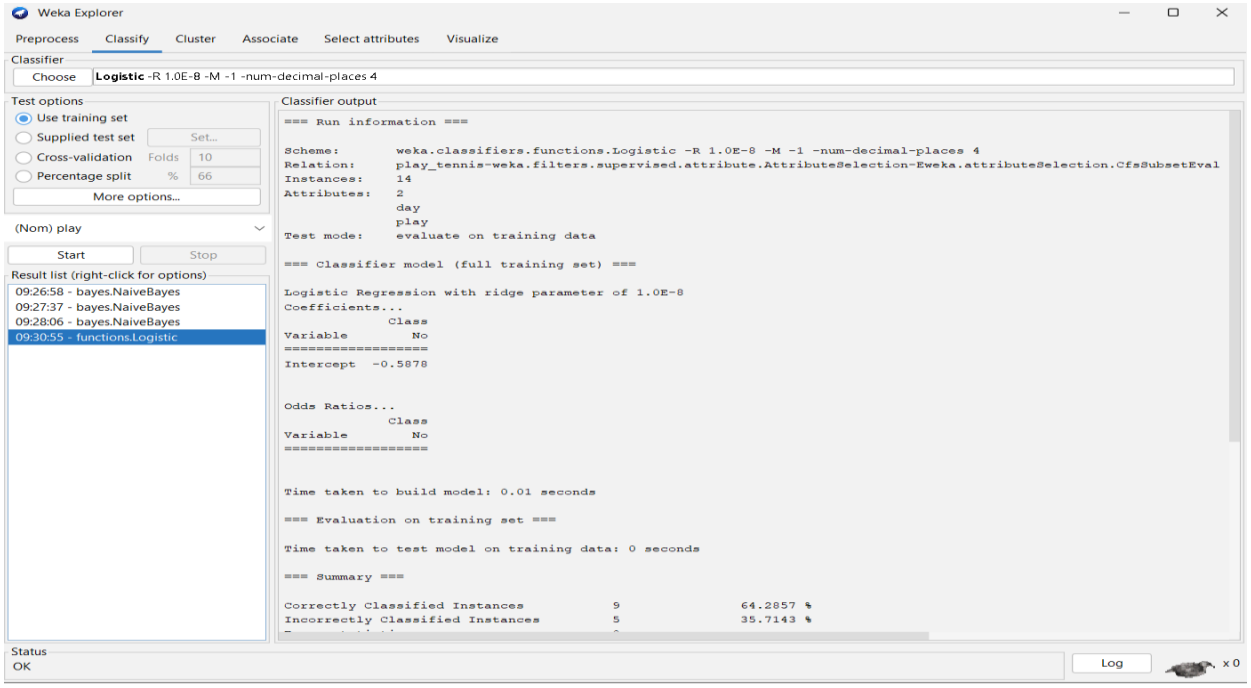
2. Choose Classifier → e.g., J48, NaiveBayes → set options
3. Select Test options → Cross-validation / Use training set / Supplied test set
4. Click Start → wait for classifier to run
5. In Result list (left panel) → right-click the classifier → Visualize threshold curve
6. Select class index (for which class you want the curve) → click OK
7. Threshold curve window opens → analyze ROC, area under curve, etc.



viii). Apply Logistic Regression model to classify.

- Open **Explorer** → **Classify** tab
- Click **Choose** → `functions → Logistic`
- Set **Test options** → `Use training set` / `Supplied test set` / `Cross-validation`
- Click **Start** → wait for model to run
- Check **Classifier output** for results, classification accuracy, and statistics.





ix). Measure the log likelihood of the clusters of training data. (Consider large data set.)

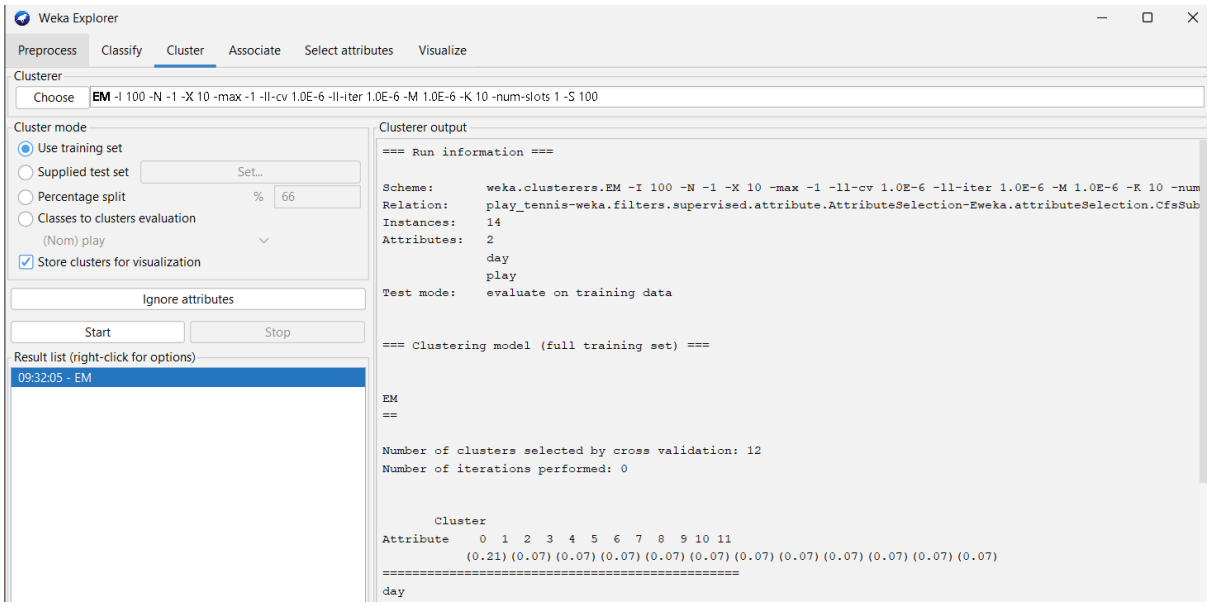
- Open **Explorer** → **Cluster** tab
- Click **Choose** → select a clustering algorithm (e.g., `EM`)

