

## EXERCISE 2 : PREPROCESSING ON DATASETS USING WEGA TOOL

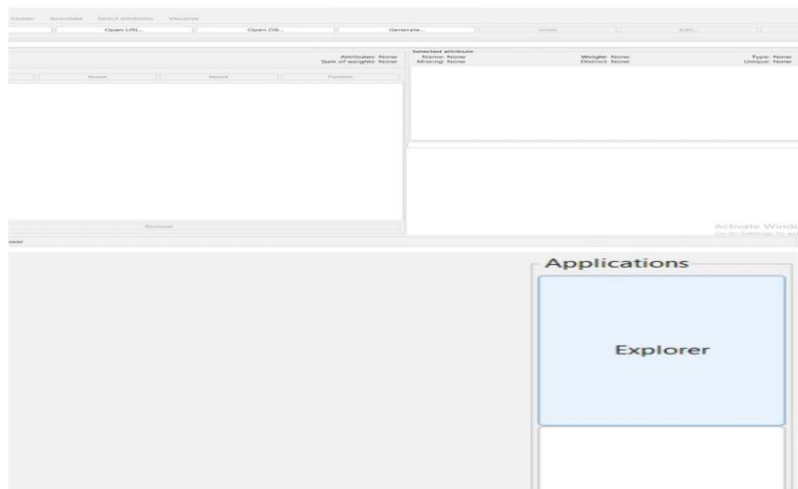
### Procedure :

Performing the following Pre-processing process on predefined datasets like weather.numeric.arff and weather.nominal.arff

#### **a) Loading the Data**

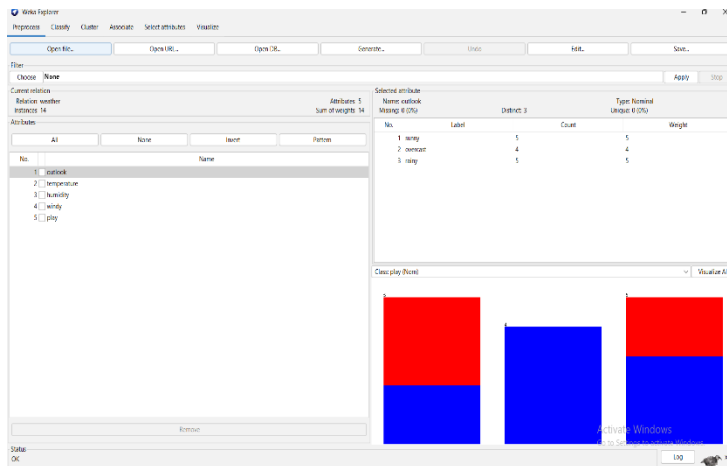
##### **Step 1: Open WEKA Explorer**

- Launch WEKA and navigate to the Explorer interface.



##### **Step 2: Load a Dataset**

- Click on the "Preprocess" tab.
- Click the "Open file..." button to invoke a file chooser dialog.
- Navigate to the directory where the dataset is located.
- Select a dataset file (e.g., " weather.numeric.arff" or "weather.nominal.arff ") and open it.
- Load the weather.numeric.arff" or "weather.nominal.arff dataset and Perform it



