EXPERIMENT NO: 2

Aim: To demonstrate pre-processing on provided dataset.

Theory:

The data that is collected from the field contains many unwanted things that leads to wrong analysis.

For example, the data may contain null fields, it may contain columns that are irrelevant to the current analysis, and so on. Thus, the data must be preprocessed to meet the requirements of the type of analysis you are seeking. This is the done in the preprocessing module.

To demonstrate the available features in preprocessing, we can use the database that is provided in the installation.

Weka - Loading Data

The data can be loaded from the following sources –

- Local file system
- Web
- Database

Loading Data from Local File System

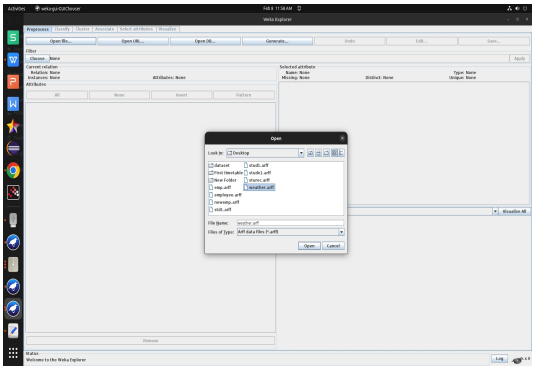
Just under the Machine Learning tabs that you studied in the previous lesson, you would find the following three buttons –

Open file ...

Open URL ...

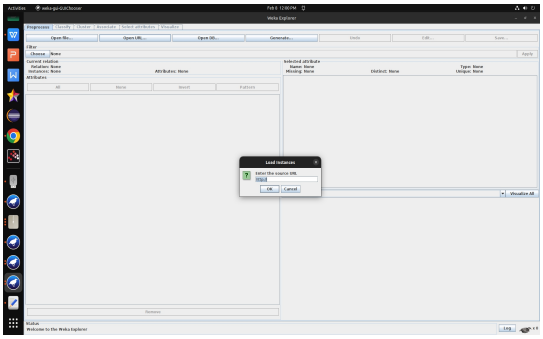
Open DB ...

Click on the **Open file** ... button. A directory navigator window opens as shown in the following screen –



Loading Data from Web

Once you click on the **Open URL ...** button, you can see a window as follows –



We will open the file from a public URL Type the following URL in the popup box –

for example: https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/weather.nominal.arff

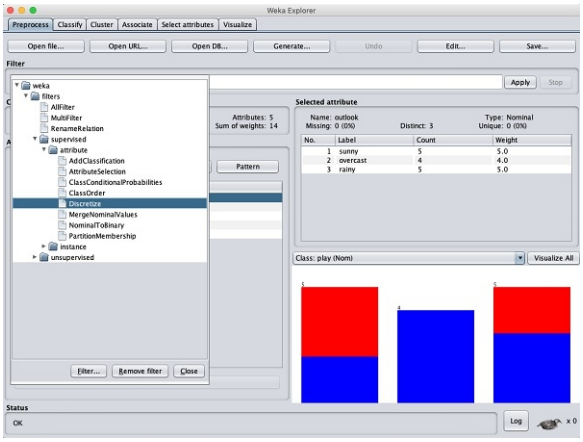
You may specify any other URL where your data is stored. The **Explorer** will load the data from the remote site into its environment.

Understanding Data

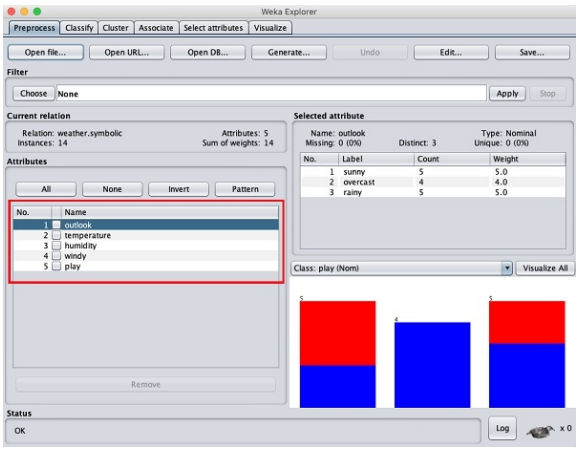
Let us first look at the highlighted Current relation sub window. It shows the name of the database that is currently loaded. You can infer two points from this sub window –
There are 14 instances - the number of rows in the table.
The table contains 5 attributes - the fields, which are discussed in the upcoming sections.
On the left side, notice the **Attributes** sub window that displays the various fields in the database.

Applying Filters

Some of the machine learning techniques such as association rule mining requires categorical data.
To illustrate the use of filters, we will use **weather-numeric.arff** database that contains two **numeric** attributes - **temperature** and **humidity**.
We will convert these to **nominal** by applying a filter on our raw data. Click on the **Choose** button in the **Filter** subwindow and select the following filter –
weka→filters→supervised→attribute→Discretize

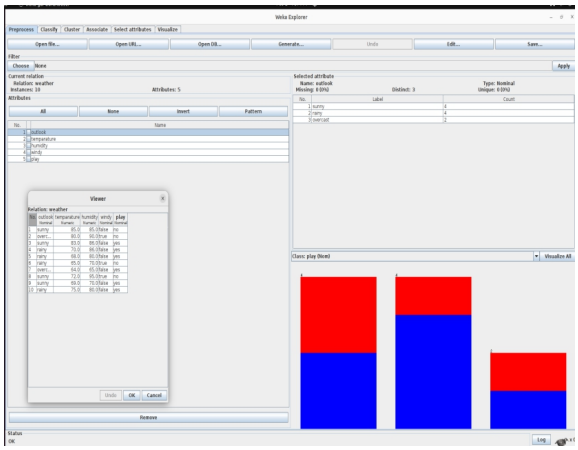


Click on the **Apply** button and examine the **temperature** and/or **humidity** attribute. You will notice that these have changed from numeric to nominal types.



