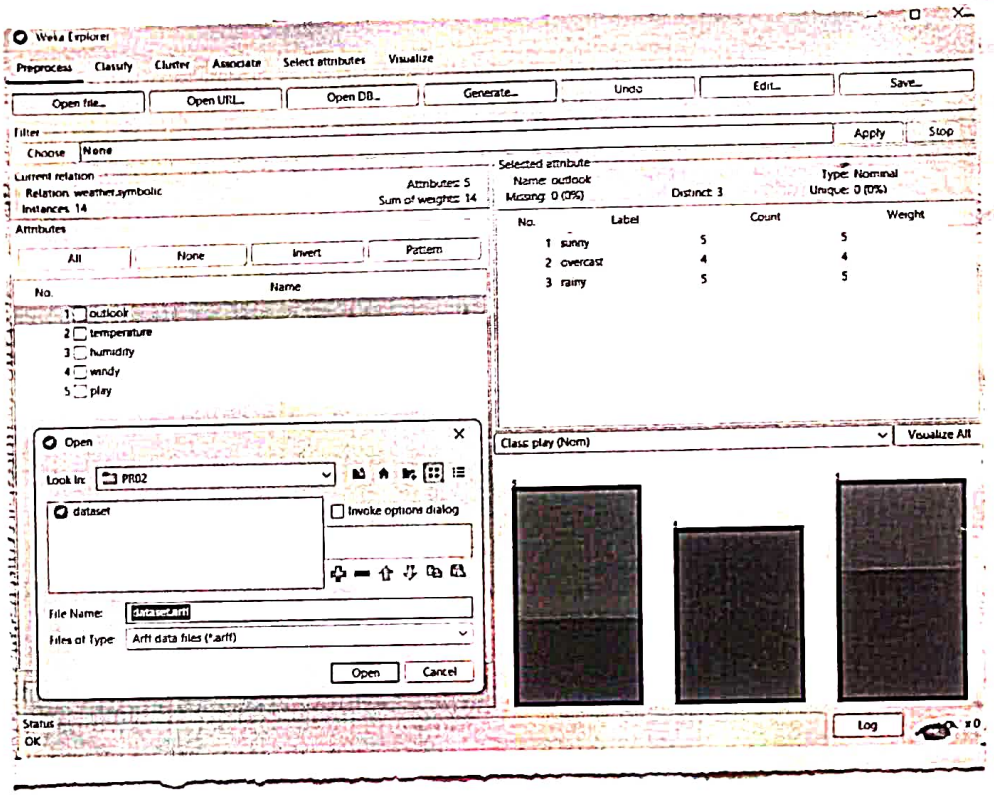
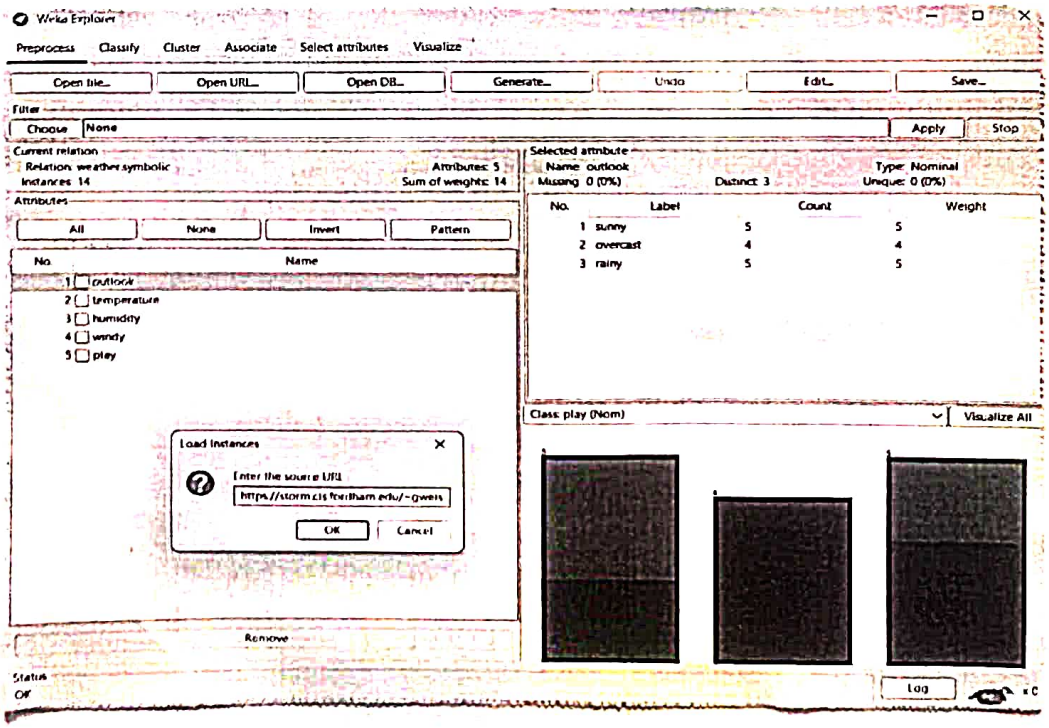
Practical No. 02