### Step 3: Inspect Dataset Attributes

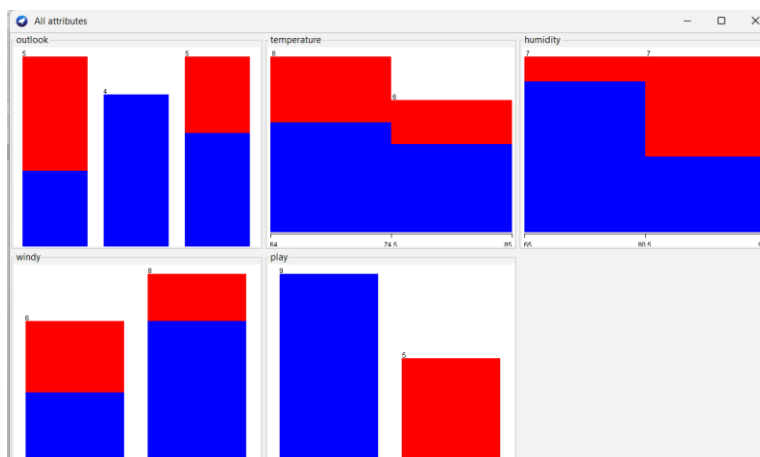
- Observe the left panel in the "Preprocess" tab where attributes are listed.
- Confirm that the base relation name (the dataset name) matches the current working relation name, indicating the dataset is loaded correctly.

### Step 4: View Attribute Statistics

- Click on an attribute in the left panel to view its statistics in the bottom panel.
- For categorical attributes, inspect the frequency distribution of each category.
- For continuous attributes, review statistics such as minimum, maximum, mean, and standard deviation.

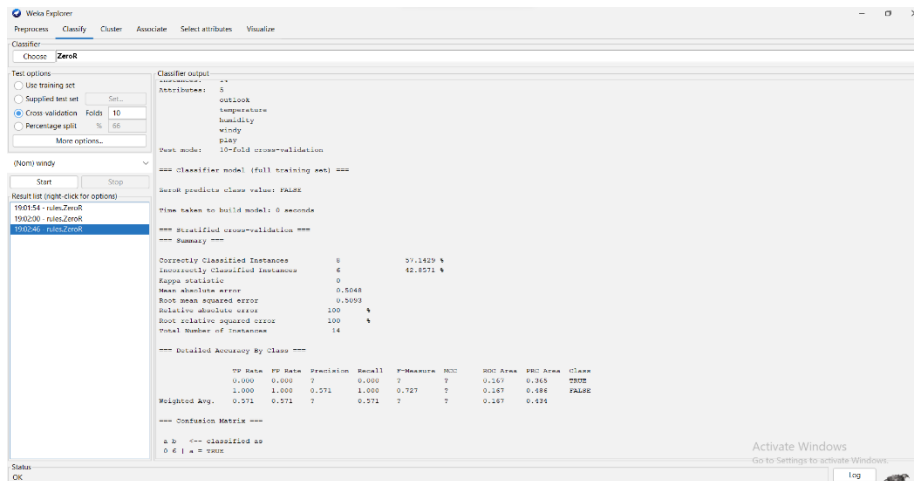
### Step 5: Visualize Attribute Statistics

- Select a continuous attribute (e.g., "temperature") to visualize its distribution.
- Use the "Visualize" button or panel to see the attribute's statistical chart.
- Choose a nominal attribute (e.g., "play") and visualize its frequency distribution.



## Step 6: Explore Attribute Cross-Tabulation

- In the visualization panel, note the default cross-tabulation (e.g., "humidity" vs. "play").
- Use the drop-down lists to select different attributes for cross-tabulation (e.g., change "humidity" to "outlook").
- Observe how the visualization changes based on selected attributes, allowing for analysis of the relationship between them.



## Step 7: Analyze and Conclude

- Use the insights gained from the visualizations and statistics to understand the dataset's characteristics.
- Make notes of any interesting patterns or anomalies for further analysis.

