

# DATA MINING & VISUALIZATION

## LAB 01

1. Experiment to be conducted using WEKA tool:

1	outlook	temperature	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

1. Preprocess and Classify panels
2. Draw the histogram to show how the values of the play class occurs for each value of the outlook attribute
3. Derive minimum and maximum values, mean, and standard deviation
4. Perform operations such as filter, delete, invert, Pattern, Undo, Edit, search, Select, Conversions etc
5. Build the decision tree and analyze the weather data.
6. Examine the Output , classification error and Kappa statistics
7. Visualize threshold curve
8. Apply Logistic Regression model to classify



□ **Preprocess and Classify Panels:**

- Open WEKA and load the dataset by clicking on "Open file" in the "Explorer" window and selecting your file.
- In the "Preprocess" panel, explore the data attributes and remove any unnecessary ones. You can also handle missing values and normalize data if needed.

□ **Draw Histogram for the 'Outlook' Attribute:**

- In the "Preprocess" panel, click on the "Visualize All" button or select "Visualize" on the attribute panel to see a histogram.
- Select "Outlook" as the attribute and "Play" as the class to view how the "Play" class varies with different "Outlook" values.

□ **Derive Minimum, Maximum, Mean, and Standard Deviation:**

- In the "Preprocess" panel, click on the attribute, and WEKA will show basic statistics like minimum, maximum, mean, and standard deviation for numeric attributes.

□ **Filter, Delete, Invert, Pattern, and Other Operations:**

- **Delete:** Remove unnecessary attributes.
- **Invert:** Invert the selected attributes.
- **Pattern:** Search for specific patterns in data if needed.

□ **Build a Decision Tree:**

- Go to the "Classify" panel.
- Select a decision tree algorithm like J48 from the list of classifiers.
- Set the test options (e.g., use 10-fold cross-validation).
- Click "Start" to build and analyze the decision tree model.

□ **Examine Output, Classification Error, and Kappa Statistics:**

- Once the model runs, check the output for classification accuracy, confusion matrix, and Kappa statistic (which measures agreement).

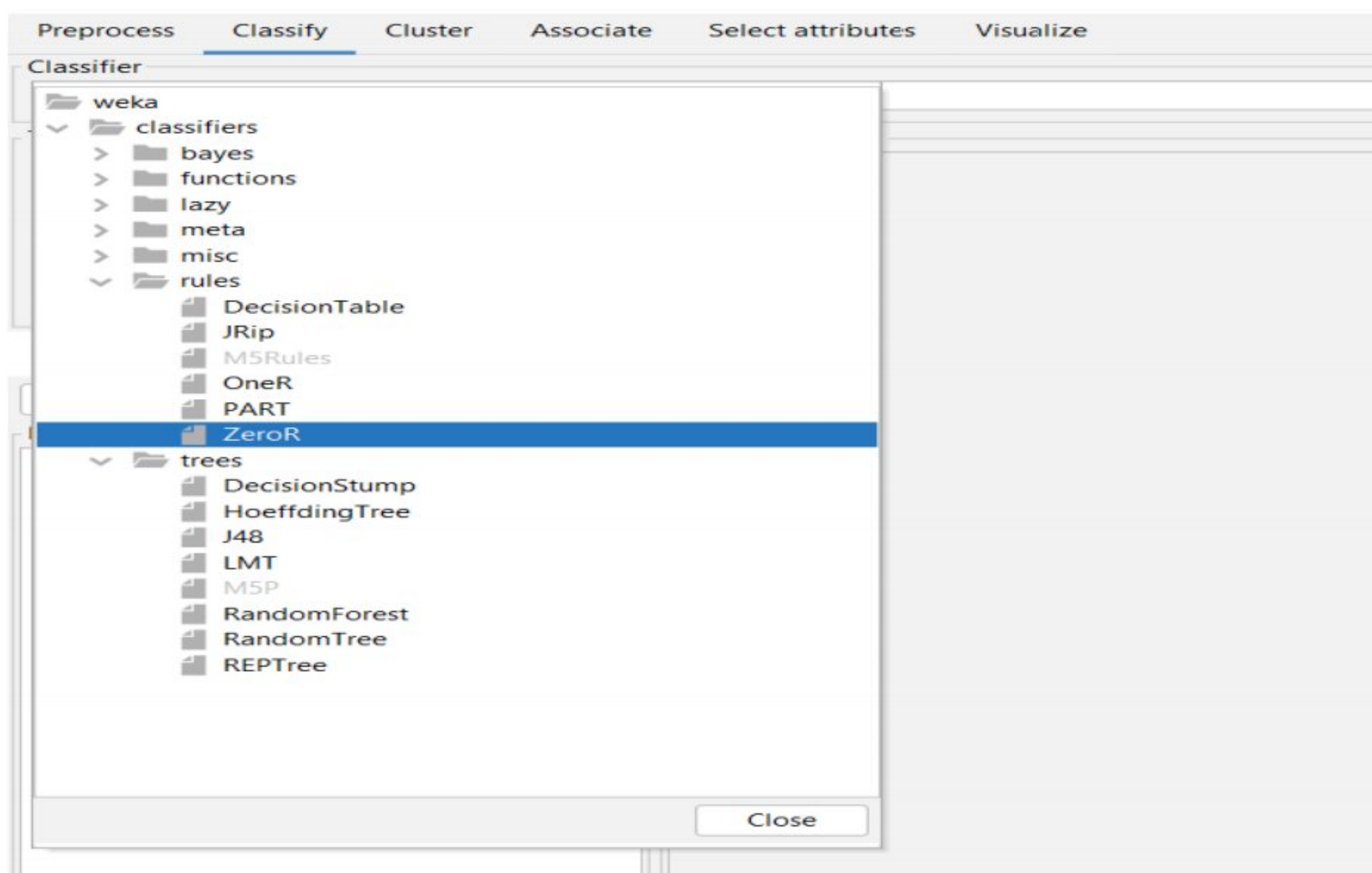
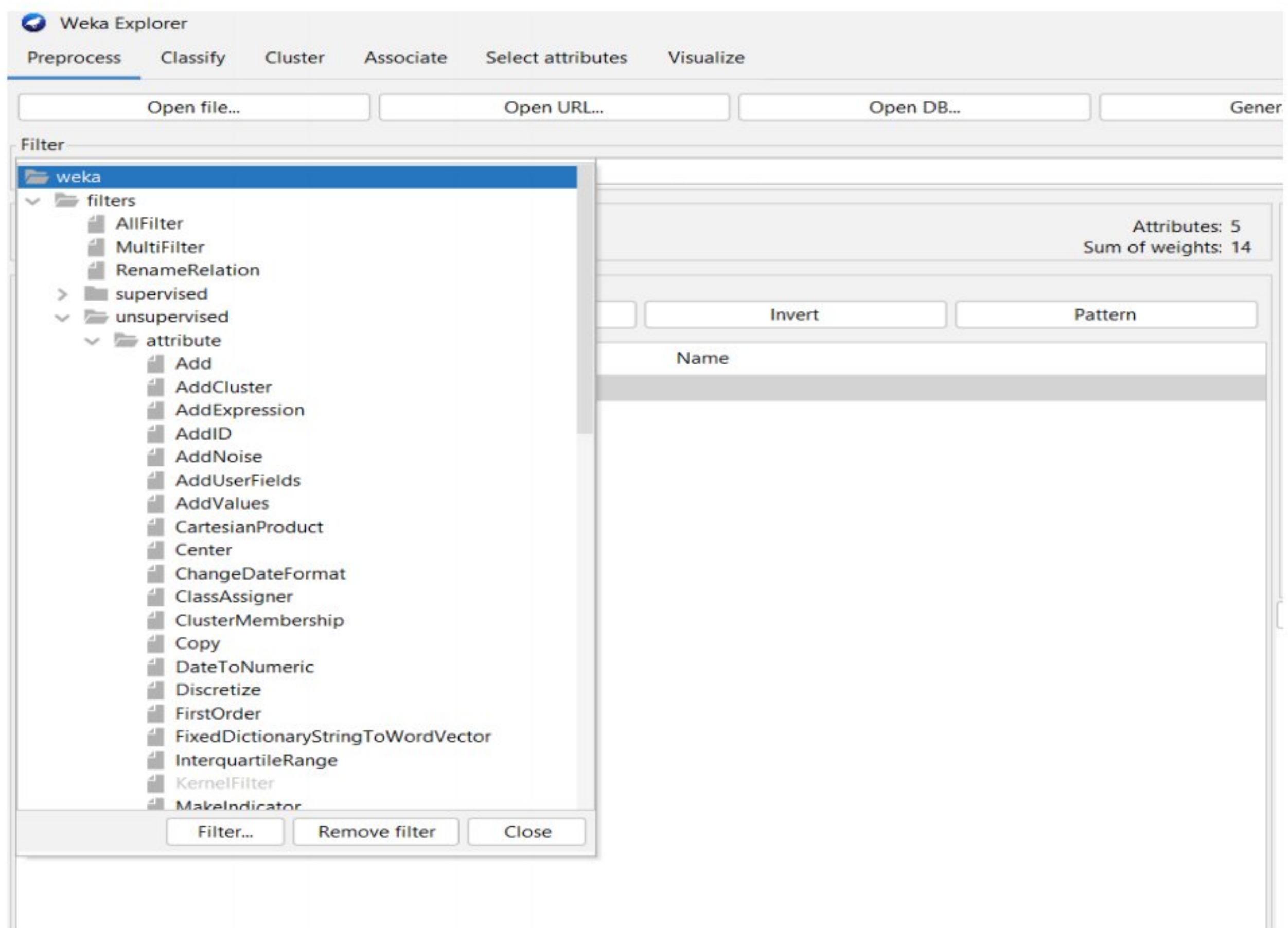
□ **Visualize Threshold Curve:**

