LAB PROGRAM – 1

1. Experiment to be conducted using WEKA tool:

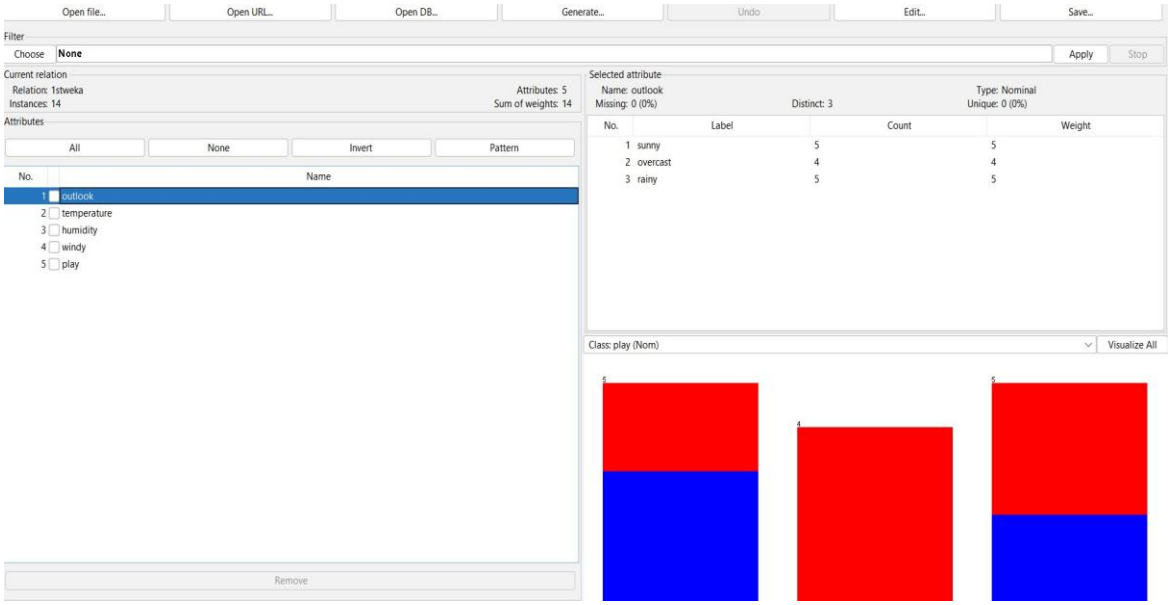
1	outlook	temperatu	humidity	windy	play
2					
3	sunny	85	85	FALSE	no
4	sunny	80	90	TRUE	no
5	overcast	83	86	FALSE	yes
6	rainy	70	96	FALSE	yes
7	rainy	68	80	FALSE	yes
8	rainy	65	70	TRUE	no
9	overcast	64	65	TRUE	yes
10	sunny	72	95	FALSE	no
11	sunny	69	70	FALSE	yes
12	rainy	75	80	FALSE	yes
13	sunny	75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

a) Preprocess and Classify panels
b) Draw the histogram to show how the values of the play class occurs for each value of the outlook attribute
c) Derive minimum and maximum values, mean, and standard deviation
d) Perform operations such as filter, delete, invert, Pattern, Undo, Edit, search, Select,Conversions etc
e) Build the decision tree and analyze the weather data.
f) Examine the Output , classification error and Kappa statistics
g) Visualize threshold curve
h) Apply Logistic Regression model to classify
i) Measure the log likelihood of the clusters of training data. (Consider large data set.)