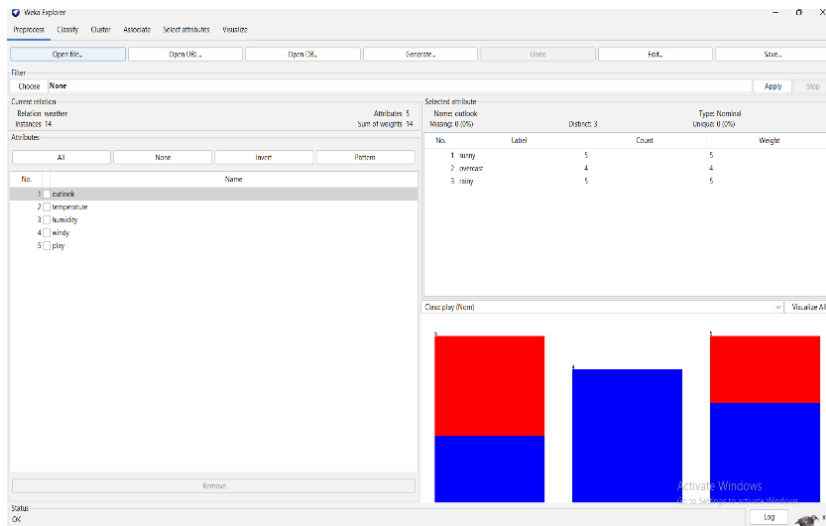
## b) Selecting or Filtering Attributes :

### Load a Dataset:

Open Weka GUI.

Go to the "Explorer" panel.

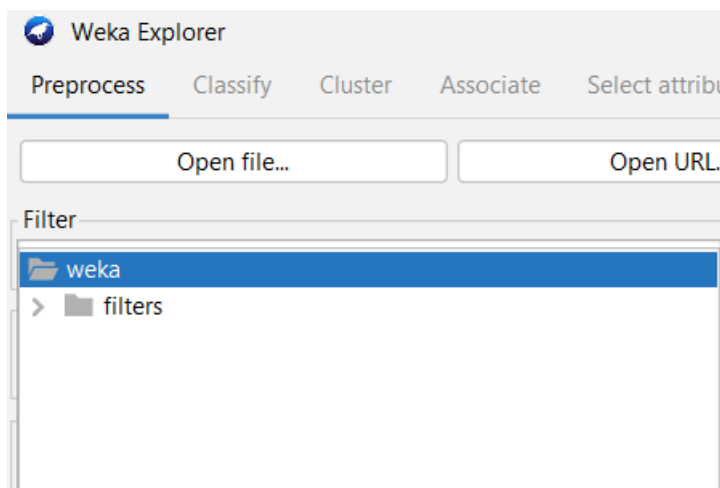
Click "Open file..." and load a dataset (e.g., "labor.arff" or "weather.nominal.arff").



## Navigate to the 'Filter' Panel:

In the Explorer, go to the "Preprocess" tab.

Below the dataset list, find the "Choose" button under the "Filter" section.



## Implement 'Remove' Filter:

Click on "Choose" and select "weka.filters.unsupervised.attribute.Remove".

In the filter's options, set the attribute to remove (e.g., "windy"). This can be done by specifying the attribute index.

Apply the filter by clicking "Apply".

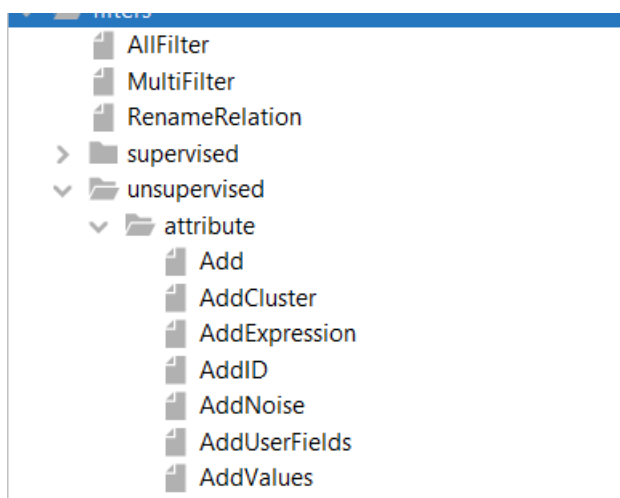
- Unsupervised
- OrdinalToNumeric
- PartitionedMultiFilter
- PKIDiscretize
- PrincipalComponents
- RandomProjection
- RandomSubset
- Remove
- RemoveByName
- RemoveType
- RemoveUseless
- RenameAttribute
- RenameNominalValues
- Reorder
- ReplaceMissingValues
- ReplaceMissingWithUserConstant
- ReplaceWithMissingValue
- SortLabels
- Standardize
- StringToNominal
- StringToWordVector
- SwapValues
- TimeSeriesDelta
- TimeSeriesTranslate

### Apply 'Add' Filter:

Choose "weka.filters.unsupervised.attribute.Add".