- In the "Classify" panel, after running the model, right-click on the result and select "Visualize threshold curve."
- Choose the class to visualize the curve, showing how varying thresholds impact classification.

□ **Apply Logistic Regression Model:**

- In the "Classify" panel, select "Logistic" under the list of classifiers to apply a logistic regression model.
- Run the model, and review the output to analyze the performance.





2. To draw a histogram that shows how the values of the "play" class occur for each value of the "outlook" attribute in Weka:

**Step 1: Load Data**

**Open Weka Explorer** and load dataset.

**Step 2: Visualize Data**

**Go to the Preprocess Tab:** Here you can see all your attributes.

**Select the "outlook" Attribute:** Click on the "outlook" attribute to highlight it.

**Step 3: Visualize with the Histogram**

**Visualize:** Click on the "Visualize" button.

**Choose Attribute:** In the visualization window, select the "outlook" attribute on the X-axis and the "play" class on the Y-axis (you may need to switch to the "Class" view to focus on class distributions).

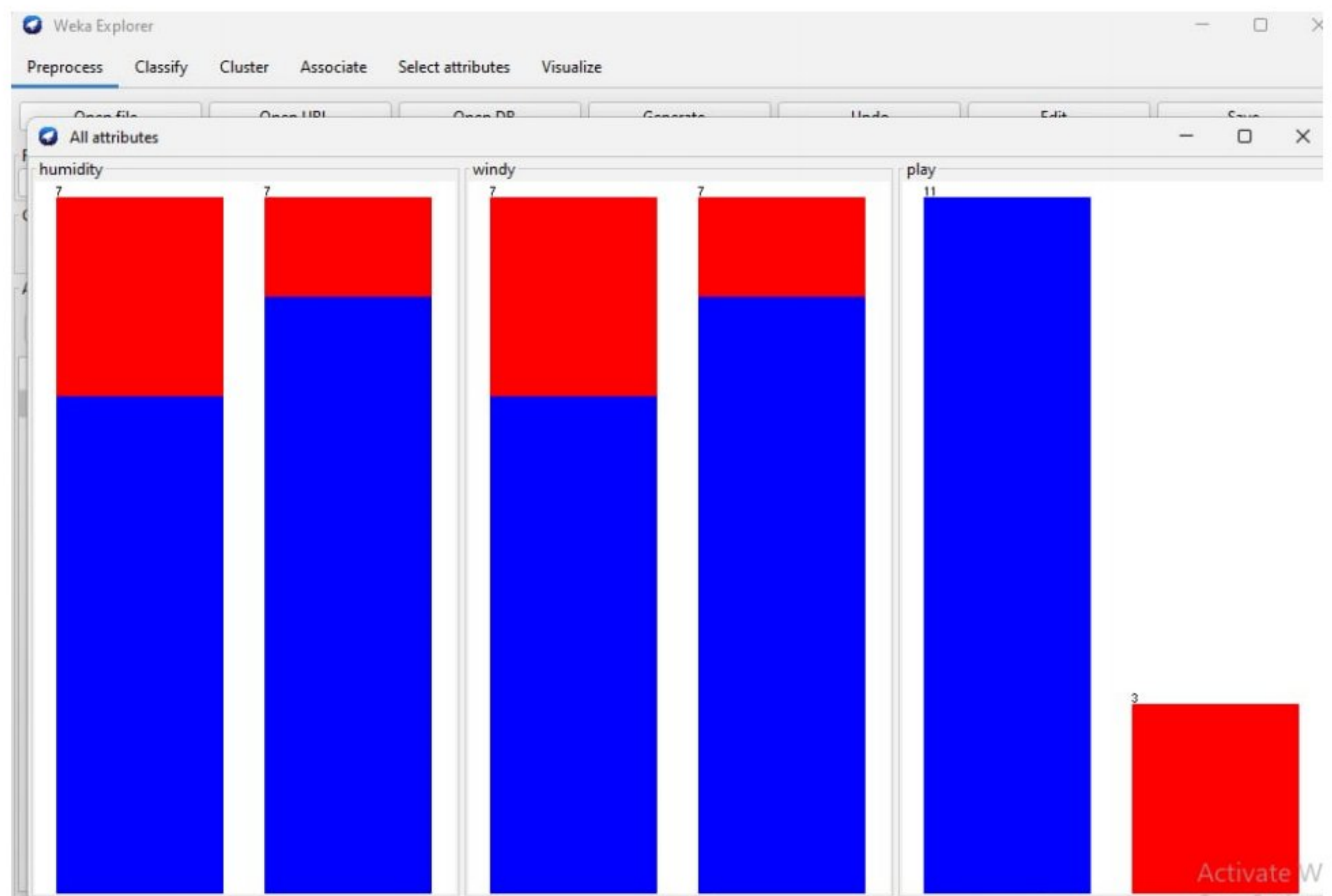
**Plot:** Weka will generate a histogram that shows the counts of "play" outcomes for each "outlook" category.

**Step 4: Customize the Visualization (if needed)**

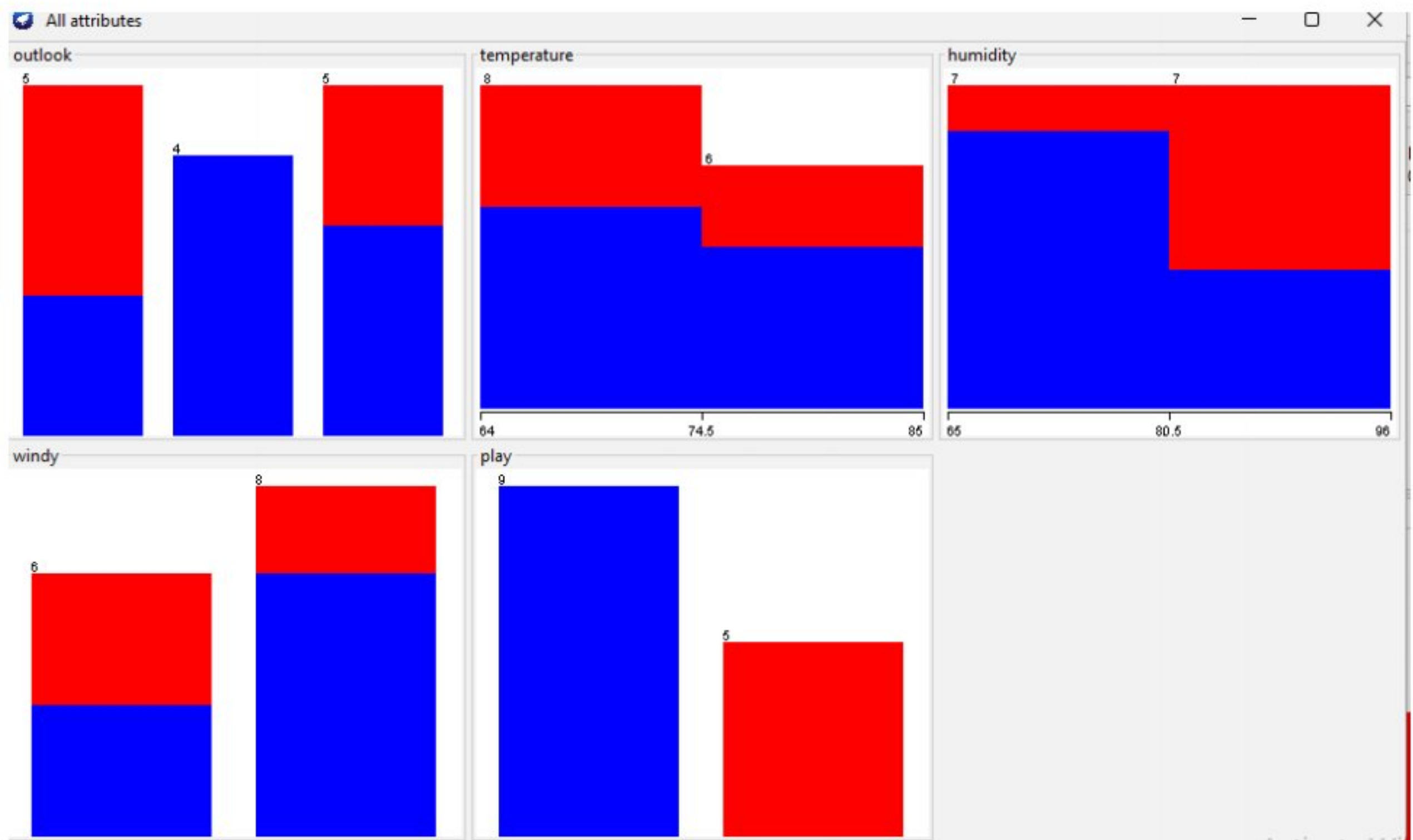
You can adjust colors, add labels, or change the style in the visualization window to make the histogram clearer.