a. Preprocess and Classify panels

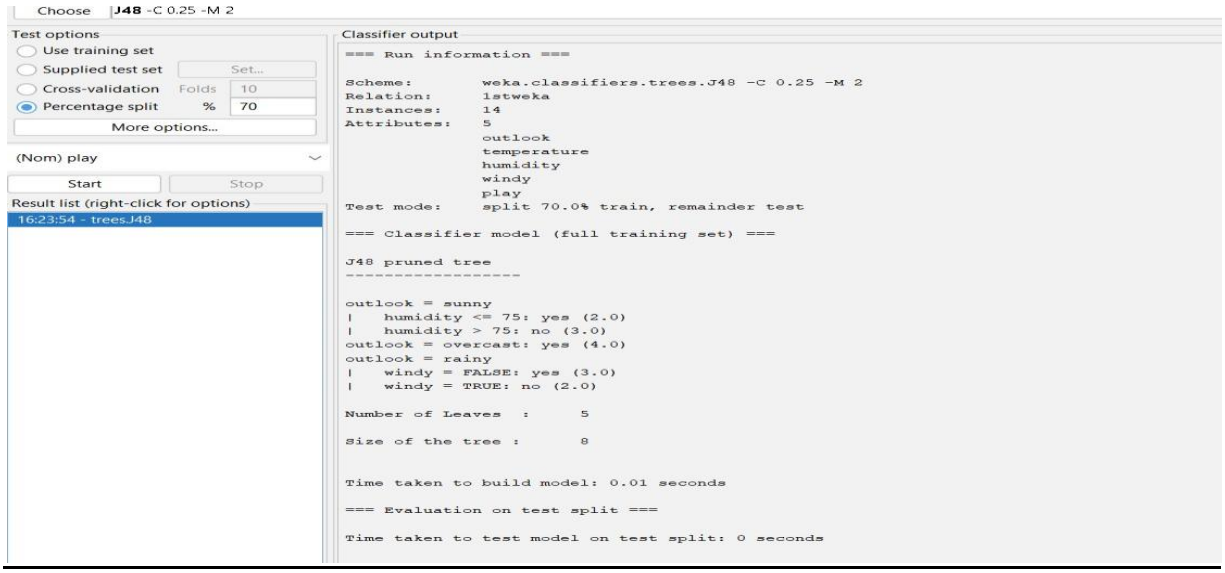
Preprocess Panel Steps :

- Load Dataset: Start in the Preprocess tab. Click Open file to load your dataset (typically in .arff or .csv format).
- View Attributes: See a summary of attributes (features) on the right. Click each attribute to view its unique values and distribution.
- Apply Filters: Use the Choose button in the Filters section to apply data preprocessing techniques, such as normalization, discretization, or missing value handling.
- Modify Data: Use attribute-specific options like removing or renaming attributes, as needed for your analysis.



Classify Panel Steps:

- Choose Classifier: Go to the Classify tab. Click Choose to select a classification algorithm, such as J48 for decision trees or NaiveBayes.
- Set Test Options: Select a testing method under Test options (e.g., cross-validation or percentage split).
- Run Classification: Click Start to train and test the model. Weka will display results, including accuracy, confusion matrix, and other metrics, in the Classifier output section at the bottom.



Choose **J48** -C 0.25 -M 2

Test options
- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 70

More options...