Aim: To demonstrate pre-processing on provided dataset.





Date :

Practical No. 02

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.

Load data from Local File System :→

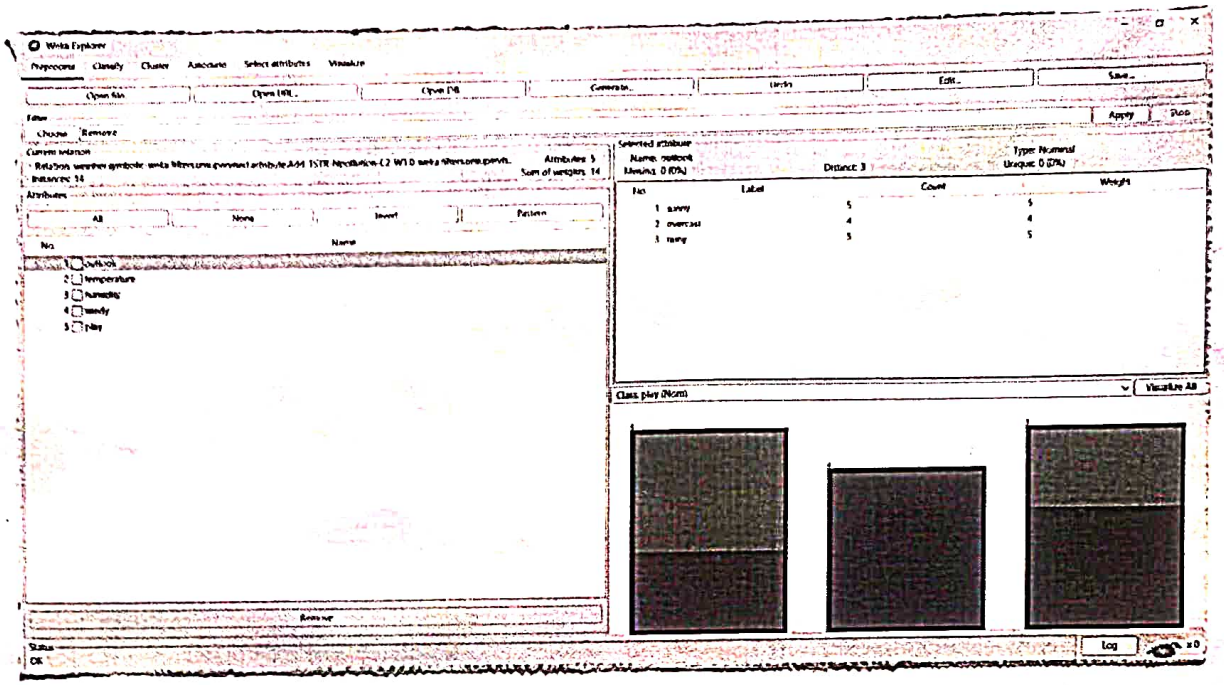
Just under the Machine learning tabs that you studies in the previous lesson, you would find the following three buttons –