Set the attribute's name and its values (e.g., attributeName=NewAttribute, nominalLabels=value1,value2).

Apply the filter.

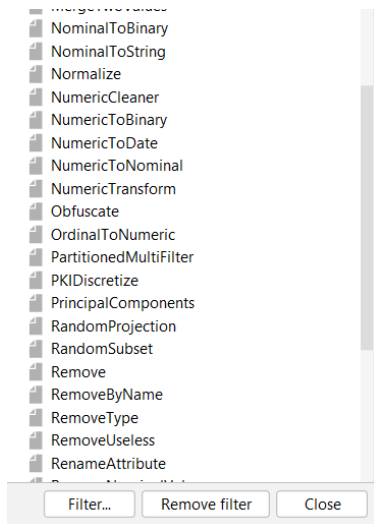


### Normalize Numeric Attributes:

Select "weka.filters.unsupervised.attribute.Normalize".

Specify the numeric attributes to normalize (you can use the attribute indices).

Apply the filter.

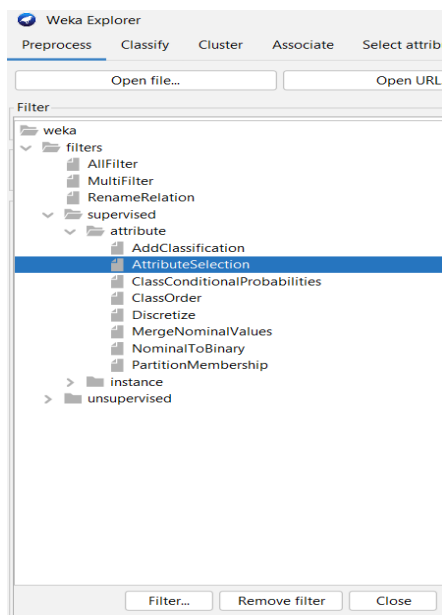


### Attribute Selection:

Choose "weka.filters.supervised.attribute.AttributeSelection".

Specify the attribute selection method and search method in the filter options (e.g., use "CfsSubsetEval" and "BestFirst").

Apply the filter.



### Save the Modified Dataset:

After each filter operation, save the modified dataset.

Click "Save..." in the "Preprocess" tab and choose the desired file format and location to save the dataset.

## C ) Discretization

### **Numeric to Nominal Discretization:**

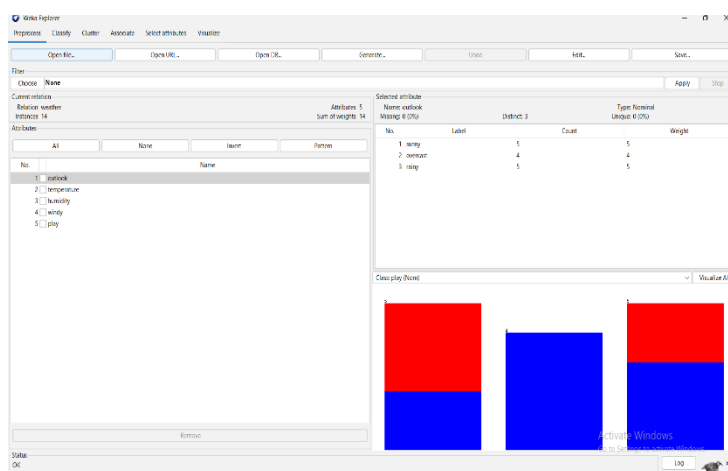
#### **Load the Dataset:**

Open the Weka Explorer by clicking on the "Explorer" button.