(Nom) play

Start | Stop

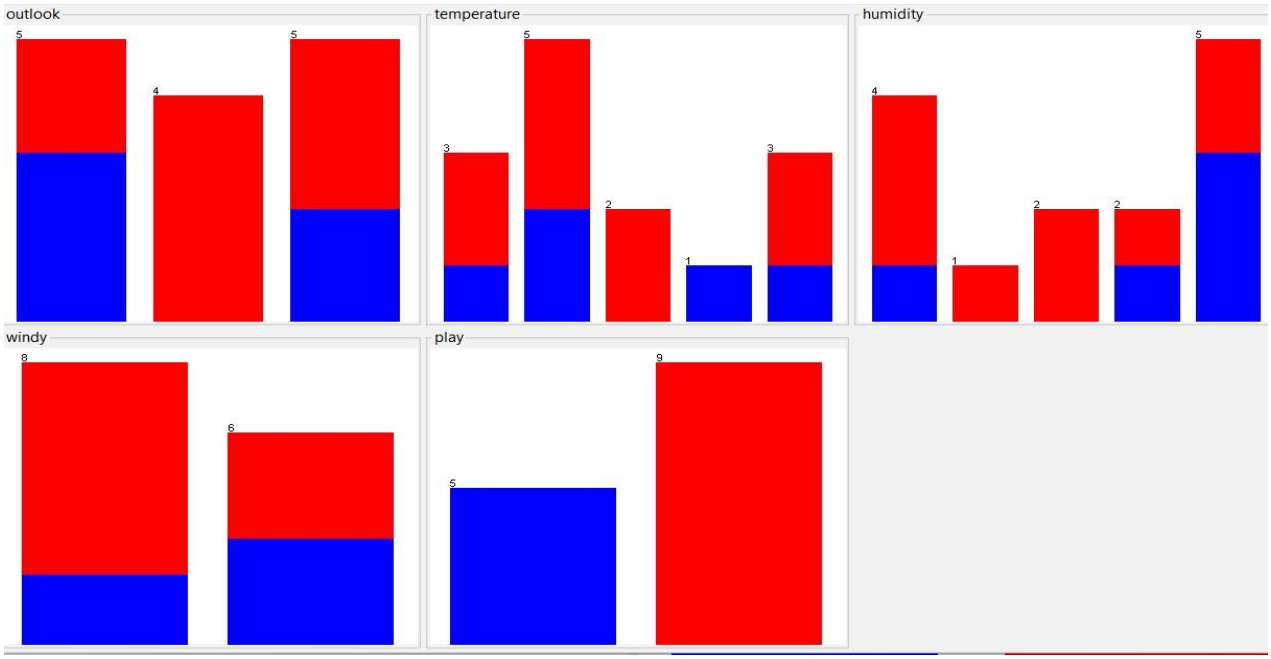
Result list (right-click for options)

16:23:54 - trees.J48

Classifier output

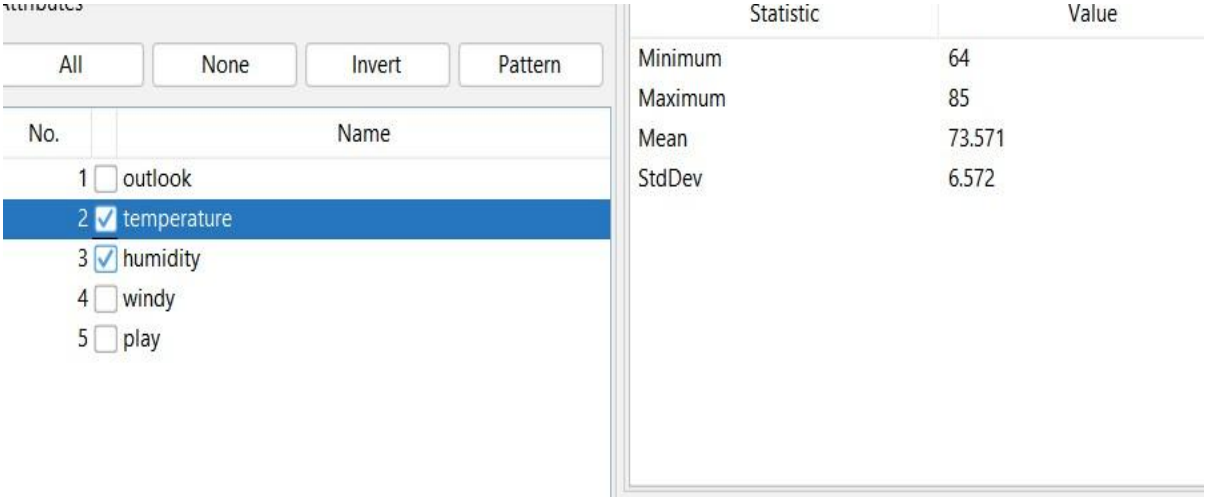
```
Number of Leaves  :     5

Size of the tree :      8


Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances           1               25      %
Incorrectly Classified Instances         3               75      %
Kappa statistic                         -0.5
Mean absolute error                      0.75
Root mean squared error                  0.866
Relative absolute error                150      %
Root relative squared error            164.3168 %
Total Number of Instances                4

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.500    0.000      0.000    0.000      -0.577   0.250     0.500     no
                 0.500    1.000    0.333      0.500    0.400      -0.577   0.250     0.417     yes
Weighted Avg.    0.250    0.750    0.167      0.250    0.200      -0.577   0.250     0.458

=== Confusion Matrix ===

 a b   <-- classified as
 0 2 | a = no
 1 1 | b = yes
```