**Step 5: Analyze the Histogram**

Look for patterns in how "play" values change across different "outlook" values.







3. To derive the minimum and maximum values, mean, and standard deviation for an attribute in Weka:

#### Step 1: Load Data

**Open Weka:** Start the Weka GUI and load dataset.

#### Step 2: Go to the Preprocess Tab

**Select the Preprocess Tab:** This is where you can view your dataset and its attributes.

#### Step 3: Select the Attribute

**Select an Attribute:** Click on the attribute you want to analyze (e.g., a numeric attribute).

#### Step 4: View Statistical Information

**Statistics Panel:** After selecting the attribute, look at the bottom of the Preprocess tab. You'll see a panel that displays various statistics.

**Locate the Statistics:** This panel will show:

- **Minimum:** The smallest value of the attribute.
- **Maximum:** The largest value of the attribute.
- **Mean:** The average value of the attribute.
- **Standard Deviation:** This measures the dispersion of the values.

Clusterer output	
Number of iterations performed: 2	
	Cluster
Attribute	0
	(1)
=====	
outlook	
sunny	6
overcast	5
rainy	6
[total]	17
temperature	
mean	73.5714
std. dev.	6.3326
humidity	
mean	81.6429
std. dev.	9.9111
windy	
TRUE	7
FALSE	9
[total]	16
play	
yes	10
no	6
[total]	16
Time taken to build model (full training data) : 0 seconds	
=== Model and evaluation on training set ===	

4. In Weka, various data operations such as filtering, deleting attributes, inverting selections, and more can be performed in the Preprocess tab:

### 1. Filter

#### Apply Filter:

- Click on the "Choose" button under the "Filter" section.
- Select a filter from the list (e.g., Remove, Normalize, etc.).
- Configure the filter settings as needed and click "Apply."

### 2. Delete Attributes or Instances

#### Delete Attribute:

- Select the attribute in the Attributes list.
- Click the "Remove" button (or right-click and select "Remove").

#### Delete Instances:

- Select the instances in the data table.
- Right-click and choose "Remove" to delete the selected instances.

### 3. Invert Selection

To invert your selection (e.g., if you've selected certain attributes or instances), use the "Invert Selection" option found in the right-click context menu.

### 4. Pattern (Search for Patterns)

Use the "Select" feature to find specific patterns in your data.



- You can apply the "Select attributes" tab to look for attributes that are significantly correlated with the class.

## 5. Undo

To undo the last action, click the "Undo" button located at the top of the Preprocess tab.

## 6. Edit

### Edit Instances:

- Select an instance in the table and use the "Edit" button to modify its values directly.

## 7. Search

### Search for Attributes or Instances:

- Use the search box above the attributes list or the instances table to quickly find specific attributes or instances.

## 8. Select

### Select Attributes:

- You can manually select attributes by holding the Ctrl or Shift key while clicking on them, or use "Select All" from the right-click menu.

### Select Instances:

- Similar to attributes, you can select instances using the same method in the instances table.

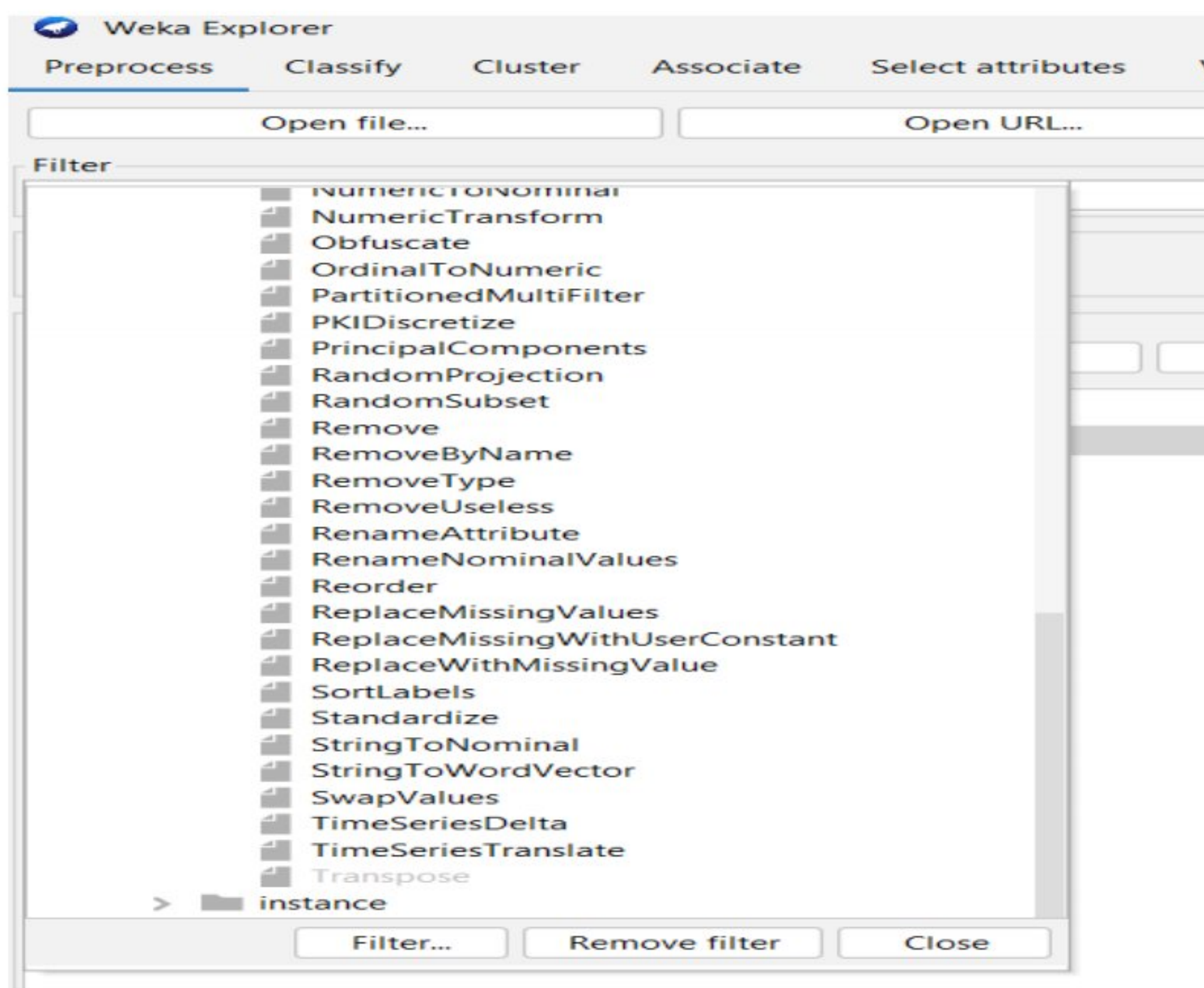
## 9. Conversions

### Convert Numeric to Nominal:

- Use the "NumericToNominal" filter to convert numeric attributes to nominal.