In the "Explorer" window, go to the "Preprocess" tab.

Click the "Open file..." button to open a file dialog.

Navigate to the location of the "weather.numeric.arff" dataset, select it, and click "Open" to load the dataset into Weka.



#### **Apply the Discretize Filter:**

Scroll down to the "Filter" section in the "Preprocess" tab.

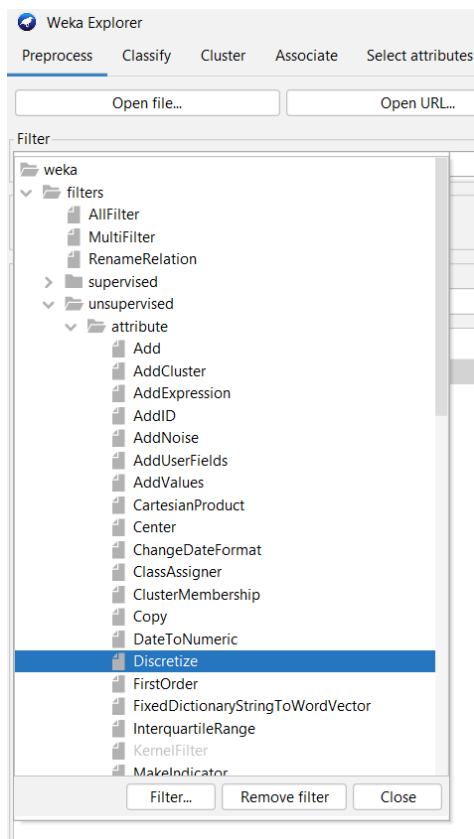
Click the "Choose" button to open the filter selection dialog.

Navigate to "weka.filters.unsupervised.attribute" and select "Discretize" from the list to choose the Discretize filter.

Next to the "Choose" button (where the Discretize filter is now displayed), there's an icon or box for filter options. Click it to configure the Discretize filter.

In the configuration window, you can specify the attributes to discretize. For example, you might see an option like -R first-last (to select all attributes) or you can specify individual indices of the numeric attributes you wish to discretize (e.g., "temperature," "humidity"). Use the attribute indices or actual names if the interface supports it.

Set the number of bins for discretization or configure other options according to your requirements.

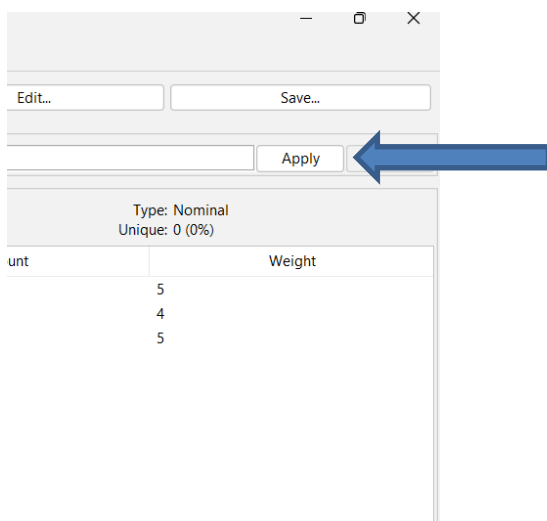


Options might include specifying the binning method, whether to use equal-width or equal-frequency bins, and whether to treat missing values as a separate category.

Click "OK" to close the configuration window and apply your settings to the Discretize filter.

### Apply the Filter

With the Discretize filter configured, click "Apply" in the "Filter" section to apply the discretization to your dataset.