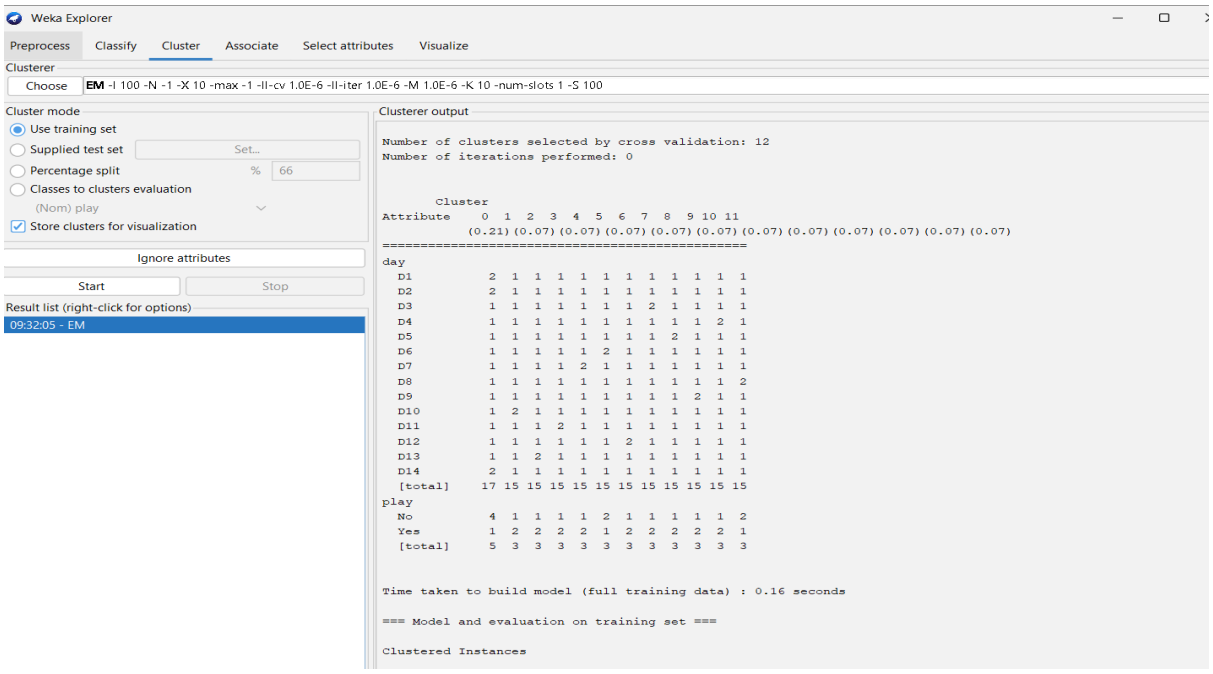
- Click the algorithm name → set **options** (e.g., number of clusters, max iterations)

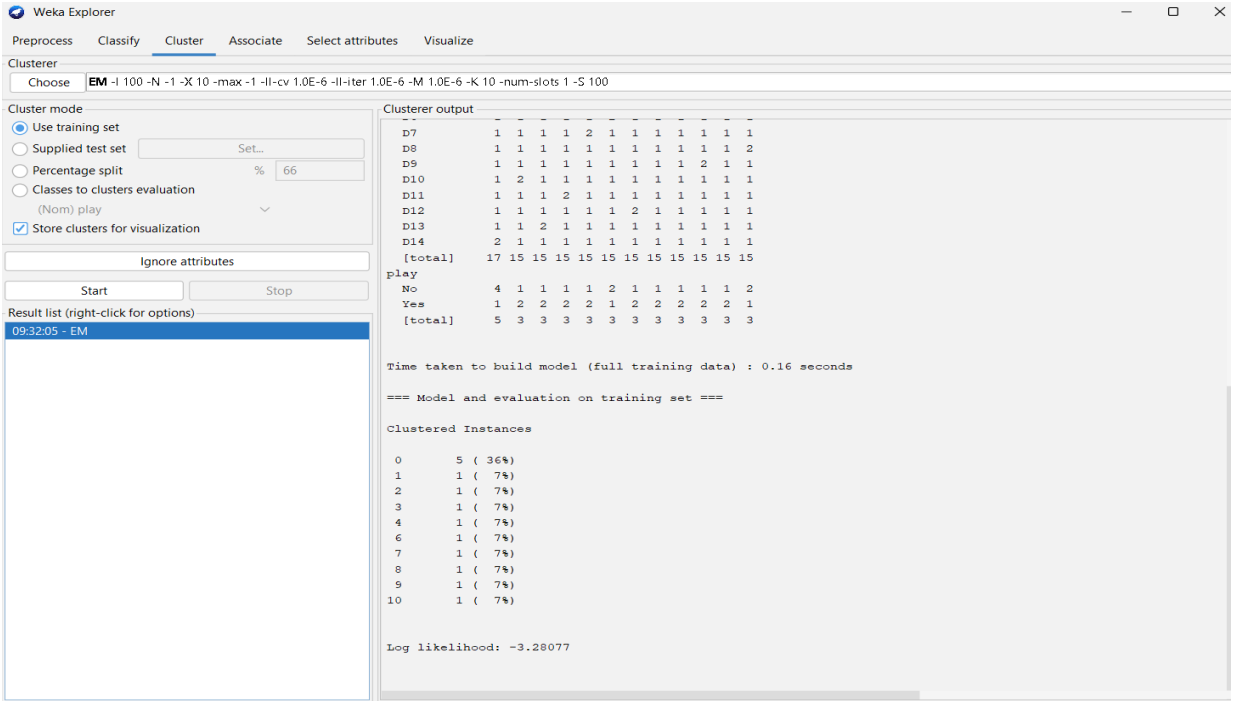
- Select **Cluster mode** → `Use training set.`

- Click **Start** → wait for clustering to complete.

- In **Result list / output panel**, check **Log likelihood** of clusters displayed in the results.