### Convert Nominal to Numeric:

- Use the "NominalToNumeric" filter if you need to perform numerical operations on nominal attributes.





5. To build and analyze a decision tree for weather data in Weka:

### **Preprocess the Data**

1. **Inspect Attributes:** Check the attributes to ensure they are in the correct format (nominal for categorical data).
2. **Handle Missing Values:** If there are any missing values, you can use filters like "ReplaceMissingValues" or remove instances/attributes as needed.
3. **Select Attributes:** Optionally, you can use the "Select attributes" feature to find relevant features.

### **Build the Decision Tree**

1. **Go to the Classify Tab:** Click on the "Classify" tab.
2. **Choose Classifier:**
  - Click on the "Choose" button and select trees → J48 (Weka's implementation of the C4.5 algorithm).
3. **Set Parameters** (if needed):
  - Right-click on J48 to set options (like pruning or confidence levels).
4. **Set Test Options:**
  - Choose "Cross-validation" (e.g., 10-fold) or "Percentage split" to evaluate the model.

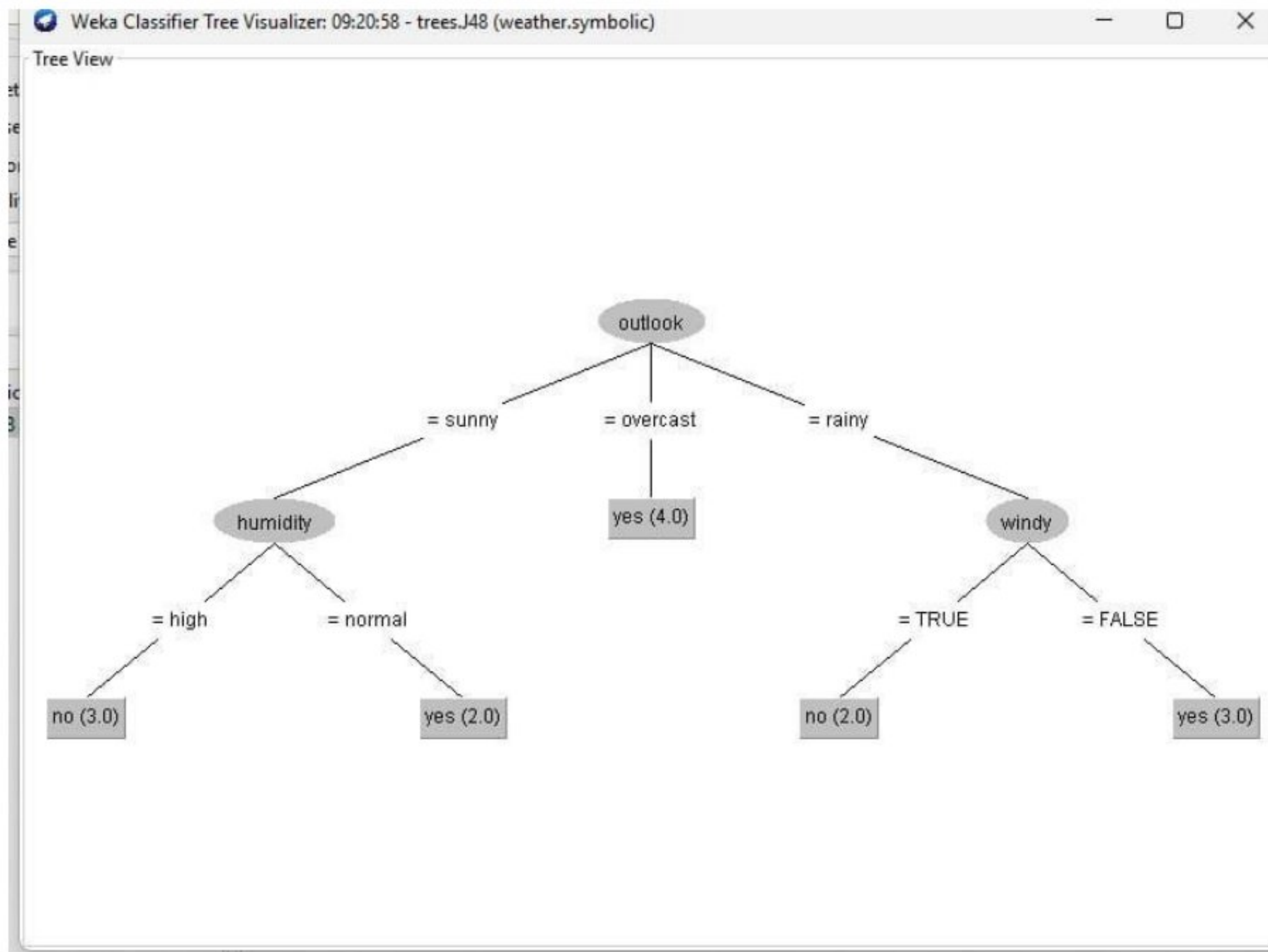
### **Run the Classifier**

1. **Start Classification:** Click on the "Start" button to build the decision tree.
2. **View Results:** Once the classification is complete, Weka will show the results in the output window, including accuracy, confusion matrix, and detailed performance metrics.