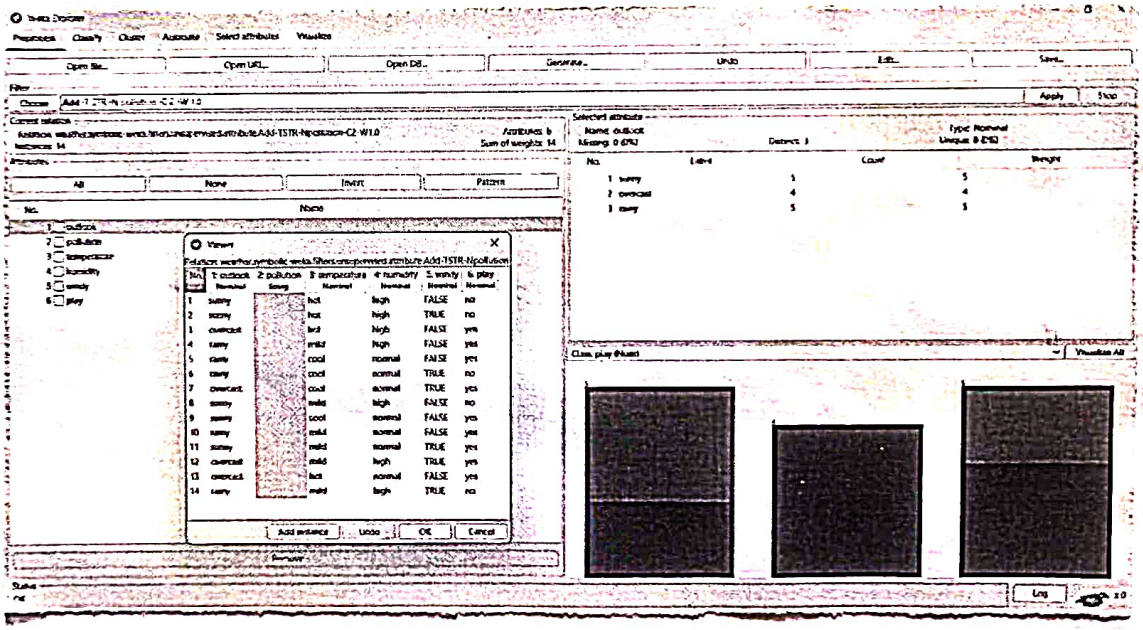
open file...
open URL...
open DB...

click on the open file ... button. A directory navigator window opens as shown in following screen –

Load data from web :→

Once you click on the open URL... button, you Page No. see a window as follows :





Date :

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 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 field, which are discussed in the upcoming sections.

On the left side, notice the Attributes sub window that display the various fields in the database.