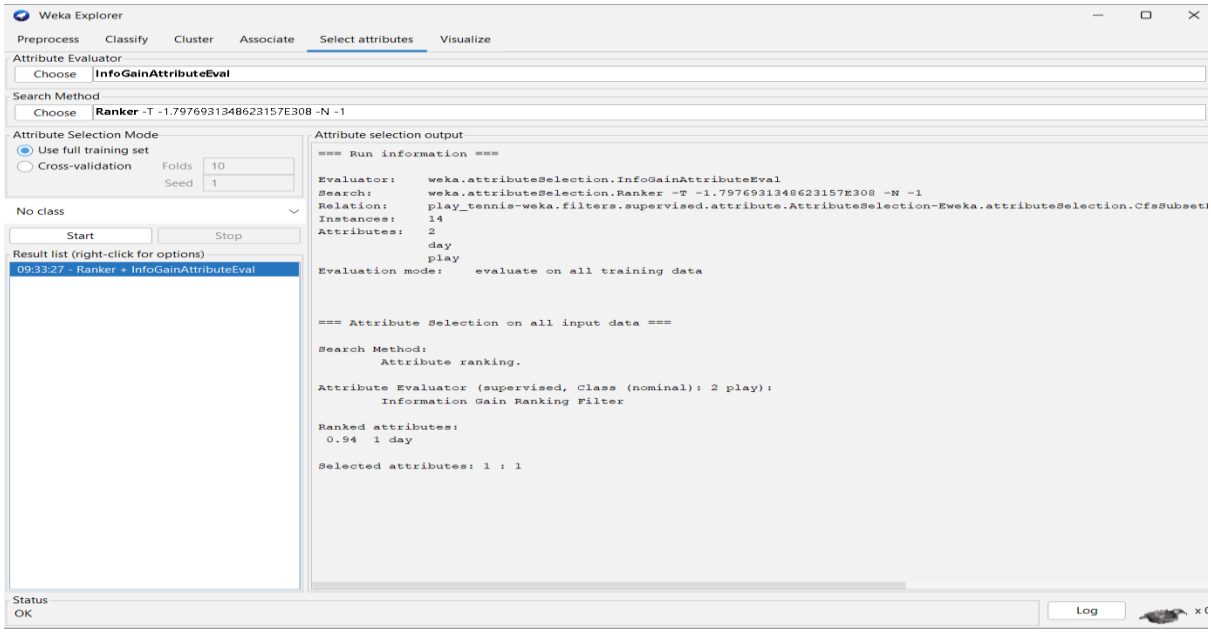






x). Derive Information gain.

1. Open Explorer → Select attributes tab
2. Click Attribute Evaluator → Choose → InfoGainAttributeEval
3. Click Search Method → Choose → Ranker
4. Click Start → wait for evaluation
5. Check Result list / output panel → displays Information Gain for each attribute.



xi). Build Decision Tree on Humidity attribute. Also demonstrate decision tree after analysis of

a. Sunny and Overcast dataset

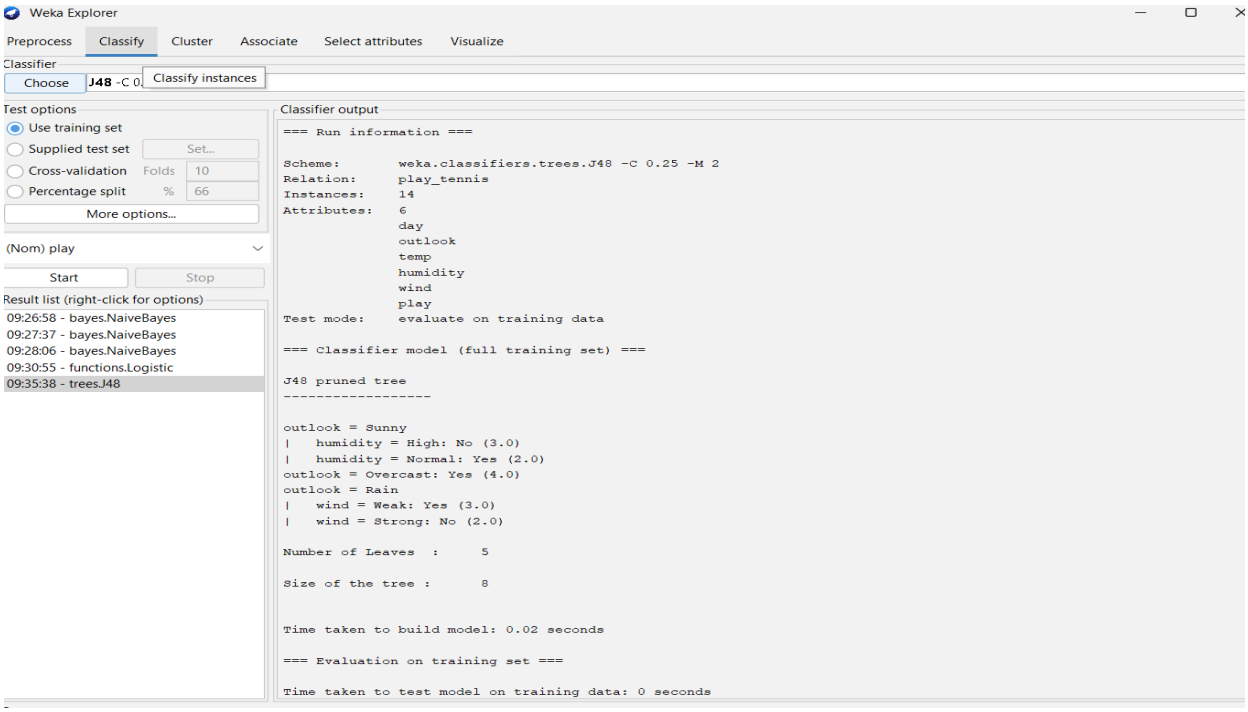
b. Sunny, Overcast and Rainy Data set.