### **Analyze the Decision Tree**

1. **Visualize the Tree:**
  - After running the classifier, click on the "Visualize tree" button (if available) to open a graphical representation of the decision tree.
2. **Interpret the Tree:**
  - Analyze the tree structure to understand how decisions are made based on the attributes. Each node represents a decision based on an attribute, leading to leaf nodes that represent class labels.
  -





6. To examine the output, classification error, and Kappa statistics in Weka after building a decision tree (or any classifier):

### Step 1: Review the Output Window

After running your classifier, the output window will display several key statistics and metrics.

### Step 2: Key Metrics to Examine

1. **Correctly Classified Instances:** This tells you how many instances were correctly classified.
2. **Incorrectly Classified Instances:** This shows how many instances were misclassified.
3. **Classification Accuracy:** Calculated as:

$$\text{Accuracy} = \frac{\text{Correctly Classified Instances}}{\text{Total Instances}} \times 100$$

This percentage indicates the overall effectiveness of your model.

4. **Kappa Statistic:**
  - The Kappa statistic measures the agreement between the predicted and actual classifications while adjusting for chance. A value closer to 1 indicates strong agreement, while a value closer to 0 indicates poor agreement.



- The output will show the Kappa value, which you should examine for understanding the reliability of your classifier.

#### 5. **Confusion Matrix:**

- This matrix provides a detailed breakdown of the true positive, true negative, false positive, and false negative counts for each class.
- Analyze the confusion matrix to see where your model is making mistakes and which classes are being confused with each other.

### Step 3: Classification Error

- **Classification Error Rate:**

- This is derived from the incorrectly classified instances and can be calculated as:

$$\text{Error Rate} = \frac{\text{Incorrectly Classified Instances}}{\text{Total Instances}} \times 100$$
  

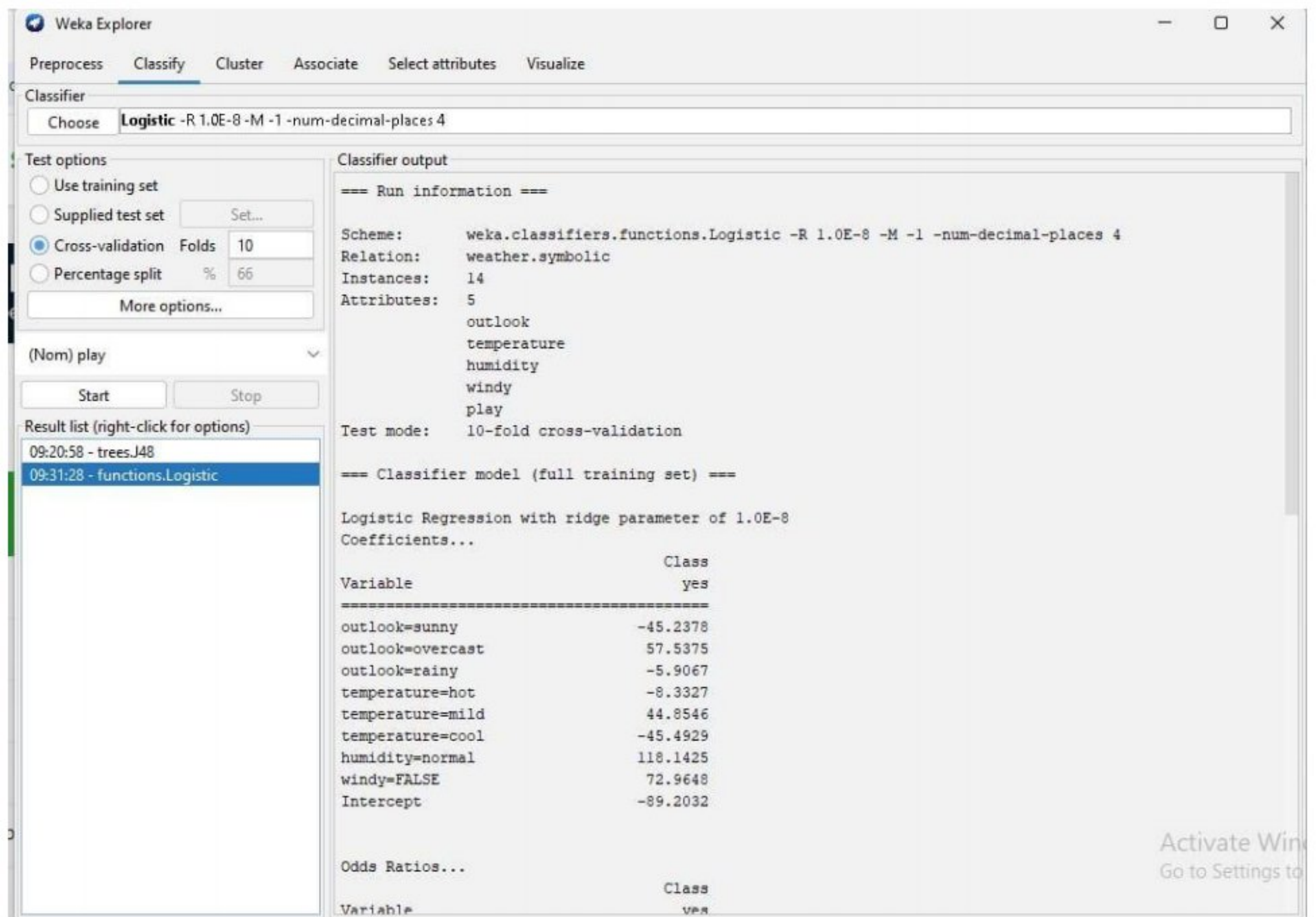
$$\text{Error Rate} = \frac{\text{Total Instances} - \text{Correctly Classified Instances}}{\text{Total Instances}} \times 100$$

- This metric helps you understand the proportion of instances that were misclassified, providing insight into model performance.

### Step 4: Summary

1. **Look for Trends:** Pay attention to which classes have high misclassification rates in the confusion matrix. This can help identify patterns or specific areas where the model may need improvement.
2. **Compare Kappa and Accuracy:** If your accuracy is high but Kappa is low, this might suggest that the model is biased towards a certain class.