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



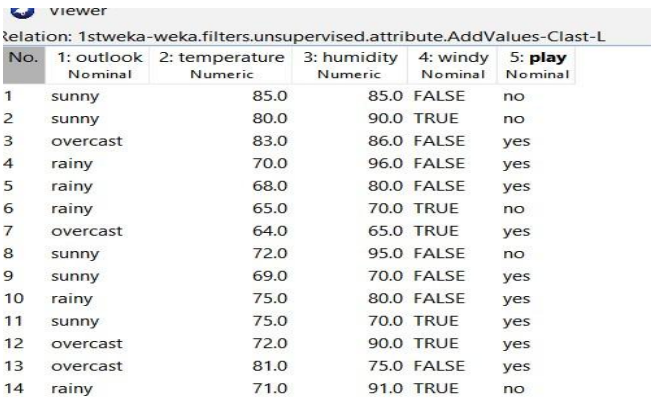
c. Derive minimum and maximum values, mean, and standard deviation

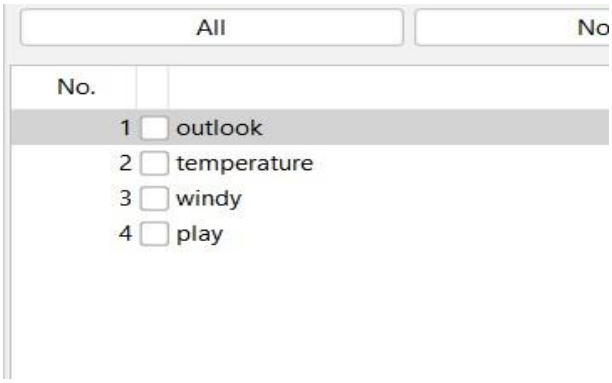
- Go to the **Preprocess** tab and load your dataset.
- Select the attribute for which you want to calculate statistics in the **Attributes** panel on the right.
- At the bottom of the **Preprocess** tab, find the **Selected attribute** summary box.
- In this summary box, Weka displays statistics for the selected attribute, including **Minimum**, **Maximum**, **Mean**, and **Standard Deviation**.
- Repeat for any other attributes as needed to view their respective statistics.