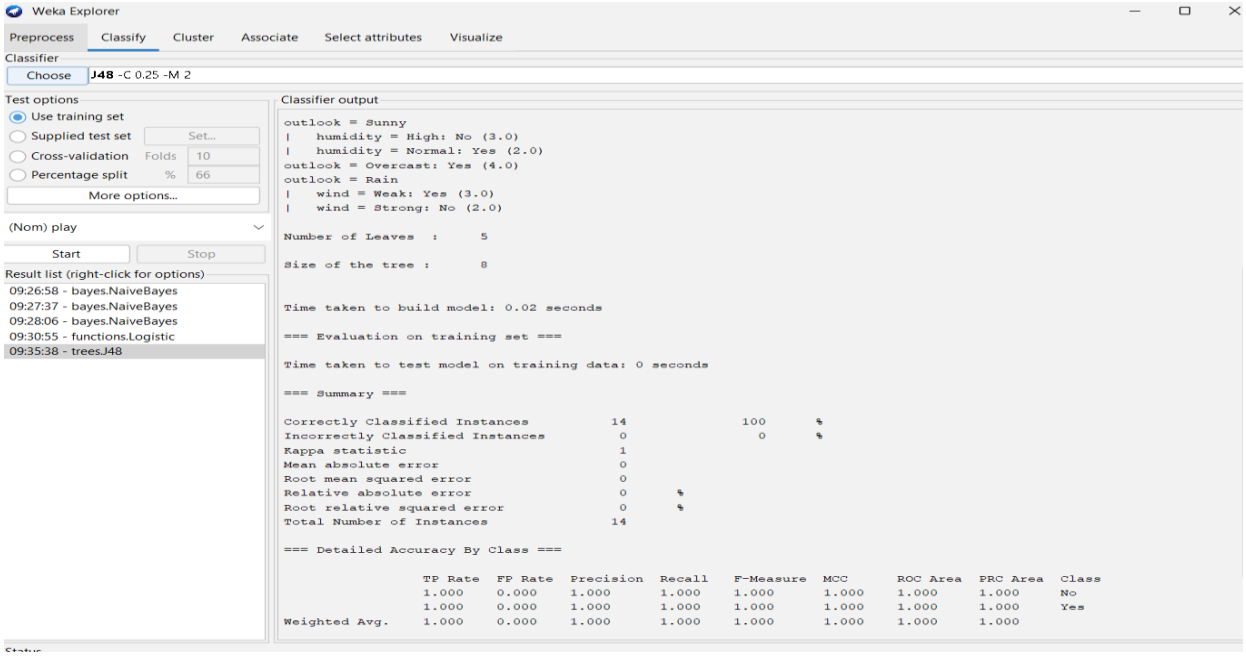
a) Sunny and Overcast dataset

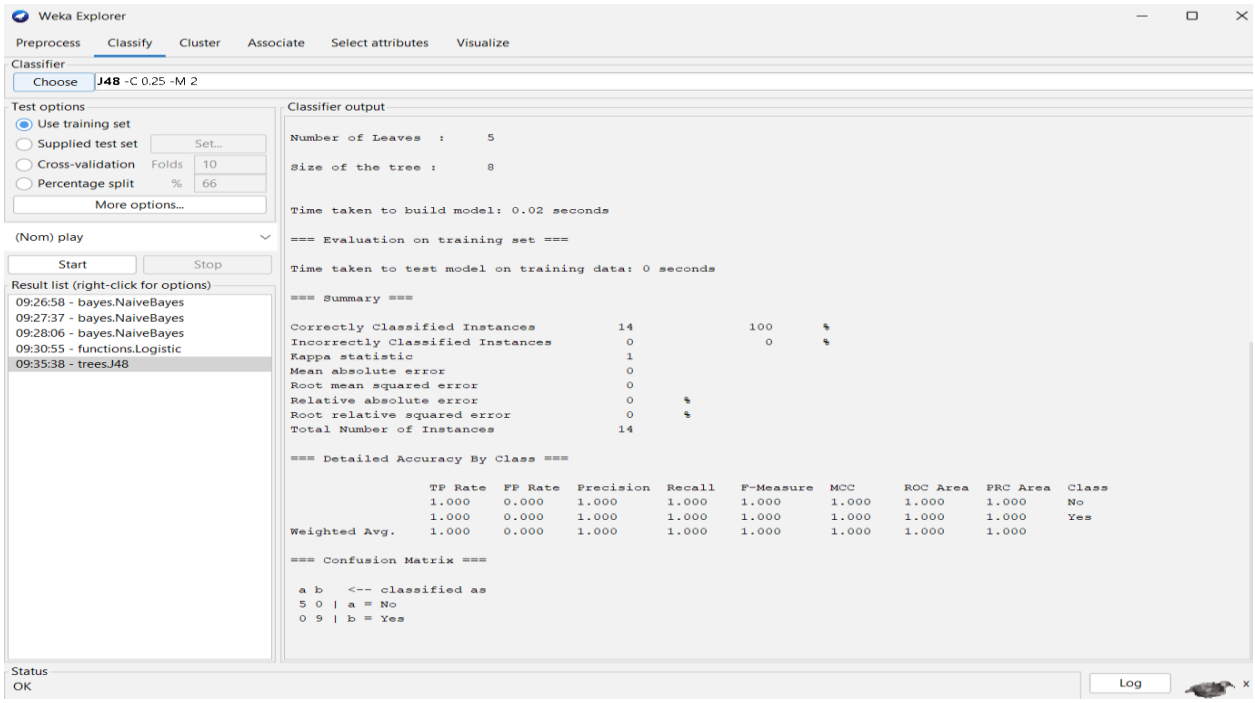
1. Preprocess → Filter → Choose → `RemoveWithValues` → set attribute `Outlook` → keep `Sunny` and `Overcast` → Apply
2. Classify → Choose → `trees → J48`
3. Test options → select dataset (training or cross-validation)
4. Start → view **Decision Tree** in output panel

b) Sunny, Overcast, and Rainy dataset

1. Preprocess → Reset dataset (all instances)
2. Classify → Choose → `trees → J48`
3. Test options → select dataset
4. Start → view **Decision Tree.**