7. To visualize a threshold curve (also known as a ROC curve) in Weka, follow these steps:

### Step 1: Load Data

**Open Weka:** Launch the Weka GUI.

**Explorer:** Click on the "Explorer" button.

**Open File:** Load your dataset.

### Step 2: Build a Classifier

**Go to the Classify Tab:** Click on the "Classify" tab.

**Choose Classifier:** Select a classifier, such as trees → J48 or any other suitable classifier.

**Set Test Options:** Choose "Cross-validation" or "Percentage split" to evaluate the model.

**Run the Classifier:** Click on the "Start" button.

### Step 3: Generate ROC Curve

**Look for ROC Options:** After running the classifier, in the result output area, look for an option that mentions "Plot ROC" or similar.

**Plot ROC Curve:**

- Click on the "Visualize threshold curve" button or check if there's a "Plot ROC" option depending on the classifier and its configuration.

**Review the ROC Curve:**



- The ROC curve will show the True Positive Rate (sensitivity) against the False Positive Rate at various threshold settings.

•

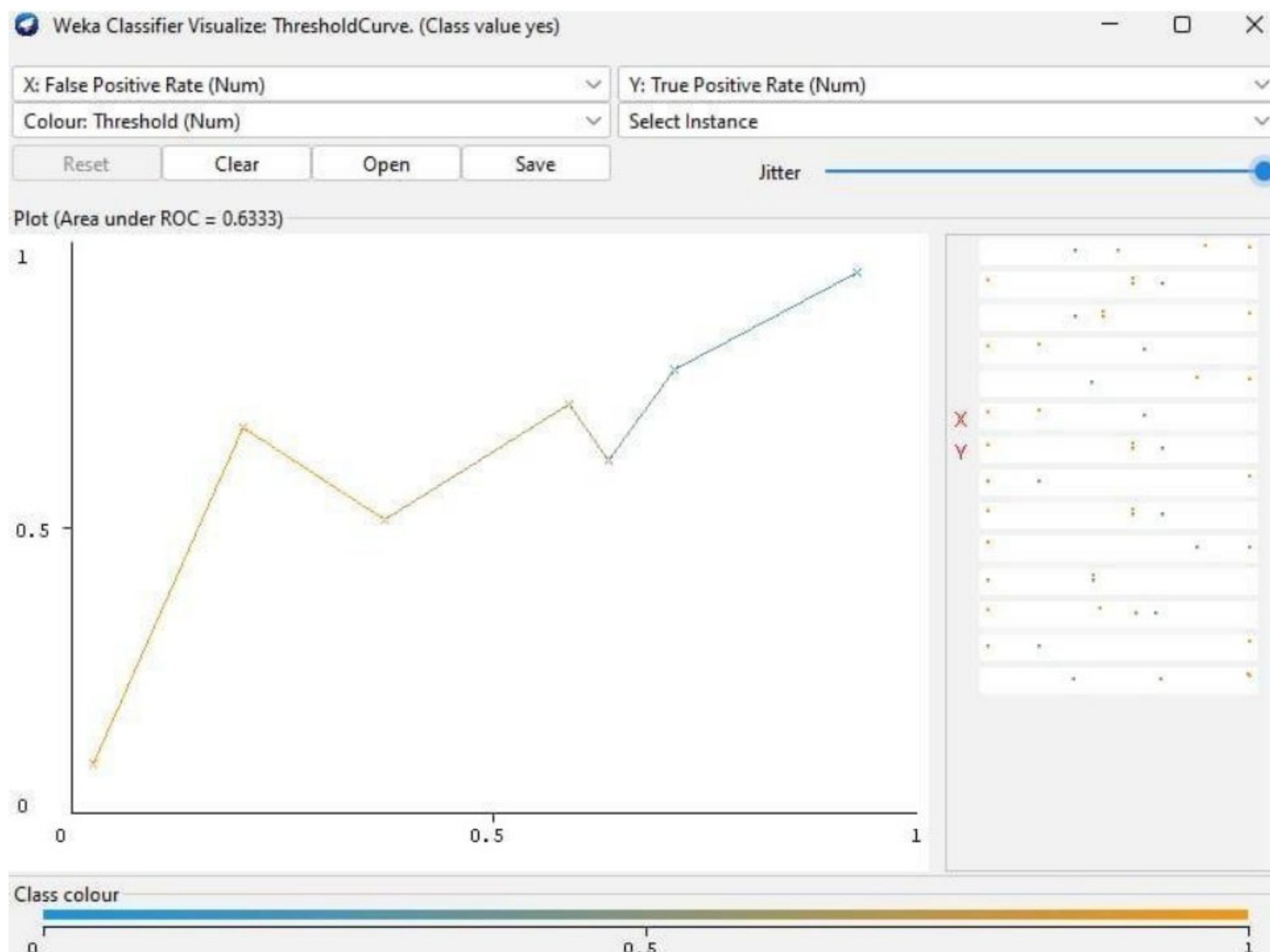
#### Step 4: Analyze the ROC Curve

##### Interpret the Curve:

- The area under the curve (AUC) represents the classifier's ability to distinguish between classes. AUC values range from 0 to 1, where:
  - 0.5 indicates no discrimination (equivalent to random guessing).
  - Values closer to 1 indicate excellent discrimination.

##### Threshold Selection:

- You can adjust the threshold for classification based on the ROC curve to optimize the balance between sensitivity and specificity according to your needs.



## 8. Build the Logistic Regression Model

**Go to the Classify Tab:** Click on the "Classify" tab.

##### Choose Classifier:

- Click on the "Choose" button and navigate to functions → Logistic.



## Set Test Options:

- Choose "Cross-validation" (e.g., 10-fold) or "Percentage split" to evaluate the model.