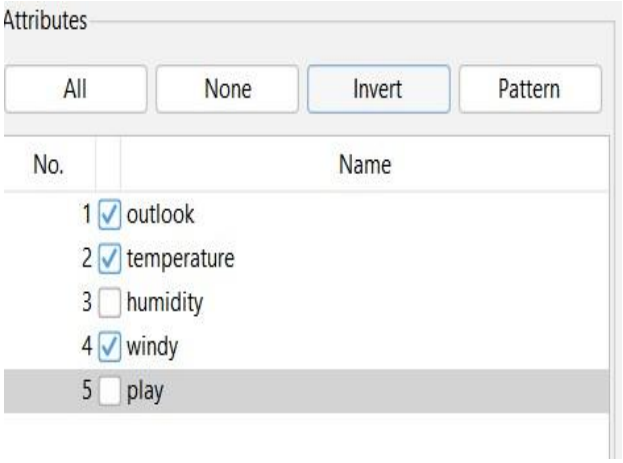


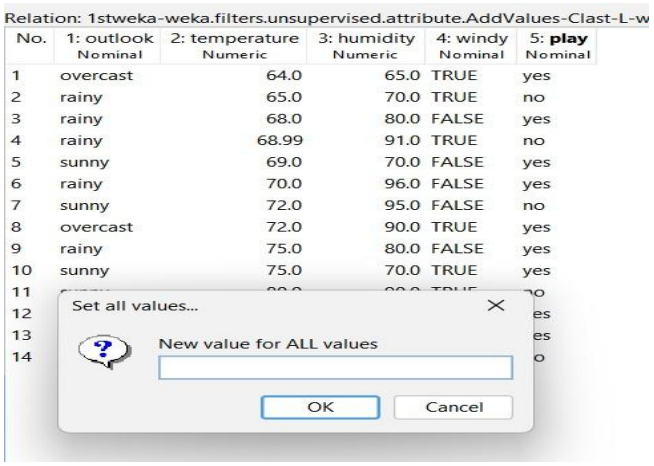
d. Perform operations such as filter, delete, invert, Pattern, Undo, Edit, search, Select, Conversions etc

- Filter: Go to Preprocess → Filters and choose filters for data preprocessing, such as normalization or discretization.
- Delete/Invert: Select attributes in the Attributes panel and use Remove or Invert Selection to manage selections.
- Pattern/Search: Use filters like StringToWordVector in Filters to search or match patterns within text attributes.
- Undo/Edit: After modifications, click Undo at the bottom to revert changes; for attribute-specific edits, use the Edit button.
- Conversions: Select Filters → unsupervised → attribute → NumericToNominal or similar to convert attribute types as needed.





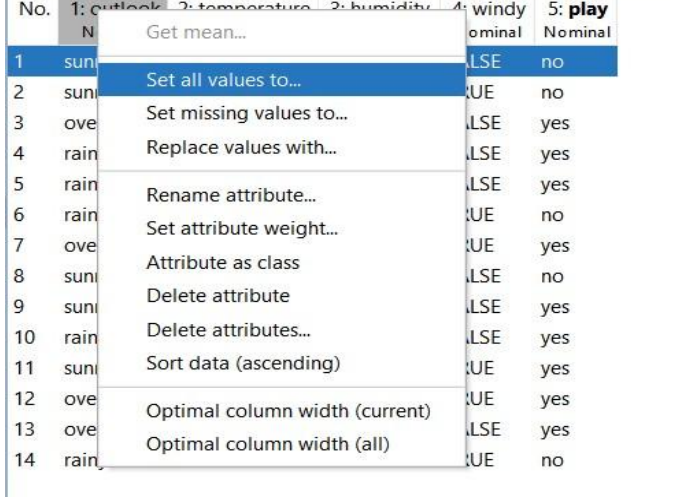




elation: 1stweka-weka.filters.unsupervised.attribute.AddValues-Clast-L-weka.filters.unsupervised.attribute.StringToWordVe

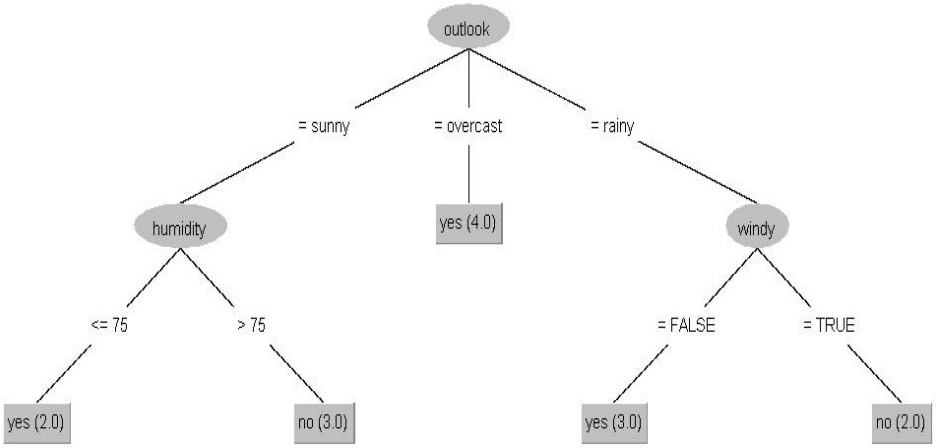
No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