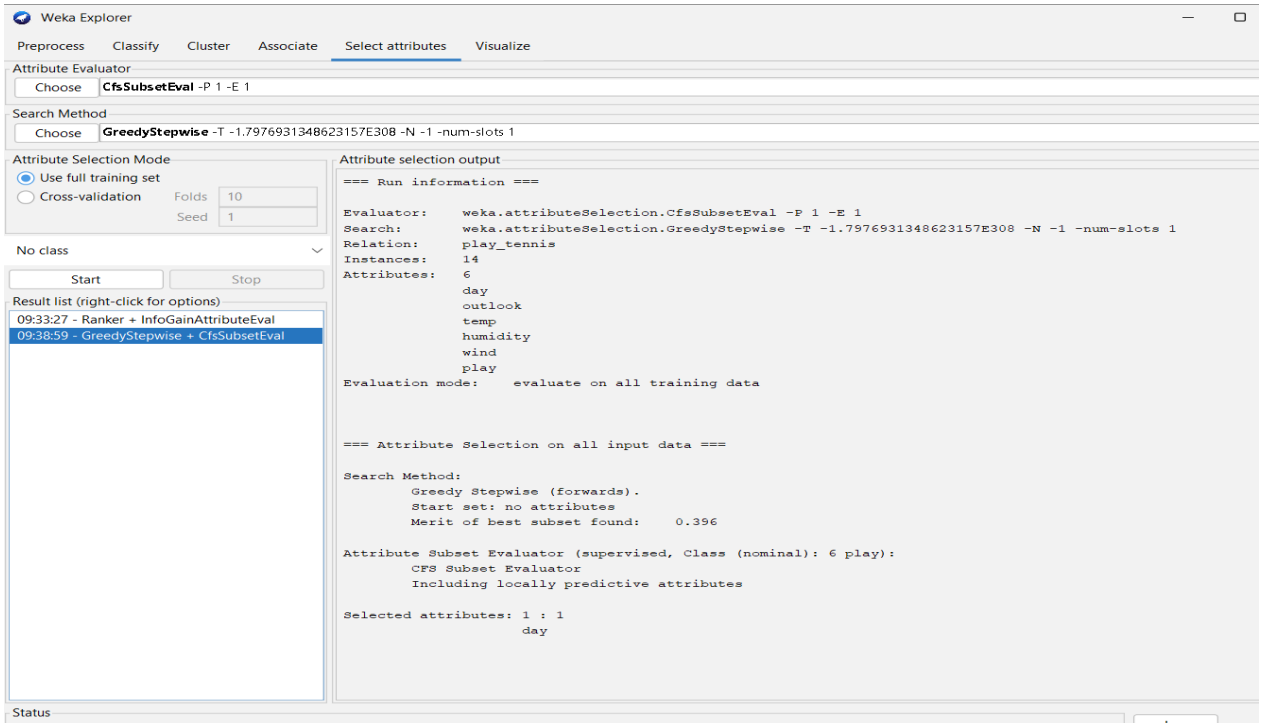




xii). Compute Gini Index representing with respect to Temperature, Humidity, and Windy attributes.

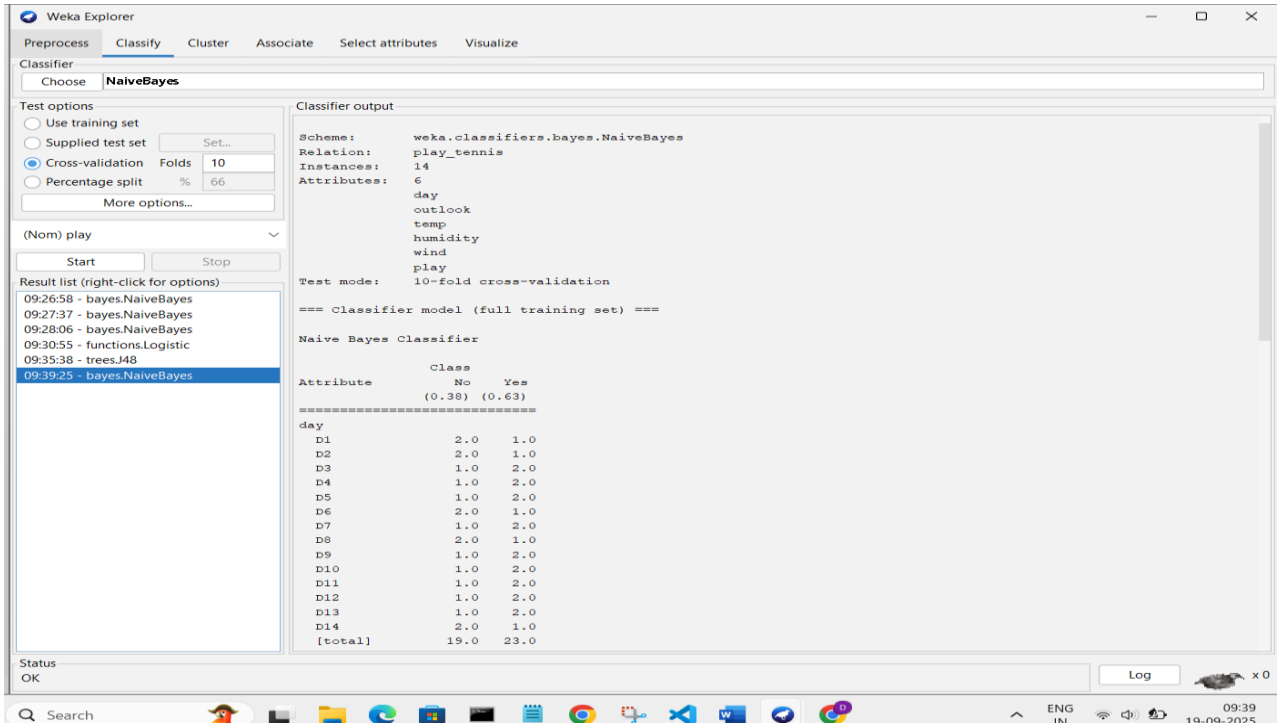
1. Select Attributes → Attribute Evaluator → weka.attributeSelection.GiniIndex (if available, or use weka.attributeSelection.CfsSubsetEval)
2. Search Method → Ranker