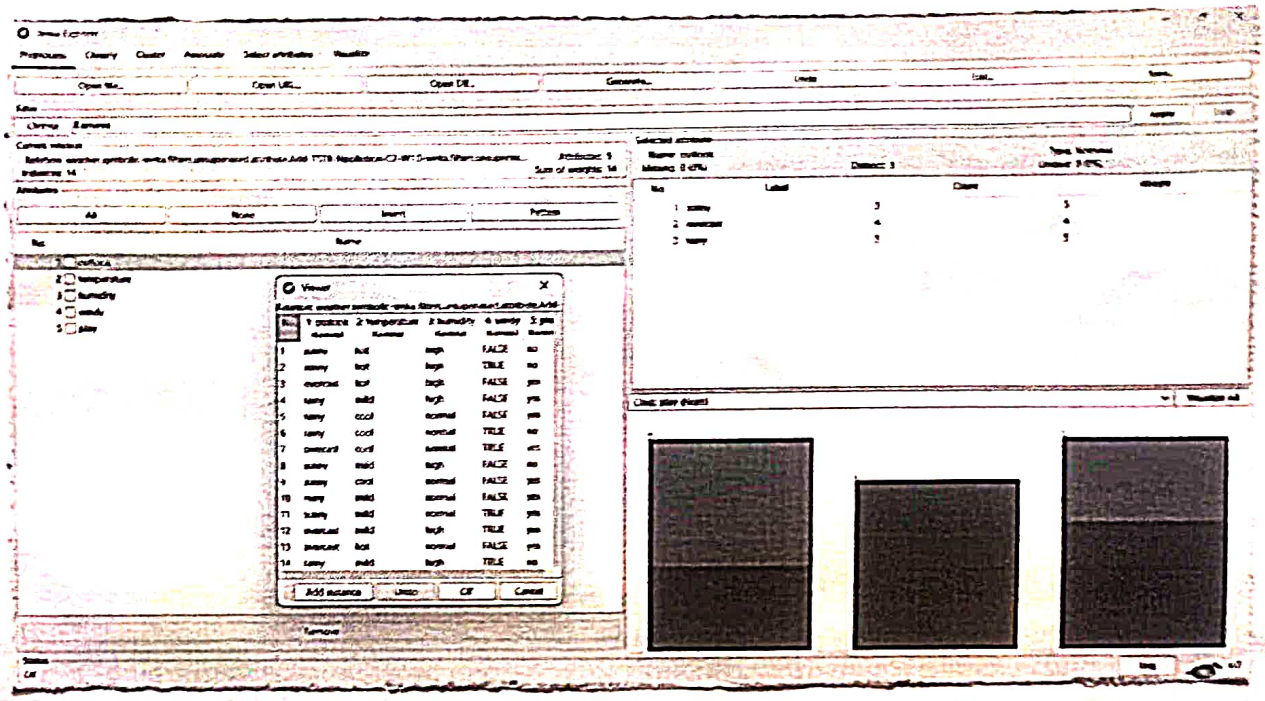
Applying Filters : →

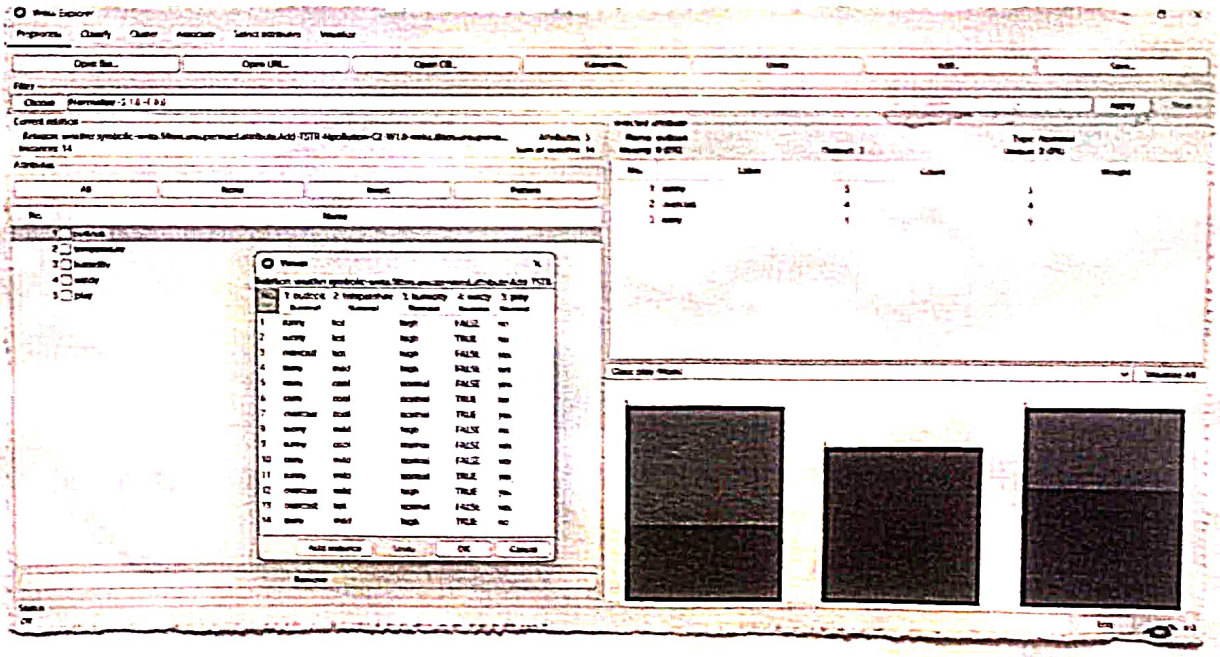
Some of the machine learning techniques such as associations rule mining required categorical data. To illustrate the use of filters, we will use weather-numeric.arff database that contain two numeric attributes - temperature and humidity.

We will convert these to nominal by applying a filters on our raw data, click on the choose button in the filter subwindow and select the following filters -

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.





Date :

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

(1) Add : To add attribute. It is present in unsupervised data. In that we can enter attribute index, type, data format, nominal label values.

(2) Remove : To remove attributes, select them and click on the remove button at the bottom. The selected attributes would be removed from db.

(3) Normalization :- It is in unsupervised data. It is good technique to use when you do not know the distribution of your data. You can normalized all of the your dataset.

Result :

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

Viva Questions :→

(1) Define preprocessing Technique ?

→ Preprocessing tools in WEKA are called filters. The preprocess retrives data from a file, SQL db or URL. Some preprocessing are Add, remove, normalization & etc..

(2) What are Data preprocessing steps ?

→ Step 1 : open weka
Step 2 : load dataset from URL, SQL database or file.
Step 3 : click on filters
Step 4 : click on supervised
Step 5 : click on attribute
Step 6 : AttributeSelection or select the preprocessing technique ie. add, remove or etc.
Step 7 : click on Apply.

Date :

③ what are the 3 stages of data processing?

→ Stage 1: Loading data or dataset

Stage 2: understanding Data

Stage 3: Apply Filters on dataset and save.