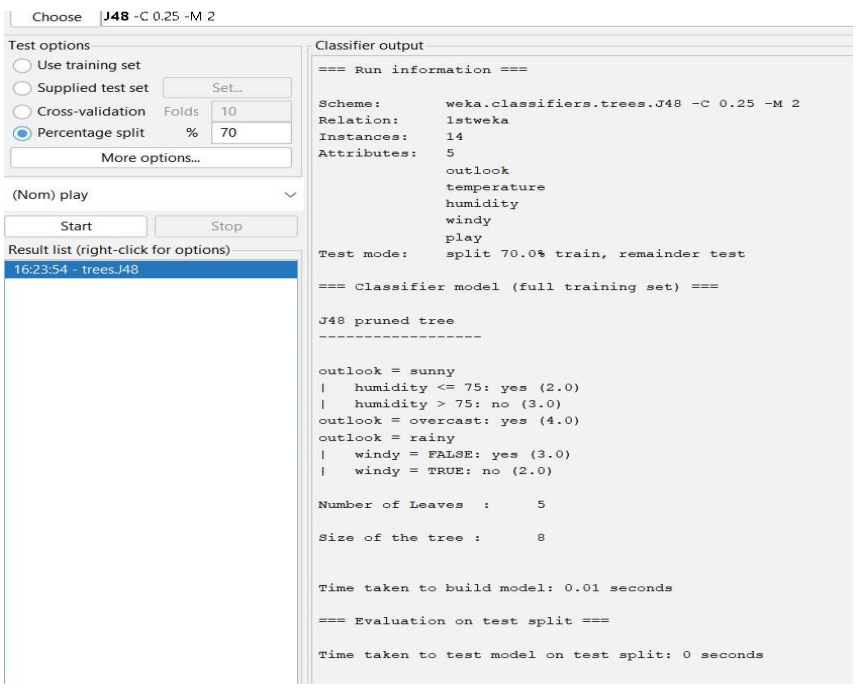
Relation: 1stweka



e. Build the decision tree and analyze the weather data

- Go to Classify tab , click choose , and select J48 under trees.
- Set test options to evaluate the model, then click start.
- Right click on J48 trees and select visualize the decision tree.





f. Examine the Output , classification error and Kappa statistics

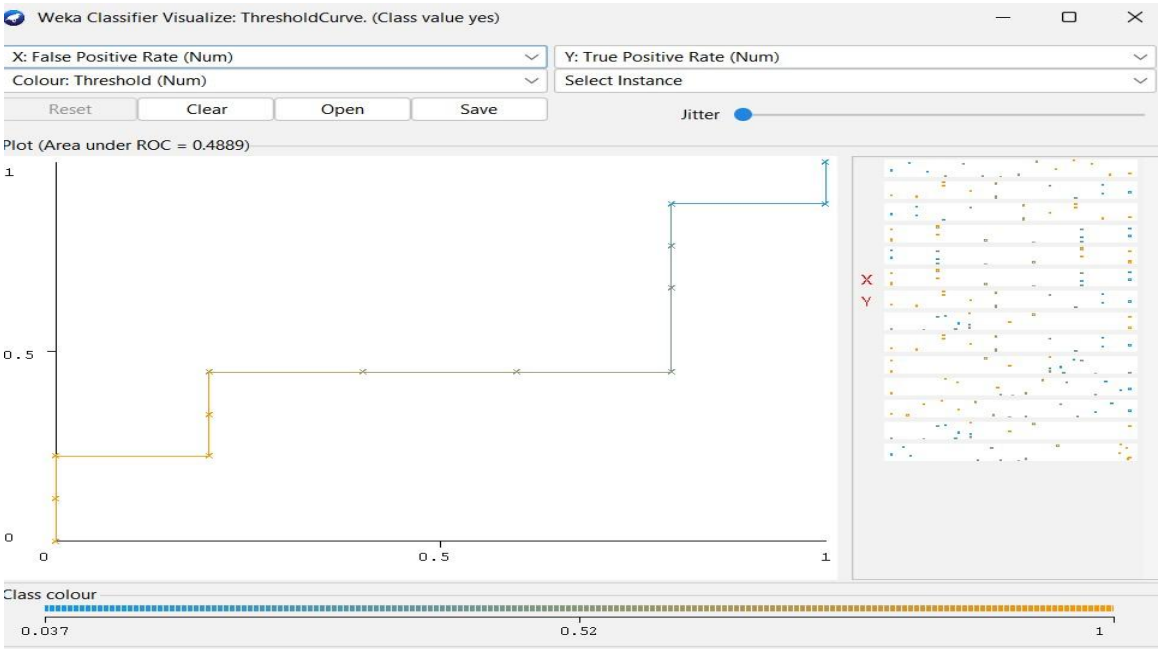
- After running your classifier in the Classify tab, go to the Classifier output section at the bottom.
- Find Correctly Classified Instances and Incorrectly Classified Instances to view classification accuracy and error rates.