Like this we can apply following preprocessing on the data set.

1) Add
2) Remove
3) Normalization



Add Pre-Processing Technique:

Procedure:

1) Start Programs Weka-3-4 Weka-3-4

2) Click on **explorer.**

3) Click on **open file.**

4) Select **Weather.arff** file and click on open.

5) Click on **Choose button** and select the **Filters option.**

6) In Filters, we have **Supervised** and **Unsupervised data.**

7) Click on **Unsupervised data.**

8) Select the attribute **Add.**

9) A new window is opened.

10) In that we enter attribute index, type, data format, nominal label values for **Climate.**

11) Click on **OK.**

12) Press the **Apply button,** then a new attribute is added to the Weather Table.

13) **Save** the file.

14) Click on the **Edit button,** it shows a new Weather Table on Weka.

Weather Table after adding new attribute CLIMATE:

Viewer

Relation: weather-weka.filters.unsupervised.attribute.Add-T...

No.	CLIMATE Nominal	outlook Nominal	temparature Numeric	humidity Numeric	windy Nominal	play Nominal
1		sunny	85.0	85.0	false	no
2		overc...	80.0	90.0	true	no
3		sunny	83.0	86.0	false	yes
4		rainy	70.0	86.0	false	yes
5		rainy	68.0	80.0	false	yes
6		rainy	65.0	70.0	true	no
7		overc...	64.0	65.0	false	yes
8		sunny	72.0	95.0	true	no
9		sunny	69.0	70.0	false	yes
10		rainy	75.0	80.0	false	yes

Undo OK Cancel

Remove Pre-Processing Technique:

Procedure:

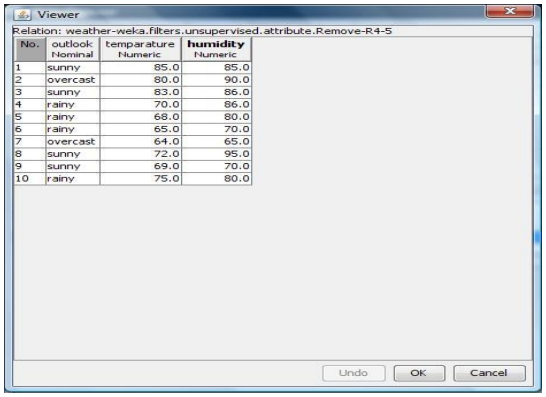
1) Start Programs Weka-3-4 Weka-3-4

2) Click on explorer.

3) Click on open file.

4) Select Weather.arff file and click on open.
5) Click on Choose button and select the Filters option.
6) In Filters, we have Supervised and Unsupervised data.
7) Click on Unsupervised data.
8) Select the attribute Remove.
9) Select the attributes windy, play to Remove.
10) Click Remove button and then Save.
11) Click on the Edit button, it shows a new Weather Table on Weka.

Weather Table after removing attributes WINDY, PLAY:



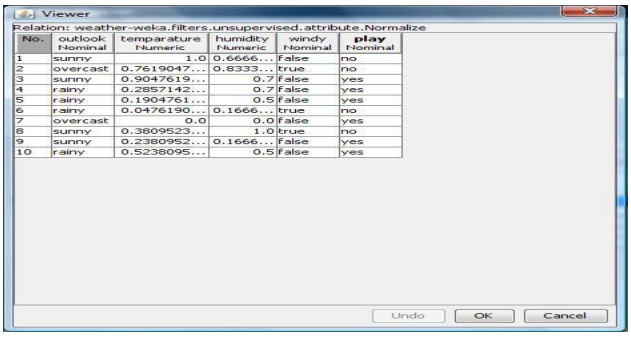
No.	outlook Nominal	temparature Numeric	humidity Numeric
1	sunny	85.0	85.0
2	overcast	80.0	90.0
3	sunny	83.0	86.0
4	rainy	70.0	86.0
5	rainy	68.0	80.0
6	rainy	65.0	70.0
7	overcast	64.0	65.0
8	sunny	72.0	95.0
9	sunny	69.0	70.0
10	rainy	75.0	80.0

Normalize Pre-Processing Technique:

Procedure:

1) Start Programs Weka-3-4 Weka-3-4
2) Click on explorer.
3) Click on open file.
4) Select Weather.arff file and click on open.
5) Click on Choose button and select the Filters option.
6) In Filters, we have Supervised and Unsupervised data.
7) Click on Unsupervised data.
8) Select the attribute Normalize.
9) Select the attributes temparature, humidity to Normalize.
10) Click on Apply button and then Save.
11) Click on the Edit button, it shows a new Weather Table with normalized values on Weka.

Weather Table after Normalizing TEMPARATURE, HUMIDITY:



No.	outlook Nominal	temparature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	1.0	0.6666...	false	no
2	overcast	0.7619047...	0.8333...	true	no
3	sunny	0.9047619...	0.7	false	yes
4	rainy	0.2857142...	0.7	false	yes
5	rainy	0.1904761...	0.5	false	yes
6	rainy	0.0476190...	0.1666...	true	no
7	overcast	0.0	0.0	false	yes
8	sunny	0.3809523...	1.0	true	no
9	sunny	0.2380952...	0.1666...	false	yes
10	rainy	0.5238095...	0.5	false	yes

Result:

The pre-processing on the given data set is executed.

Viva Questions

1. Define Preprocessing Technique**?**
2. What are Data preprocessing steps**?**
3. What are the 3 stages of data processing**?**