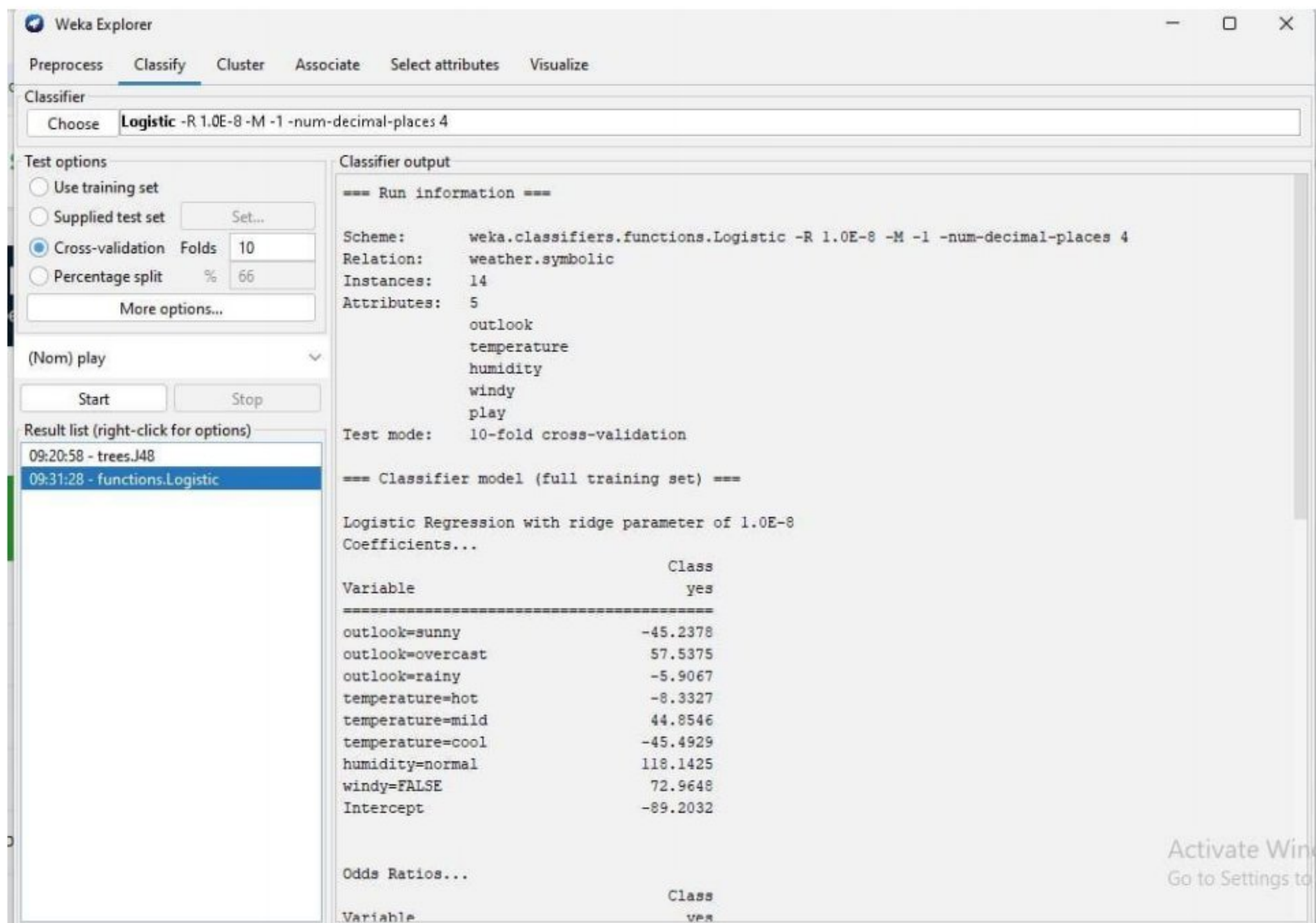
**Run the Classifier:** Click on the "Start" button to build and evaluate the model.

## Analyze the Output

**Review Results:** After running the classifier, check the output window for key statistics:

- **Correctly Classified Instances:** Number and percentage of instances classified correctly.
- **Incorrectly Classified Instances:** Number and percentage of misclassifications.
- **Kappa Statistic:** Measures agreement between predicted and actual classifications.
- **Confusion Matrix:** Detailed breakdown of true positives, false positives, true negatives, and false negatives.

**Coefficients:** The output will also include the coefficients for each attribute, which can help in understanding the influence of each feature on the classification.



The screenshot shows the Weka Explorer application window. The 'Classify' tab is active. In the 'Classifier' section, 'Logistic -R 1.0E-8 -M -1 -num-decimal-places 4' is selected. Under 'Test options', 'Cross-validation' is chosen with 'Folds' set to 10. The 'Start' button is visible. The 'Classifier output' window is open, showing the following information:

```
=== Run information ===  
Scheme:      weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4  
Relation:    weather.symbolic  
Instances:   14  
Attributes:  5  
              outlook  
              temperature  
              humidity  
              windy  
              play  
Test mode:   10-fold cross-validation  
  
=== Classifier model (full training set) ===  
  
Logistic Regression with ridge parameter of 1.0E-8  
Coefficients...  
  
Variable          Class  
=====
```

Variable	Class
outlook=sunny	-45.2378
outlook=overcast	57.5375
outlook=rainy	-5.9067
temperature=hot	-8.3327
temperature=mild	44.8546
temperature=cool	-45.4929
humidity=normal	118.1425
windy=FALSE	72.9648
Intercept	-89.2032

```
Odds Ratios...  
  
Variable          Class  
=====
```



9. To measure the log likelihood of clusters in a large dataset using Weka:

### Apply Clustering Algorithm

1. **Go to the Cluster Tab:** Click on the "Cluster" tab.
2. **Choose Clustering Algorithm:**
  - Click on the "Choose" button and select EM (for Expectation-Maximization).
3. **Set Parameters:** Optionally adjust parameters for the EM algorithm, like the number of clusters or the maximum iterations.
4. **Run the Clustering:** Click the "Start" button to perform clustering on your dataset.

### Step 4: Review Output

1. **Check the Log Likelihood:** Once the clustering completes, the output window will display several statistics, including:
  - **Log Likelihood:** This value indicates how well the model fits the data. Higher log likelihood values generally indicate a better fit.

### Step 5: Interpret Log Likelihood

- **Understanding Log Likelihood:**
  - Log likelihood is the logarithm of the likelihood function, which measures how probable the observed data is under the model.
  - If comparing different models, a higher log likelihood suggests a better model fit.

The screenshot shows the Weka Clusterer window with the EM algorithm selected. The 'Cluster mode' section has 'Use training set' selected. The 'Clusterer output' section displays the following data:

Attribute	Cluster
outlook	
sunny	6
overcast	5
rainy	6
[total]	17
temperature	
mean	73.5714
std. dev.	6.3326
humidity	
mean	81.6429
std. dev.	9.9111
windy	
TRUE	7
FALSE	9
[total]	16
play	
yes	10
no	6
[total]	16

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

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

Clustered Instances