## Nominal to Numeric Discretization:

### Load the Dataset:

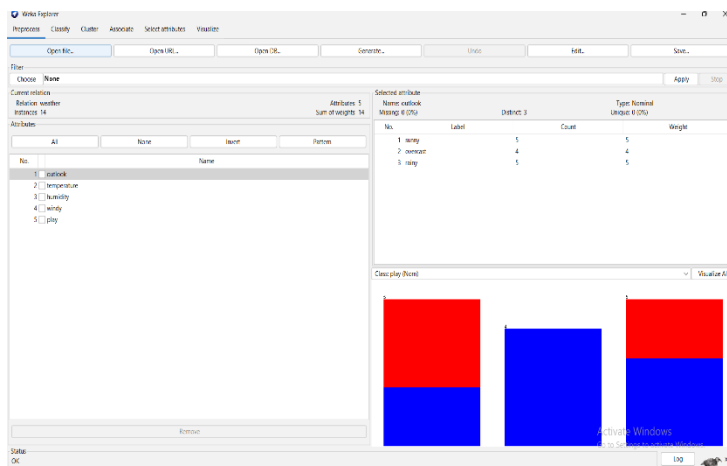
In the "Preprocess" tab, click the "Open file..." button.

Locate and select the "weather.nominal.arff" file, then click "Open" to load the dataset.

Select the Appropriate Filter:

Scroll to the "Filter" section.

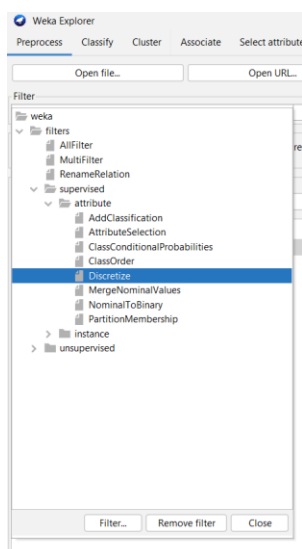
Click "Choose" to open the filter selection dialog.



Navigate to "weka.filters.unsupervised.attribute" and select "NominalToNumeric" from the list. This filter converts nominal attributes to numeric by assigning a unique number to each category within a nominal attribute.

### Configure the Filter:

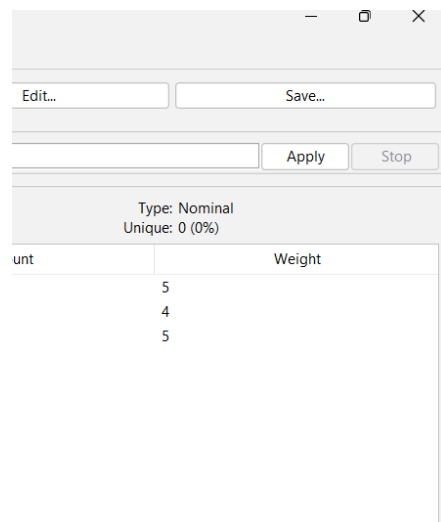
Next to the "Choose" button, click the box or icon to open the filter's configuration window. Here, you can select the specific nominal attributes you wish to convert to numeric.



If your dataset uses attribute names, you may need to check the attribute indices in the dataset and specify them. By default, it may apply to all nominal attributes. There might not be a need to set the number of bins since this process isn't discretization but rather a direct conversion, where each category of a nominal attribute is assigned a unique numeric value.

### **Apply the Filter:**

Click "OK" to confirm the filter settings.



Click "Apply" to convert the selected nominal attributes to numeric.

### **Save the Modified Dataset:**

After conversion, click the "Save" button in the "Preprocess" tab.

Choose a save location and file name, such as "weather-numeric-discretized.arff" (although "discretized" might not be the most accurate term for this process; "weather-nominal-to-numeric.arff" could be more appropriate).

Click "Save" to export the dataset.

## **d) Handling Missing Values:**

### **Introduce Missing Values:**

In the data section of the ARFF file, identify the "outlook" (if it's a numeric dataset, "outlook" might be a misunderstanding, as "outlook" is typically nominal; ensure you're editing the correct file or attribute) and "temperature" columns.