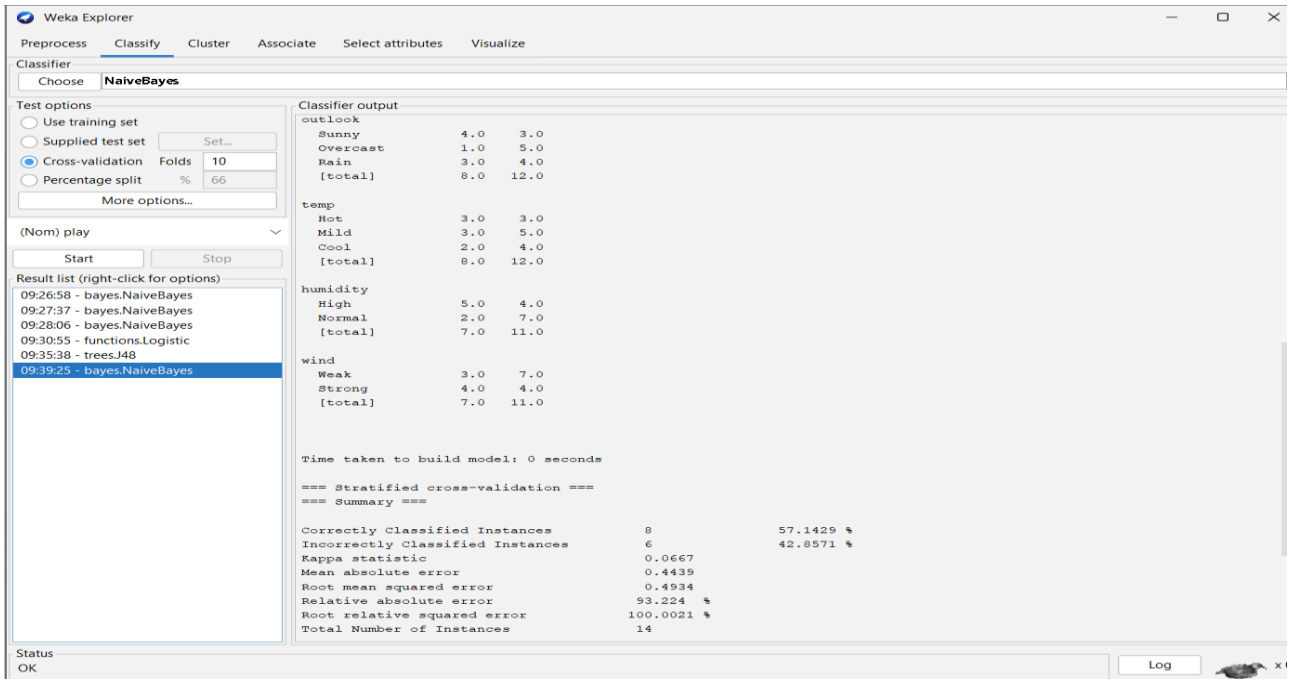
3. Start → check **Gini Index** values for each attribute.

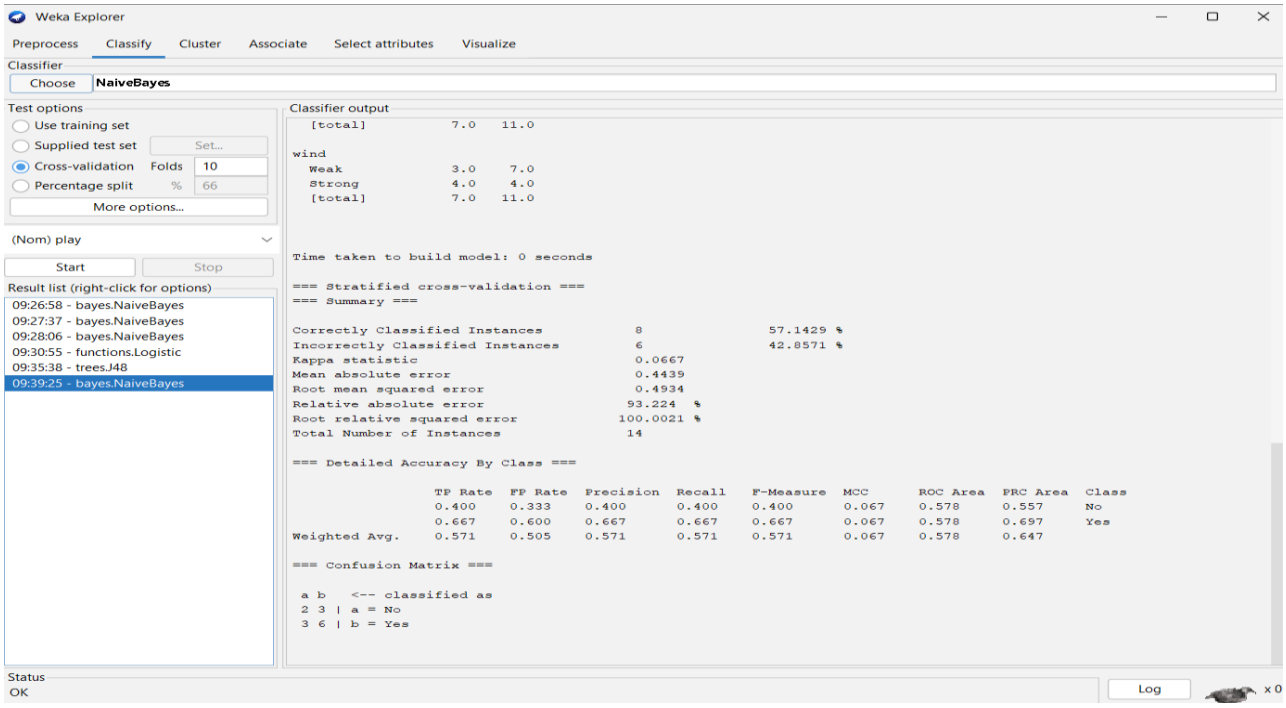


xiii), Obtain the Prediction of Play 'Yes' as well as 'No' for an unknown instance.

1. Classify → Choose → bayes → NaiveBayes
2. Test options → Use training set / Cross-validation
3. Start → in output → check **Conditional probabilities** table → find Play=No for Outlook=Rainy.

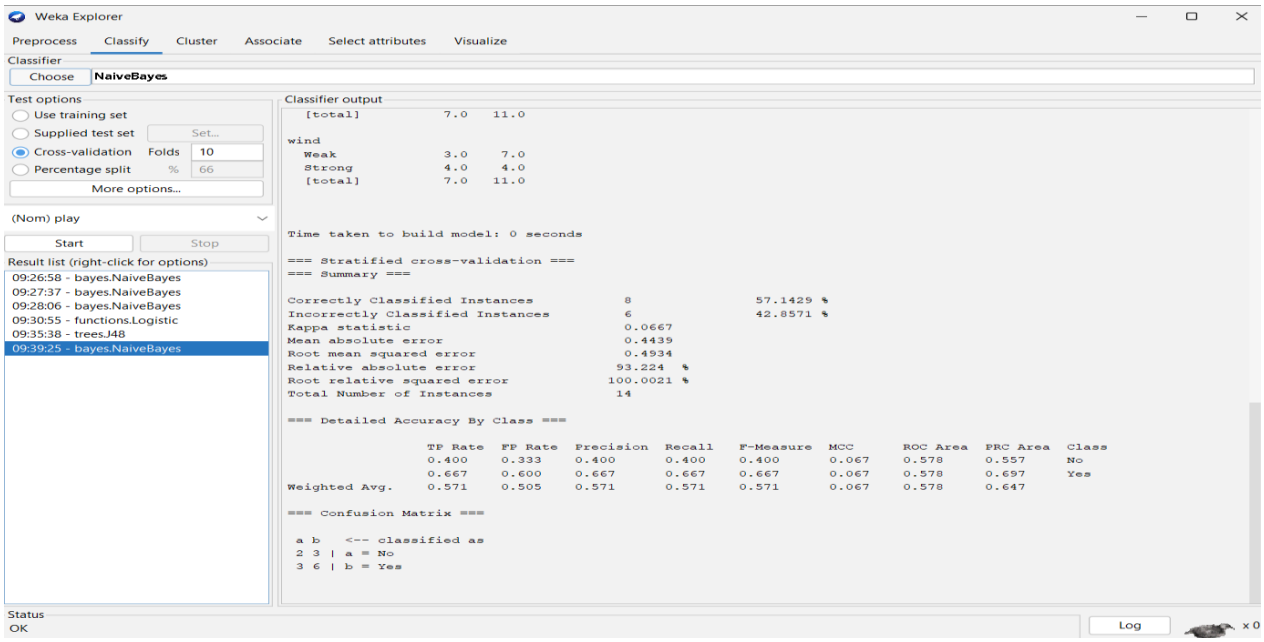

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
09:26:58 - bayes.NaiveBayes
09:27:37 - bayes.NaiveBayes
09:28:06 - bayes.NaiveBayes
09:30:55 - functions.Logistic
09:35:38 - trees.J48
09:39:25 - bayes.NaiveBayes
Classifier output
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: play_tennis
Instances: 14
Attributes: 6
day
outlook
temp
humidity
wind
play
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class
Attribute No Yes
(0.38) (0.63)
==========
day
D1 2.0 1.0
D2 2.0 1.0
D3 1.0 2.0
D4 1.0 2.0
D5 1.0 2.0
D6 2.0 1.0
D7 1.0 2.0
D8 2.0 1.0
D9 1.0 2.0
D10 1.0 2.0
D11 1.0 2.0
D12 1.0 2.0
D13 1.0 2.0
D14 2.0 1.0
[total] 19.0 23.0
Status
OK
Log
x 0
Search
ENG
09:39


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
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)
09:26:58 - bayes.NaiveBayes
09:27:37 - bayes.NaiveBayes
09:28:06 - bayes.NaiveBayes
09:30:55 - functions.Logistic
09:35:38 - trees.J48
09:39:25 - bayes.NaiveBayes
Classifier output
outlook
Sunny 4.0 3.0
Overcast 1.0 5.0
Rain 3.0 4.0
[total] 8.0 12.0
temp
Hot 3.0 3.0
Mild 3.0 5.0
Cool 2.0 4.0
[total] 8.0 12.0
humidity
High 5.0 4.0
Normal 2.0 7.0
[total] 7.0 11.0
wind
Weak 3.0 7.0
Strong 4.0 4.0
[total] 7.0 11.0
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 8 57.1429 %
Incorrectly Classified Instances 6 42.8571 %
Kappa statistic 0.0667
Mean absolute error 0.4439
Root mean squared error 0.4934
Relative absolute error 93.224 %
Root relative squared error 100.0021 %
Total Number of Instances 14
Status
OK
Log