- Look for Kappa statistic in the output, which indicates the agreement between predicted and actual classes adjusted for chance, helping assess model reliability.

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances           6               42.8571 %
Incorrectly Classified Instances         8               57.1429 %
Kappa statistic                         -0.1429
Mean absolute error                      0.4699
Root mean squared error                  0.5738
Relative absolute error                 98.6821 %
Root relative squared error            116.299  %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.400    0.556    0.286      0.400    0.333      -0.149   0.489     0.404     no
                 0.444    0.600    0.571      0.444    0.500      -0.149   0.489     0.744     yes
Weighted Avg.    0.429    0.584    0.469      0.429    0.440      -0.149   0.489     0.623

=== Confusion Matrix ===

 a b   <-- classified as
 2 3 | a = no
 5 4 | b = yes
```

g. Visualize threshold curve

- In the Classify tab, check More options and enable Output predictions to include probability estimates.
- Run your classifier (e.g., J48 or NaiveBayes), then right-click the resulting entry in the Result list and select Visualize threshold curve.
- Choose the class you want to analyze from the dropdown, and the threshold curve will display,



h.Apply Logistic regression model to classify

- In the Classify tab, click Choose and select Logistic under functions.
- Set Test options (e.g., cross-validation or percentage split), then click Start to run the classifier.
- Review the Classifier output for accuracy, coefficients, and other metrics that indicates **i. Measure the log likelihood of the clusters of training data. (Consider large data set.)**
- To measure the log likelihood of clusters in Weka:
- Go to the Cluster tab, click Choose, and select EM (Expectation-Maximization) from Clusterers.
- Configure EM options if needed, then click Start to apply clustering to your dataset.
- In the Clusterer output section, find the Log likelihood value, which indicates the model's fit to the training data**.**

```
=== Run information ===

Scheme:       weka.classifiers.trees.LMT -I -1 -M 15 -W 0.0
Relation:     1stweka-weka.filters.unsupervised.attribute.AddValue
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    10-fold cross-validation

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

Logistic model tree
------------------
: LM_1:11/11 (14)

Number of Leaves  :     1

Size of the Tree :      1
LM_1:
Class no :
-6.95 +
[outlook=sunny] * 0.65 +
[outlook=overcast] * -2.82 +
[temperature] * 0.02 +
[humidity] * 0.06 +
[windy=TRUE] * 1.38

Class yes :
6.95 +
[outlook=sunny] * -0.65 +
[outlook=overcast] * 2.82 +
[temperature] * -0.02 +
[humidity] * -0.06 +
```

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

Load Data: Load your dataset from the *Preprocess* tab.

Apply Clustering Algorithm:

- Go to the *Cluster* tab.
- Select EM as the clustering algorithm (other algorithms like SimpleKMeans do not provide log likelihood).
- Set parameters as needed, then click *Start*.

View Log Likelihood:

- In the output, Weka provides the log likelihood of the model on the dataset after clustering is complete.