Replace some of the values in these columns with '?' to signify missing values. For a numeric dataset, "outlook" might not exist, so you may focus on "temperature" or another relevant attribute.

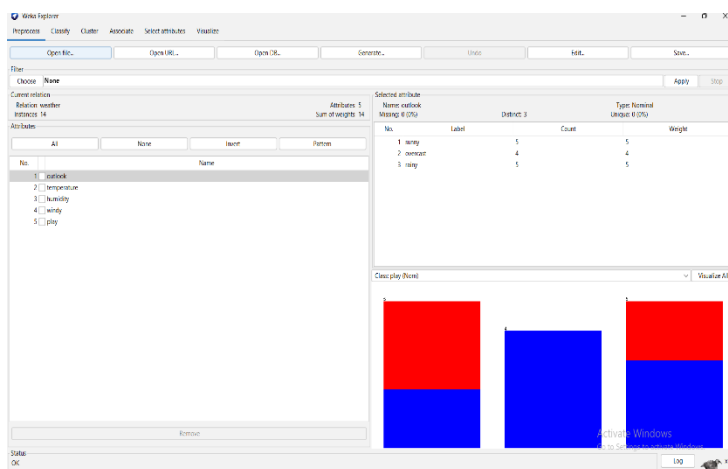
Save the file with a new name, e.g., "weather-numeric-missing.arff".

## Load the Edited Dataset in Weka:

Open Weka Explorer.

Click "Open file..." and select the "weather-numeric-missing.arff" file to load it.

Check for Missing Values:

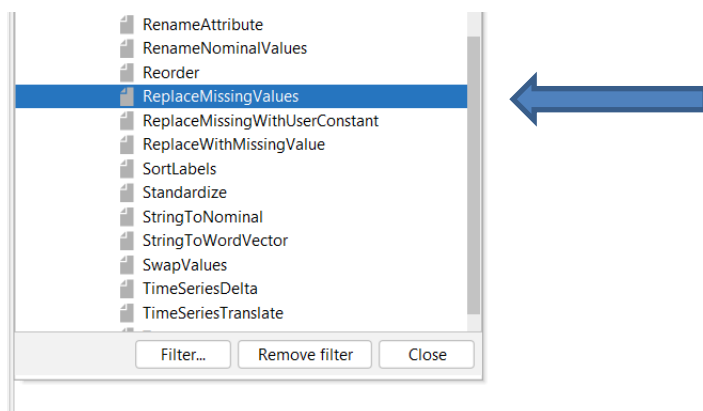


In the "Preprocess" tab, you can view a summary for each attribute in the dataset by selecting it. Missing values are indicated in the attribute summaries.

## Apply the "ReplaceMissingValues" Filter:

Scroll to the "Filter" section.

Click "Choose" and navigate to "weka. filters.unsupervised.attribute.ReplaceMissingValues".



Without further configuration, this filter will replace missing values throughout the entire dataset. For numeric attributes like "temperature", it replaces missing values with the mean

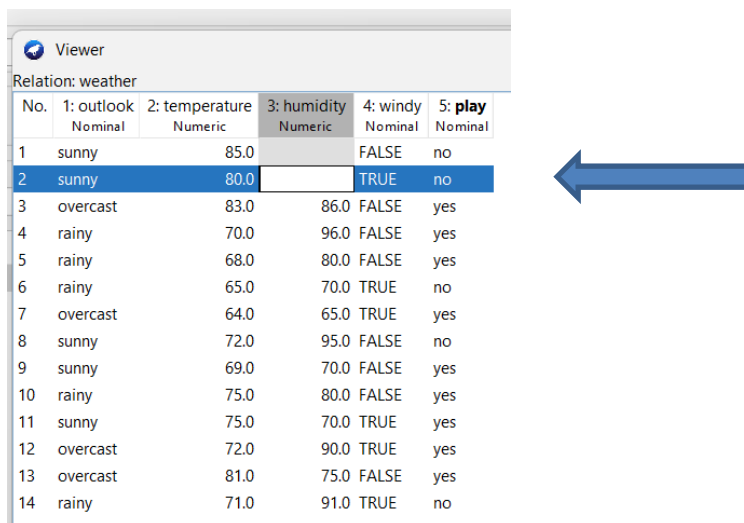
value of the attribute. For nominal attributes, it replaces missing values with the mode (most common value) of the attribute.