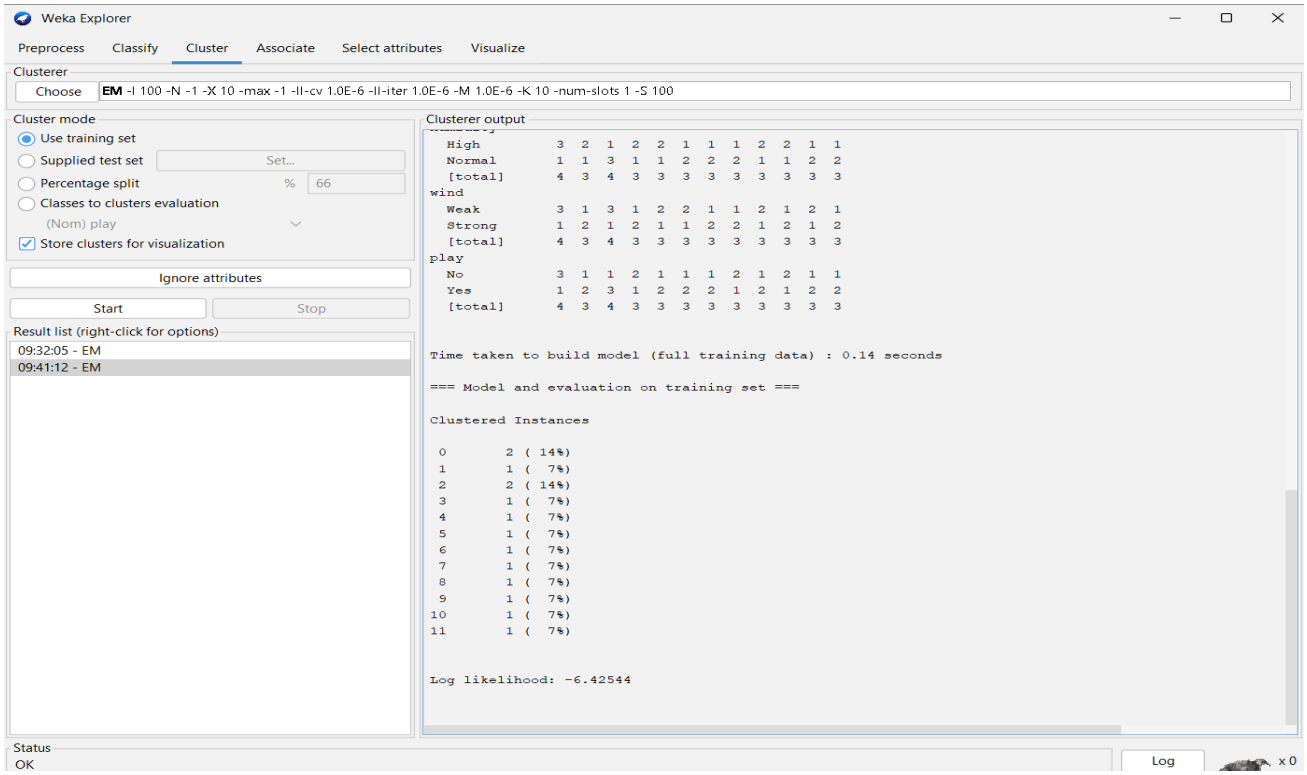
xvi). Apply Naïve Bayes Classifier to the Weather play data set and derive the probability for play no given outlook rainy.

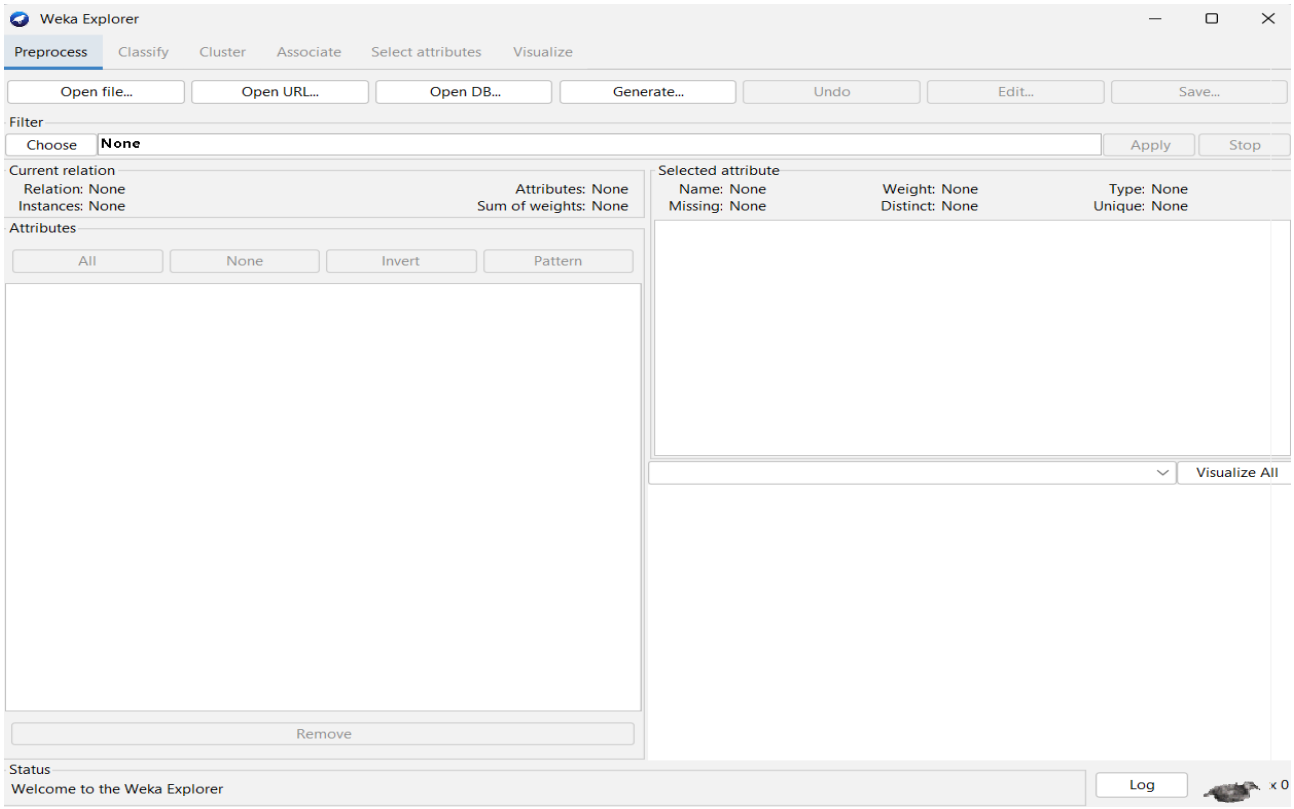
1. Classify → Choose → bayes → NaiveBayes
2. Test options → Use training set / Cross-validation
3. Start → in output → check **Conditional probabilities** table → find Play=No for Outlook=Rainy.



xv) Classification → Clustering → Class-to-Cluster Evaluation → Classification on unlabeled data.

1. Preprocess → Remove **Play** attribute → Save as unlabeled dataset
2. Cluster → Choose → EM or SimpleKMeans → Start → get cluster assignments
3. Class → Cluster Evaluation → Classify → load original labeled dataset → select cluster assignments → Evaluate
4. Classify on unlabeled dataset (without Play) → Choose classifier → Start → view predictions
5. Prepare **analysis report** → include classification accuracy, clusters, class-to-cluster mapping, and prediction.




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: None
Instances: None
Attributes: None
Sum of weights: None
Selected attribute
Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None
Attributes
All
None
Invert
Pattern
Visualize All
Remove
Status
Welcome to the Weka Explorer
Log
x 0