Click "Apply" to execute the filter.

### Examine the Data for Replacement:

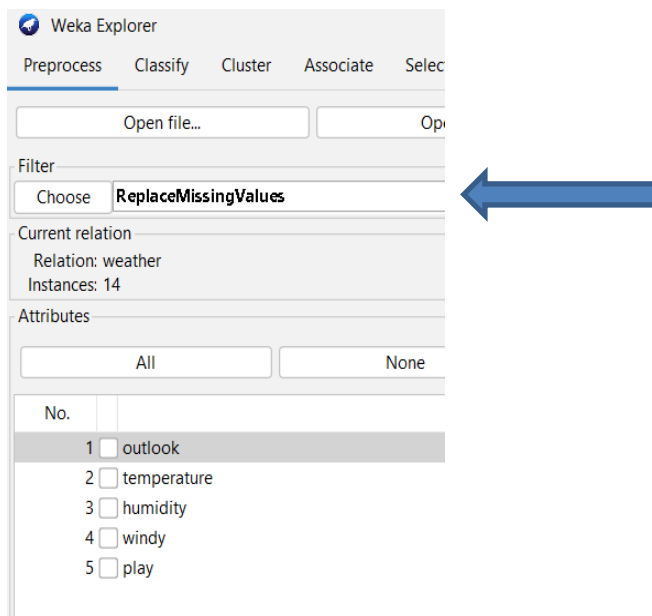
After applying the filter, you can select the previously edited attributes in the "Preprocess" tab to observe the changes. The summary for each attribute should now show zero missing values, and the replacements made by the filter according to the rules mentioned above.

### Before Applying **ReplaceMissingValues**:



No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: play Nominal
1	sunny	85.0		FALSE	no
2	sunny	80.0		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

### Use **ReplaceMissingValues** :



Weka Explorer

Preprocess   Classify   Cluster   Associate   Select

Open file...   Open project...

Filter

Choose **ReplaceMissingValues**

Current relation

Relation: weather

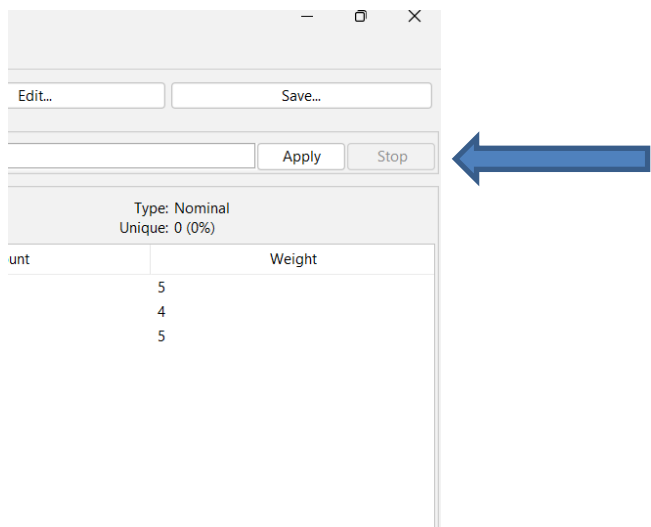
Instances: 14

Attributes

All   None

No.   ☐ outlook  
☐ temperature  
☐ humidity  
☐ windy  
☐ play

To apply ReplaceMissingValues:



After Apply ReplaceMissingValues:

Viewer

relation: weather-weka.filters.unsupervised.attribute.ReplaceMissingValu

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

Result: